Jean-Michel Josselin
Benoît Le Maux

Statistical Tools for Program Evaluation

Methods and Applications to Economic Policy, Public Health, and Education

Springer

Statistical Tools for Program Evaluation

Jean-Michel Josselin • Benoît Le Maux

Statistical Tools for Program Evaluation

Methods and Applications to Economic Policy, Public Health, and Education

Springer

Jean-Michel Josselin
Faculty of Economics
University of Rennes 1
Rennes, France

Benoît Le Maux
Faculty of Economics
University of Rennes 1
Rennes, France

ISBN 978-3-319-52826-7 ISBN 978-3-319-52827-4 (eBook)
DOI 10.1007/978-3-319-52827-4

Library of Congress Control Number: 2017940041

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

We would like to express our gratitude to those who helped us and made the completion of this book possible.

First of all, we are deeply indebted to the Springer editorial team and particularly Martina BIHN whose support and encouragement allowed us to finalize this project.

Furthermore, we have benefited from helpful comments by colleagues and we would like to acknowledge the help of Maurice BASLÉ, Arthur CHARPENTIER, Pauline CHAUVIN, Salah GHABRI, and Christophe TAVÉRA. Of course, any mistake that may remain is our entire responsibility.

In addition, we are grateful to our students who have been testing and experimenting our lectures for so many years. Parts of the material provided here have been taught at the Bachelor and Master levels, in France and abroad. Several students and former students have been helping us improve the book. We really appreciated their efforts and are very grateful to them: Erwan AUTIN, Benoît CARRÉ, Aude DAILLÈRE, Kristýna DOSTÁLOVÁ, and Adrien VEZIE.

Finally, we would like to express our sincere gratefulness to our families for their continuous support and encouragement.

Contents

Part I Identifying the Context of the Program

Part II Ex ante Evaluation

Part III Ex post Evaluation

Statistical Tools for Program Evaluation: Introduction and Overview

1

1.1 The Challenge of Program Evaluation

The past 30 years have seen a convergence of management methods and practices between the public sector and the private sector, not only at the central government level (in particular in Western countries) but also at upper levels (European commission, OECD, IMF, World Bank) and local levels (municipalities, cantons, regions). This "new public management" intends to rationalize public spending, boost the performance of services, get closer to citizens' expectations, and contain deficits. A key feature of this evolution is that program evaluation is nowadays part of the policy-making process or, at least, on its way of becoming an important step in the design of public policies. Public programs must show evidence of their relevance, financial sustainability and operationality. Although not yet systematically enacted, program evaluation intends to grasp the impact of public projects on citizens, as comprehensively as possible, from economic to social and environmental consequences on individual and collective welfare. As can be deduced, the task is highly challenging as it is not so easy to put a value on items such as welfare, health, education or changes in environment. The task is all the more demanding that a significant level of expertise is required for measuring those impacts or for comparing different policy options.

The present chapter offers an introduction to the main concepts that will be used throughout the book. First, we shall start with defining the concept of program evaluation itself. Although there is no consensus in this respect, we may refer to the OECD glossary which states that evaluation is the "*process whereby the activities undertaken by ministries and agencies are assessed against a set of objectives or criteria*." According to Michael Quinn Patton, former President of the American Evaluation Association, program evaluation can also be defined as "*the systematic collection of information about the activities, characteristics, and outcomes of programs, for use by people to reduce uncertainties, improve effectiveness, and make decisions*." We may also propose our own definition of the concept: program evaluation is a process that consists in collecting, analyzing, and using information

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_1

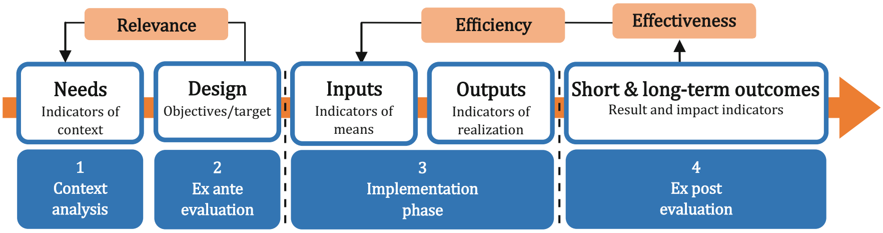


Fig. 1.1 Program evaluation frame

to assess the relevance of a public program, its effectiveness and its efficiency. Those concepts are further detailed below. Note that a distinction will be made throughout the book between a program and its alternative and competing strategies of implementation. By strategies, we mean the range of policy options or public projects that are considered within the framework of the program. The term program, on the other hand, has a broader scope and relates to the whole range of steps that are carried out in order to attain the desired goal.

As shown in Fig. 1.1, a program can be described in terms of needs, design, inputs and outputs, short and long-term outcomes. Needs can be defined as a desire to improve current outcomes or to correct them if they do not reach the required standard. Policy design is about the definition of a course of action intended to meet the needs. The inputs represent the resources or means (human, financial, and material) used by the program to carry out its activities. The outputs stand for what comes out directly from those activities (the intervention) and which are under direct control of the authority concerned. The short-term and long-term outcomes stand for effects that are induced by the program but not directly under the control of the authority. Those include changes in social, economic, environmental and other indicators.

Broadly speaking, the evaluation process can be represented through a linear sequence of four phases (Fig. 1.1). First, a context analysis must gather information and determine needs. For instance, it may evidence a high rate of school dropout among young people in a given area. A program may help teachers, families and children and contribute to prevent or contain dropout. If the authority feels that the consequences on individual and collective welfare are great enough to justify the design of a program, and if such a program falls within their range of competences, then they may wish to put it forward. Context analysis relies on descriptive and inferential statistical tools to point out issues that must be addressed. Then, the assessment of the likely welfare changes that the program would bring in to citizens is a crucial task that uses various techniques of preference revelation and measurement.

Second, ex-ante evaluation is interested in setting up objectives and solutions to address the needs in question. Ensuring the relevance of the program is an essential part of the analysis. Does it make sense within the context of its environment? Coming back to our previous example, the program can for instance consist of alternative educational strategies of follow-up for targeted schoolchildren, with various projects involving their teachers, families and community. Are those strategies consistent with the overall goal of the program? It is also part of this stage to define the direction of the desired outcome (e.g., dropout reduction) and, sometimes, the desired outcome that should be arrived at, namely the target (e.g., a reduction by half over the project time horizon). Another crucial issue is to select a particular strategy among the competing ones. In this respect, methods of ex-ante evaluation include financial appraisal, budget impact analysis, cost benefit analysis, cost effectiveness analysis and multi-criteria decision analysis. The main concern is to find the most efficient strategy. Efficiency can be defined as the ability of the program to achieve the expected outcomes at reasonable costs (e.g., is the budget burden sustainable? Is the strategy financially and economically profitable? Is it cost-effective?)

Third, during the implementation phase, it is generally advised to design a monitoring system to help the managers follow the implementation and delivery of the program. Typical questions are the following. Are potential beneficiaries aware of the program? Do they have access to it? Is the application and selection procedure appropriate? Indicators of means (operating expenditures, grants received, number of agents) and indicators of realization (number of beneficiaries or users) can be used to measure the inputs and the outputs, respectively. Additionally, a set of management and accounting indicators can be constructed and collected to relate the inputs to the outputs (e.g., operating expenditures per user, number of agents per user). Building a well documented data management system is crucial for two reasons. First, those performance indicators can be used to report progress and alert managers to problems. Second, they can be used subsequently for ex-post evaluation purposes.

Last, the main focus of ex post evaluation is on effectiveness, i.e. the extent to which planned outcomes are achieved as a result of the program, ceteris paribus. Among others, methods include benchmarking, randomized controlled experiments and quasi-experiments. One difficulty is the time frame. For instance, the information needed to assess the program's outcomes is sometimes fully available only several years after the end of the program. For this reason, one generally distinguishes the short-term outcomes, i.e. the immediate effects on individuals' status as measured by a result indicator (e.g., rate of dropout during mandatory school time) from the longer term outcomes, i.e. the environmental, social and economic changes as measured by impact indicators (e.g., the impact of dropout on unemployment). In practice, ex post evaluation focuses mainly on short-term outcomes, with the aim to measure what has happened as a direct consequence of the intervention. The analysis also assesses what the main factors behind success or failure are.

We should come back to this distinction that we already pointed out between efficiency and effectiveness. Effectiveness is about the level of outcome per se and whether the intervention was successful or not in reaching a desired target. Depending on the policy field, the outcome in question may differ greatly. In health, for instance, the outcome can relate to survival. In education, it can be school completion. Should an environmental program aim at protecting and restoring watersheds, then the outcome would be water quality. An efficiency analysis on the other hand has a broader scope as it relates the outcomes of the intervention to its cost.

Note also that evaluation should not be mistaken for monitoring. Roughly speaking, monitoring refers to the implementation phase and aims to measure progress and achievement all along the program's lifespan by comparing the inputs with the achieved outputs. The approach consists in defining performance indicators, routinely collect data and examine progress through time in order to reduce the likelihood of facing major delays or cost overruns. While it constitutes an important step of the intervention logic of a program, monitoring is not about evaluating outcomes per se and, as such, will be disregarded in the present work.

The remainder of the chapter is as follows. Section 1.2 offers a description of the tools that can be used to assess the context of a public program. Sections 1.3 and 1.4 are about ex-ante and ex-post evaluations respectively. Section 1.5 explains how to use the book.

1.2 Identifying the Context of the Program

The first step of the intervention logic is to describe the social, economic and institutional context in which the program is to be implemented. Identifying needs, determining their extent, and accurately defining the target population are the key issues. The concept of "needs" can be defined as the difference, or gap, between a current situation and a reasonably desired situation. Needs assessment can be based on a cross-sectional study (comparison of several jurisdictions at one specific point in time), a longitudinal study (repeated observations over several periods of time), or a panel data study (both time and individual dimensions are taken into account). Statistical tools which are relevant in this respect are numerous. Figure 1.2 offers an illustration.

First, a distinction is made between descriptive statistics and inferential statistics. Descriptive statistics summarizes data numerically, graphically or with tables. The main goal is the identification of patterns that might emerge in a sample. A sample is a subset of the general population. The process of sampling is far from straightforward and it requires an accurate methodology if the sample is to adequately represent the population of interest. Descriptive statistical tools include measures of central tendency (mean, mode, median) to describe the central position of observations in a group of data, and measures of variability (variance, standard deviation) to summarize how spread out the observations are. Descriptive statistics does not claim to generalize the results to the general population. Inferential statistics on the other hand relies on the concept of confidence interval, a range of values which is likely to include an unknown characteristic of a population. This population parameter and the related confidence interval are estimated from the sample data. The method can also be used to test statistical hypotheses, e.g., whether the population parameter is equal to some given value or not.

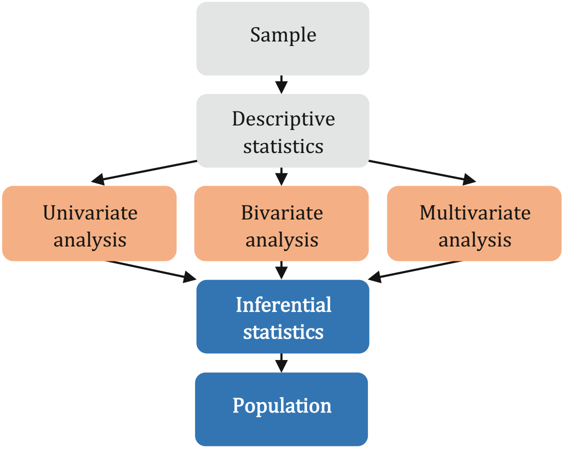


Fig. 1.2 Statistical methods at a glance

Second, depending on the number of variables that are examined, a distinction is made between univariate, bivariate and multivariate analyses. Univariate analysis is the simplest form and it examines one single variable at a time. Bivariate analysis focuses on two variables per observation simultaneously with the goal of identifying and quantifying their relationship using measures of association and making inferences about the population. Last, multivariate analyses are based on more than two variables per observation. More advanced tools, e.g., econometric analysis, must be employed in that context. Broadly speaking, the approach consists in estimating one or several equations that the evaluator think are relevant to explain a phenomenon. A dependent variable (explained or endogenous variable) is then expressed as a function of several independent variables (explanatory or exogenous variables, or regressors).

Third, program evaluation aims at identifying how the population would fare if the identified needs were met. To do so, the evaluator has to assess the indirect costs (negative externalities) as well as benefits (direct utility, positive externalities) to society. When possible, these items are expressed in terms of equivalent money-values and referred to as the willingness to pay for the benefits of the program or the willingness to accept its drawbacks. In other cases, especially in the context of health programs, those items must be expressed in terms of utility levels (e.g., quality adjusted life years lived, also known as QALYs). Several methods exist with their pros and cons (see Fig. 1.3). For instance, stated preference methods (contingent valuation and discrete choice experiment) exploit specially constructed questionnaires to elicit willingness to pay. Their main shortcoming is the failure to properly consider the cognitive constraints and strategic behavior of the agents participating in the experiment, leading to individuals' stated preferences that may not totally reflect their genuine preferences. Revealed preference methods use information from related markets and examine how agents behave in the face of real choices (hedonic-pricing and travel-cost methods). The main advantage of those methods is that they imply real money transactions and, as such, avoid the

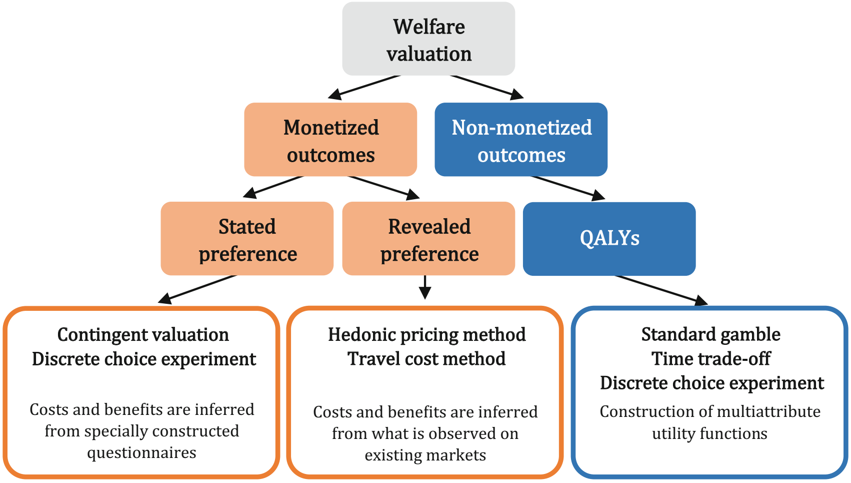

Welfare valuation
Monetized outcomes
Non-monetized outcomes
Stated preference
Revealed preference
QALYs
Contingent valuation
Discrete choice experiment
Costs and benefits are inferred from specially constructed questionnaires
Hedonic pricing method
Travel cost method
Costs and benefits are inferred from what is observed on existing markets
Standard gamble
Time trade-off
Discrete choice experiment
Construction of multiattribute utility functions

Fig. 1.3 Estimation of welfare changes

potential problems associated with hypothetical responses. They require however a large dataset and are based on sets of assumptions that are controversial. Last, health technology assessment has developed an ambitious framework for evaluating personal perceptions of the health states individuals are in or may fall into. Contrary to revealed or stated preferences, this valuation does not involve any monetization of the consequences of a health program on individual welfare.

Building a reliable and relevant database is a key aspect of context analysis. Often one cannot rely on pre-existing sources of data and a survey must be implemented to collect information from some units of a population. The design of the survey has its importance. It is critical to be clear on the type of information one needs (individuals and organizations involved, time period, geographical area), and on how the results will be used and by whom. The study must not only concern the socio economic conditions of the population (e.g., demographic dynamics, GDP growth, unemployment rate) but must also account for the policy and institutional aspects, the current infrastructure endowment and service provision, the existence of environmental issues, etc. A good description of the context and reliable data are essential, especially if one wants to forecast future trends (e.g., projections on users, benefits and costs) and motivate the assumptions that will be made in the subsequent steps of the program evaluation.

1.3 Ex ante Evaluation Methods

Making decisions in a non-market environment does not mean the absence of budget constraint. In the context of decisions on public projects, there are usually fixed sectoral (healthcare, education, etc.) budgets from which to pick the resources required to fund interventions. Ex ante evaluation is concerned with designing public programs that achieve some effectiveness, given those budget constraints. Different forms of evaluation can take place depending on the type of outcome that is analyzed. It is therefore crucial to clearly determine the program's goals and objectives before carrying out an evaluation. The goal can be defined as a statement of the desired effect of the program. The objectives on the other hand stand for specific statements that support the accomplishment of the goal.

Different strategies/options can be envisaged to address the objectives of the program. It is important that those alternative strategies are compared on the basis of all relevant dimensions, be it technological, institutional, environmental, financial, social and economic. Among others, most popular methods of comparison include financial analysis, budget impact analysis, cost benefit analysis, cost effectiveness analysis and multi-criteria decision analysis. Each of these methods has its specificities. The key elements of a financial analysis are the cost and revenue forecasts of the program. The development of the financial model must consider how those items interact with each other to ensure both the sustainability (capacity of the project revenues to cover the costs on an annual basis) and profitability (capacity of the project to achieve a satisfactory rate of return) of the program. Budget impact analysis examines the extent to which the introduction of a new strategy in an existing program affects the authority's budget as well as the level and allocation of outcomes amongst the interventions (including the new one). Cost benefit analysis aims to compare cost forecasts with all social, economic and environmental benefits, expressed in monetary terms. Cost effectiveness analysis on the other hand focuses on one single measure of effectiveness and compares the relative costs and outcomes of two or more competing strategies. Last, multi-criteria decision analysis is concerned with the analysis of multiple outcomes that are not monetized but reflect the several dimensions of the pursued objective. Financial flows may be included directly in monetary terms (e.g., a cost, an average wage) but other outcomes are expressed in their natural unit (e.g., success rate, casualty frequency, utility level).

Figure 1.4 underlines roughly the differences between the ex ante evaluation techniques. All approaches account for cost considerations. Their main difference is with respect to the outcome they examine.

Financial Analysis Versus Cost Benefit Analysis A financial appraisal examines the projected revenues with the aim of assessing whether they are sufficient to cover expenditures and to make the investment sufficiently profitable. Cost benefit analysis goes further by considering also the satisfaction derived from the consumption of public services. All effects of the project are taken into account, including social, economic and environmental consequences. The approaches are thereby different, but also complementary, as a project that is financially viable is not necessarily economically relevant and vice versa. In both approaches, discounting can be used to compare flows occurring at different time periods. The idea is based on the principle that, in most cases, citizens prefer to receive goods and services now rather than later.

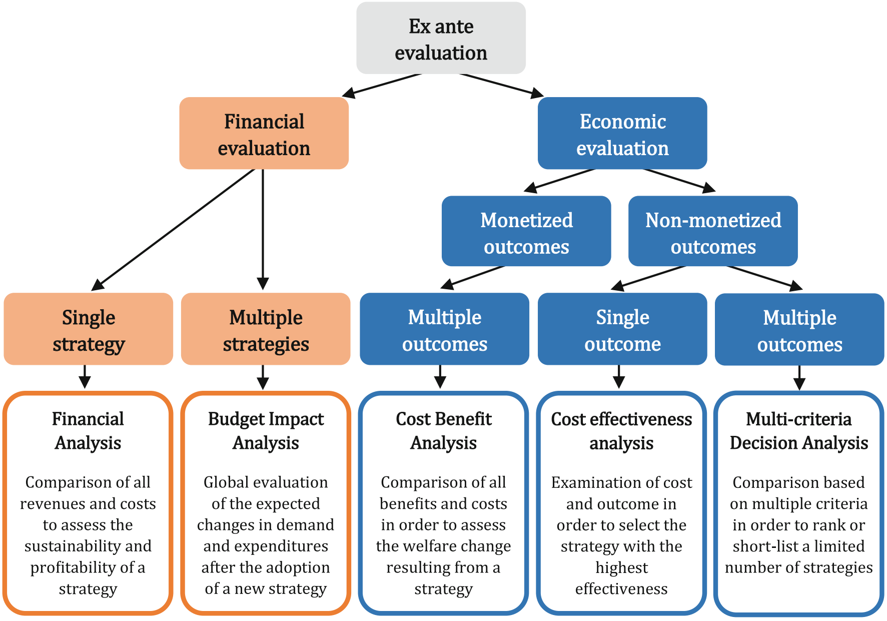

Ex ante evaluation
Financial evaluation
Economic evaluation
Monetized outcomes
Non-monetized outcomes
Single strategy
Multiple strategies
Multiple outcomes
Single outcome
Multiple outcomes
Financial Analysis
Comparison of all revenues and costs to assess the sustainability and profitability of a strategy
Budget Impact Analysis
Global evaluation of the expected changes in demand and expenditures after the adoption of a new strategy
Cost Benefit Analysis
Comparison of all benefits and costs in order to assess the welfare change resulting from a strategy
Cost effectiveness analysis
Examination of cost and outcome in order to select the strategy with the highest effectiveness
Multi-criteria Decision Analysis
Comparison based on multiple criteria in order to rank or short-list a limited number of strategies

Fig. 1.4 Ex ante evaluation techniques

Budget Impact Versus Cost Effectiveness Analysis Cost effectiveness analysis selects the set of most efficient strategies by comparing their costs and their outcomes. By definition, a strategy is said to be efficient if no other strategy or combination of strategies is as effective at a lower cost. Yet, while efficient, the adoption of a strategy not only modifies the way demand is addressed but may also divert the demand for other types of intervention. The purpose of budget impact analysis is to analyze this change and to evaluate the budget and outcome changes initiated by the introduction of the new strategy. A budget impact analysis measures the evolution of the number of users or patients through time and multiplies this number with the unit cost of the interventions. The aim is to provide the decision-maker with a better understanding of the total budget required to fund the interventions. It is usually performed in parallel to a cost effectiveness analysis. The two approaches are thus complementary.

Cost Benefit Versus Cost Effectiveness Analysis Cost benefit analysis compares strategies based on the net welfare each strategy brings to society. The approach rests on monetary measures to assess those impacts. Cost effectiveness analysis on the other hand is a tool applicable to strategies where benefits can be identified but where it is not possible or relevant to value them in monetary terms (e.g., a survival rate). The approach does not sum the cost with the benefits but, instead, relies on pairwise comparisons by valuing cost and effectiveness differences. A key feature of the approach is that only one benefit can be used as a measure of effectiveness.

For instance, quality adjusted life years (QALYs) are a frequently used measure of outcome. While cost effectiveness analysis has become a common instrument for the assessment of public health decisions, it is far from widely used in other fields of collective decisions (transport, environment, education, security) unlike cost benefit analysis.

Cost Benefit Versus Multi-criteria Decision Analysis Multi-criteria decision analysis is used whenever several outcomes have to be taken into account but yet cannot be easily expressed in monetary terms**.** For instance, a project may have major environmental impacts but it is found difficult to estimate the willingness to pay of agents to avoid ecological and health risks. In that context, it becomes impossible to incorporate these elements into a conventional cost benefit analysis. Multi-criteria decision analysis overcomes this issue by measuring those consequences on numerical scales or by including qualitative descriptions of the effects. In its simplest form, the approach aims to construct a composite indicator that encompasses all those different measurements and allows the stakeholders' opinions to be accounted for. Weights are assigned on the different dimensions by the decision-maker. Cost benefit analysis on the other hand does not need to assign weights. Using a common monetary metric, all effects are summed into a single value, the net benefit of the strategy.

1.4 Ex post Evaluation

Demonstrating that a particular intervention has induced a change in the level of effectiveness is often made difficult by the presence of confounding variables that connect with both the intervention and the outcome variable. It is important to keep in mind that there is a distinction between causation and association. Imagine for instance that we would like to measure the effect of a specific training program, (e.g., evening lectures) on academic success among students at risk of school failure. The characteristics of the students, in particular their motivation and abilities, are likely to affect their grades but also their participation in the program. It is thereby the task of the evaluator to control for those confounding factors and sources of potential bias. As shown in Fig. 1.5., one can distinguish three types of evaluation techniques in this matter: randomized controlled experiment, benchmarking analysis and quasi-experiment.

Basically speaking, a controlled experiment aims to reduce the differences among users before the intervention has taken place by comparing groups of similar characteristics. The subjects are randomly separated into one or more control groups and treatment groups, which allows the effects of the treatment to be isolated. For example, in a clinical trial, one group may receive a drug while another group may receive a placebo. The experimenter then can test whether the differences observed between the groups on average (e.g., health condition) are caused by the intervention or due to other factors. A quasi-experiment on the other hand controls for the differences among units after the intervention has taken place.

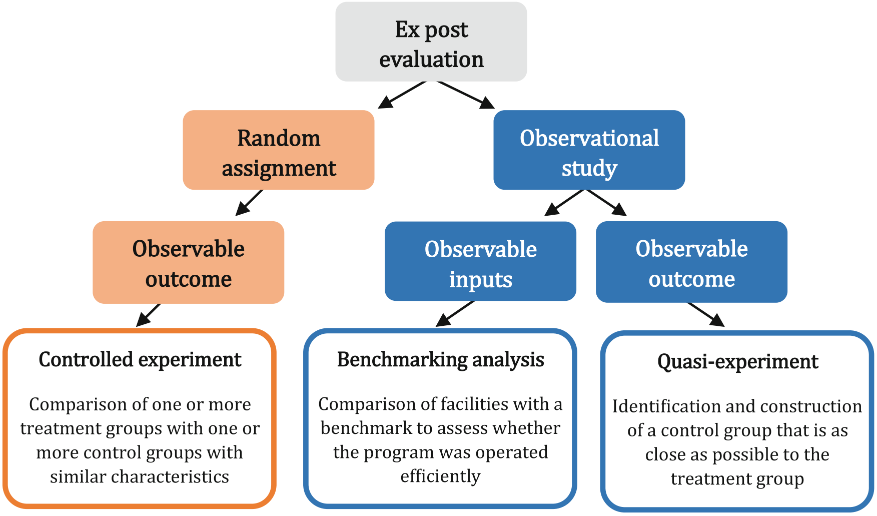

Ex post evaluation
Random assignment
Observational study
Observable outcome
Observable inputs
Observable outcome
Controlled experiment
Comparison of one or more treatment groups with one or more control groups with similar characteristics
Benchmarking analysis
Comparison of facilities with a benchmark to assess whether the program was operated efficiently
Quasi-experiment
Identification and construction of a control group that is as close as possible to the treatment group

Fig. 1.5 Ex-post evaluation techniques

It does not attempt to manipulate or influence the environment. Data are only observed and collected (observational study). The evaluator then must account for the fact that multiple factors may explain the variations observed in the variable of interest. In both types of study, descriptive and inferential statistics play a determinant role. They can be used to show evidence of a selection bias, for instance when some members of the population are inadequately represented in the sample, or when some individuals select themselves into a group.

The main goal of ex post evaluation is to answer the question of whether the outcome is the result of the intervention or of some other factors. The true challenge here is to obtain a measure of what would have happened if the intervention did not take place, the so-called counterfactual. Different evaluation techniques can be put in place to achieve this goal. As stated above, one way is through a randomized controlled experiment. Other ways include difference-in-differences, propensity score matching, regression discontinuity design, and instrumental variables. All those quasi-experimental techniques aim to prove causality by using an adequate identification strategy to approach a randomized experiment. The idea is to estimate the counterfactual by constructing a control group that is as close as possible to the treatment group.

Another important aspect to account for is whether the program has been operated in the most effectual way in terms of input combination and use. Often, for projects of magnitude, there are several facilities that operate independently in their geographical area. Examples include schools, hospitals, prisons, social centers, fire departments. It is the task of the evaluator to assess whether the provision of services meets with management standards. Yet, the facilities involved in the implementation process may face different constraints, specific demand settings and may have chosen different organizational patterns. To overcome those issues, one may rely on a benchmarking analysis to compare the cost structure of the facilities with that of a given reference, the benchmark.

Choosing which method to use mainly depends on the context of analysis. For instance, random assignment is not always possible legally, technically or ethically. Another problem with random assignment is that it can demotivate those who have been randomized out, or generate noncompliance among those who have been randomized in. In those cases, running a quasi-experiment is preferable. In other cases, the outcome in question is not easily observable and one may rely instead on a simpler comparison of outputs, and implement a benchmarking analysis. The time horizon and data availability thus also determine the choice of the method.

1.5 How to Use the Book?

The goal of the book is to provide the readers with a practical guide that covers the broad array of methods previously mentioned. The brief description of the methodology, the step by step approach, the systematic use of numerical illustrations allow to become fully operational in handling the statistics of public project evaluation.

The first part of the book is devoted to context analysis. It develops statistical tools that can be used to get a better understanding of problems and needs: Chap. 2 is about sampling methods and the construction of variables; Chap. 3 introduces the basic methods of descriptive statistics and confidence intervals estimation; Chap. 4 explains how to measure and visualize associations among variables; Chap. 5 describes the econometric approach and Chap. 6 is about the estimation of welfare changes.

The second part of the book then presents ex ante evaluation methods: Chap. 7 develops the methodology of financial analysis and details several concepts such as the interest rate, the time value of money or discounting; Chap. 8 includes a detailed description of budget impact analysis and extends the financial methodology to a multiple demand structure; Chaps. 9, 10 and 11 relate to the economic evaluation of the interventions and successively describe the methodology of cost benefit analysis, cost-effectiveness analysis, and multi-criteria decision analysis, respectively. Those economic approaches offer a way to compare alternative courses of action in terms of both their costs and their overall consequences and not on their financial flows only.

Last but not least, the third part of this book is about ex post evaluation, i.e. the assessment of the effects of a strategy after its implementation. The key issue here is to control for all those extra factors that may affect or bias the conclusion of the study. Chapter 12 introduces follow up by benchmarking. Chapter 13 explains the experimental approach. Chapter 14 details the different quasi-experimental techniques (difference-in-differences, propensity score matching, regression discontinuity design, and instrumental variables) that can be used when faced with observational data.

We have tried to make each chapter as independent of the others as possible. The book may therefore be read in any order. Readers can simply refer to the table of contents and select the method they are interested in. Moreover, each chapter contains bibliographical guidelines for readers who wish to explore a statistical tool more deeply. Note that this book assumes at least a basic knowledge of economics, mathematics and statistics. If you are unfamiliar with the concept of inferential statistics, we strongly recommend you to read the first chapters of the book.

Most of the information that is needed to understand a particular technique is contained in the book. Each chapter includes its own material, in particular numerical examples that can be easily reproduced. When possible, formulas in Excel are provided. When Excel is not suitable anymore to address specific statistical issues, we rely instead on R-CRAN, a free software environment for statistical computing and graphics. The software can be easily downloaded from internet. Codes will be provided all along the book with dedicated comments and descriptions. If you have questions about R-CRAN like how to download and install the software, or what the license terms are, please go to https://www.r-project.org/.

Bibliographical Guideline

The book provides a self-contained introduction to the statistical tools required for conducting evaluations of public programs, which are advocated by the World Bank, the European Union, the Organization for Economic Cooperation and Development, as well as many governments. Many other guides exist, most of them being provided by those institutions. We may name in particular the Magenta Book and the Green Book, both published by the HM Treasury in UK. Moreover, the reader can refer to the guidance document on monitoring and evaluation of the European Commission as well as its guide to cost benefit analysis and to the evaluation of socio-economic development. The World Bank also offers an accessible introduction to the topic of impact evaluation and its practice in development. All those guides present the general concepts of program evaluation as well as recommendations. Note that the definition of "program evaluation" used in this book is from Patton (2008, p. 39).

Bibliography

European Commission. (2013). *The resource for the evaluation of socio-economic development*.

European Commission. (2014). *Guide to cost-benefit analysis of investment projects*.

European Commission. (2015). G*uidance document on monitoring and evaluation*.

HM Treasury. (2011a). *The green book. Appraisal and evaluation in Central Government*.

HM Treasury. (2011b). *The magenta book. Guidance for evaluation*.

Patton, M. Q. (2008). *Utilization focused evaluation* (4th ed.). Saint Paul, MN: Sage.

World Bank. (2011). *Impact evaluation in practice*.

Part I

Identifying the Context of the Program

Sampling and Construction of Variables 2

2.1 A Step Not to Be Taken Lightly

Building a reliable and relevant database is a key aspect of any statistical study. Not only can misleading information create bias and mistakes, but it can also seriously affect public decisions if the study is used for guiding policy-makers. The first role of the analyst is therefore to provide a database of good quality. Dealing with this can be a real struggle, and the amount of resources (time, budget, personnel) dedicated to this activity should not be underestimated.

There are two types of sources from which the data can be gathered. On one hand, one may rely on pre-existing sources such as data on privately held companies (employee records, production records, etc.), data from government agencies (ministries, central banks, national institutes of statistics), from international institutions (World Bank, International Monetary Fund, Organization for Economic Co-operation and Development, World Health Organization) or from non-governmental organizations. When such databases are not available, or if information is insufficient or doubtful, the analyst has to rely instead on what we might call a homemade database. In that case, a survey is implemented to collect information from some or all units of a population and to compile the information into a useful summary form. The aim of this chapter is to provide a critical review and analysis of good practices for building such a database.

The primary purpose of a statistical study is to provide an accurate description of a population through the analysis of one or several variables. A variable is a characteristic to be measured for each unit of interest (e.g., individuals, households, local governments, countries). There are two types of design to collect information about those variables: census and sample survey. A census is a study that obtains data from every member of a population of interest. A sample survey is a study that focuses on a subset of a population and estimates population attributes through statistical inference. In both cases, the collected information is used to calculate indicators for the population as a whole.

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_2

Since the design of information collection may strongly affect the cost of survey administration, as well as the quality of the study, knowing whether the study should be on every member or only on a sample of the population is of high importance. In this respect, the quality of a study can be thought of in terms of two types of error: sampling and non-sampling errors. Sampling errors are inherent to all sample surveys and occur because only a share of the population is examined. Evidently, a census has no sampling error since the whole population is examined. Non-sampling errors consist of a wide variety of inaccuracies or miscalculations that are not related to the sampling process, such as coverage errors, measurement and nonresponse errors, or processing errors. A coverage error arises when there is non-concordance between the study population and the survey frame. Measurement and nonresponse errors occur when the response provided differs from the real value. Such errors may be caused by the respondent, the interviewer, the format of the questionnaire, the data collection method. Last, a processing error is an error arising from data coding, editing or imputation.

Before deciding to collect information, it is important to know whether studies on a similar topic have been implemented before. If this is to be the case, then it may be efficient to review the existing literature and methodologies. It is also critical to be clear on the objectives, especially on the type of information one needs (individuals and organizations involved, time period, geographical area), and on how the results will be used and by whom. Once the process of data collection has been initiated or a fortiori completed, it is usually extremely costly to try and add new variables that were initially overlooked.

The construction of a database includes several steps that can be summarized as follows. Section 2.2 describes how to choose a sample and its size when a census is not carried out. Section 2.3 deals with the various ways of conceiving a questionnaire through different types of questions. Section 2.4 is dedicated to the process of data collection as it details the different types of responding units and the corresponding response rates. Section 2.5 shows how to code data for subsequent statistical analysis.

2.2 Choice of Sample

First of all, it is very important to distinguish between the target population, the sampling frame, the theoretical sample, and the final sample. Figure 2.1 provides a summary description of how these concepts interact and how the sampling process may generate errors.

The target population is the population for which information is desired, it represents the scope of the survey. To identify precisely the target population, there are three main questions that should be answered: who, where and when? The analyst should specify precisely the type of units that is the main focus of the study, their geographical location and the time period of reference. For instance, if the survey aims at evaluating the impact of environmental pollution, the target population would represent those who live within the geographical area over which

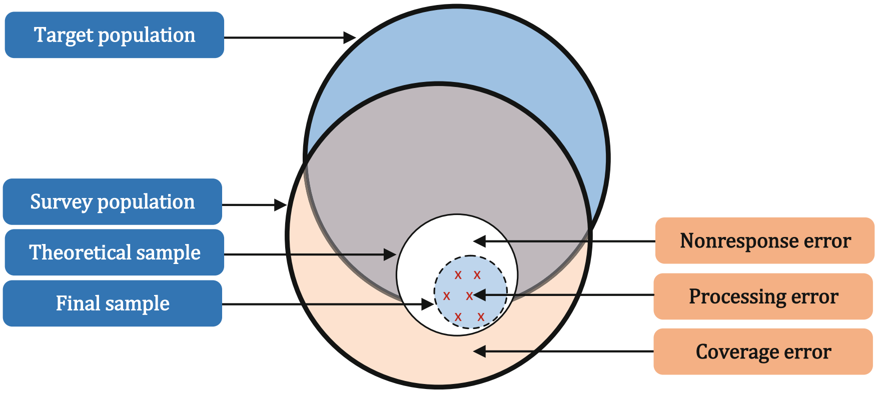


Fig. 2.1 From the target population to the final sample

the pollution is effective or those who may be using the contaminated resource. If the survey is about the provision of a local public good, then the target population may be the local residents or the taxpayers. As to a recreational site, or a better access to that site, the target population consists of all potential users. Even at this stage carefulness is required. For instance, a local public good may generate spill-over effects in neighboring jurisdictions, in which case it may be debated whether the target population should reach beyond local boundaries.

Once the target population has been identified, a sample that best represents it must be obtained. The starting point in defining an appropriate sample is to determine what is called a survey frame, which defines the population to be surveyed (also referred to as survey population, study population or target population). It is a list of all sampling units (list frame), e.g., the members of a population, which is used as a basis for sampling. A distinction is made between identification data (e.g., name, exact address, identification number) and contact data (e.g., mailing address or telephone number). Possible sampling frames include for instance a telephone directory, an electoral register, employment records, school class lists, patient files in a hospital, etc. Since the survey frame is not necessarily under the control of the evaluator, the survey population may end up being quite different from the target population (coverage errors), although ideally the two populations should coincide.

For large populations, because of the costs required for collecting data, a census is not necessarily the most efficient design. In that case, an appropriate sample must be obtained to save the time and, especially, the expense that would otherwise be required to survey the entire population. In practice, if the survey is well-designed, a sample can provide very precise estimates of population parameters. Yet, despite all the efforts made, several errors may remain, in particular nonresponse, if the survey fails to collect complete information on all units in the targeted sample. Thus, depending on survey compliance, there might be a large difference between the theoretical sample that was originally planned and the final sample. In addition to these considerations, several processing errors may finally affect the quality of the database.

A sample is only a portion of the survey population. A distinction is consequently made between the population parameter, which is the true value of the population attribute, and the sample statistic, which is an estimate of the population parameter. Since the value of the sample statistic depends on the selected sample, the approach introduces variability in the estimation results. The computation of a margin of error $\pm e$ is therefore crucial. It yields a confidence interval, i.e. a range of values, which is likely to encompass the true value of the population parameter. It is a proxy for the sampling error and an important issue with sampling design is to minimize this confidence interval.

How large should a sample be? Unfortunately, there is no unique answer to this question since the optimal size can be thought of in terms of a tradeoff between precision requirements ($\pm e$) and operational considerations such as available budget, resources and time. Yet, an indicative formula provides the minimum size of a sample. It is based on the calculation of a confidence interval for a proportion. As an illustration, assume that one wishes to estimate the portion of a population that has a specific characteristic, such as the share of males. The true population proportion is denoted π and the sample proportion is denoted p. Since π is unknown, we can only use the characteristics of the sample to compute a confidence interval. Assume for instance that we find $p = 45\%$ (i.e. 45 percent of the sample units are male) and calculate a margin of error equal to $e = 3\%$. The analyst can specify a range of values $45\,\% \pm 3\%$ in which the population parameter π is likely to belong, i.e. the confidence interval is $[42\%, 48\%]$. Statistical precision can thus be thought of as how narrow the confidence interval is.

The formula for calculating a margin of error for a proportion is:

$$e = z_{\alpha} \times \sqrt{\frac{p(1-p)}{n_0}}$$

Three main factors determine the magnitude of the confidence interval. First, the higher is the sample size n_0, the lower is the margin of error e. At first glance, one should then try to maximize the sample size. However, since the margin of error decreases with the square root of the sample size, there is a kind of diminishing returns to increasing sample size. Concurrently, the cost of survey administration is likely to increase linearly with n_0. There is consequently a balance to find between those opposing effects. Second, a sample should be as representative as possible of the population. If the population is highly heterogeneous, the possibility of drawing a non-representative sample is actually high. In contrast, if all members are identical, then the sample characteristics will perfectly match the population, whatever the selected sample is. Imagine for instance that $\pi = 90\%$, i.e. most individuals in the population are males. In that case, if the sample is randomly chosen, the likelihood of selecting a non-representative sample (e.g., only females) is low. On the contrary, if the gender attribute is equally distributed ($\pi = 50\%$), then this likelihood is high. Since the population variance $\pi(1-\pi)$ is unknown, the sample variance $p(1-p)$ will serve as a proxy for measuring the heterogeneity in the population. The higher is $p(1-p)$, the lower is the precision of the sample estimate. Third, the z_α statistic allows to compute a margin of error with a $(1-\alpha)$ confidence level, which corresponds to the probability that the confidence interval calculated from the sample encompasses the true value of the population parameter. The sampling distribution of p is approximately normally distributed if the population size is sufficiently large. The usually accepted risk is $\alpha = 5\%$ so that the confidence level is 95%. The critical value $z_{5\%} = 1.96$ is computed with a normal distribution calculator.

Let us now consider the formula for the margin of error from a different perspective. Suppose that instead of computing e, we would like to determine the sample size n_0 that achieves a given level of precision, hence keeping the margin of error at the given level e. The equation can be rewritten:

$$n_0 = z_{5\%}^2 \times \frac{p(1-p)}{e^2}$$

Table 2.1 highlights the relationship between the parameters. For instance, when the proportion p is 10% and the margin of error e is set to 5%, the required sample size is $n_0 = 138$. If we want to reach a higher precision, say $e = 1\%$, then we have to survey a substantially higher number of units: $n_0 = 3457$. Of course, the value of p is unknown before the survey has been implemented. Yet, the maximum of the sample variance $p(1-p)$ is obtained for $p = 50\%$. For that value of the proportion, and in order to achieve a level of precision $e = 1\%$, one should survey at least $n_0 = 9604$ units, and $n_0 = 384$ to achieve $e = 5\%$.

The sample size also depends on the size of the target population, denoted N hereafter. Below approximately $N = 200{,}000$, a finite population correction factor has to be used:

Table 2.1 Sample size for an estimated proportion

Proportion		Margin of error			
p (%)	1–p (%)	0.5%	1%	5%	10%
10	90	13,830	3457	138	35
20	80	24,586	6147	246	61
30	70	32,269	8067	323	81
40	60	36,879	9220	369	92
50	50	38,416	9604	384	96
60	40	36,879	9220	369	92
70	30	32,269	8067	323	81
80	20	24,586	6147	246	61
90	10	13,830	3457	138	35

$$e = z_{5\%} \times \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Solving for n yields:

$$n \times \frac{N-1}{N-n} = z_{5\%}^2 \times \frac{p(1-p)}{e^2}$$

$$n \times \frac{N-1}{N-n} = n_0,$$

$$n = \frac{n_0 N}{n_0 + N - 1}.$$

For instance, while we were previously suggesting a sample size of $n_0 = 384$ to ensure a margin of error of 5%, now, with the new formula, and if the population size is $N = 500$, we have:

$$n = \frac{384 \times 500}{384 + 500 - 1} \approx 217$$

Table 2.2 provides an overview of the problem. Those figures provide a useful rule of thumb for the analyst. For a desired level of precision e, the lower is the population size N, the lower is the number n of units to survey. Those results, however, have to be taken with caution. What matters at the end is common sense. For instance, according to Table 2.2, if $N = 1000$ the analyst should survey $n = 906$ units to ensure a margin of error of 1%. In that case, sampling would virtually be equivalent to a census, in statistical terms but also in budget and organizational terms. Moving to a less stringent 5% margin of error would provide a much more relevant and tractable number of units to survey.

In practice, most polling companies survey from 400 to 1000 units. For instance, the NBC News/Wall Street Journal conducted in October 2015 a public opinion poll

Table 2.2 Target population and sample size

Population	Margin of error			
N	0.5%	1%	5%	10%
50	50	50	44	33
100	100	99	80	49
500	494	475	217	81
1000	975	906	278	88
2000	1901	1655	322	92
5000	4424	3288	357	94
10,000	7935	4899	370	95
100,000	27,754	8763	383	96

relating to the 2016 United States presidential election (a poll is a type of sample survey dealing mainly with issues of public opinions or elections). A number of 1000 sampling units were interviewed by phone. Most community satisfaction surveys rely on similar sample sizes. For instance, in 2011, the city of Sydney, Australia, focused on a series of $n = 1000$ telephone interviews to obtain a satisfaction score related to community services and facilities. Smaller cities may instead focus on $n = 400$ units. At a national level, sample sizes reach much larger values. To illustrate it, in 2014, the American Community Survey selected a sample of about 207,000 units from an initial frame of 3.5 million addresses. According to our rule of thumb, this would yield a rather high precision, approximately $e = 0.2\%$.

The choice of sample size also depends on the expected in-scope proportion and response rate. First, it is possible that despite all efforts coverage errors exist and that a number of surveyed units do not belong to the target population. On top of these considerations, the survey may fail to reach some sampling units (refusals, noncontacts). To guarantee the desired level of precision, one needs therefore to select a sample larger than predicted by the theory, using information about the expected in-scope and response rates. More specifically, the following adjustment can be implemented:

$$\text{Adjusted sample size} = \frac{n}{\text{Expected response rate} \times \text{Expected in-scope rate}}$$

Suppose for instance that the in-scope rate estimated from similar surveys or pilot tests is 91%. Assume also that the expected response rate is 79%. When $n = 1000$, the adjusted sample size is:

$$\text{Adjusted sample size} = \frac{1000}{0.91 \times 0.79} = 1391$$

A crucial issue here is that once the expected in-scope and response rates have been defined ex ante, their values should serve as a target during the data collection process. A response rate or in-scope rate lower than the desired values will result in a sample size that does not ensure anymore the precision requirement. For instance, in the case of the American Community Survey, if we fictitiously assume an ex-post response rate of 25% and in-scope rate of 85%, which can be realistic in some cases (if not in this particular one), then the margin of error increases from e = 0.2% to 0.5%.

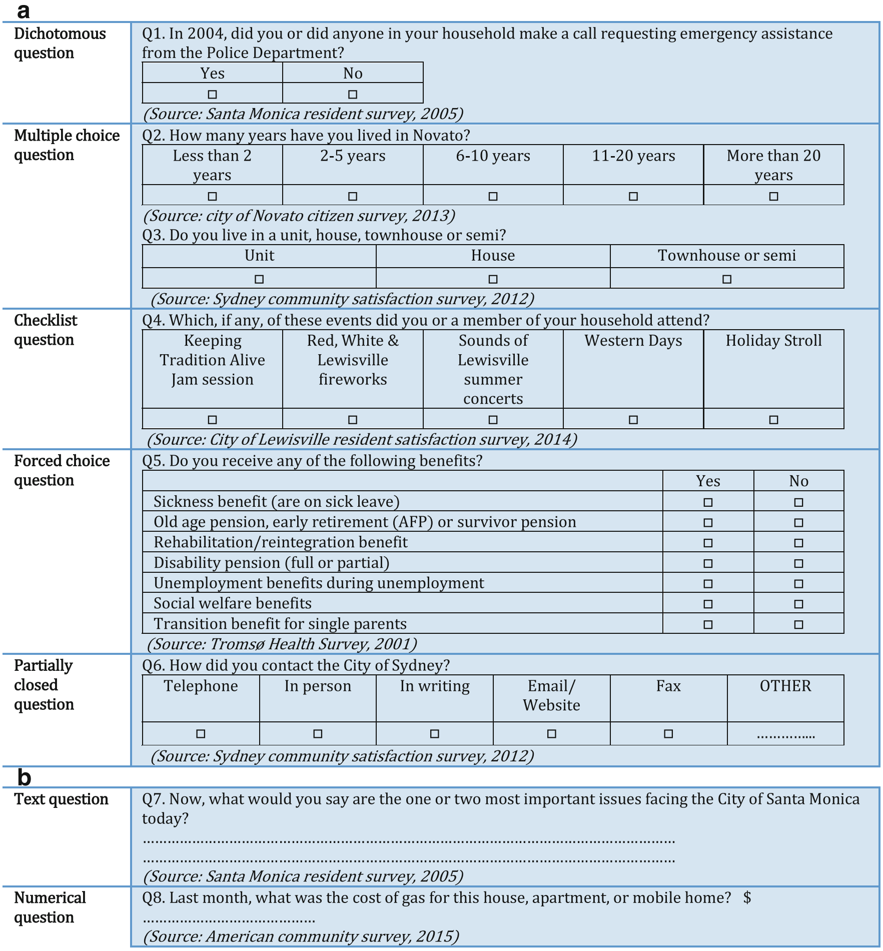
To conclude, whether one chooses a higher or lower sample size (or equivalently, a higher or lower precision) mainly depends on operational constraints such as the budget, but also the time available to conduct the entire survey and the size of the target population. First, there are direct advantages and disadvantages to using a census to study a population. On the one hand, a census provides a true measure of the population but also detailed information about sub-groups within the population, which can be useful if heterogeneity matters. On the other hand, a sample generates lower costs both in staff and monetary terms and is easier to control and monitor. Second, the time needed to collect and process the data increases with the sample size. Thus, with a sample survey of realistic size, the results are generally available in less time and can still be representative of the population. Third, the population size is also a determinant factor. If the population is small, a census is always preferable. In contrast, for large populations, accurate results can be obtained from reasonably small samples. In any case, the next step now consists in conceiving the questionnaire that will be proposed to respondents.

2.3 Conception of the Questionnaire

A questionnaire is a set of questions designed to elicit information upon a subject, or sequence of subjects, from a respondent. Given its impact on data quality, the questionnaire design plays a central role. The purpose of a survey is to obtain sincere responses from the respondent. One main principle applies in this matter: one should start on the basis that most people do not want to spend time on a survey, and if they do, it could be that they actually are not satisfied with the policy under evaluation, which may be non-representative of the population as a whole. Nonresponses should be minimized as much as possible. This can be done by explaining why the survey is carried out, by keeping it quick and by telling the respondents that the results will be communicated once finalized. Those three rules are even truer nowadays since people are frequently required to participate in surveys in many fields.

An important aspect of questionnaire design is the type of response formats. There are two categories of questions: open-ended versus close-ended. Close-ended questions request the respondent to choose one or several responses among a predetermined set of options. While they limit the range of respondents' answers on the one hand, they require less time and effort for both the interviewer and the participant on the other hand. In contrast, open-ended questions do not give respondents options to choose from. Thereby, they allow them to use their own words and to include more information, including their feelings and understanding of the problem.

Examples of close-ended and open-ended questions are provided in Fig. 2.2. Dichotomous questions (also referred to as two-choice questions) are the simplest version of a close-ended question. They propose only two alternatives to the respondent. Multiple choice questions propose strictly more than two alternatives and ask the respondent to select one single response from the list of possible choices. Checklist questions (or check-all questions) allow a respondent to choose more than one of the alternatives provided. Forced choice questions are similar to checklist questions, although the respondent is required to provide an answer (e.g., yes–no) for every response option individually. Partially closed questions provide an alternative "Other, please specify", followed by an appropriately sized answer box. This type of question is useful when it is difficult to list all possible alternatives or when responses cannot be fully anticipated. Last, open-ended questions can be of two forms, either text or numerical.



a

Dichotomous question

Q1. In 2004, did you or did anyone in your household make a call requesting emergency assistance from the Police Department?

Yes	No
□	□

(Source: Santa Monica resident survey, 2005)

Multiple choice question

Q2. How many years have you lived in Novato?

Less than 2 years	2-5 years	6-10 years	11-20 years	More than 20 years
□	□	□	□	□

(Source: city of Novato citizen survey, 2013)

Q3. Do you live in a unit, house, townhouse or semi?

Unit	House	Townhouse or semi
□	□	□

(Source: Sydney community satisfaction survey, 2012)

Checklist question

Q4. Which, if any, of these events did you or a member of your household attend?

Keeping Tradition Alive Jam session	Red, White & Lewisville fireworks	Sounds of Lewisville summer concerts	Western Days	Holiday Stroll
□	□	□	□	□

(Source: City of Lewisville resident satisfaction survey, 2014)

Forced choice question

Q5. Do you receive any of the following benefits?

	Yes	No
Sickness benefit (are on sick leave)	□	□
Old age pension, early retirement (AFP) or survivor pension	□	□
Rehabilitation/reintegration benefit	□	□
Disability pension (full or partial)	□	□
Unemployment benefits during unemployment	□	□
Social welfare benefits	□	□
Transition benefit for single parents	□	□

(Source: Tromsø Health Survey, 2001)

Partially closed question

Q6. How did you contact the City of Sydney?

Telephone	In person	In writing	Email/ Website	Fax	OTHER
□	□	□	□	□	....

(Source: Sydney community satisfaction survey, 2012)

b

Text question

Q7. Now, what would you say are the one or two most important issues facing the City of Santa Monica today?

....

....

(Source: Santa Monica resident survey, 2005)

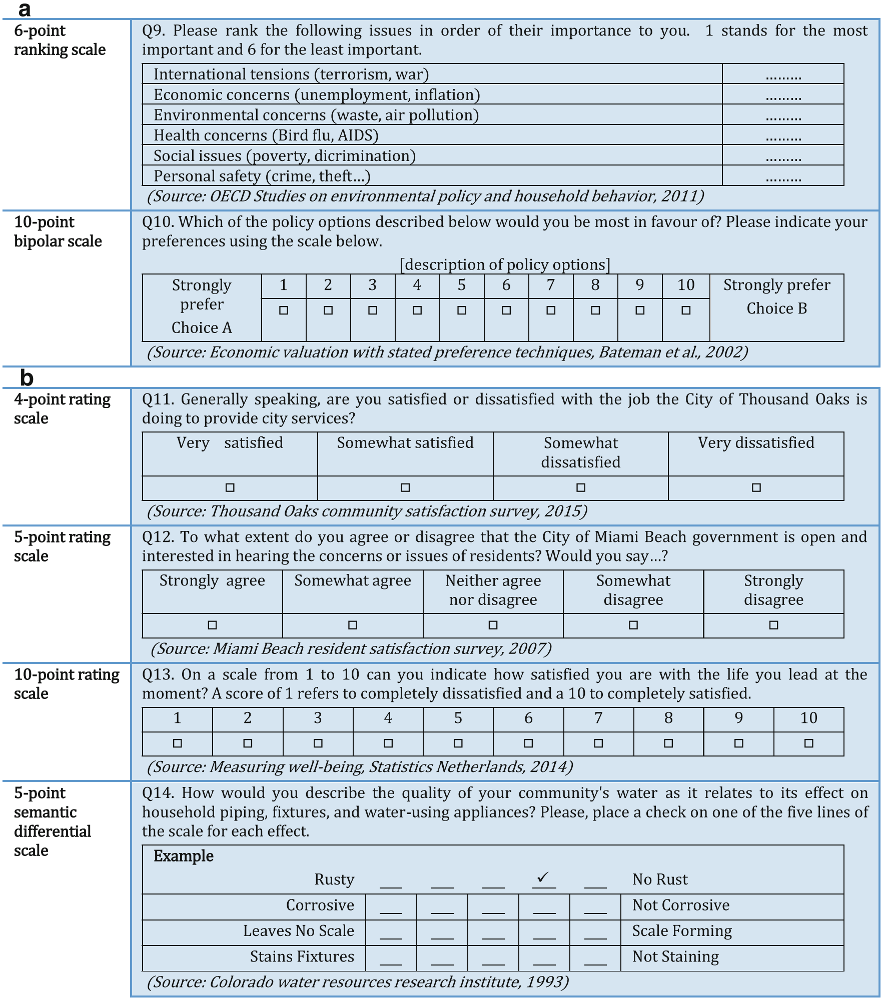
Numerical question

Q8. Last month, what was the cost of gas for this house, apartment, or mobile home? $ ....

(Source: American community survey, 2015)

Fig. 2.2 Close-ended and open-ended questions. (**a**) close-ended questions and (**b**) open-ended questions

Another widely used format is the scale question, which asks the respondent to grade the response on a given range of options (see Fig. 2.3). These questions can be grouped into two subcategories: ranking questions and rating questions. Ranking questions offer several options and request the respondent to rank them from most important to least important on a ranking scale (where 1 is the most important, 2 is the second most important, and so on) or a bipolar scale (where respondents have to rate the intensity of their preference). Respondents thus compare each item to each other. A ranking scale has the inconvenient to force the respondent to make one item worse or better than another, when they actually could be indifferent between them. They also require a significant cognitive effort. Pairwise comparisons overcome these problems through the use of bipolar scales. When the number of



a

6-point ranking scale

Q9. Please rank the following issues in order of their importance to you. 1 stands for the most important and 6 for the least important.

International tensions (terrorism, war)	....
Economic concerns (unemployment, inflation)	....
Environmental concerns (waste, air pollution)	....
Health concerns (Bird flu, AIDS)	....
Social issues (poverty, dicrimination)	....
Personal safety (crime, theft...)	....

(Source: OECD Studies on environmental policy and household behavior, 2011)

10-point bipolar scale

Q10. Which of the policy options described below would you be most in favour of? Please indicate your preferences using the scale below.

[description of policy options]

Strongly prefer Choice A	1	2	3	4	5	6	7	8	9	10	Strongly prefer Choice B
	□	□	□	□	□	□	□	□	□	□	

(Source: Economic valuation with stated preference techniques, Bateman et al., 2002)

b

4-point rating scale

Q11. Generally speaking, are you satisfied or dissatisfied with the job the City of Thousand Oaks is doing to provide city services?

Very satisfied	Somewhat satisfied	Somewhat dissatisfied	Very dissatisfied
□	□	□	□

(Source: Thousand Oaks community satisfaction survey, 2015)

5-point rating scale

Q12. To what extent do you agree or disagree that the City of Miami Beach government is open and interested in hearing the concerns or issues of residents? Would you say...?

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
□	□	□	□	□

(Source: Miami Beach resident satisfaction survey, 2007)

10-point rating scale

Q13. On a scale from 1 to 10 can you indicate how satisfied you are with the life you lead at the moment? A score of 1 refers to completely dissatisfied and a 10 to completely satisfied.

1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□

(Source: Measuring well-being, Statistics Netherlands, 2014)

5-point semantic differential scale

Q14. How would you describe the quality of your community's water as it relates to its effect on household piping, fixtures, and water-using appliances? Please, place a check on one of the five lines of the scale for each effect.

Example

Rusty	___	___	___	✓	___	No Rust
Corrosive	___	___	___	___	___	Not Corrosive
Leaves No Scale	___	___	___	___	___	Scale Forming
Stains Fixtures	___	___	___	___	___	Not Staining

(Source: Colorado water resources research institute, 1993)

Fig. 2.3 Scale questions. (**a**) ranking questions and (**b**) rating questions

alternatives is large, it is possible to ask the respondents to choose one single item through a multiple choice question, for instance:

Let's assume for a moment that the Santa Monica Police Department hired another officer and assigned that officer to your neighborhood. Which of the following five items should be the single highest priority for a new police officer assigned to your neighborhood?

1. *Working with local kids to prevent gangs and youth crime,*
2. *Patrolling on foot in your local neighborhood,*
3. *Working with local residents and neighborhood groups to help prevent crime,*
4. *Patrolling in police cars in your local neighborhood,*

5. *Patrolling near the schools in your neighborhood.*

(Source: Santa Monica Resident Survey, 2005)

The other category of scale question is the rating question, which requires the respondents to rate their answer, independently of other options, on a rating scale (also referred to as a Likert scale). Usually, this type of scale contains equal numbers of positive and negative positions, which creates a less biased measurement. Often, it is preferable not to propose a neutral position in the middle, as otherwise the respondents could choose this category to save time or hide their preference. Last, semantic differential scales ask the respondents to choose between two opposite positions, with bipolar adjectives at each end. Such a scale allows to include several dimensions in a single question, but also demands higher cognitive effort from the respondent.

The sequencing of questions is as important as the questions themselves. It should be designed to encourage the respondent to complete and maintain interest in the questionnaire. It is usually advised to follow the following sequence. First, an introductory section should give the title of the survey and introduce the authority under which the survey is conducted, its purpose, and the general contents of the questionnaire. What is included is crucial in securing the participation of respondents. This section usually contains general instructions for the interviewer and respondents, provides reassurances about confidentiality and states the expected length of the survey. It requests the respondent's cooperation and stresses the importance of his/her participation. It explains how the survey data will be used and includes contact information. Finally, this section may include the signature of the person in charge of the authority under which the survey is conducted.

The sequence of questions should be as logical as possible. For instance, the first questions should be easy to answer. Sensitive questions should not be placed at the beginning of the questionnaire, but introduced at a point where the respondent is more likely to feel comfortable answering them. The first questions are generally about things respondents do or have experienced, the so-called behavior questions. Knowledge questions can be included to better assess whether the respondent knows the topic. Those types of question are then followed by opinion questions, which ask what the respondents think about a specific item. Motive questions require the respondents to evaluate why they behave in a particular manner. Personal and confidential questions as well as questions about socio-economic status are located at the end of the questionnaire. One should not forget to include an open-ended question at the end, so that the respondents have the possibility to express themselves, as well as an acknowledgement to thank the respondent.

Between each part of the questionnaire it is important to use transitional statements to explain that a new topic will be examined. In addition, several rules have to be obeyed with respect to question writing. Spelling, style and grammar should be carefully checked, otherwise it would devalue the organization that implements or orders the study. It is also recommended to minimize the length of the questionnaire. The greater is the number of questions, the less time the respondents spend, on average, answering each question. There is a point at which survey completion rates start to drop off, usually after 5–8 min (i.e. 15–20 questions—one web page—one sheet of paper). Do not ask open-ended questions unless necessary. Use the same scales over the questionnaire. Regroup similar questions as follows:

Now, please rate each of the following possible problems in Santa Monica on a scale of 1 to 5. Use a 1 if you feel the problem in NOT serious at all, and a 5 if you feel it is a VERY serious problem in Santa Monica:

1. *Traffic congestion* □ 1 □ 2 □ 3 □ 4 □ 5
2. *The affordability of housing for low income families and seniors* □ 1 □ 2 □ 3 □ 4 □ 5
3. *Gang violence* □ 1 □ 2 □ 3 □ 4 □ 5
4. *The number of homeless people in the city* □ 1 □ 2 □ 3 □ 4 □ 5
5. *Lack of parking* □ 1 □ 2 □ 3 □ 4 □ 5

(Source: 2005 Santa Monica resident survey)

Another point is to define and choose carefully the time horizon. For instance, depending on the context, the question "How many times per year do you take the bus" may not be enough specific and "per year" should be replaced by "per week". Avoid using terms such as "regularly" or "often", which do not convey the same meaning for all respondents. Instead, an appropriate time horizon should be offered, e.g.:

How often do you suffer from headaches?

1. *Rarely or never*
2. *Once or more a month*
3. *Once or more a week*
4. *Daily*

(Source: 2001 Tromsø Health Survey)

Perhaps it is obvious, but simple and clear questions are better than long questions, with complex words, abbreviations, acronyms, or sentences that are difficult immediately to understand. Define the technical terms if necessary. Do not ask negatively worded questions like "Should the City not invest in energy efficiency for municipal buildings?" Avoid double-barreled questions that ask two or more questions in a row. Do not use confusing terms or vague concepts. For instance, when asked "how much do you pay per year in taxes?" respondents may not know what is meant by "taxes", whether it is income taxes, property taxes, national or local taxes. Finally, there is always the risk of a framing effect when phrasing a question. For instance, questions like "Don't you think that the city needs to cut the grass around our schools?" may induce yea-saying bias. Prefer instead a question like "to what extent do you agree or disagree that…". Such a question should also specify the cost and/or additional increase in taxes. Check also that the response options do not force people to answer in a way that they do not wish to. Questions must propose all the relevant options. One should open the question with an item "other", if one is not sure about the exhaustiveness of the options. Last, each item should be totally independent from the others.

Not only respondents but also public decision-makers or experts in the field should be consulted to provide insight into the type of information that is required. Meetings and focus-groups can help identify issues and concerns that are important. Whether it is a new questionnaire or a set of questions that have been used before, it is also essential to test it before the survey is implemented. This stage represents an opportunity to check whether the interviewers and respondents understand the questions, whether the survey retains the attention of respondents and whether it is sufficiently short. In a first step, an informal pilot test can be implemented using a number of colleagues. While they may be familiar with the questionnaire and will tend to answer the questions more quickly, they will also be more likely to pick up errors than the respondents themselves. The next stage for the questionnaire writer is to implement a larger scale pilot test on a subsample of the target population, but also on specific subgroups of the population that may have difficulties with particular questions. A pilot test of 30–100 cases is usually sufficient to discover the major problems in a questionnaire. The questionnaire should be administered in the same manner as planned for the main survey. A minimum of 30 observations also yields the possibility for the questionnaire writer to implement a preliminary statistical analysis, in order to assess whether the survey is suitable to achieve the objectives of the study.

2.4 Data Collection

Data collection is any process whose purpose is to acquire information. When it has been decided that a census is not preferable over a sample survey, the first stage consists in selecting a subset of units from the population. There are two kinds of methods in this respect: non-probability and probability sampling. Whether one chooses the first or the second mainly depends on the availability of a survey frame, i.e. a list of each unit in the population. If a survey frame is not available, then one can implement a probability sampling, i.e. select randomly a sample from that list. By definition, probability sampling is a procedure in which each unit of the population has a fixed probability of being selected from the sample. Reliable inferences can then be made about the population. If a survey frame is not available, then one has to rely instead on subjective and personal judgment in the process of sample selection, i.e. on non-probability sampling. The procedure is usually simpler and cheaper to implement, but also more likely to be subject to bias. Hence, whether one chose an approach or another depends on the availability of a survey frame and how one values the sampling error against the cost of survey administration.

Common methods of probability sampling are simple random sampling, systematic sampling, stratified sampling and cluster sampling. We shall consider them successively. With simple random sampling, each unit is chosen randomly using a random number table or a computer-generated random number. Such sampling is done without replacement, i.e. the procedure should avoid choosing any unit more than once. Systematic sampling is a method that selects units at regular intervals. In a first step, all units in the survey frame are numbered from 1 to N. Second, a periodic interval $k = N/n$ is calculated, where n represents the desired sample size. Third, a starting point is randomly selected between 1 and k. Fourth, every k th unit after the random starting point is selected. For instance, assume that the survey frame contains $N = 10{,}000$ units and that we would like to sample $n = 400$ units. The sampling interval is $k = N/n = 25$. Then a random number between 1 and 25, say 12, is selected. The units that are selected are 12, $12+25 = 37$, $37+25 = 62$, etc. Stratified sampling is a method by far superior to simple random and systematic sampling because it may significantly improve sampling precision and reduce the costs of the survey. It is used when the survey frame can be divided into non-overlapping subgroups, called strata, according to some variable whose information is available ex ante (e.g., males/females, age categories, income categories). The approach consists in drawing a separate random sample from each stratum and then to combine the results. Specifically, the population N is divided into m groups with N_i units in group i, $i = 1, \ldots, m$. If the desired sample size is n and for a proportional (N_i/N) allocation of units between groups, one should then survey nN_i/N units in each group i. Systematic or simple random sampling is then used to select a sufficient number of units from each stratum. Finally, cluster sampling randomly selects subgroups of the population. In contrast with stratified sampling, the subgroups are not based on the population attributes, but rather on independent subdivisions, or clusters, such as geographical areas, districts, factories, schools. Clusters i, $i = 1, \ldots, M$ of size N_i must be mutually exclusive and together they must encompass the entire population: $\sum_{i=1}^{M} N_i = N$. The first step amounts to drawing randomly m clusters amongst the M. Then two possibilities arise. Either one surveys all units in each selected cluster, in which case the method is referred to as “one-stage cluster sampling”, or one selects a random sample from each cluster, which is the “two-stage cluster sampling”. One advantage of the procedure is that it may significantly reduce the cost of collection for instance if personal interviews are conducted and the geographical zones are spread out. One difficulty, however, is that the selected clusters may be non-representative of the population.

Methods of non-probability sampling encompass convenience sampling, judgment sampling, volunteer sampling, quota sampling, and snowball sampling. Convenience sampling, also referred to as haphazard sampling, is the most common approach. As can be deduced from the name, it consists in selecting a sample because it is convenient to do so. Typical examples include surveying people in a street, at a subway stop, at a crowded place. The approach is based on the assumption that the population is equally distributed from one geographical zone to the other. If not, then some bias may occur. Judgment sampling selects the sample based on what is thought to be a representative sample. For instance, one may decide to draw the entire sample from one “typical” city or “representative” street. The approach may results in several biases, and is generally used for exploratory studies only. Volunteer sampling selects the respondent on the basis of their willingness to participate voluntarily in the survey. Here again, the approach is subject to many bias. In particular, self-selection may produce a sample of highly motivated (pro or against the project) individuals and neglect average or less contrasting views. It is however often used when one needs to survey people with a particular disease or health condition. Quota sampling is usually said to be the non-probability equivalent of stratified sampling. In both cases, one has to identify representative strata that characterize the population. Information about the true population attributes (available from other sources such as a national census) can be used to guarantee that each subgroup is proportionally represented. Then convenience, volunteer, or judgment sampling is used to select the required number of units from each stratum. The procedure may save a lot of time as one would typically stop to survey people with a particular characteristic once the quota has been reached. For instance, assume one would like to survey $n = 400$ units. If we have an equal share of males and females in the population, one should survey only 200 males and 200 females. Last, snowball sampling is recommended when one needs to survey people with a particular but not frequent characteristic. The approach identifies initial respondents who are then used to refer on to other individuals. Again, it may generate several biases. It is generally used when one wants to survey hard-to-reach units at a minimum cost, such as the deprived, the socially stigmatized, or the users of a specific public service.

Once the sampling procedure has been selected, one has to start the collection of data. The basic methods are self-enumeration, telephone interview, and personal interview. The characteristics of the target population and whether a frame is easily available strongly influence the choice of the method, which can be paper or computer based. Self-enumeration requires the respondents to answer the questionnaire without the assistance of an interviewer. This method of data collection is easy to administer and is typically suited to large samples or when some questions are highly personal or sensitive and easier to complete in private. Respondents should be sufficiently motivated and educated, so that they do not skip or misinterpret information. The response rate can be very low, and one may have to contact several times the respondents to remind them to complete the questionnaire. Personal interview requires the respondents to answer the questionnaire with the assistance of an interviewer, at home, at work, or at a public place. The method yields high response rates but it can however be expensive and thereby more suited to smaller sample sizes. Another issue is that the interviewers may have to reschedule the interview until the respondent is present or has time. Last, telephone interviews offer good response rates at reasonable costs since the interviewers do not need to travel, and the interview can be rescheduled more easily than with personal interviews. It is also easier to control the quality of the interviewing process if it is recorded.

The type of questions may strongly influence the choice of the collection method. If complex questions are asked, then personal or telephone interviews are preferable. In contrast, if questions concern highly personal or sensitive issues, self-enumeration is preferable. The nature of the sample units is also important. For instance, if people need assistance (e.g., children or distressed people), personal interviews are more relevant. For example, in the case of child's health condition, the sample unit can be the child's family. Within this sample unit, one may distinguish between the unit of reference (the child who provides the information) and the reporting unit (one of the parents carrying out the information).

When personal or telephone interviews are chosen as methods for data collection, it is important to prepare the interviewers. They should be informed that the questionnaire has been carefully prepared to minimize potential biases, that they should not improvise, nor influence the respondents. Every question should be asked, in the order presented, exactly as worded. They should be provided with a manual that contains guidelines. These guidelines should also contain answers to the most common questions that respondents may ask, as displayed in Table 2.3. Interviewers must be honest about the length of the interview. Questions that are misunderstood or misinterpreted should be repeated. Personal interviewers should have official badges or documents in case a respondent ask them to prove they are a legitimate representative of the public sector. Last, if a person still refuses to answer the questionnaire, it should be recorded as "refusal".

It is important to assess the performance of data collection during the survey process itself. In this matter, many rates can be computed. Figure 2.4 provides an

Table 2.3 Examples of questions and answers during interviews

Usual question of the respondent	Standard answer by the interviewer
Why did you pick me?	*By selecting a few people like you, we are able to reduce the costs associated with collecting information because we do not have to get responses from everybody. On average, the data collected will be representative of the population because respondents have been selected randomly.*
Who is going to see my data?	*All information collected is highly confidential and will be seen only by the survey staff. Your answers will be used only for the production of anonymous statistics.*
Why should I participate? How will you use my answers?	*The purpose of this survey is to find out your views on ________. Your input in this study will provide useful information and help improving public services.*
I do not have the time right now	*The questionnaire consists of ____ short questions and will not take more than ____ minutes of your time. Your responses are very important. If you are very busy now, please tell me when I can reach you again.*
I do not see how I can help you; I really don't know the topic.	*We are interested in your opinions and experiences, not in what information you may or may not have. In a study of this type, there are no right or wrong answers to questions.*
Who is behind this?	*This study is supervised by ______. The purpose is to collect information that will be helpful in improving public services.*

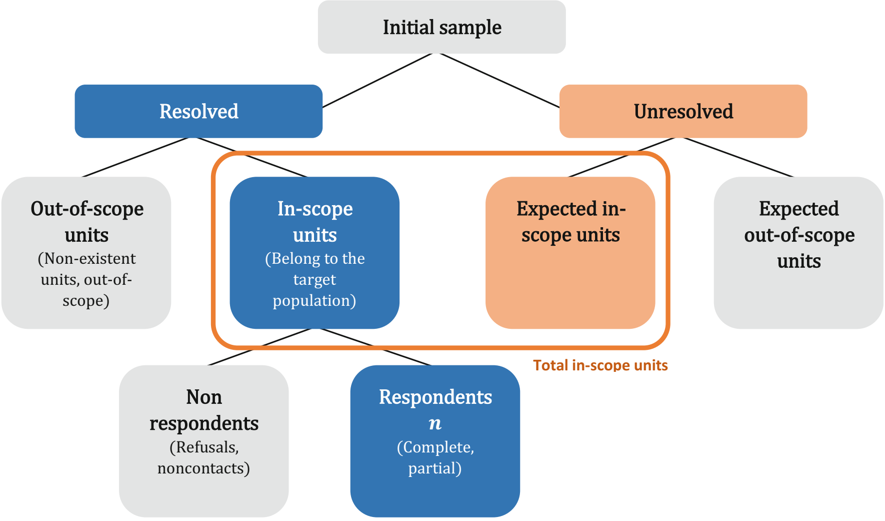


Fig. 2.4 From the initial sample to the responding units

illustration of these concepts. A first rate is based on the proportion of resolved records:

$$\text{Resolved rate} = \frac{\text{Number of resolved units}}{\text{Initial sample}}$$

This rate is defined as the ratio of the number of resolved units to the total number of sampling units. A unit is categorized as resolved if it has a determinate status, i.e. if the unit is either in-scope (complete, partial, refusal, noncontact) or out-of-scope.

A crucial issue is that some units may not belong to the target population so that they are out-of-scope. The following indicator estimates the extent of the phenomenon:

$$\text{In-scope rate} = \frac{\text{Number of in-scope units}}{\text{Number of resolved units}}$$

Using this proportion, it is also possible to approximate the expected number of in-scope units among the resolved and unresolved units:

$$\text{Expected number of in-scope units} = \text{In-scope rate} \times \text{Initial sample}$$

The assumption underlying this expectation is that the in-scope rate can be extrapolated to the whole sample.

Another indicator of interest is the response rate, namely the number of respondents (either complete or partial response) divided by the total number of sample units that are in-scope (resolved and unresolved) for the survey. Since the latter is unknown during the collection process, the previous formula is used for the denominator:

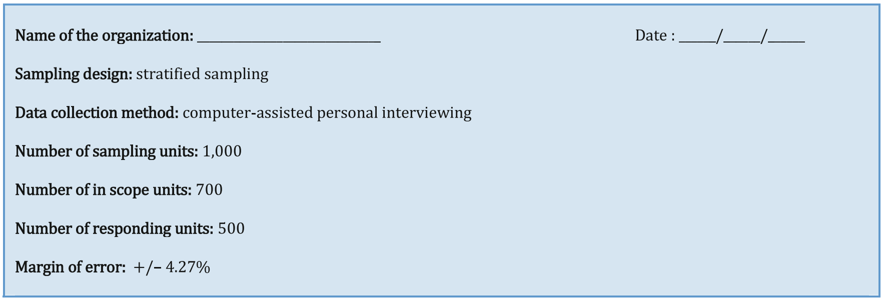
$$\text{Response rate} = \frac{\text{number of responding units}}{\text{expected number of in-scope units}}$$

Once the data has been collected, it is common to provide the following information at the beginning of a survey study: (1) the sampling design and data collection method, (2) the number of sampling units, (3) the number of in-scope units, (4) the number of responding units, and (5) the margin of error, as illustrated in Fig. 2.5.

In Fig. 2.5, a sample of 1000 units has been gathered via stratified sampling and computer-assisted personal interviewing. Assume that after one week of data collection, we have 600 resolved units among which 300 units are in-scope. This yields a resolved rate of 600/1000 = 60% and an in-scope rate equal to 300/600 = 50%. The expected total number of in-scope units is thus 1000 × 50% = 500. Suppose now that among the 300 units that are in scope, 200 units responded to the survey (either complete or partial response). Then the response rate is 200/500 = 40%. Now imagine that survey completion occurs after 3 weeks. This means that one finally gets 1000 resolved units. Among these units, suppose that 700 units are in-scope and that 500 units responded to the survey. If the target population size is $N = 10{,}000$, the margin of error can be obtained using the formula described in Sect. 2.2:

$$\mathrm{e} = 1.96 \times \sqrt{\frac{50\%(1-50\%)}{500}} \times \sqrt{\frac{10000-500}{10000-1}}$$

This yields a margin of error of approximately 4.27%.



Name of the organization: ______________________ Date : _____/_____/_____

Sampling design: stratified sampling

Data collection method: computer-assisted personal interviewing

Number of sampling units: 1,000

Number of in scope units: 700

Number of responding units: 500

Margin of error: +/− 4.27%

Fig. 2.5 Typical header for a survey study

2.5 Coding of Variables

Coding is the process of converting textual information into numbers or other symbols that can be counted and tabulated. This step is essential as it determines the final variables that will be used for subsequent analyses. To better implement it, one should understand what a database is. In statistics, it is a computer file (e.g., Excel file) made of rows i and columns j, where rows stand for the responding units, and columns for the variables. This framework is illustrated in Table 2.4 where each x_{ij} represents the value assigned by respondent i to variable j.

In this section, we propose to explain how a database is coded using the questions from Figs. 2.2 and 2.3. Table 2.5 shows what the final database looks like for a selected set of questions. For closed questions, codes are generally established before the survey takes place. The categories may be split into one or several variables depending on the nature of the questions.

For dichotomous questions, such as question Q1 (*"In 2004, did you or did anyone in your household make a call requesting emergency assistance from the Police Department?"*), the coding involves the creation of one single variable:

$$Q1 = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}$$

Since their values belong to the set $\{0, 1\}$, such variables are also called binary variables.

For multiple choice questions, two cases arise depending on whether there is a clear ordering of the variables. First, when the options of answer can be ordered, one can build a unique variable using a scale relevant to the investigated topic. For instance, question Q2 (*"How many years have you lived in Novato?"*) can be coded as:

Table 2.4 A usual database format

Responding unit/individual	Variable 1	Variable 2	...	Variable j	...
1	x_{11}	x_{12}	...	x_{1j}	...
2	x_{21}	x_{22}	...	x_{2j}	...
3	x_{31}	x_{32}	...	x_{3j}	...
...	...	...	...	...	...
i	x_{i1}	x_{i2}		x_{ij}	...
...	...	...	...	...	...
$n-2$	$x_{n-2,1}$	$x_{n-2,2}$	...	$x_{n-2,j}$	...
$n-1$	$x_{n-1,1}$	$x_{n-1,2}$	...	$x_{n-1,j}$	...
n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,j}$	...

Table 2.5 Examples of coding

(a) Questions Q1–Q7

			Q3	*Q3*	*Q4*	*Q4*	*Q6*	*Q7*	*Q7*
Ind	*Q1*	*Q2*	*unit*	*house*	*jam*	*fireworks*	*centre*	*school*	*unemp*
1	1	4	0	1	0	0	0	0	0
2	0	3	0	0	0	0	0	0	1
3	1	1	1	0	1	0	0	0	0
4	1	3	1	0	1	1	1	1	1
. . .	0	5	0	0	1	1	0	1	0

(b) Questions Q8–Q14

		Q9	*Q9*			*Q14*
Ind	*Q8*	*International tensions*	*Economic concerns*	*Q11*	*Q12*	*Corrosive*
1	30	5	1	4	2	5
2	63	4	3	1	3	4
3	140	3	4	3	5	1
4	37	6	5	2	4	3
. . .	0	2	3	3	1	2

$$Q2 = \begin{cases} 1 \text{ if}''Less\ than\ 2\ years'' \\ 2 \text{ if}''2-5\ years'' \\ 3 \text{ if}''6-10\ years'' \\ 4 \text{ if}''11-20\ years'' \\ 5 \text{ if}''More\ than\ 20\ years'' \end{cases}$$

Second, when there is no intrinsic ordering to the options, one has to split them into separate variables. This is the case for instance with question Q3 *(“Do you live in a unit, house, townhouse or semi?”)*:

$$Q3|unit = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, Q3|house = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, \ldots$$

In a similar manner, checklist questions, whether they are forced or not, imply a transformation of each category into one specific variable. For instance, with question Q4 (“Which, if any, of these events did you or a member of your household attend?”), we have:

$$Q4|jam = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, Q4|fireworks = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, \ldots$$

Note that forced choice questions like Q5 (“Do you receive any of the following benefits?”) can also be treated as a series of dichotomous questions, with for instance:

$$Q5|sickness\ benefit = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}$$

For text and partially closed questions, coding is more difficult as it requires interpretation and personal judgment to recode them into close-ended questions. The survey staff has to detect whether some items appear frequently. If it seems to be the case, then different categories should be created to take into account the heterogeneity in the responses. Afterwards, the coding can be done either manually or by computer. For instance, to question Q6 (*"How did you contact the City of Sydney?"*), the option "*other*" may have been selected frequently by people who wrote down "*response centre*", which does not belong to the initial set of categories. If the number of occurrences is large enough, this item should be included as a new item among the categories of question Q6:

$$Q6|response\ centre = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}$$

Similarly, to question Q7 ("Now, what would you say are the one or two most important issues facing the City of Santa Monica today?"), one has to detect first which items have been frequently raised (e.g., quality of schools, unemployment rate, environmental concerns, crime rate), and create a category for each of them:

$$Q7|schools = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, \ Q7|unemployment = \begin{cases} 1 \text{ if}''yes'' \\ 0 \text{ if}''no'' \end{cases}, \ldots$$

Coding is much simpler when it comes to numerical questions. We may directly use the question as it is. To illustrate, for question Q8 ("Last month, what was the cost of gas for this house, apartment, or mobile home?"), we have:

$$Q8 = declared\ cost\ of\ gas$$

For ranking order questions, the data set should include a column for each item being ranked. For instance, for question Q9 from Fig. 2.3 *("Please rank the following issues in order of their importance to you.")* we have:

$$Q9|international\ tensions = score\ obtained$$

$$Q9|Economic\ concerns = score\ obtained$$

$$\ldots$$

For any given respondent, each ranked item has a unique value, and once an item has reached a score, that score cannot be employed anymore. Notice also that this type of question may return different results depending on the completeness and relevance of the list of items being ranked. Thus, these scores should be analyzed with caution.

Instead of using a ranking scale, one may prefer to use a bipolar scale or a rating scale. With a bipolar scale, and any rating question, it is common to code the values using 1, 2, 3, etc. For instance, let us consider question Q10 (*"Which of the policy options described below would you be most in favour of?"*). The coding has been implicitly made since the respondent must choose a value between 1 and 10. For question Q11 (*"Generally speaking, are you satisfied or dissatisfied with the job the City of Thousand Oaks is doing to provide city services?"*), the coding is similar, the only difference being that a 4-point rating scale should be used:

$$Q11 = \begin{cases} 4 \text{ if} {''Very\ satisfied''} \\ 3 \text{ if} {''Somewhat\ satisfied''} \\ 2 \text{ if} {''Somewhat\ dissatisfied''} \\ 1 \text{ if} {''Very\ dissatisfied''} \end{cases}$$

For question Q12 ("To what extent do you agree or disagree that the City of Miami Beach government is open and interested in hearing the concerns or issues of residents?"), one should use a 5-point ranking scale, where 3 represents the neutral position:

$$\mathrm{Q12} = \begin{cases} 5 \text{ if} {''Strongly\ agree''} \\ 4 \text{ if} {''Somewhat\ agree''} \\ 3 \text{ if} {''Neutral''} \\ 2 \text{ if} {''Somewhat\ disagree''} \\ 1 \text{ if} {''Strongly\ disagrees''} \end{cases}$$

Question Q13 (*"On a scale from 1 to 10 can you indicate how satisfied you are with the life you lead at the moment?"*) already provides the respondent with a 10-point rating scale.

Last, question Q14, which uses a semantic differential scale, can be separated into three variables (Corrosive, Leaves No Scale, Stains Fixtures) and recoded on a 5-point scale, for instance:

$$\mathrm{Q14}\big|Corrosive = \begin{cases} 5 \text{ if} {''check\ on\ 1st\ line''} \\ 4 \text{ if} {''check\ on\ 2nd\ line''} \\ 3 \text{ if} {''check\ on\ 3rd\ line''} \\ 2 \text{ if} {''check\ on\ 4th\ line''} \\ 1 \text{ if} {''check\ on\ 5th\ line''} \end{cases}$$

One difficulty with survey methods is that the questionnaire may contain a high number of nonresponses. They can be of two types: item nonresponse, which occurs when the respondent partially answered the questionnaire, and total or unit nonresponse, which occurs when all or almost all data for a sampling unit are missing. While the first type of error can be solved using imputation techniques, the second type generates more severe biases, especially when these nonresponses are correlated to some characteristic of the population (e.g., illiteracy). If the nonresponse rate is high, that can also impact the sample size and therefore the precision of the analysis.

The problem of total nonresponse can only be tackled ex ante. First, as already stated, if the response rate can be predicted in advance, the initial sample size should be adjusted accordingly. Second, it is possible to improve the response rate by providing higher incentives to participate, for instance by explaining the purpose of the study, by offering coupons, additional services, by using media to let citizens know that their feedbacks have been used after previous surveys.

When faced with item nonresponses, there are two possibilities. Either one excludes the item from the analysis, or replaces the missing value using imputation techniques. In the first case, the value is generally coded with a blank or NA (for Non Available). Using NA is preferable as it does not incur the risk to be mistaken with a value one has forgotten to report. In addition, spreadsheets commonly understand what NA means. For instance, the command AVERAGE from Excel yields 7 when faced with values "10, NA, 4". It should be stressed that "0" (zero) should never been used to code a missing item. This would be highly confusing since in practice, many variables can reach this value even when the item is not missing. More generally, note that characters like ";" or "/" should be avoided as most statistical packages cannot handle them properly.

Second, if one wants to replace missing values using imputation techniques, the two most common methods are deductive imputation and mean value imputation. Deductive imputation consists in using logic to deduce the missing value. Typical examples are when the sum of percentage items is less than 100%, or when a ranking question has missing values. Assume for instance that question Q9 has been filled in as follows:

Q9. Please rank the following issues in order of their importance to you. 1 stands for the most important and 6 for the least important.

1. *International tensions (terrorism, war)* 3
2. *Economic concerns (unemployment, inflation)* 2
3. *Environmental concerns (waste, air pollution)* 1
4. *Health concerns (Bird flu, AIDS)* ___
5. *Social issues (poverty, discrimination)* 5
6. *Personal safety (crime, theft...)* 4

In that case, one may quite safely attribute "6" to the missing item. However, deduction may not always be so straightforward, for instance with two or more missing values.

Mean value imputation replaces the missing value with the mean value for a given class. Assume for instance that a data set contains information about employees in a given industry and that values are missing with respect to their monthly income. Those missing values can be imputed by the average monthly income for respondents who correctly reported their remuneration and who are in the same company or geographic area. This may however reduce the sampling variance and, as such, artificially increase the sampling precision. The method should thus be used only as a last resort.

There is also always a risk that participants do not pay attention, do not read instructions, or answer randomly. Several methods exist to identify careless responders or inconsistent values. The most popular approach consists in including an attention filter where respondents are required to choose one specific answer option, sometimes regardless of their own preference:

Reading the instructions carefully is critical. If you are paying attention please choose "7" below.

1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□

This type of question is used to flag those who do not carefully read the instructions. An alternative method is to use reaction times or the duration of survey completion if the interview is computer based.

Another possibility is to identify outliers, which by definition are values that lie in the tails of a statistical distribution. In this respect, the first thing to do when checking for the quality of a database is to compute minimum and maximum values. This allows to verify whether the collected information is consistent with what one might expect. In Excel, functions *MIN* and *MAX* can be used. A more general approach is to identify values that lie outside the interquartile range. The latter is defined as $[Q_1, Q_3]$, where Q_1 is the middle value in the first half of the rank-ordered data set and Q_3 is the middle value in the second half of the rank-ordered data set. In Excel, one may use for instance the function QUARTILE(*array*, *quart*), which returns the quartile of a data set. If *quart* equals 1, the function returns the first quartile (25th percentile); if *quart* equals 2, the function returns the median value (50th percentile); If *quart* equals 3, it returns the third quartile (75th percentile).

Respondents who are flagged as outliers can be excluded from subsequent analysis, or inconsistent values be imputed using the previous techniques. One should however be careful as being an outlier is not necessarily synonymous with careless responding. Some respondents may be natural outliers, with preferences rather apart from those of more standard individuals.

Consider for instance Fig. 2.6. It provides a database constructed only for the purpose of illustrating the approach. The data correspond to a survey based on 22 citizens of a city and their satisfaction (on a 4-rating scale) about a public service, say, a response center. The city is divided into three districts whose zip codes are 700, 800 and 900, respectively. *Gender* is coded as 1 for female and 2 for male. The data have been ordered according to age (variable *Age*1). As can be seen, the minimum value for this variable is 1 and the maximum is 861. This quick glance thus points out problems in the database. Using the *quartile* function, we find $Q_1 = 34$ and $Q_3 = 74$. These values correspond to individuals 6 and 16 in the dataset. In theory, one should be suspicious about any value out of this range. For instance, we can eliminate individuals 1, 20 and 21 as their age corresponds to inconsistent values. However, individuals 2, 3, 4, 5, 17, 18, and 19 are

Individual	Age1	Gender	Zip code	Satisfaction	Age2
1	1	2	700	4	55.4
2	20	2	800	2	20
3	20	1	800	4	20
4	21	1	900	3	21
5	29	2	700	2	29
6	34	1	800	2	34
7	36	2	700	3	36
8	45	1	700	4	45
9	48	1	800	4	48
10	56	1	900	2	56
11	56	1	900	1	56
12	56	2	700	3	56
13	65	2	700	2	65
14	66	2	800	4	66
15	68	1	800	4	68
16	74	2	700	4	74
17	76	2	800	3	76
18	77	2	900	1	77
19	98	1	700	3	98
20	455	2	800	4	55.4
21	861	2	900	1	55.4
22	NA	2	700	2	55.4
MIN	1				
MAX	861				
Q1	34				
Q3	74				

Fig. 2.6 Database: example 1

natural outliers and should not be eliminated. If one wants to keep all the observations for a subsequent analysis, it is possible to replace the missing or inconsistent values either with "NA" or mean values. For instance, the average age for females (individuals 3, 4, 6, 8, 9, 10, 11, 15 and 19) is 49.6, while the average age for males (individuals 2, 5, 7, 12, 13, 14, 16, 17 and 18) is 55.4. These values can be used to recode variable *Age*1 into variable *Age*2.

Ideally, a sample survey should cover groups of the target population in proportions that match the proportions of those groups in the population itself. In practice, this may not always be the case. Due to the sampling design, to non-coverage issues or nonresponse, some groups may be over- or under-represented. In such situations, no reliable conclusions can be drawn from the sample, unless it is adjusted using raking techniques (also known as sample-balancing or iterative proportional fitting). The idea is to assign a weight to each responding unit so that the sample better matches the target population. Units that are under-represented are attributed a weight greater than 1, and those that are over-represented are attributed a weight smaller than 1.

Let us first consider a simple example with one single control variable. In Fig. 2.6, we have information about the gender of each respondent: 9 respondents are females, and 13 are males (outliers included). Assume now that we can compare the response distribution of *Gender* with the population distribution, assumed to be equally distributed between males and females:

Sample	Female: 9 (40.9%)	Male: 13 (59.1%)
Population (census)	Female: 2200 (50%)	Male: 2200 (50%)

The population includes 50% of males, while it is 59% in the sample. The males are thus over-represented in our sample. We can solve this representativeness bias by assigning adequate weights to male and female respondents:

$$Weight\Big|female = 50\%/40.9\% = 1.22$$

$$Weight\Big|male = 50\%/59.1\% = 0.85$$

The weights are obtained by dividing the population percentage by the corresponding sample percentage.

In practice, it is frequent to use several control variables. The computational approach is complex and relies on raking algorithms. To illustrate, assume now that we use both *Gender* (two categories: male, female) and *Zip code* (three categories: 700, 800, 900) to correct the representativeness bias. Combining all possibilities of gender and zip code leads to 2×3 different groups. Assume now that we have information about the distribution of *Zip code* within the target population:

Zip code	700	800	900
Sample	9 (40.9%)	8 (36.4%)	5 (22.7%)
Population	2000 (45.45%)	1200 (27.27%)	1200 (27.27%)

How can we use this information to compute the weights? Figure 2.7 illustrates the approach. Figure 2.7a contains information about the total frequencies in the sample. For instance, in our dataset, 2 females live in district 700, 4 in district 800 and 3 in district 900. Last row and column of Fig. 2.7a provide the target to attain. Since the population is equally distributed among males (50%) and females (50%), one should obtain similar proportions in the sample, i.e. $50\,\% \times 22 = 11$. Similarly, since 45.45% of the population lives in district 700, one should have $45.45\,\% \times 22 = 10$ units for this category in the sample, and 6 units for districts 800 and 900.

Raking is achieved with successive iterations until one converges to the desired set of proportions. In Fig. 2.7b, the first iteration consists in aiming at 11 for the total of males and females. Value 2.44 is obtained by multiplying the sample frequency (here 2) by 11/9. Similarly, 4.89 is the product of 4 with 11/9, and so on. The second iteration aimins at the desired set of proportions for *Zip code*. For instance, value 2.92 is obtained by multiplying 2.44 with 10/8.37. The new values obtained however affect in return the total frequency of males and females, and the process must be reiterated until one reaches convergence. As can be seen from Fig. 2.7e, after four iterations, the values are more stable. The weights are finally obtained by dividing the values of Fig. 2.7e by those of Fig. 2.7a (see Fig. 2.7f).

As can be deduced from the previous example, raking can be laborious, and one may rely instead on statistical software to assess the relevant weights. Figure 2.8

a

	700	800	900	Total	Target
Female	2	4	3	9	11
Male	7	4	2	13	11
Total	9	8	5	22	
Target	10	6	6		

b

	700	800	900	Total	Target
Female	$\mathbf{2.44} = 2 \times 11/9$	$4.89 = 4 \times 11/9$	$3.67 = 3 \times 11/9$	11	11
Male	5.92	3.38	1.69	11	11
Total	8.37	8.27	5.36	22	
Target	10	6	6		

c

	700	800	900	Total	Target
Female	$2.92 = \mathbf{2.44} \times 10/8.37$	3.55	4.11	10.57	11
Male	$7.08 = 5.92 \times 10/8.37$	2.45	1.89	11.43	11
Total	10.00	6.00	6.00	22.00	
Target	10	6	6		

d

	700	800	900	Total	Target
Female	3.04	3.69	4.27	11.00	11
Male	6.81	2.36	1.82	11.00	11
Total	9.85	6.05	6.10	22.00	
Target	10	6	6		

e

	700	800	900	Total	Target
Female	3.08	3.66	4.20	10.95	11
Male	6.92	2.34	1.80	11.05	11
Total	10.00	6.00	6.00	22.00	
Target	10	6	6		

f

	700	800	900		
Female	$3.08/2 = 1.54$	0.91	1.40		
Male	0.98	0.58	0.89		

Fig. 2.7 Raking with two variables: example 1. (**a**) Sample, (**b**) Iteration 1, (**c**) Iteration 2, (**d**) Iteration 3, (**e**) Iteration 4, (**f**) weights

provides the coding to be used in R-CRAN. Command *read . table* is used to upload the database, saved afterwards under the name "D", using the path *C : //mydata . csv*, which denotes the location of the file. The file format is *.csv*, with ";" as a separator, and can be easily created with Excel. The command *head* displays the first rows of the dataset. To use the *anesrake* function, all variables must be coded continuously (1, 2, 3, etc.) with no missing values. Variable *Zip.code* has thus been recoded from 1 to 3. On average, the level of satisfaction with respect to the public service under evaluation is 2.81. This value however does not take into account the representativity bias. The package *weights* allows to compute the proportion of males and females using *wpct(D$Gender)*, as well as how the sampling units are distributed among the districts, using *wpct(D$Zip.code)*. The next step is to specify manually the population characteristics:

$$p.gender = c(0.50, 0.50)$$

$$p.zip = c(0.454545455, 0.272727273, 0.272727273)$$

$$targets = list(p.gender, p.zip)$$

$$names(targets) = c("Gender", "Zip.code")$$

```
> D=read.table("C://mydatasampling.csv",head=TRUE,sep=";")
> head(D)
  Individual Age1 Gender Zip.code Satisfaction Age2
1          1    1      2        1            4 55.4
2          2   20      2        2            2 20.0
3          3   20      1        2            4 20.0
4          4   21      1        3            3 21.0
5          5   29      2        1            2 29.0
6          6   34      1        2            2 34.0
> mean(D$Satisfaction)
[1] 2.818182

> library(weights)
> wpct(D$Gender)
        1         2
0.4090909 0.5909091
> wpct(D$Zip.code)
        1         2         3
0.4090909 0.3636364 0.2272727

> p.gender=c(0.50,0.50)
> p.zip=c(0.454545455,0.272727273,0.272727273)
> targets=list(p.gender, p.zip)
> # important: use the same variable names of the dataset
> names(targets) =c ("Gender", "Zip.code")

> library(anesrake)
> myrake=anesrake(targets, D, caseid=D$Individual)

> D$myweights=myrake$weightvec
> head(D$myweights)
[1] 0.9845212 0.5817086 0.9182914 1.4061610 0.9845212 0.9182914
> mean(D$myweights*D$Satisfaction)
[1] 2.78211
```

Fig. 2.8 Raking with R-CRAN: example 1

It is important to use the same variable names as the dataset. The raking algorithm is implemented using the package *anesrake*. One has to specify the desired set of proportions (*targets*), the dataset (*D*), the column that contains the individual numbers (*D*$*Individual*). Finally, the weights (*myrake*$*weightvec*) are saved into the dataset *D* under the name *D*$*myweights*. Those weights are similar to those presented in Fig. 2.7f. For instance, individual 1 is a male from district 1 and as such receives a weight of 0.98. Individual 3 is a female from the second district and gets a weight of 0.91. Weights can be used to compute the average satisfaction *mean* (*D*$*myweights* $*$ *D*$*Satisfaction*). We now find 2.78.

Note that raking adjustments imply to know only the population totals of the specific variables, not all cells of a cross-table. The first step is to identify a set of variables likely to be used as control variables, and to compare them with reliable data sources (e.g., census). Some typical variables are age groups, gender, socio economic status, geographical location. When selected variables have categories with less than 5% in the sample, it is recommended to collapse them.

Bibliographical Guideline Several definitions in this chapter have been taken and modified from the OECD Glossary of Statistical Terms. This glossary contains additional definitions of key concepts and commonly used acronyms. These definitions are primarily drawn from existing international statistical guidelines and recommendations that have been prepared over the last two or three decades by international organizations (such as the United Nations, International Labor Organization, Organization for Economic Co-operation and Development, Eurostat, International Monetary Fund) working with national statistical institutes and other agencies responsible for the initial compilation and dissemination of statistical data.

Several guides are also available online, such as the “Guidelines for Designing Questionnaires for Administration in Different Modes” proposed by the United States Census Bureau, “Public Opinion Surveys as Input to Administrative Reform” by the Organization for Economic Co-operation and Development, “Designing Household Survey Samples: Practical Guidelines” by the Department of Economic and Social Affairs of the United Nations, “Survey Methods and Practices” by Statistics Canada, the “Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System” by Eurostat.

Bibliography

Eurostat. (2004). *Handbook of recommended practices for questionnaire development and testing in the European Statistical System*.

OECD. (1998). *Public opinion surveys as input to administrative reform* (SIGMA Papers, No. 25). OECD Publishing.

Statistics Canada. (2010). *Survey methods and practices*.

US Census Bureau. (2007). *Guidelines for designing questionnaires for administration in different modes*.

United Nations. (2005). *Designing household survey samples: Practical guidelines*.

Descriptive Statistics and Interval Estimation

3

3.1 Types of Variables and Methods

Statistics divide into two branches, namely descriptive statistics and inferential statistics. Descriptive statistics are used to describe and summarize the basic features of the data in a study. It is the starting point of any evaluation. The approach is not difficult to implement as it consists in simplifying often large amounts of data using tabular, graphical and numerical techniques. Yet, it remains essential if one wants to get a clear picture of what the dataset contains. Assume for instance that one has gathered data about 1000 students receiving a particular educational training. Providing extensive information about the socio-economic characteristics of each individual as well as their marks would be useless or even uninformative as it would be impossible to grasp the phenomena at stake. Instead, it is preferable to provide simple and understandable summary statistics such as information about the share of males and females, or graphs visualizing the characteristics of the sample as a whole.

Providing descriptive statistics is thus a preliminary and necessary step. It gives a snapshot of the information that has been gathered. In most cases, however, the description is done in the context of a sample survey. The conclusions depend as such on the selected sample, which introduces some uncertainty in the results. Any generalization should be ventured with care. Inferential statistics is very useful in this respect. It aims at generalizing the sample findings to the population of interest through the calculation of well-defined degrees of uncertainty. This uncertainty is accounted for through a margin of error $\pm e$, calculated with adequate probability distributions. The approach offers a way of assessing the confidence the evaluator has in drawing conclusions from the sample. For instance, based on a sample study, we can use inferential statistics to deduce the satisfaction inhabitants derive from a public service as if the survey was involving the whole population.

Whether they are inferential or descriptive, statistical tools vary depending on the type of variables that are examined (Table 3.1.). A distinction is made between a categorical variable and a numerical variable. A categorical variable, also known as

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_3

Table 3.1 Statistical variables

Type of variable	Subtype	Example
Categorical variables	Ordinal	Academic grades A, B, C,. . .
	Nominal	Political party affiliation
Numerical variables	Continuous	Income
	Discrete	Number of children

qualitative variable, describes a quality or characteristic of a data unit. This type of variable cannot be meaningfully described in terms of numbers. It may be classified as either ordinal or nominal. Ordinal variables take values that can be logically ordered or ranked. Examples include academic grades (e.g., A, B, C) or answers to scale questions (e.g., strongly agree, agree, disagree, strongly disagree). Nominal variables take values that cannot be ranked on an ordinal scale. Examples include gender, place of residence, political affiliation. Numerical variables on the other hand describe a numerically measured value. They are also referred to as quantitative variables. Two subcategories exist. A continuous variable can take any value between its minimum value and its maximum value (e.g., 20.1, 35.2, 40.3). Examples include income, time, and age. A discrete variable takes values based on a count from a set of distinct whole values (e.g., 20, 35, 40). Examples include the number of items bought by a consumer, the number of children in a family. In practice, the distinction between a continuous variable and a discrete variable is not as straightforward as one might think. For the sake of simplicity, continuous variables are often reported on a discrete scale. For instance, taxable income is often rounded to the nearest whole dollar.

This chapter aims to review the different statistical methods that can be used to describe a sample and to make inference for a larger population. Despite its apparent simplicity, one should not underestimate the importance of the task, especially in the context of public policies. Any evaluation study involves the presentation of a lot of data in a concise manner. In particular, identifying the problems or needs that a given policy must address requires informative statistics, often presented in the form of context indicators. Those statistics may for instance provide information about the economic situation of a jurisdiction (local GDP per capita, unemployment rate, public debt, etc.), socio-demographics characteristics (share of elderly people, gender, etc.) or other variables of interest (environment, health, education). Such indicators can reflect the jurisdiction's situation at a given date or over a large set of time periods. They provide information about the different aspects that are likely to influence policy evaluation. They may also evidence selection biases in the collection of information, which may in turn affect the evaluation process.

The outline of the chapter follows the classification of Table 3.2. Section 3.2 explains how to describe a database using tables. Section 3.3 is about graphical methods. Section 3.4 details the different measures of central tendency and variability. Section 3.5 explains how to describe the shape of a distribution of data. Section 3.6 introduces inferential statistics and explains how to compute

Table 3.2 Tools for describing statistical variables

	Descriptive statistics			Inferential statistics
	Tabular analysis	Numerical analysis	Graphical analysis	Use of sample statistics to infer characteristics of the population
Categorical variables	One-way and two-way tables	Absolute and relative frequencies	Bar graph, pie chart, radar chart	Confidence interval for a population proportion
Numerical variables	One-way and two-way tables	Min, max, mean, mode, median, variance, standard deviation, coefficient of variation, skewness, kurtosis	Histogram, box-plot, radar chart, line graph	Confidence interval for a population mean

Table 3.3 Frequency table

Class	Frequency	Relative frequency	Cumulative relative frequency
$\boldsymbol{x}_1$	n_1	$f_1 = \frac{x_1}{n}$	$F_1 = f_1$
$\boldsymbol{x}_2$	n_2	$f_2 = \frac{x_2}{n}$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_l$	n_l	$f_l = \frac{x_l}{n}$	$F_l = f_1 + f_2 + \ldots + f_l$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_L$	n_L	$f_L = \frac{x_L}{n}$	$F_L = f_1 + f_2 + \ldots + f_L = 1$
Total	$\boldsymbol{n}$	**1**	

confidence intervals in order to generalize the results obtained from a sample to the whole population of interest.

3.2 Tabular Displays

One important step in a statistical analysis is to show how the observations are distributed in a dataset. This can be done through what is termed a one-way table, or frequency table, which counts the number of times each value occurs. To build such a table, one needs to split the data into categories or intervals known as classes or bins. There are two rules in this respect. First, every observation must belong to one and only one class. Second, as a matter of simplicity, the classes should have the same width. Once created, the classes are listed in the first column of the one-way table and the frequencies are displayed in a second column, as depicted in Table 3.3.

Imagine that a variable x is divided into L levels, or classes, namely $x_1 \ldots x_L$. The frequency (or absolute frequency) of a particular class x_l represents the number of observations n_l that fall in the class. For example, suppose that a frequency distribution is based on a sample of 200 students. Imagine that 50 of these students got a score between 60 and 70. In that context, the interval 60–70 represents the class, and 50 is the absolute frequency.

As shown in Table 3.3, the frequency distribution can also be thought in terms of relative frequency (or percent frequency). The relative frequency f_l of a class x_l is defined as the number of occurrences n_l in the dataset divided by the total number of observations n:

$$f_l = \frac{n_l}{n}$$

The value f_l assigned to each class l represents the proportion of the total data set that belongs in the class. In our previous example, this number amounts to 25% (50/200) for the class 60–70. By construction and except for any rounding error, relative frequencies should add up to 100%. It is then possible to compute cumulative relative frequencies. The approach consists in adding up the relative frequencies from one class to the next, to give a running total. The cumulative relative frequency F_l of a class x_l is given by:

$$F_l = \sum_{p=1}^{l} f_p$$

It denotes the percentage of the observations that are lower than x_l. The cumulative frequency of the first class is the same as its relative frequency. The cumulative frequency of the second class is the sum of the first and second relative frequencies, and so on. Note that cumulative frequencies have no meaning for nominal variables as no category can be higher or lower than another.

A frequency table provides information on the distribution of one single variable at a time. With a categorical variable, the observations are already organized into different groups. With a continuous variable on the other hand, several steps are involved for specifying a set of classes. First, the range of a variable x is defined as the difference between the minimum and maximum values in the sample:

$$\text{range}(x) = \max(x) - \min(x)$$

Second, the number of classes is commonly between five and twenty. Sturges' formula is generally used:

$$L = 1 + 3.322 \times \log_{10}(n)$$

where n denotes the sample size. The number of classes increases with the sample size. Third, the class width is obtained by dividing the range of the data by the number L of classes:

$$w = \frac{\text{range}(x)}{L}$$

The next higher convenient number is chosen in case of fractional results. The final step consists in determining the class limits. From an appropriate starting point for the lower limit, the lower limit of the next class is obtained by adding the class width. The process continues until the last class is reached. Statistical softwares such as Excel or R-CRAN can be used to automatize the procedure of class specification.

To illustrate the approach, assume that we have information about the research and development (R&D) expenditures of a sample of 60 firms for a given time period (example 1). Table 3.4 provides the raw data. We have figures about R&D intensity, the number of patents assigned to each firm, whether those firms have received a government subsidy and the sector they belong to. Variable *intensity* is continuous and defined as R&D expenditures divided by value added. As it cannot take the value of a fraction, *patent* is a discrete variable. Variable *subsidy* is a nominal variable. It equals one if the firm has received a subsidy and zero otherwise. Last, variable *sector* is an ordinal variable. It classifies the manufacturing industries into three categories, coded as 3 for high-technology industries (e.g., aircraft and spacecraft), 2 for medium-technology industry (e.g., motor vehicles, trailers and semi-trailers) and 1 for low-technology industry (e.g., food products, beverages and tobacco).

The frequency distributions of the variables are provided in Table 3.5. For each variable, we have the number of elements that belong to each class as well as the relative frequencies. As can be seen, the framework is easily settled for categorical variables. One simply needs to record the number of firms that have received a subsidy or the number of firms that belong to one of the considered sectors. For instance, in Table 3.5a, we can see that the share of firms that have received a subsidy amounts to 33% (i.e. 20 firms out of 60). Table 3.5b provides information on the distribution of firms among the industrial sectors: 35% are low-technology, 40% are medium-technology and 25% are high-technology. Last column of Table 3.5b (cumulative frequency) points out that 75% of firms are not high-technology. Table 3.5c, d respectively provide frequencies for R&D intensity and patent claims.

With numerical variables, one needs to define the class intervals. The Sturges formula provides the number of class intervals:

$$L = 1 + 3.322 \times \log_{10}(60) = 6.91 \ (\approx 7 \text{ rounded off})$$

As can be seen from Table 3.5, all intervals have the same width and are continuous throughout the distribution. For variable *Intensity* the range is computed as:

Table 3.4 Raw data for example 1

Firm	Intensity	Patents	Subsidy	Sector
1	0.23	42	1	3
2	0.19	31	0	3
3	0.20	20	0	3
4	0.19	33	1	3
5	0.18	30	1	3
6	0.18	43	0	3
7	0.17	37	1	3
8	0.23	27	1	3
9	0.11	7	0	3
10	0.16	21	1	3
11	0.14	13	0	3
12	0.20	20	1	3
13	0.21	42	1	3
14	0.18	35	1	3
15	0.14	13	0	3
16	0.21	26	1	2
17	0.11	24	0	2
18	0.10	1	0	2
19	0.13	33	0	2
20	0.17	49	1	2
21	0.08	3	0	2
22	0.11	11	0	2
23	0.14	28	0	2
24	0.11	3	1	2
25	0.09	4	0	2
26	0.17	32	1	2
27	0.08	2	0	2
28	0.13	22	0	2
29	0.08	2	0	2
30	0.12	16	0	2
31	0.07	1	0	2
32	0.16	21	1	2
33	0.18	39	1	2
34	0.08	2	0	2
35	0.16	25	1	2
36	0.06	0	0	2
37	0.12	19	0	2
38	0.20	58	1	2
39	0.10	7	0	2
40	0.04	0	0	1
41	0.09	13	1	1
42	0.01	2	0	1
43	0.04	13	0	1

(continued)

Table 3.4 (continued)

Firm	Intensity	Patents	Subsidy	Sector
44	0.04	11	0	1
45	0.07	2	0	1
46	0.07	5	0	1
47	0.00	0	0	1
48	0.02	0	0	1
49	0.08	2	1	1
50	0.02	2	0	1
51	0.01	4	0	1
52	0.05	0	0	1
53	0.05	6	0	1
54	0.02	0	0	1
55	0.02	4	0	1
56	0.07	19	0	1
57	0.00	0	1	1
58	0.08	12	0	1
59	0.08	14	0	1
60	0.01	0	0	1

$$\text{range}(Intensity) = 0.23 - 0 = 0.23$$

The ratio of these numbers yields the width of class intervals:

$$w = \frac{0.23}{6.91} = 0.0333$$

This value can be rounded to 0.035. The first class interval is thus defined as $[0\%, 3.5\%]$, then it is $]3.5\%, 7\%]$, and so on. For variable *Patents*, the range is $58 - 0 = 58$. The width is computed as $58/6.91 = 8.36$, which can be rounded to 9. The class intervals are then computed as simple multiples of the width.

The main task in creating a frequency table is counting the number of units observed in each class. Excel can be very helpful in this respect. One needs to load an add-in program that is available when one installs Excel: the Analysis ToolPak. In the Excel Options, click Add-Ins, and then in the Manage box, select Excel Add-ins. In the Add-Ins available box, select the Analysis ToolPak check box, and then click OK. Once the Analysis ToolPak is loaded, the Data Analysis command is available in the Analysis group on the Data tab. It is possible to create a frequency table with the "Histogram tool". The input range corresponds to the column of the raw data one wants to analyze. The bin range stands for the class intervals. If no class number is entered, then the Histogram tool will create evenly distributed class intervals by using the minimum and maximum values in the input range as start and end points.

Table 3.5 Frequency distributions: example 1

	Frequency		**Relative frequency (%)**
(a) Subsidy			
Yes	20		33
No	40		67
Total	60		100
	Frequency	**Relative frequency (%)**	**Cumulative frequency (%)**
(b) Industrial sectors			
Low-tech	21	35	35
Medium-tech	24	40	75
High-tech	15	25	100
Total	60	100	
(c) R&D intensity			
0–3.5%	9	15	15
3.5–7%	10	17	32
7–10.5%	11	18	50
10.5–14%	11	18	68
14–17.5%	6	10	78
17.5–21%	11	18	97
21–24.5%	2	3	100
Total	60	100	
(d) Patent claims			
0–9	26	43	43
9–18	9	15	58
18–27	11	18	77
27–36	7	12	88
36–45	5	8	97
45–54	1	2	98
54–63	1	2	100
Total	60	100	

Once frequency tables are created, the analysis can move to the next step with the construction of two-way tables (also known as contingency tables or cross-tabulations). Table 3.6 illustrates the approach. Entries in the cells of a two-way table can be displayed as absolute frequencies (Table 3.6a) or as relative frequencies (Table 3.6b). The classes of variable x label the rows while the classes of variable y label the columns. Scalars L and M denote the number of classes for the row variable and the column variable, respectively. Entries in the body of the table (n_{lm}) are termed joint frequencies. Entries in the total row $(n_{.m}, m = 1 \ldots M)$ and total column $(n_{l.}, l = 1 \ldots L)$ are referred to as marginal frequencies or marginal distributions. This term is not to be mistaken with that of a conditional distribution. Marginal frequencies provide information about the distribution of a variable in the whole dataset. It corresponds to the values reported in the one-way table. Conditional frequencies on the other hand relate to the distribution of one

Table 3.6 Two-way tables

(a) Absolute frequencies							
	y_1	y_2	...	y_m	...	y_M	Total
$\boldsymbol{x}_1$	n_{11}	n_{12}	...	n_{1m}	...	n_{1M}	$n_{1.}$
$\boldsymbol{x}_2$	n_{21}	n_{22}	...	n_{2m}	...	n_{2M}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
$\boldsymbol{x}_l$	n_{l1}	n_{l2}	...	n_{lm}	...	n_{lM}	$n_{l.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
$\boldsymbol{x}_L$	n_{L1}	n_{L2}	...	n_{Lm}	...	n_{LM}	$n_{L.}$
Total	$\boldsymbol{n}_{.1}$	$\boldsymbol{n}_{.2}$		$\boldsymbol{n}_{.m}$		$\boldsymbol{n}_{.M}$	$\boldsymbol{n}$
(b) Relative frequencies							
	y_1	y_2	...	y_m	...	y_M	Total
$\boldsymbol{x}_1$	f_{11}	f_{12}	...	f_{1m}	...	f_{1M}	$f_{1.}$
$\boldsymbol{x}_2$	f_{21}	f_{22}	...	f_{2m}	...	f_{2M}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
$\boldsymbol{x}_{\mathrm{l}}$	$f_{\mathrm{l}1}$	$f_{\mathrm{l}2}$	...	f_{lm}	...	f_{lM}	$f_{\mathrm{l.}}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
$\boldsymbol{x}_{\mathrm{L}}$	$f_{\mathrm{L}1}$	$f_{\mathrm{L}2}$	...	f_{Lm}	...	f_{LM}	$f_{\mathrm{L.}}$
Total	$\boldsymbol{f}_{.1}$	$\boldsymbol{f}_{.2}$		$\boldsymbol{f}_{.\mathrm{m}}$		$\boldsymbol{f}_{.\mathrm{M}}$	1
(c) Conditional distribution of x given the value y_m of y							
Class			Frequency		Relative frequency		
$\boldsymbol{x}_1$			n_{1m}		f_{1m}		
$\boldsymbol{x}_2$			n_{2m}		f_{2m}		
$\vdots$			$\vdots$		$\vdots$		
$\boldsymbol{x}_l$			n_{lm}		f_{lm}		
$\vdots$			$\vdots$		$\vdots$		
$\boldsymbol{x}_L$			n_{Lm}		f_{Lm}		
Total			$\boldsymbol{n}_{.m}$		$\boldsymbol{f}_{.m}$		

Table 3.7 Marginal and joint distributions: example 1

		Subsidy		
		No	Yes	Total
Sector	Low-tech	18	3	**21**
	Medium-Tech	16	8	**24**
	High-tech	6	9	**15**
	Total	**40**	**20**	**60**

variable for a given value or class of the other variable. For instance, Table 3.6c provides the conditional distribution of x given the value y_m of y.

To illustrate, consider the categorical variables of example 1. First, one needs to calculate the appropriate frequency counts. To do so, all the 60 firms must be examined. In Table 3.7, the marginal distributions of *Subsidy* and *Sector* appear at the right and bottom margins of the two-way table, respectively. The numbers reported in the body of the table represent the joint frequencies. In practice, they can be computed using pivot tables in Excel. To insert a pivot table, one needs to execute the following steps. On the Insert tab, click PivotTable and then select the variables to be analyzed. The relevant fields must be dragged to the different areas. With numerical variables, it is possible to specify the class intervals by right-clicking any cell inside the first row or first column and by selecting “Group”.

Two-way tables are useful for displaying data and identifying trends. In Table 3.7 for instance, it can be seen that among the 40 firms that did not receive a subsidy, 18 belong to the low-technology sector, 16 to the medium-technology sector and only 6 to the high technology sector. This shows evidence of a selection bias in the way the grant was assigned. This may farther affect the method of analysis. In this example, it may be misleading directly to compare the targeted group (or treatment group), those who received a subsidy, with the control group (those who did not). Evaluating the role the subsidy has on patent claims is thus not as straightforward as one might think.

3.3 Graphical Representations

Visual displays are useful in many occasions. Graphs for instance can be easier to read than a table, especially when the reader has a limited knowledge of statistics. They can summarize the key features of a set of data in a very efficient manner with minimum loss of information. A variety of methods exists. Among the most popular we can name bar graphs, circle graphs, histograms, scatter plots, line graphs and radar charts. Ideally, those graphs should convey information that would not be readily understandable if it was provided in the main text or in a table. They can be generated by statistical packages at the analyst’s convenience.

Graphs should be used with care. A badly conceived figure may not accurately reflect the true nature of the data. Designing good charts is all the more challenging that their conception does not only depend on the knowledge one has of the data, but also on how the reader will apprehend each graphical element. Several rules exist in this respect: (1) there is a title for the graph as well as for the axes; (2) a legend is included if different symbols or colors are used; (3) the scales defining the axes are apparent on the axes and, unless otherwise specified, they must encompass the whole set of observations; (4) units are visible (e.g., percentage, thousand dollars); (5) as they may distort the visual representation of the data, 3D, shadow and other fancy effects should be used with extreme caution; (6) colors should be avoided if the document has to be printed in black and white; (7) the sources of data must be specified either in the text or in the title; (8) unless otherwise specified, the source of the graph as well as the data is included in the title (e.g., “Source: Author’s contribution based on data from. . .” or “Source: OECD, 2016”) and, when necessary, the sources are integrated as new items in the bibliography.

Bar graphs are used for displaying the distribution of categorical variables (Fig. 3.1). For each class there is a bar and the height of each bar represents the frequency of the class. The frequency can be expressed in absolute, relative or joint values. The bars may be drawn to be contiguous or detached. The bars are uniform

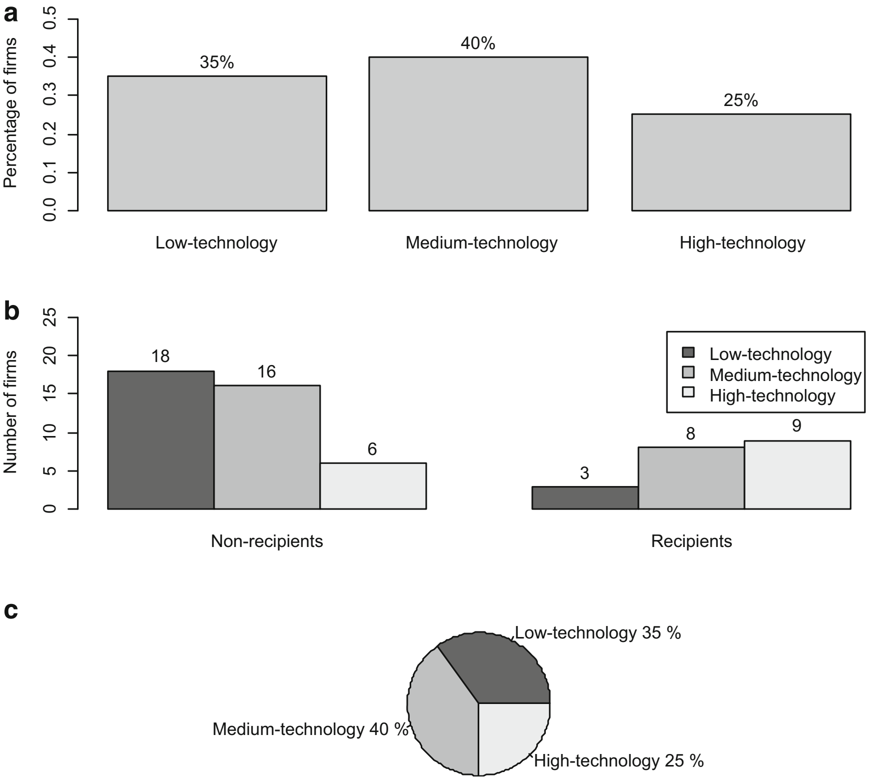


Fig. 3.1 Plotting categorical variables: example 1. (**a**) Bar graph: relative frequencies. (**b**) Bar graph: joint frequencies. (**c**) Circle graph: relative frequencies

in width. This is an important feature as a larger bar could make the reader think that there might be a higher number of occurrences than there actually is. Figure 3.1a provides an illustration using data from Table 3.4. Frequency percentages or numbers can be included at the top of the bars. With nominal variables, the categories are generally rearranged so that the bars grade sequentially from the most frequent category to the less frequent. This is not the case with ordinal variables as the sequence of categories has a meaning. When it comes to joint distributions, multiple bar graphs can appear on the same figure, as illustrated in Fig. 3.1b.

Circle graphs (or pie charts) offer another way of summarizing the distribution of categorical variables. They take the form of a circle divided into a series of portions. Each portion represents a particular class. Figure 3.1c provides an example. When faced with subgroups (e.g., countries), the diameter of the circle can also be linked to the number of observations in each subgroup in order to reflect their size. Circle graphs are visually simpler than any other type of graph. Yet, they should be used sparingly. They are often considered as the worst way to convey information. They become unreadable when the circle is divided into more than six portions. With respect to ordinal variables, bar graphs are usually more appropriate than pie charts as they allow the natural ordering of the categories to be visualized. Moreover, it is difficult to visually assess the different sizes of the portions and, thereby, the fractions that are associated with them. This may in return affect the reader's judgment.

Histograms provide information on the distribution of numerical variables. A histogram looks somewhat like a bar chart. One difference is that the observations need to be grouped into classes. The classes must be indicated by increasing order on the horizontal axis. For each class, a vertical bar whose length is the frequency of that class is drawn. This frequency can be expressed in absolute, relative or conditional values. The process of constructing the classes is the same as with tables (Sturges formula). Statistical packages usually offer those classes automatically. Figure 3.2a provides an example using information from Table 3.5c about

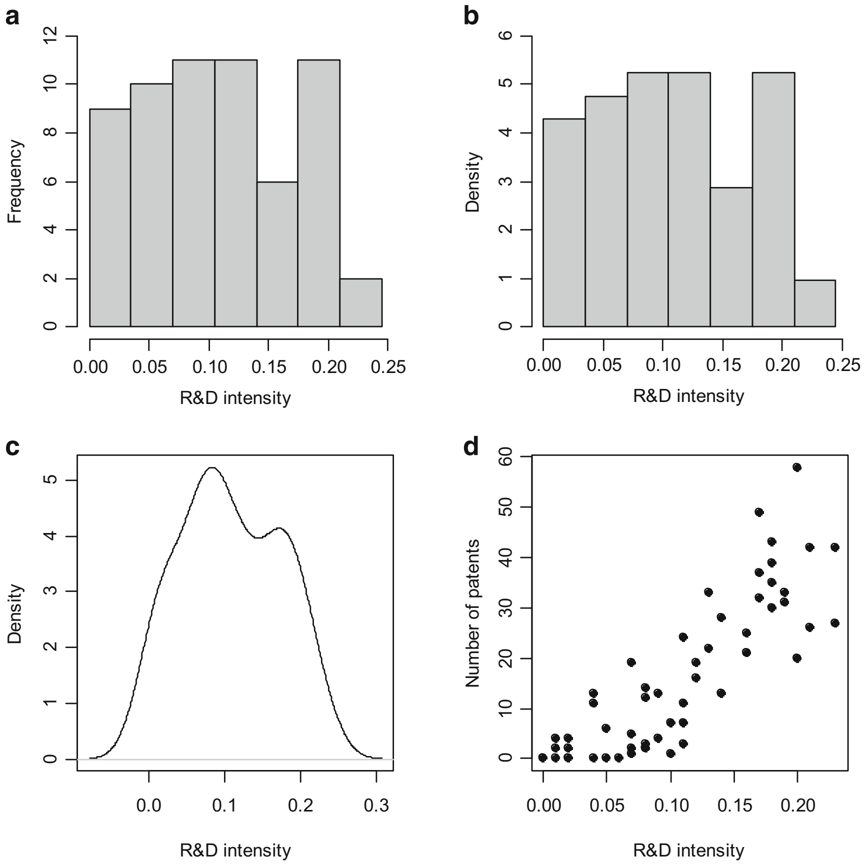


Fig. 3.2 Plotting numerical variables: example 1. (**a**) Histogram: absolute frequencies. (**b**) Histogram: adjusted frequencies. (**c**) Kernel density plot. (**d**) Scatter plot

R&D intensity. The length of the first bar corresponds to the absolute frequency of the first class, that is $n_1 = 9$. The length of the second bar amounts to $n_2 = 10$ and so on.

It is also possible that the height of the bars in a histogram does not represent the number of observations or the relative frequencies. In that case, the length stands for an adjusted number so that the area of each bar amounts to the frequency (absolute or relative) of each class. The total area of the bars then corresponds either to the total number of observations or to 100%. Adjustments are made as follows. For a given class l, one needs to divide the frequency of the class by the width w_l of the class. We have:

$$\text{Adjusted } n_l = \frac{n_l}{w_l};\ \text{Adjusted} f_l = \frac{f_l}{w_l}$$

For instance, in the case of a relative frequency distribution, the area of each bar is obtained by the multiplication of the bar's height (f_l/w_l) with its width (w_l), i.e. amounts to the relative frequency (f_l) of the class. Hence, the total area of the bars is given by $f_1 + f_2 + \ldots + f_L = 1$.

Figure 3.2b illustrates the approach. For the first class on the left (0, 3.5%), the relative frequency amounts to 15% (see Table 3.5c). The adjustment is as follows:

$$\text{Adjusted} f_1 = \frac{f_1}{w_1} = \frac{15\%}{3.5\% - 0\%} \approx 4.28$$

This value corresponds to the height of the first bar in Fig. 3.2b. As can be seen, the re-scaling does not change the shape of the histogram. The approach is particularly useful when class intervals are unequal or when one wants to get an approximation of a probability density function, in order to find the probability that a random variable falls into a particular range of values. Probability density functions are often used to model large sets of data. By construction, the total area underneath a probability density curve is 1 because the probability that an observation falls within the minimum value and the maximum value is 100%.

Histograms provide information about the likelihood that a particular event occurs. The higher the bar of a class, the more likely that particular class is to be observed. Yet, it should be stressed that the shape of a histogram depends on the number of class intervals and the way they are constructed. We may have for instance as many classes as there are values, or only one class that would encompass the whole set of observations. Kernel density plots offer an alternative way to visualize the distribution of numerical variables. It consists in using nonparametric techniques to approximate the shape of the distribution. Figure 3.2c provides an example. The approach yields a curve that is closely related to the shape of Fig. 3.2b in a smoother and continuous way.

Scatter plots are used to display the distribution of two numerical variables. This type of graph has two dimensions: a horizontal dimension for the x -variable and a vertical dimension for the y -variable. Each observation has two coordinates which

Table 3.8 Population of five cities from 1950 to 2010: example 2

Year	City A	City B	City C	City D	City E
1950	6035	12,301	4351	1987	8864
1960	6666	13,190	4759	2195	9317
1970	8126	14,715	4904	2307	10,090
1980	10,921	16,743	4953	2812	11,481
1990	11,480	18,313	5525	3107	12,974
2000	12,681	19,636	5808	4175	15,816
2010	12,936	22,123	6416	4612	17,470

indicate the position of the observation with respect to the horizontal and vertical axes. A symbol, generally a bullet point, represents the observation at the intersection of the two coordinates. Figure 3.2d provides an illustration. The way in which the points are distributed indicates whether there is a relationship, in the form of a line or a curve for instance, between x and y. Figure 3.2d hints that on average the number of patent claims increases as R&D intensity increases.

Line graphs offer a way to display the behavior of one or several variables over time. Time appears on the horizontal axis and the variable(s) of interest on the vertical axis. Each observation has two coordinates corresponding to the position of the observation in time and the value it achieves. The symbols representing the observations are usually connected by segments to highlight changes over time. When faced with subgroups (e.g., countries), one curve for each subgroup is displayed, and the symbol used is specific to the subgroup. Table 3.8, for example, provides information on the population growth of five (fictitious) cities from 1950 to 2010. From Fig. 3.3, one can see that the population is constantly increasing, but at different rates depending on the city.

Finally, radar charts allow the distributions of several variables to be displayed simultaneously. Also known as web charts, spider charts or star charts, they are used to compare the performance of one or more units (e.g., individuals, cities, drugs). Variables are displayed on separate axes. Each axis extends outward from the center of the chart and has the same size as the other axes. Observations are represented by a symbol on the corresponding axis. These symbols are connected by segments in order to highlight the specificities of each unit. For example, radar charts offer a useful way for presenting multivariate clinical data. To illustrate, Table 3.9 provides information on the side effects of two competing drugs, gathered from two different samples. Each observation corresponds to the proportion of patients who suffered from side effects. Figure 3.4 illuminates the differences between treatments. While drug A is more likely to induce headaches, drug B is prone to cause nausea.

All the graphs displayed in this section have been created with R-CRAN. The source codes for the figures of example 1 are provided in Fig. 3.5, those of examples 2 and 3 in Fig. 3.7.

Figure 3.5 starts with the *read.table* command. It reads a file in table format (saved as a *.csv* file on disc *C*:) and creates a data frame *D* in the R-CRAN environment. The command *head*() returns the first parts of the table, which are

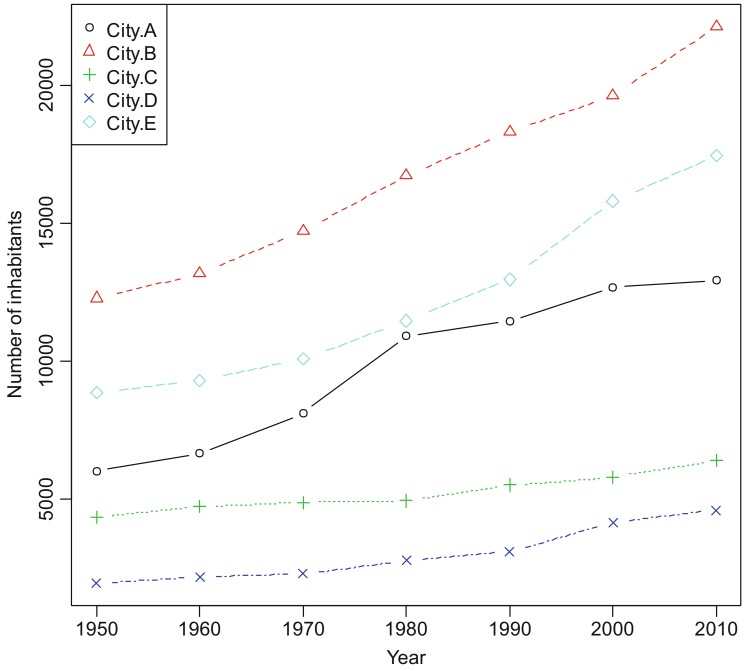


Fig. 3.3 Line graph of the population of five jurisdictions: example 2

Table 3.9 Proportion of patients who suffered from side effects: example 3

	Nausea (%)	Vomiting (%)	Diarrhea (%)	Headache (%)	Rash (%)
Drug A	13	1	1	10	4
Drug B	15	2	3	5	5

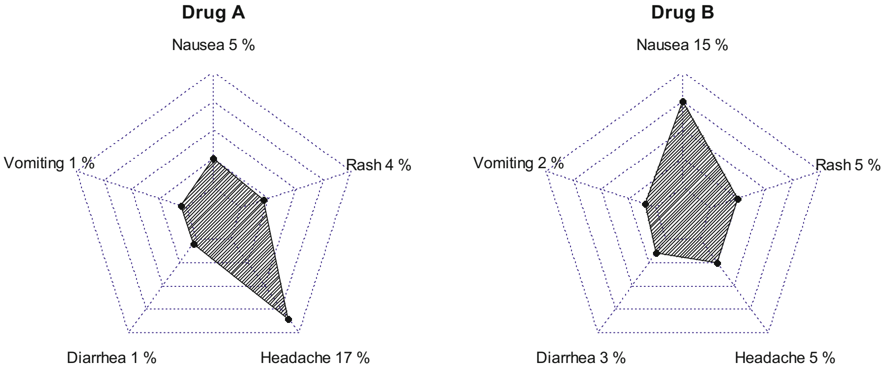


Fig. 3.4 Radar chart: example 3

```
> D=read.table("C://mydata1.csv",head=TRUE,sep=";")
> head(D)
  Firm Intensity Patents Subsidy Sector
1    1      0.23      42       1      3
2    2      0.19      31       0      3
3    3      0.20      20       0      3
4    4      0.19      33       1      3
5    5      0.18      30       1      3
6    6      0.18      43       0      3

> # BAR GRAPH: RELATIVE FREQUENCIES
> table(D$Sector)
 1  2  3
21 24 15
> relat.freq=prop.table(table(D$Sector))
> rownames(relat.freq)=c("Low-technology","Medium-technology",
+ "High-technology")
> myplot=barplot(relat.freq,main="1.1 Bar graph: relative
+ frequencies",ylab="Percentage of firms",ylim=c(0,0.5))
> text(myplot,relat.freq,c("35%","40%","25%"),pos=3)

> # BAR GRAPH: JOINT FREQUENCIES
> joint.freq=table(D$Sector,D$Subsidy)
> colnames(joint.freq)=c("Non-recipients","Recipients")
> rownames(joint.freq)=c("Low-technology","Medium-technology",
+ "High-technology")
> joint.freq

                    Non-recipients Recipients
  Low-technology                18          3
  Medium-technology             16          8
  High-technology                6          9

> myplot=barplot(joint.freq,beside=TRUE,legend=rownames(joint.freq),
+ main="1.2 Bar graph: joint frequencies",
+ ylab="Number of firms",ylim=c(0,25))
> text(myplot,joint.freq,joint.freq,pos=3)

> # CIRCLE GRAPH: RELATIVE FREQUENCIES
> myplot=pie(relat.freq,labels=paste(rownames(relat.freq),
+ relat.freq*100,"%"), col=gray.colors(3),main="1.3 Circle
+ graph: relative frequencies")

> # HISTOGRAM: ABSOLUTE FREQUENCIES
> myclasses=c(0,0.035,0.07,0.105,0.14,0.175,0.21,0.245)
> hist(D$Intensity,breaks=myclasses, main="2.1 Histogram:
+ absolute frequencies", xlab="R&D intensity",col="gray",ylim=c(0,12))

> # HISTOGRAM: ADJUSTED FREQUENCIES
> hist(D$Intensity,breaks=myclasses, freq=FALSE, main="2.2
+ Histogram: adjusted frequencies",
+ xlab="R&D intensity",col="gray",ylim=c(0,6))

> # KERNEL DENSITY PLOT
> plot(density(D$Intensity),main="2.3 Kernel density plot",
+ xlab="R&D intensity")

> # SCATTER PLOT
> plot(D$Patents~D$Intensity,main="2.4 Scatter plot",
+ xlab="R&D intensity",ylab="Number of patents",pch=19)
```

Fig. 3.5 Graphs with R-CRAN: example 1

Table 3.10 Main graph options in R-CRAN[a]

Options	Definition
type	Character string giving the type of plot desired. The following values are possible: “p” for points, “l” for lines, “b” for both points and lines, “c” for empty points joined by lines, “o” for over-plotted points and lines, “s” and “S” for stair steps and “h” for histogram-like vertical lines. Finally, “n” does not produce any points or lines.
xlim	The *x* limits of the plot.
ylim	The *y* limits of the plot.
main	Main title for the plot.
sub	Subtitle for the plot.
xlab	Label for the x axis.
ylab	Label for the y axis.
col	The colors for lines and points. Multiple colors can be specified so that each point can be given its own color: see Fig. 3.6.
bg	Vector of background colors for open plot symbols.
pch	Vector of plotting characters or symbols: see Fig. 3.6.
cex	Numerical vector giving the amount by which plotting characters and symbols should be scaled relative to the default.
lty	Vector of line types: 1 = solid, 2 = dashed, etc.
lwd	Vector of line widths.

[a] Additional information is available using “ ?***plot . default*** ” in the R-CRAN console

similar to the first rows of Table 3.4. This is a quick way to check that you are working with the right file. To create Fig. 3.1a one needs to build a frequency table using the class intervals defined in the previous section. This is done thanks to *table* (*D*$*Sector*) which provides the absolute frequency of the variable *Sector*. The dollar sign ($) in R-CRAN is used to identify each component of a variable of interest (here *Sector*) in database *D*. The relative frequencies are obtained with *prop . table* (). Those frequencies are saved in an additional item called *relat .freq*. By using the function *rownames* we are able to replace the codes (1,2,3) of variable *Sector* by text names ("Low-technology", "Medium-technology", "High-technology"). The command *c*() is frequently used in R-CRAN to combine elements into a vector. A *barplot* is then created and saved under the name *myplot* for a subsequent use.

As can be seen, several options are included inside the *barplot* command, such as *main* for the main title, *ylab* for the name of the vertical axis, or *ylim* for the limits of the vertical axis. As they are similar from one graph to another, those options are detailed in Table 3.10. The command *text* is used to display the percentage values at the top of the bars. The vector *c*("35 %", "25 %", "40 %") is drawn such that the coordinates correspond to *myplot* (horizontal axis) and *relat .freq* (vertical axis). Option $pos = 3$ indicates that the text should be placed above the specified coordinates.

Figure 3.1b is constructed in a similar manner excepted that the focus is now on two variables and their joint frequencies. The frequency distribution is obtained using the *table* command and saved under the name *joint .freq*. The names of the columns and rows are specified accordingly. The command *barplot* draws the graph and a legend is included to help the reader differentiate the different colors.

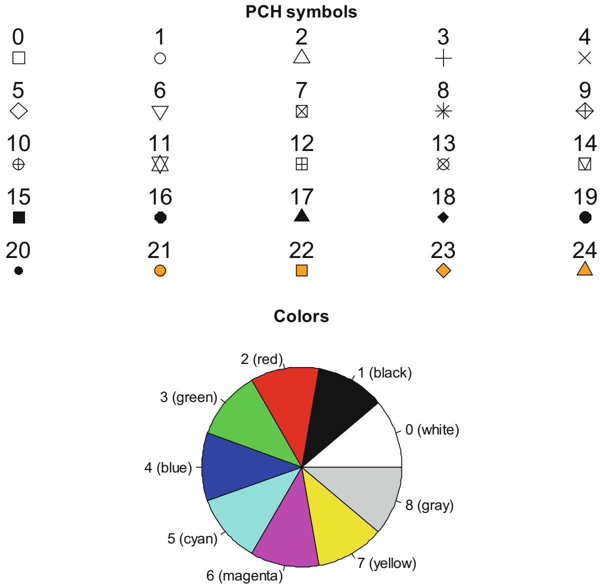
Command *beside* = *TRUE* allows to avoid the piling-up of bars. More information is available by typing " *?barplot* ", " *?legend* " or " *?plot . default* " in the R-CRAN console.

In Fig. 3.5, the *pie* command is used to draw the circle graph of Fig. 3.1c. The item *relat . freq* stands for the relative frequency of variable *sector*. It was previously created with function *table*. The values of the pie slices are included in the graph with *labels*. The command *paste* is very useful in this matter. It glues several objects together, here the name of the sectors (*rownames*(*relat . freq*)) with the value of the relative frequencies (*relat . freq* ∗ 100), and the % sign.

Figure 3.5 finally provides the codes for Fig. 3.2. First, a variable *myclasses* is created to specify the class intervals. The command *hist* is then used to draw the histogram of absolute frequencies (Fig. 3.2a). The option *breaks* uses the vector *myclasses* to define the bins. R-CRAN determines those classes automatically when this option is not entered. The inclusion of the option *freq* = *FALSE* in the *hist* command generates the adjusted frequencies of Fig. 3.2b. Last, function *plot* (*density*()) allows the probability density function to be drawn in a smooth and continuous way as in the Kernel representation of Fig. 3.2c. The command *plot* is also used to draw the scatter plot of Fig. 3.2d. Using the ~ sign, the command specifies the variable on the vertical axis (here *D*$*Patents*) as a function of the variable on the horizontal axis (here *D*$*Intensity*). The term *pch* = 19 specifies the type of symbol (see also Fig. 3.6).

To obtain the line graph of Fig. 3.3, one makes use of the *matplot* function as in Fig. 3.7. First, the database is uploaded under the name *E* in the R-CRAN environment. Notice that the column "years" (see Table 3.8) is not included as a variable but as the name of the rows. Several options are specified in the *matplot* command:

Fig. 3.6 *Colors* and *symbols* used in R plot



E denotes the data, $type = "b"$ indicates that both a line and a symbol are plotted; $pch = 1:5$ defines the type of symbol (symbol 1 for the first jurisdiction, symbol 2 for the second, and so on); $col = 1:5$ stands for the color (color 1 for the first jurisdiction, color 2 for the second, and so on); $xaxt = "n"$ suppresses plotting of the axis. The *xaxt* command is used because we need to put years as the x -axis. This is done with the command *Axis* where $side = 1$ specifies the axis to be modified (the horizontal axis), $at = c(1, 2, 3, 4, 5, 6, 7)$ represents the horizontal coordinates, and $rownames(E)$ are the new labels.

Loaded with the command *library*, the package *fmsb* provides several functions for medical and health data analysis. Figure 3.7 codes the *radarchart* function. The raw data are uploaded with the *read.table* command. Two vectors, *MIN* and *MAX*, are created. They specify the lower and upper limits of the axes. We need to draw two radar charts in the same box in order to compare the side-effects of the two

```
> # LINE GRAPH

> E=read.table("C://mydata2.csv",head=TRUE,sep=";",row.names="Year")
> E
     City.A City.B City.C City.D City.E
1950   6035  12301   4351   1987   8864
1960   6666  13190   4759   2195   9317
1970   8126  14715   4904   2307  10090
1980  10921  16743   4953   2812  11481
1990  11480  18313   5525   3107  12974
2000  12681  19636   5808   4175  15816
2010  12936  22123   6416   4612  17470

> matplot(E,type="b",pch=1:5,col=1:5,xaxt="n",xlab="Year",
+ ylab="Number of inhabitants")
> Axis(side=1,at=c(1,2,3,4,5,6,7),labels=rownames(E))
> legend("topleft",legend=colnames(E),pch=1:5,col=1:5)

> # RADAR CHART

> F=read.table("C://mydata3.csv",head=TRUE,sep=";",
+ row.names="Treatment")
> F
        Nausea Vomiting Diarrhea Headache Rash
Drug A    0.05     0.01     0.01     0.17 0.04
Drug B    0.15     0.02     0.03     0.05 0.05

> library(fmsb)
>
> MAX=c(0.2,0.2,0.2,0.2,0.2)
> MIN=c(0,0,0,0,0)

> # par(mar=c(bottom,left,top,right),mfrow=c(n of rows, n of col))
> par(mar=c(1,1,2,1),mfrow=c(1,2))

> labelA=paste(colnames(F),F[1,]*100,"%")
> labelB=paste(colnames(F),F[2,]*100,"%")

> radarchart(rbind(MAX,MIN,F[1,]),title="Drug A",
+ pdensity=c(40),vlabel=labelA)
>radarchart(rbind(MAX,MIN,F[2,]),title="Drug B",
+ pdensity=c(40),vlabel=labelB)
```

Fig. 3.7 Graphs with R-CRAN: examples 2 and 3

treatments. To do so, the command $par(mar = c(1, 1, 2, 1))$ modifies the margins using a numerical vector of the form $c(bottom, left, top, right)$ that gives the number of lines to be specified on the four sides of the plot. The term $mfrow = c(1, 2)$ specifies the number of boxes/graphs to be drawn (1 row and 2 columns). Extend the graph window in the R-CRAN console so that all the text appears on the screen if you wish to copy for further use. Two additional vectors are created, *labelA* and *labelB*. They are used to rename the axes so that they include the value of each observation. The *paste* command is used in this respect: $colnames(F)$ stands for the names of the side effects; $F[1,]*100$ and $F[2,]*100$ are the relative frequencies observed for drug A and drug B, respectively; "%" indicates the percentage sign. The command *radarchart* is then used to plot the graph. The *rbind* command allows the data frame to include the vectors of maximum and minimum values. Option $pdensity = 40$ specifies the filling density of polygons and *vlabel* indicates the names of the variables.

3.4 Measures of Central Tendency and Variability

It is often easier to interpret data when they are presented graphically rather than as a table. Another option to summarize information is to rely on numerical indicators. Two types of indicators are used in this respect: measures of central tendency and measures of variability. The first approach consists in condensing the set of data into a single value that represents the middle or center of the distribution. The usual measures include the mode, the median, and the mean. Measures of variability are used to describe dispersion in a set of data. Typical examples are the range, the interquartile range, the variance, the standard deviation and the coefficient of variation.

It is important to remember that a difference exists between a sample and a population. While the population includes all of the elements from a set of data, a sample is only a subset of it. Any measure that results from a sample is referred to as a statistic. Measures that refer to population attributes are termed parameters. For instance, in example 1, the average number of patent claims is a sample statistic. If one were able to survey the whole set of firms, it would be a population parameter. Upper-case letters are often used to denote population parameters (e.g., N =population size), while lower-case letters refer to sample statistics (e.g., n =sample size). By convention, specific symbols are used to represent certain statistics and parameters. For the sample,$\bar{x}$ is commonly used for the mean, s for the standard deviation, p for a proportion. For the population, μ is employed for the mean, σ for the standard deviation and π for a proportion.

The mode, or modal value, is the most frequently occurring value in a set of data. Sometimes there are multiple modes. Graphically, it shows up on a bar graph or a histogram as the highest column. To find the mode, one can also make use of a frequency table. For example, from Table 3.5a, we can see that the mode of *Subsidy* is “No” while for variable *Sector* it is “medium-tech” (Table 3.5b). With numerical variables, the data must be grouped into classes. In this case, the mode is also referred to as the modal interval. For instance, in Table 3.5c, variable *Intensity* is characterized by three modes: [7%, 10.5%], [10.5%,14%] and [17.5%,21%]. With continuous variables, the mode can also be expressed as the center of the class.

The mean (or arithmetic average) is the sum of the values of each observation in a dataset divided by the number of observations. Depending on whether we are dealing with the population or with a sample, the formulas are written as:

$$\mu = \frac{1}{N}\sum\nolimits_{i=1}^{N} X_i \text{ (population)}$$

and

$$\bar{x} = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i \text{ (sample)}$$

where N is the size of the population, n the sample size, X_i and x_i denotes the i th observation in the population or the sample. For instance, using information from Table 3.4, the mean of *intensity* is computed as:

$$\bar{x} = \frac{1}{60}(0.23 + 0.19 + \ldots + 0.08 + 0.01) = 0.109$$

When observations are arranged in ascending order, the median is the value that falls in the middle of the data. This measure is for instance used to provide information about the distribution of household income. The median income is the value that divides the income distribution into two equal groups, half having income above that amount, and half having income below it. If there is an odd number of observations, the median value is simply the middle value. Imagine for instance that we have information about the grade of eleven students:

$$\underbrace{11;\ 17;\ 21;\ 34;\ 42;}_{\text{Lower half}}\ \mathbf{51};\ \underbrace{67;\ 71;\ 82;\ 93;\ 99}_{\text{Upper half}}$$

The median is 51. If there is an even number of observations, the median value is the mean of the two middle values. For instance, using data from Table 3.4, the value of *Intensity* can be rearranged by increasing order:

0; 0; 0.01; 0.01; 0.01; 0.02; 0.02; 0.02; 0.02; 0.04; 0.04; 0.04; 0.05; 0.05; 0.06; 0.07; 0.07; 0.07; 0.07; 0.08; 0.08; 0.08; 0.08; 0.08; 0.08; 0.08; 0.09; 0.09; 0.1; **0.1**; **0.11**; 0.11; 0.11; 0.11; 0.12; 0.12; 0.13; 0.13; 0.14; 0.14; 0.14; 0.16; 0.16; 0.16; 0.17; 0.17; 0.17; 0.18; 0.18; 0.18; 0.18; 0.19; 0.19; 0.2; 0.2; 0.2; 0.21; 0.21; 0.23; 0.23

The median in that case is equal to (0.1+0.11)/2=0.105.

The mode, the mean and the median provide different types of information. The mean uses every value in the data to provide a central measure. It is thereby sensitive to extreme values or outliers. The median and the mode, on the other hand, are not affected by outliers. The median provides a useful and simple description of central tendency: half of the values are below it. The mode is the only measure of central tendency that can be used for nominal variables. It is the most frequently occurring value or class in a set of data. Yet, this does not mean that it always stands for the center of the distribution. It is also possible that more than one mode exist for the same variable. More important, the mode of a numerical variable depends on how the class intervals are constructed.

Measures of dispersion indicate how spread out the observations are. The range is calculated as the highest value minus the lowest value. It provides a simple measure but is very sensitive to outliers. Other alternatives exist. The interquartile range for instance is the difference between the upper and lower quartiles:

$$IQR = Q_3 - \ Q_1$$

In descriptive statistics, the quartiles are the three points that divide a ranked set of values into four equal groups. The first quartile Q_1 is defined as the median of the lower half of the data set. About 25% of the observations lie below Q_1 and about 75% lie above. The second quartile Q_2 is the median of the data. The third quartile Q_3 is the middle value of the upper half of the data set. About 75% of the numbers in the data set lie below Q_3 and about 25% lie above.

Consider the previous distribution of marks. In Excel and R-CRAN, when there is an odd number of observations, the approach consists in including the median in both halves:

$$(\underbrace{11;\ 17;\ \mathbf{21};\ \mathbf{34};\ 42;\ \mathbf{51}}_{\text{Lower half}});\ (\underbrace{\mathbf{51};\ 67;\ \mathbf{71};\ \mathbf{82};\ 93;\ 99}_{\text{Upper half}})\ \text{(Case 1)}$$

The interquartile range is then computed as:

$$Q_1 = \frac{21+34}{2} = 27.5; Q_3 = \frac{71+82}{2} = 76.5; IQR = 76.5 - 27.5 = 49$$

On the other hand, when there is an even number of observations, the approach consists in splitting the data set exactly in half. Two cases arise depending on the number of observations in those two halves. Consider the following distribution:

$$\underbrace{11;\ 17;\ \mathbf{21};\ \mathbf{34};\ 42;\ 51}_{\text{Lower half}};\ \underbrace{67;\ 71;\ \mathbf{82};\ \mathbf{93};\ 94;\ 99}_{\text{Upper half}}\ \text{(Case 2)}$$

There is a total of 12 data points. Both halves also contain an even number of observations. The lower quartile is 25% of the third data value (21) plus 75% of the fourth value (34) while the upper quartile is 75% of the ninth value (82) plus 25% of the tenth value (93):

$$Q_1 = \frac{1}{4}\ 21 + \frac{3}{4}\ 34 = 30.75; Q_3 = \frac{3}{4}82 + \frac{1}{4}93 = 84.75; IQR = 54$$

Let us now examine another distribution with 10 observations:

$$\underbrace{11;\ 17;\ \mathbf{21};\ \mathbf{34};\ 42}_{\text{Lower half}};\ \underbrace{51;\ \mathbf{67};\ \mathbf{71};\ 82;\ 93}_{\text{Upper half}}\ \text{(Case 3)}$$

There is now an odd number of observations in both halves. The lower quartile is 75% of the third data value (21) plus 25% of the fourth value (34) while the upper quartile is 25% of the seventh value (67) plus 75% of the eighth value (71):

$$Q_1 = \frac{3}{4}\,21 + \frac{1}{4}\,34 = 24.25; Q_3 = \frac{1}{4}67 + \frac{3}{4}71 = 70; IQR = 45.75$$

In Excel, the corresponding formula is $QUARTILE(array, quart)$ where *array* denotes the range of data values for which one wants to calculate the specified quartile and *quart* is an integer representing the required quartile. In R-CRAN, the command *summary* offers those quartiles automatically.

To illustrate the methodology, let us examine variable *Intensity* from Table 3.4. We use the previous rearrangement of the observations (by increasing order) and now split them exactly in two halves with 30 observations in each. This corresponds to case 2 (even number of observations in both halves).

(0; 0; 0.01; 0.01; 0.01; 0.02; 0.02; 0.02; 0.02; 0.04; 0.04; 0.04; 0.05; 0.05; **0.06**; **0.07**; 0.07; 0.07; 0.07; 0.08; 0.08; 0.08; 0.08; 0.08; 0.08; 0.08; 0.09; 0.09; 0.1; 0.1); (0.11; 0.11; 0.11; 0.11; 0.12; 0.12; 0.13; 0.13; 0.14; 0.14; 0.14; 0.16; 0.16; 0.16; **0.17**; **0.17**; 0.17; 0.18; 0.18; 0.18; 0.18; 0.19; 0.19; 0.2; 0.2; 0.2; 0.21; 0.21; 0.23; 0.23)

The first quartile is computed as:

$$Q_1 = \frac{1}{4} \times 0.06 + \frac{3}{4} \times 0.07 = 0.0675$$

For the third quartile, we have:

$$Q_3 = \frac{3}{4} \times 0.17 + \frac{1}{4} \times 0.17 = 0.17$$

The interquartile range of *Intensity* is thus $IQR = 0.17 - 0.0675 = 0.1025$.

Two other common measures of variability are the variance and the standard deviation. For the population, the variance is expressed as the sum of the squares of the differences between each observation and the mean, divided by the population size:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}X_i^{\,2} - \mu^2 \text{ (population)}$$

For the sample, the formula is slightly different. We have:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^{\,2} - \frac{n}{n-1}\bar{x}^2 \text{ (sample)}$$

The standard deviation is expressed as the square root of the variance. It can be interpreted as the approximate average distance from the mean:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum\nolimits_{i=1}^{N} (X_i - \mu)^2} \text{ (population)}$$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2} \text{ (sample)}$$

Using information from Table 3.4, the variance of *Intensity* is computed as:

$$s^2 = \frac{1}{60-1}\left[(0.23 - 0.109)^2 + \cdots + (0.01 - 0.109)^2\right] \approx 0.00427$$

The standard deviation amounts to $\sqrt{0.00427} \approx 0.065$.

The standard deviation expresses the dispersion of data in the same unit as the original observations. If all values are the same and thus equal to the mean, then the standard deviation is null. The concept should not be mistaken with that of the standard error (*se*). The latter denotes the standard deviation of the sampling distribution of a statistic (such as a mean, proportion, etc.) and is used when one wants to make statistical inference about a defined population.

As we have seen, the formulas for the sample standard deviation and variance differ from population formulas. When calculating a sample variance, we divide by $(n - 1)$ instead of dividing by N. The reason behind is that we need to ensure that the sample variance is an unbiased estimator of the population variance. Simply put, the value of a statistic is likely to be above or below the true value of the population parameter. The mean of the sampling distribution could however converge to the population parameter. If this is to be the case, the sample statistic is said to be an unbiased estimator of the population parameter. While this holds true for the sample mean (the mean of the distribution of sample means is the mean of the population), this is not the case mathematically for the variance. By construction, the mean of the distribution of sample variances is not the variance of the population. Using a factor $(n - 1)$ to weight the sample variance however ensures that this condition is verified.

The coefficient of variation offers a relative measure of dispersion:

$$c_v = \frac{\sigma}{\mu} \text{ (population)}; \widehat{c}_v = \frac{s}{\bar{x}} \text{ (sample)}$$

The coefficient of variation is usually expressed in percentage and it is very useful as it allows distributions with different scales to be compared. In Table 3.11, for instance, we have information about the per capita expenditures of a set of local jurisdictions. The standard deviation points out an important heterogeneity with respect to social assistance. The average distance from the mean amounts approximately to $42. If one were to reduce disparities among jurisdictions, this category of

Table 3.11 Expenditures per capita: example 4

Variable	Mean	Standard deviation	Coefficient of variation (%)
Social assistance	\$226	\$42	18.58
Education	\$90	\$26	28.89
Culture	\$22	\$5	22.73
Police	\$39	\$13	33.33
Fire protection	\$67	\$27	40.30

Table 3.12 Variance and standard deviation in R-CRAN and Excel

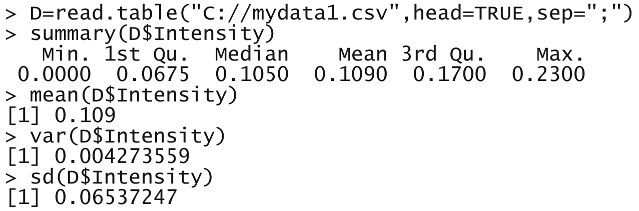
Context	Formula	R-CRAN	Excel
Sample variance	$s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$	*var*(*x*)	*VAR* or *VAR . S* depending on the version
Population variance	$\sigma^2 = \frac{1}{N}\sum\limits_{i=1}^{N}(X_i - \mu)^2$	(*length*(*x*) − 1)/ *length*(*x*)* *var*(*x*)	*VAR . P*
Sample standard deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{Nn}(x_i - \bar{x})^2}$	sd(x)	*STDEV* or *STDEV . S* depending on the version
Population standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$	sqrt((length(x)-1)/ length(x))* sd(x)	*STDEV . P*

expenditures would appear as a priority, all the more so that it represents by far the main expenditure item. Yet, proportionally speaking, jurisdictions are more unequally distributed with regard to fire protection. The coefficient of variation amounts to 40.30%, which means that the jurisdictions differ by about 40% from the average level. Understanding those disparities is also at the heart of public evaluation.

Table 3.12 provides a summary of the built-in functions available in Excel and R-CRAN. As can be seen, the names can be misleading. For instance, in both R-CRAN and Excel the command *var* denotes the variance of the sample, and not the variance of the population. Figure 3.8 illustrates the method in R-CRAN using data from example 1. The *summary* command provides the min, the max as well as the quartiles of variable *Intensity*. Commands *mean*, *var*, and *sd* offer the mean, the variance and the standard deviation, respectively.

3.5 Describing the Shape of Distributions

A fundamental task in describing the nature of a statistical variable is to portray the shape of its distribution. In this context, the normal or Gaussian distribution is commonly used as a point of reference. It describes a class of distributions that are described by two parameters: the mean μ and the standard deviation σ. The probability density functions shown in Fig. 3.9 provide illustrations of normal



```
> D=read.table("C://mydata1.csv",head=TRUE,sep=";")
> summary(D$Intensity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0675  0.1050  0.1090  0.1700  0.2300
> mean(D$Intensity)
[1] 0.109
> var(D$Intensity)
[1] 0.004273559
> sd(D$Intensity)
[1] 0.06537247
```

Fig. 3.8 Central tendency and variability in R-CRAN: example 1

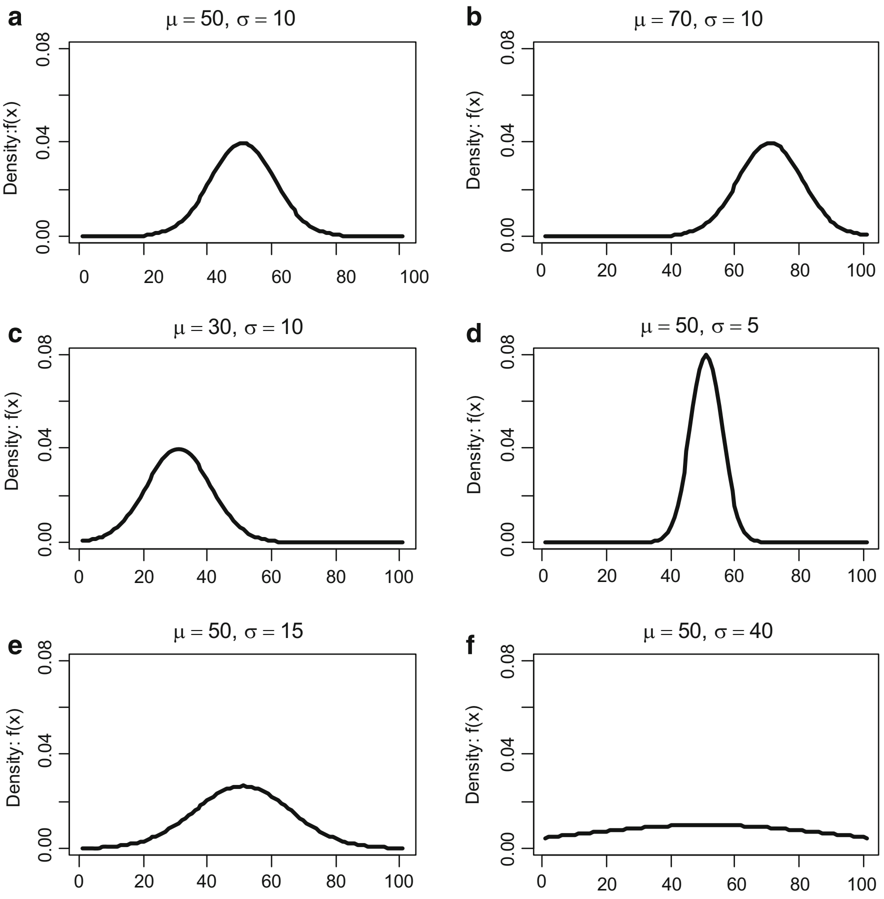


Fig. 3.9 Examples of normal distributions

distributions. There is only one mode, which is equal to the mean and the median. The shape of those distributions is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distributions are extremely important because they describe many real events. They also play an essential role in inferential statistics.

A shown in Fig. 3.9, a normal distribution is characterized by a symmetric bell-shaped curve. Most observations are clustered around the center of the distribution. The shape of the curve differs depending on the values of μ and σ. When the mean parameter μ increases from 50 (Fig. 3.9a) to 70 (respectively 30), then the distribution shifts to the right-hand (respectively left-hand) side of the graph (Figure 3.9b) (respectively Figure 3.9c). When the standard deviation decreases from 10 (Figure 3.9a) to 5 (Figure 3.9d), then the mass of the distribution is more concentrated around the mean. When the standard deviation increases, the distribution is more and more dispersed (Fig. 3.9e, f).

Two indicators are used to compare the distribution of a variable with the shape of a normal curve: skewness and kurtosis. The sample skewness (or Fisher-Pearson coefficient of skewness) measures the asymmetry of the distribution:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \quad \text{(sample)}$$

By construction, a normal distribution has a skewness of 0. If $g_1 = 0$, the observations are evenly distributed on both sides of the mean, implying a symmetric distribution. If $g_1 < 0$, the left tail of the distribution is longer and the mass of the observations is concentrated on the right. In this case, the distribution is said to be left-skewed, left-tailed, or skewed to the left. If $g_1 > 0$, the right tail is longer and the mass of the distribution is concentrated on the left. The distribution is said to be right-skewed, right-tailed, or skewed to the right.

In statistics, the denominator of a ratio frequently serves as a weight to normalize a measure. This is typically the case with the skewness coefficient. What matters in the previous formula is actually the numerator. We can see that it is defined as the sum of cubes of differences with the mean. Unlike the variance which measures a sum of positive distances, the skewness coefficient accounts for the fact that an observation can be positioned either to the left or to the right of the mean value. The numerator is thus composed of negative and positive values. If those negative values outweigh the positive ones, then the left tail of the distribution is longer, as illustrated in Fig. 3.10a. In that case, the mean lies toward the direction of skew, i.e. on the left, relative to the median. In contrast, if some observations attain high values, the distribution is right-skewed, as shown in Fig. 3.10b. The mean lies to the right of the median because those extreme values serve to compute the mean, and not the median. A symmetric distribution is characterized by observations that are equally distributed around the central position (Fig. 3.10c).

The skewness formula may vary from one statistical software to another. For instance Excel uses:

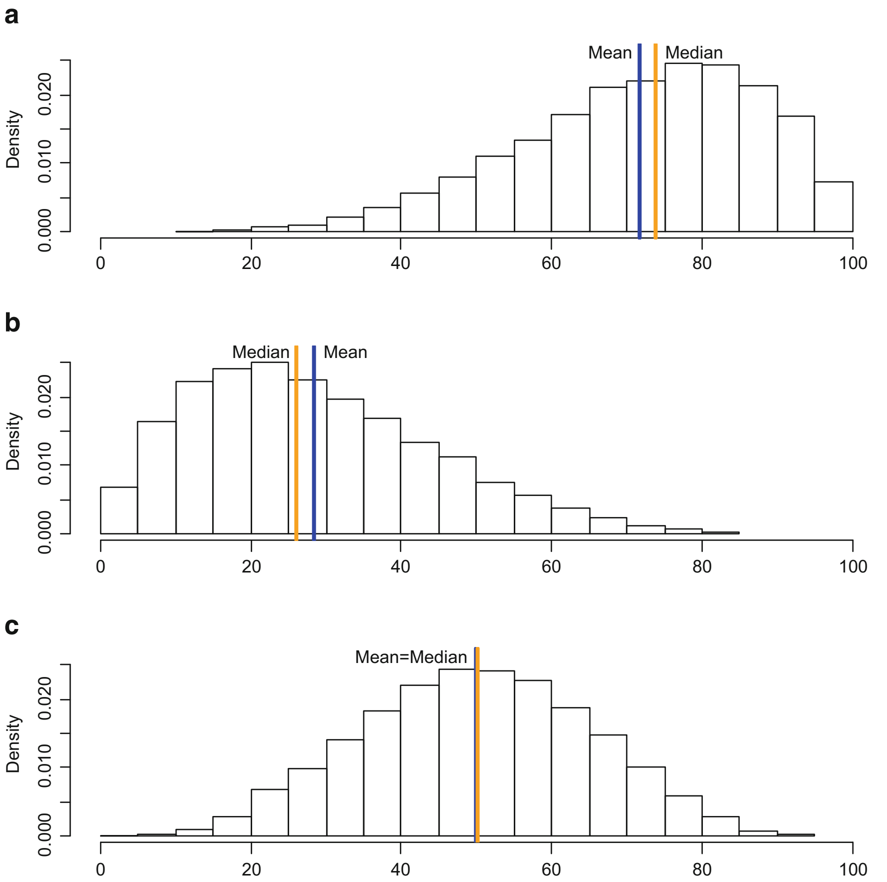


Fig. 3.10 Examples of asymmetric and symmetric distributions. (**a**) Left-tailed distribution: skewness = −0.6. (**b**) Right-tailed distribution: skewness = 0.6. (**c**) Symmetric distribution: skewness = 0

$$G_1 = \frac{n^2}{(n-1)(n-2)} g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s}\right]^3 \text{ (sample)}$$

Choosing one or another measure is of less importance. They converge as the sample size increases.

Note that it is possible to approximate the asymmetry of a distribution by computing what is termed the median skewness coefficient. It is a measure based on the difference between the sample mean and median:

$$\text{median skewness coefficient} = \frac{3 \times (\text{mean} - \text{median})}{\text{standard deviation}}$$

Skewness is positive (respectively negative) if mean is greater (respectively lower) than median. With this measure, one may however reach conclusions that are different from g_1 and G_1. The fact that the mean is right of the median does not necessarily imply that the distribution is right skewed as defined with g_1 and G_1.

Another measure that is frequently used to characterize the shape of a distribution is the kurtosis:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \text{ (sample)}$$

It assesses whether a distribution is heavy-tailed or light-tailed relative to a normal distribution. By construction, a normal distribution has a kurtosis of 3. If $g_2 > 3$, the distribution is said to be leptokurtik (or heavy-tailed) and characterized by a high degree of peakedness. As shown in Fig. 3.11a, observations are mainly distributed around the mode. If $g_2 < 3$, the distribution is described as platykurtik (or light-tailed). It is characterized by a high degree of flatness, as depicted in Fig. 3.11b. Last, if $g_2 = 3$ the distribution is mesokurtic. The observations are dispersed around the mean like any normal distribution (Fig. 3.11c). The kurtosis measure is sometimes normalized and computed as "$g_2 - 3$". This measure is termed excess kurtosis. For instance, it is the one that is used in Excel.

Another useful tool for describing the shape of a distribution is the box plot. This graph depicts a variable through its quartiles by representing the distances between the minimum, the lower quartile Q_1, the median, the upper quartile Q_3 and the maximum. An illustration is provided in Fig. 3.12. A rectangle is drawn so that the left side of the box corresponds to Q_1 and the right side to Q_3. When box plots are displayed horizontally as in Fig. 3.12, the width of the rectangle represents the interquartile range and, as such, contains 50% of the observations. The height of the rectangle does not represent anything in particular. Observations outside the rectangle are considered as extremes values. Inside the box, a vertical line is drawn representing the median. The minimum and maximum are described by two vertical lines positioned at the extremities of the graph. Box plots can be drawn either horizontally or vertically.

Despite its simplicity, a box plot provides detailed information about the shape of a distribution. To illustrate, Fig. 3.13 gives the box plots associated with the histograms of Figs. 3.10 and 3.11. For vertically displayed boxes, height gives an indication of the variance and the line inside the box provides a measure of central tendency. The lines extending vertically from the box, known as whiskers, indicate variability outside the upper and lower quartiles. The whiskers can tell us whether the sample is skewed, either to the left or to the right. When the distribution is symmetric, long whiskers (compared to the box length) mean that the distribution is heavy-tailed (leptokurtic distribution). If, on the other hand, the whiskers are shorter than the box, then the distribution is short-tailed (platykurtic distribution).

Let us now consider the data of Table 3.4 (example 1). Figure 3.14 provides the code to be used in R-CRAN to construct the box plots of *Intensity* and *Patents*. The *boxplot* command is used to draw the graphs of Fig. 3.15. Both distributions are

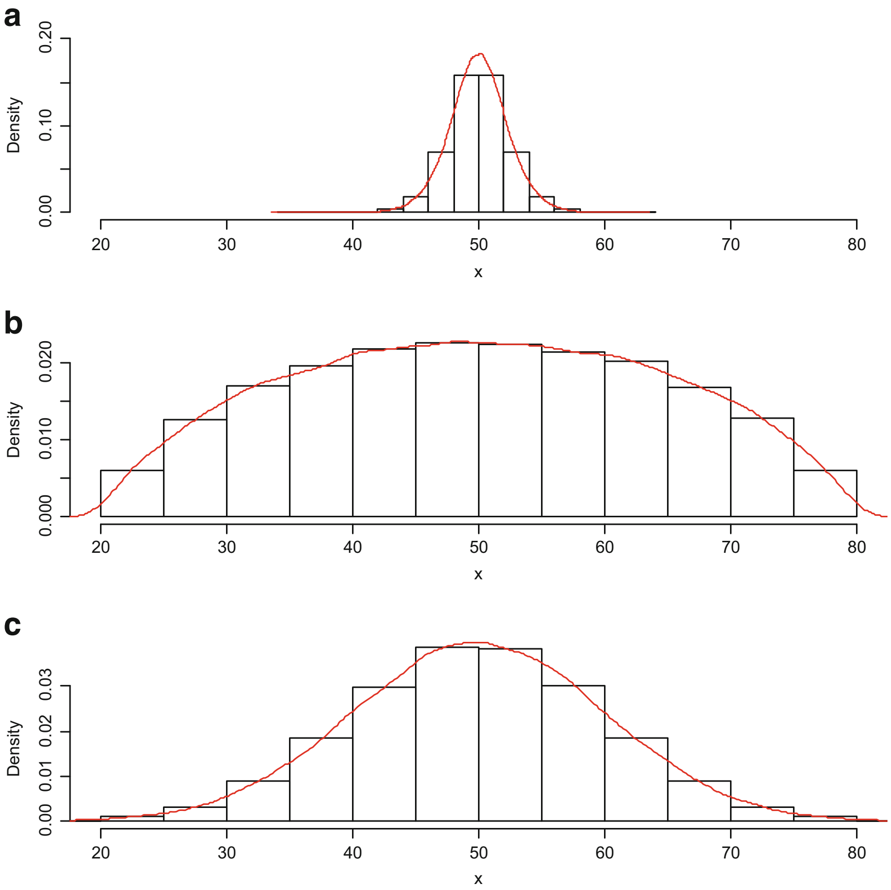


Fig. 3.11 Examples of heavy-tailed and light-tailed distributions. (**a**) Leptokurtic distribution: kurtosis = 3.72. (**b**) Platykurtic distribution: kurtosis = 2.06. (**c**) Mesokurtic distribution: kurtosis = 3

right-skewed. There is some inequality in the distribution of firms with respect to those variables. The package *moments* is uploaded so that we can compute the skewness and kurtosis. The numerical results confirm the graphical analysis. The mean is higher than the median. Both skewness coefficients are positive and the excess kurtosis coefficients are lower than 3, indicating platykurtic distributions.

Note that box plots produced by statistical packages may differ from those presented above. Extreme values are sometimes removed from the graph to draw attention to the values that are unusually far away from the distribution. In that case, outliers are represented as circles or asterisks beyond the bounds of the whiskers. Hence, the whiskers do not extend to the most extreme data points.

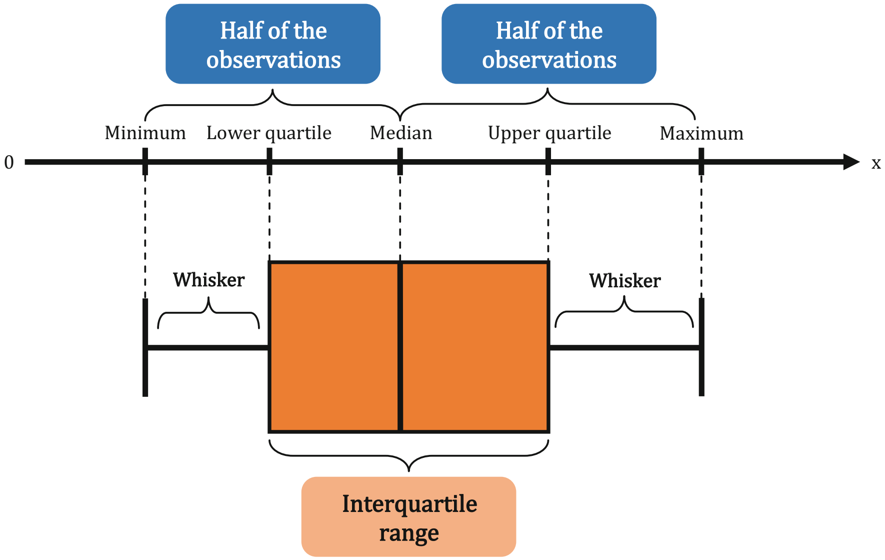


Fig. 3.12 How to read a box plot

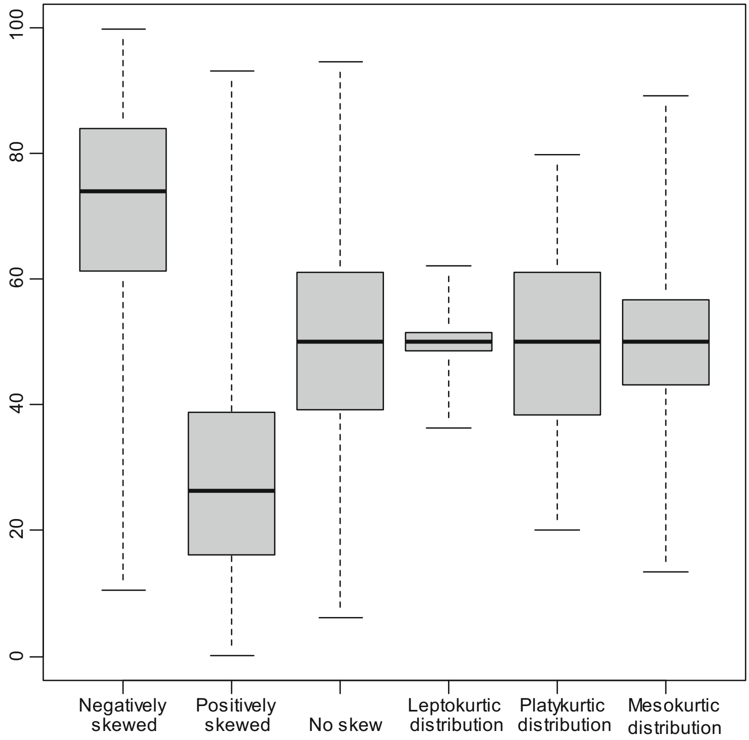


Fig. 3.13 Box plot for visualizing density (examples of Figs. 3.10 and 3.11)

```
> par(mar=c(2,2,2,1))
> par(mfrow = c(1,2))

> D=read.table("C://mydata1.csv",head=TRUE,sep=";")
> boxplot(D$Intensity,col="grey",main="R&D intensity")
> boxplot(D$Patents,col="grey",main="Number of patents")

> library(moments)
> mean(D$Intensity)
[1] 0.109
> median(D$Intensity)
[1] 0.105
> skewness(D$Intensity)
[1] 0.07895221
> kurtosis(D$Intensity)
[1] 1.91517

> mean(D$Patents)
[1] 15.85
> median(D$Patents)
[1] 13
> skewness(D$Patents)
[1] 0.7711708
> kurtosis(D$Patents)
[1] 2.672576
```

Fig. 3.14 Box plots with R-CRAN: example 1

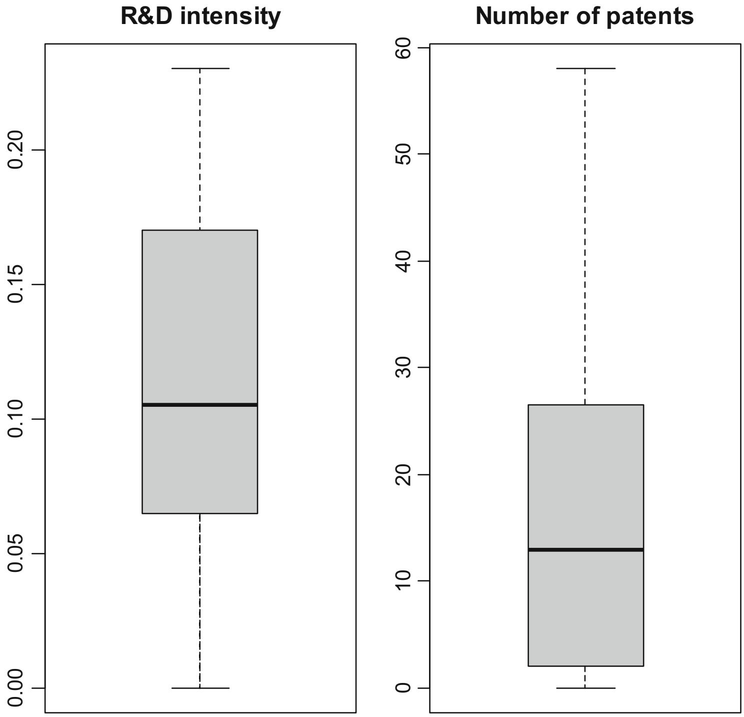


Fig. 3.15 Box plots for example 1

3.6 Computing Confidence Intervals

The computation of a confidence interval is crucial if one wants to generalize the sample findings to the population of interest. The approach consists in providing a range of values $\pm e$ above and below the sample statistic which is likely to contain the parameter of the population we are studying. The width of the interval depends on the desired confidence level but also on the probability distribution assumed behind observed data. Often, one chooses a 95% confidence level, so that 95% of the estimated intervals would include the true parameter.

To understand the principle of inferential statistics, let us consider a simple illustration (example 5). Assume that a radar unit is used to measure the speed of cars on a highway. This variable is assumed to be normally distributed. The sample standard deviation is found to be $s = 16km/hr$ while the sample mean is $\bar{x} = 128km/hr$. We can use this information to estimate the probability function for each value x of X:

$$f(x) = \frac{1}{16\sqrt{2\pi}} e^{-\frac{(x-128)^2}{2\times 16^2}}$$

This is the probability density of a normal distribution (see Sect. 3.5). A graph of it is provided in Fig. 3.16. If we want to calculate the probability that a car picked at random is travelling at less than x km/hr, we need to compute the area beneath the density curve that lies to the left of x, and thus compute the following integral:

$$\Pr\{X < x\} = F(x) = \int_{-\infty}^{x} \frac{1}{16\sqrt{2\pi}} e^{-\frac{(t-128)^2}{2\times 16^2}} dt$$

where F denotes the cumulative distribution function. For instance, in Fig. 3.16, this area amounts to 91.5% when $x = 150$ (area displayed in grey). To compute this probability we can implement $pnorm(150, mean = 128, sd = 16)$ in R-CRAN. This equivalently means that the probability that a car is travelling at more than 150 km/h is 8.5%.

The previous approach can be extended to compute the confidence interval of the mean speed. Assume that $x_1, x_2, \ldots x_n$ are n independent observations drawn from a population with mean μ and variance σ^2. By definition, if the random variables are independent, the variance of their sum is equal to the sum of their variances:

$$\text{Var}(x_1 + x_2 \ldots + x_n) = n\sigma^2$$

It follows that the mean $\bar{x}$ is normally distributed with mean μ and variance σ^2/n. Given that the population characteristics are unknown, we can use $\bar{x}$ and s^2/n to approximate this distribution. We can then compute the probability that the mean falls inside a given interval. A distinction is thus made between the estimated density probability function $f(x|\bar{x}, s^2)$ of the variable and the estimated probability density function $f(x|\bar{x}, se^2)$ of its mean, with $se = \sqrt{s^2/n}$. The standard error se represents the standard deviation of the sample distribution of the mean.

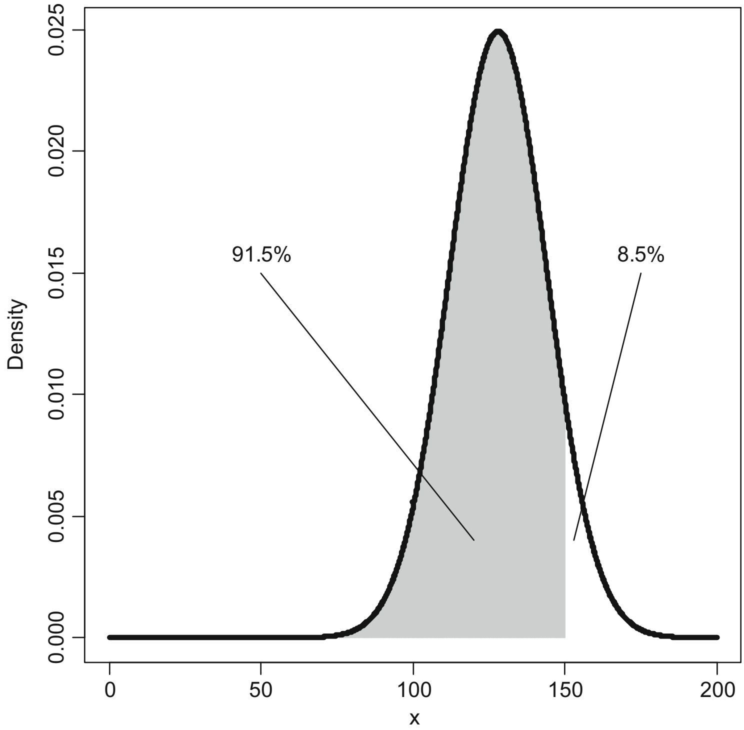


Fig. 3.16 Estimated probability density function: example 5

Computing a confidence interval is not straightforward since we need to compute the area below the curve of the density function. The task is however made easier by the use of statistical tables. Those tables provide information about the cumulative probabilities for a set of distributions. To illustrate, Fig. 3.17 provides the statistical table of the standard normal distribution, which by definition is the distribution of a normal random variable with zero mean and unit variance. This probability distribution function is usually symbolized by a lowercase Phi (ϕ). We have:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The cumulative distribution function (uppercase Phi) of the standard normal distribution is given by:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

	$F(x) = Pr\{X < x\}$									
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	**0.8461**	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	**0.9147**	**0.9162**	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	**0.9750**	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Fig. 3.17 Cumulative probabilities of the standard normal distribution

Assume for example that we would like to know the probability that X is lower than 1.02 (=1.0+0.02). In Fig. 3.17, one would select 1.00 in row and 0.02 in column to find a probability equal to 84.61%. The other way round, we can also compute a value x such that $\Phi(x)$ is equal to some probability p. For instance, if $p = 97.50\%$, we can see that x amounts to 1.96. This means equivalently that the probability that X is higher than 1.96 is 2.5%.

Figure 3.20 is built with the Excel command *NORMDIST*$(x, 0, 1, TRUE)$ where x denotes the variable, "0" denotes the mean, "1" is the standard deviation, and "TRUE" is to ensure that the function returns the cumulative distribution function. In R-CRAN the function *pnorm*$(x, 0, 1)$ yields an equivalent output. For instance, *pnorm*(1.96,0, 1) yields 0.9750021.

We do not need to construct a statistical table for each form of normal distribution. As a matter of fact, every normal distribution is a variant of the standard normal distribution. Formally, the cumulative distribution function of a normal distribution with mean μ and standard deviation σ is given by:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi(z)$$

where z is called a standard score or z-score. Observations have to be weighted by the standard deviation and centered by subtracting the mean so that the distribution has a mean of zero and a standard deviation of one. In statistical terms, we need to standardize the variable of interest:

$$z = \frac{x - \mu}{\sigma}$$

Coming back to example 5 (Fig. 3.16), x is equal to 150, the sample mean to 128, and the sample standard deviation to 16. We thus have:

$$z = \frac{150 - 128}{16} = 1.375$$

Using Fig. 3.17, we get:

$$F(1.375) \in [91.47\%, 91.62\%] \approx 91.5\%$$

Figure 3.18 illustrates the approach. The probability density function is now centered on 0. The probability that x is lower than 1.375 is displayed in grey. Figs. 3.16 and 3.18 are strictly equivalent. The only difference is the scale of the axes.

Every normal random variable X can be transformed into a z-score via the previous formula. If the mean $\bar{x}$ is used, rather than a single value, then the standard score should be divided by the relevant standard deviation $\sigma/\sqrt{n}$:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Assume now that we would like to compute a confidence interval such that the probability that μ lies in this interval is 95%. As described in Fig. 3.19, we need to compute the upper and lower bounds of the interval so that the shaded areas at either end of the distribution are equal to 2.5%. From Fig. 3.17 we know that those limits amount to –1.96 and +1.96, respectively. We thus have:

$$\Pr\left\{-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right\} = 95\%$$

which can be rewritten as:

$$\Pr\left\{\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

The confidence interval of the population mean is therefore:

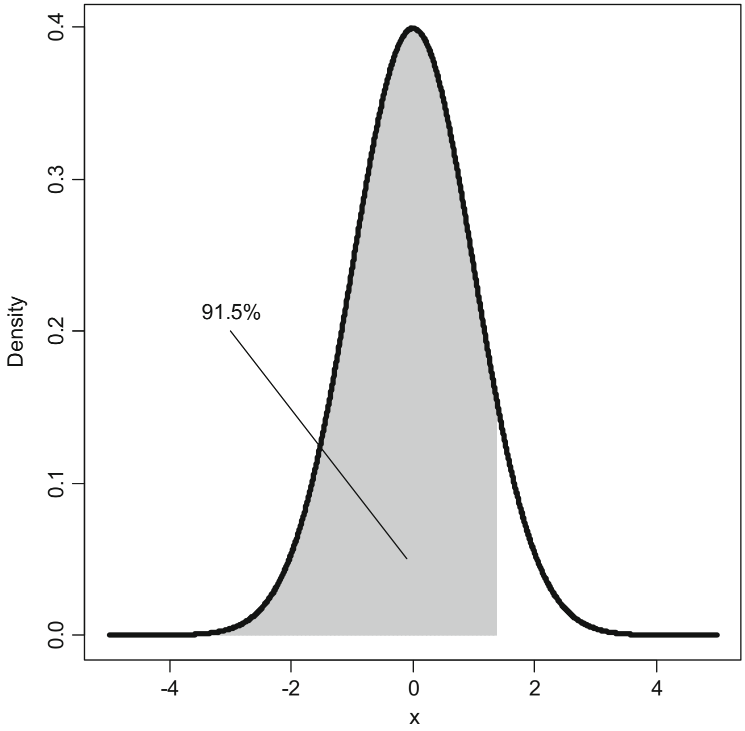


Fig. 3.18 The standard normal probability distribution: example 5

$$\left[\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right]$$

The margin of error e is given by $1.96 \times \sigma/\sqrt{n}$. It is a measure of how closely one expects the sample results to represent the entire population being studied. The greater the dispersion σ around the mean, the less certain we are about the actual population mean, and the larger is the confidence interval. Similarly, the lower the sample size, the less confidence we have in the sample statistic and the larger is the confidence interval.

In practice, the population standard deviation σ is unknown. We rely instead on the sample standard deviation s. In addition, the standard deviation is not the same depending on whether we estimate a sample mean or a sample proportion. Last, the formula for computing a confidence interval depends both on the size of the sample and on the size of the population. Table 3.13 illustrates those differences.

First, when the sample size is approximately lower than 100, a Student distribution must be used to compute the confidence interval of the mean. The Student distribution (or t-distribution) is leptokurtic, i.e. has a higher kurtosis than the normal distribution, which thereby yields higher confidence intervals. Fig. 3.20 provides the cumulative probabilities of this distribution. This table has been

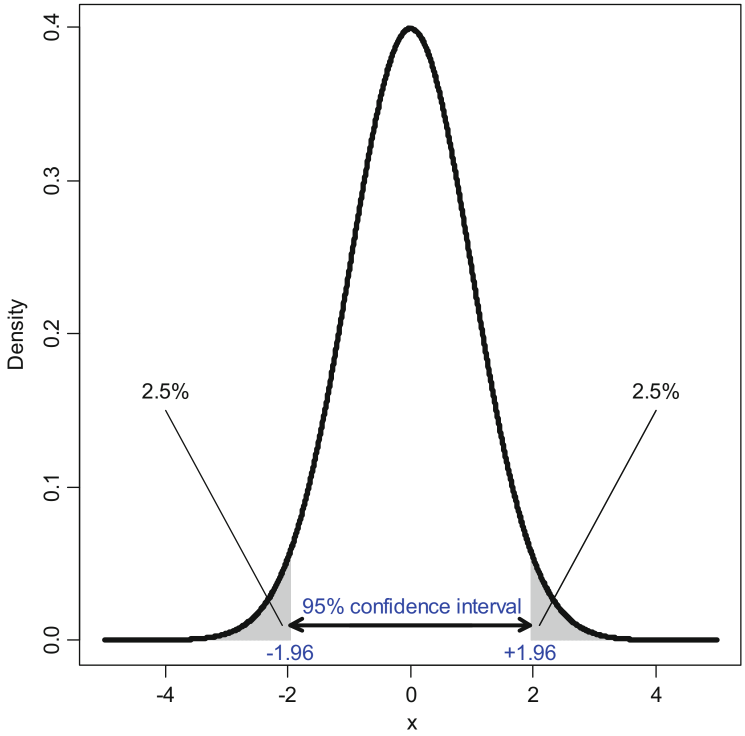


Fig. 3.19 Standard normal curve showing the 95% confidence interval

Table 3.13 Confidence interval of a mean and a proportion

Statistic	Standard error	Critical value	Margin of error
(a) Large population (N>200,000) and large sample (n>100)			
Mean	$\frac{s}{\sqrt{n}}$	z obtained from a standard normal distribution (1.96 for a 95% confidence interval)	$z \times \frac{s}{\sqrt{n}}$
Proportion	$\sqrt{\frac{(p)(1-p)}{n}}$	z obtained from a standard normal distribution (1.96 for a 95% confidence interval)	$z \times \sqrt{\frac{(p)(1-p)}{n}}$
(b) Large population (N>200,000) and small sample (n<100)			
Mean	$\frac{s}{\sqrt{n}}$	t obtained from a Student distribution with $n-1$ degrees of freedom	$t \times \frac{s}{\sqrt{n}}$
Proportion	$\sqrt{\frac{(p)(1-p)}{n}}$	z obtained from a standard normal distribution (1.96 for a 95% confidence interval)	$z \times \sqrt{\frac{(p)(1-p)}{n}}$
(c) Small population (N<200,000) and small sample (n<100)			
Mean	$\frac{s}{\sqrt{n}}$	t obtained from a Student distribution with $n-1$ degrees of freedom	$z \times \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
Proportion	$\sqrt{\frac{(p)(1-p)}{n}}$	z obtained from a standard normal distribution (1.96 for a 95% confidence interval)	$z \times \sqrt{\frac{(p)(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$

	$F(x) = Pr\{X < x\}$							
Degrees of freedom	**0.60**	**0.70**	**0.80**	**0.85**	**0.90**	**0.95**	**0.975**	**0.995**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	63.657
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	9.925
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	5.841
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	4.604
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	4.032
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	3.707
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	3.499
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	3.355
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	3.250
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	3.169
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	3.106
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	3.055
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	3.012
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.977
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.947
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.921
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.898
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.878
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.861
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.845
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.831
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.819
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.807
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.797
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.787
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.779
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.771
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.763
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.756
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.750
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.704
59	0.254	0.527	0.848	1.046	1.296	1.671	2.001	2.662
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.660
100	0.254	0.526	0.845	1.042	1.290	1.660	1.984	2.626
∞	**0.253**	**0.524**	**0.842**	**1.036**	**1.282**	**1.645**	**1.960**	**2.576**

Fig. 3.20 Cumulative probabilities of the *t*-distribution

generated with the *TINV* function from Excel (in R-CRAN one would use the function *qt*). The most important column is displayed in red. It provides the *t*-values to be used in order to compute a 95% confidence interval (2.5% on both sides of the distribution). As can be seen, those *t*-values are all higher than 1.96. They depend on the number of degrees of freedom. Basically speaking, degrees of freedom relate the precision of the estimate to the sample size. When we estimate the mean, we need to have a least one observation to do so. Any additional observation is a bonus that will improve the precision of the estimator. The number of degrees of freedom in that case is $n - 1$. For instance, if we were estimating the equation of a line, we would need at least two observations to do so. The number of degrees of freedom would be $n - 2$. In Fig. 3.20, the higher the number of degrees of freedom, the closer is the critical value to 1.96. Second, when the population size (N) is small, below approximately $N = 200{,}000$, a finite population correction factor (CF) has to be used, in order to weight the margin of error e:

$$CF = \sqrt{\frac{N-n}{N-1}}$$

Turning back to example 1, let us compute manually the confidence intervals for the population mean of variable *Intensity* and the proportion of firms that have received a subsidy (see Table 3.4 for the corresponding raw data). The sample size is $n = 60$ firms. We know that the mean of *Intensity* is $\bar{x} = 0.109$ and its standard deviation is $s = 0.065$. The standard error se is computed as:

$$se(Intensity) = \frac{s}{\sqrt{n}} = \frac{0.065}{\sqrt{60}}$$

The critical t-value is obtained from Fig. 3.20 for $n - 1 = 60 - 1 = 59$ degrees of freedom. We have $t = 2.001$. Assume that the population size is $N = 5{,}000$. A correction factor is thus required. We obtain the following margin of error:

$$e(Intensity\ ) = t \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = 2.001 \times \frac{0.065}{\sqrt{60}} \times \sqrt{\frac{5000-60}{5000-1}} \approx 0.017$$

The confidence interval of the mean for variable *Intensity* is thus:

$$0.109 \pm 0.017$$

For variable *Subsidy*, we know that a proportion 20/60 (≈33.3%) of the firms received a subsidy (see Table 3.5). From Table 3.13, we know that the standard error is:

$$se(Subsidy) = \sqrt{\frac{(p)(1-p)}{n}} = \sqrt{\frac{(20/60) \times (40/60)}{60}}$$

The critical z-value amounts to 1.96. We have:

$$e(Subsidy) = 1.96 \times \sqrt{\frac{(20/60) \times (40/60)}{60}} \times \sqrt{\frac{5000-60}{5000-1}} \approx 11.8\%$$

The confidence interval for the average number of recipients is:

$$33.3\% \pm 11.8\%$$

Figure 3.21 details those computations in R-CRAN. Functions $qt(0.975, n-1)$ and $qnorm(0.975)$ are used to find the critical values t and z, respectively.

It should be stressed that the formulas presented in Table 3.13 can be used if and only if the data are normally distributed. This may hold true in many situations, but there are some cases however where we have to rely on nonparametric techniques to avoid bias. The bootstrapping method allows the estimation of the sampling distribution of almost any statistic by using random sampling methods. Picking data

```
> D=read.table("C://mydata1.csv",head=TRUE,sep=";")

> N=5000
> n=60
> CF=sqrt((N-n)/(N-1))
> CF
[1] 0.9940813

> # R&D Intensity
> se=sd(D$Intensity)/sqrt(n)
> se
[1] 0.008439549
> t=qt(0.975,n-1)
> t
[1] 2.000995
> e=t*se*CF
> e
[1] 0.01678755
> c(mean(D$Intensity)-e,mean(D$Intensity)+e)
[1] 0.09221245 0.12578755

> # Subsidy
> se=((20/60*40/60)/60)^0.5
> se
[1] 0.06085806
> z=qnorm(0.975)
> z
[1] 1.959964
> e=z*se*CF
> e
[1] 0.1185736
> c(mean(D$Subsidy)-e,mean(D$Subsidy)+e)
[1] 0.2147597 0.4519070
```

Fig. 3.21 Computing confidence intervals with R-CRAN: example 1

from the sample, the method successively creates a large number of subsamples (known as bootstrap samples) for the purpose of approximating the sampling distribution. Those subsamples are the same size as the initial sample and created with replacement. Assume for instance that we have a sample of 10 observations numbered from 1 to 10. We can create a set of subsamples by selecting randomly 10 observations from the sample. For instance we may have:

Bootstrap sample 1: (9 7 8 2 1 9 10 2 7 9);
Bootstrap sample 2: (2 9 2 2 6 9 6 1 5 9);
Bootstrap sample 3: (1 3 8 2 1 6 6 4 1 10); and so on.

For each bootstrap sample, a point estimate (such as the mean, the median, etc.) is computed. The distribution of those bootstrap statistics is then used to compute a confidence interval (usually a 95% percentile confidence interval) for the relevant population parameter.

Let us consider again variables *Intensity* and *Subsidy*. Figure 3.22 provides the codes to bootstrap confidence intervals with R-CRAN. First, we need to specify the statistic that we estimate, here the mean. We could use the same approach to compute the confidence interval of the median. We create *myfunction*, defined by two entries: a database (*data*) and a random index (*i*) for the bootstrap sample. In this function, a subsample denoted *data*2 is generated and used for computing the mean of each bootstrap sample. Once the function is created, and the data uploaded, the *boot* command (from the *boot* package) is used to create the bootstrap statistics.

```
> # Bootstrap
> myfunction=function(data,i){data2=data[i,];mean(data2)}
> library(boot)

> # Bootstrap R&D Intensity
> mysample=data.frame(D$Intensity)
> myboot=boot(mysample,myfunction,R=10000)
> CI=quantile(myboot$t,c(0.025,0.975))
> CI
      2.5%      97.5%
0.09283333 0.12533750
> hist(myboot$t,main="15.1 R&D Intensity",freq=FALSE,col="grey",
+ ylim=c(0,50))
> abline(v=CI,col="red")

> # Bootstrap Subsidy
> mysample=data.frame(D$Subsidy)
> myboot=boot(mysample,myfunction,R=10000)
> CI=quantile(myboot$t,c(0.025,0.975))
> CI
     2.5%     97.5%
0.2166667 0.4500000
> hist(myboot$t,main="15.2 Subsidy",freq=FALSE,col="grey",ylim=c(0,7))
> abline(v=CI,col="red")
```

Fig. 3.22 Bootstrap confidence intervals with R-CRAN: example 1

The *boot* command uses both the original sample (*mysample*) and the function *myfunction* to generate randomly $R = 10{,}000$ observations named *myboot*$*t*. Histograms of those statistics are provided in Fig. 3.23. They are constructed in the same way as the adjusted frequencies of Fig. 3.2b.

The computation of the bootstrap intervals is based on the values of the percentiles (function *quantile*). A percentile is a measure indicating a threshold in a frequency distribution. The *k*th percentile is the value that splits the observations into two groups: the lower group contains *k* percent of the observations while the upper group contains the rest of the observations. For instance, the 2.5th percentile is the value which marks off the lowest 2.5% of the observations from the rest, the 25th percentile is the same as Q_1, the 50th percentile is the same as the median, the 75th percentile is the same as Q_3, and the 97.5th percentile contains all but 2.5% of the observations. Within the 2.5th and 97.5th percentiles lie 95% of the values, which gives us a 95% confidence interval.

For variable *Intensity*, the bootstrap yields a confidence interval equal to $[0.092, 0.125]$. Due to the random nature of the *boot* process, and depending on the number of bootstrap samples randomly generated, each bootstrapping is likely to produce slightly different results. For variable *Subsidy*, the confidence interval amounts to $[0.216, 0.450]$. Those percentile confidence intervals are displayed in red on Fig. 3.23 using the *abline* command. Option *v* specifies the *x*-values for a vertical line. Note that the approach does not offer to correct for the population size, which is of less importance as the correction factor, by decreasing *e*, would only reduce the size of the interval and, therefore, increase the confidence we have in the statistic.

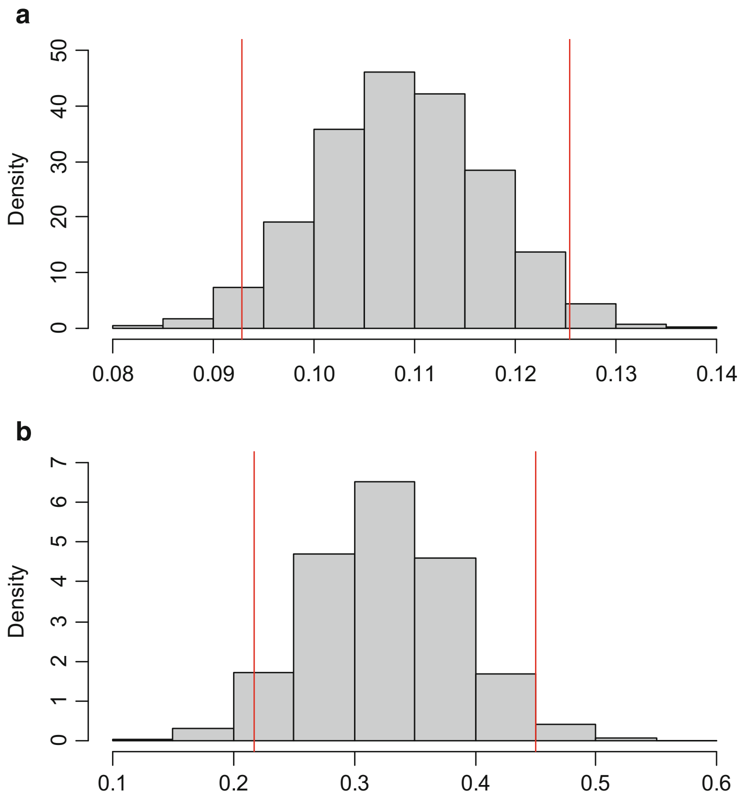


Fig. 3.23 Histograms of the bootstrap statistics. (**a**) R&D intensity. (**b**) Subsidy

Bibliographical Guideline

Many textbooks deal with the concepts that have been presented in this chapter. We may cite in particular Johnson and Bhattacharyya (2009) and Anderson et al. (2014). Those manuals provide an introduction to statistics (organization and description of data, the normal distribution, inference, etc.) with real-world examples. The reader can also refer to Richardson (2012), which offers a description of the numerous techniques that can be used to display information. Osborn (2006) covers the basic biostatistics, descriptive statistics, and inferential statistics that are unique to health information management. Last, many statistical terms are defined online. In this respect, the reader may have a look at the websites of the Australian Bureau of Statistics and the OECD. They both offer a glossary of the most important statistical terms.

References

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran J. J. (2014). *Essentials of statistics for business and economics*. Cengage Learning.

Johnson, R. A., & Bhattacharyya, G. K. (2009). *Statistics: Principles and methods*. New York: Wiley.

Osborn, C. E. (2006). *Statistical applications for health information management*. Jones & Bartlett Learning.

Richardson, G. T. (2012). *Illustrations: Everybody's complete and practical guide*. Springer Science & Business Media.

Measuring and Visualizing Associations 4

4.1 Identifying Relationships Between Variables

In its simplest form, a statistical report provides information about a series of variables by separately computing for each of them summary statistics, displaying frequencies, and estimating intervals for population parameters. In a more sophisticated manner, it can also describe the relationships that potentially exist between the variables in question. A distinction is thus made between a univariate analysis, which involves describing the distribution of each variable separately, and a bivariate analysis that is used to highlight associations between pairs of variables. While the former type of analysis serves to better understand the descriptive context of a public program, the latter is particularly useful when one wants to further interpret the data in order to motivate a particular policy intervention.

Finding the causes of a problem and knowing why a particular policy should be implemented is essential to the evaluative approach. Consider a program aiming at reducing poverty in developing countries. The identification of needs is made difficult by the multidimensional nature of the problems people experience. Poverty may be thought in terms of economic growth, income inequalities, unemployment, malnutrition, poor sanitation, lack of infrastructures, lack of education and political instability. Those items are interrelated and must be examined together. What are the relationships linking health, education and economic growth? What is the priority? Is it to feed children, to educate people, or to promote industrial activities? To have a clear picture of the relationships at stake is essential if one wants to attain desirable goals in an efficient manner. One tool in the statistician's toolbox that can serve this purpose is bivariate analysis. We may find for instance that with a higher level of education, people would receive higher wages and be more careful with their health. In this context, education would become a mean to achieve a general development agenda.

Formally, one of the goals of statistical studies is the identification of cause-and-effect relationships. The idea is to examine whether a variable X has an impact on a variable Y:

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_4

$$X \rightarrow Y$$

In that context, X represents what is termed an independent variable (also known as exogenous or explanatory variable) while Y is called a dependent variable (endogenous or explained variable). The dependent variable Y represents the output or outcome whose variation is being studied and that is thought to be influenced by factors such as X. An independent variable on the other hand is a factor that the evaluator thinks is causing a variation in the dependent variable. Examples include the demand Y of consumers for the product of a firm, which depends on the selling price X of the good; or air pollution Y that varies with GDP per capita X. When causation is found, it means that a change in one variable directly causes a change in the other variable.

Demonstrating that a particular variable X (the independent variable) has an effect on some outcome of interest Y (the dependent variable) is generally made difficult by the presence of confounding variables Z that may connect with both the dependent variable and the independent variable. There are two types of situations. First, there may be a causal relationship between X and Y that is also affected by a third variable Z. This yields the following path analysis diagram:

$$\begin{tikzcd}
X \arrow[r] & Y \\
Z \arrow[u] \arrow[ur] &
\end{tikzcd}$$

In that case, the effect of X on Y can be underestimated or overestimated depending on how Z affects X and Y. Second, it is possible that a spurious relationship exists between X and Y:

$$\begin{tikzcd}
X & Y \\
Z \arrow[u] \arrow[ur] &
\end{tikzcd}$$

Variables Y and X have no direct causal connection and yet some association is highlighted. For instance, if one examines the spending behavior of local governments without controlling for the population size, a comparison of their expenditures levels with car ownership would make one conclude that a positive relationship exists between those variables: an increase in the population (Z) is likely to induce an increase in the total number of vehicles (X), and will generate in the meantime a higher demand for public spending (Y). Variable Y is thereby artificially associated with variable X. Under this framework, the term "association" denotes a relationship between two variables that renders them statistically dependent. It does not, however, imply causation.

Correlation and association are closely terms. Both concepts imply that two variables vary according to some common pattern. Yet, association is more general than correlation. The latter can be considered as a special case of association, where the relationship between the variables is linear. The scatter plots of Fig. 4.1

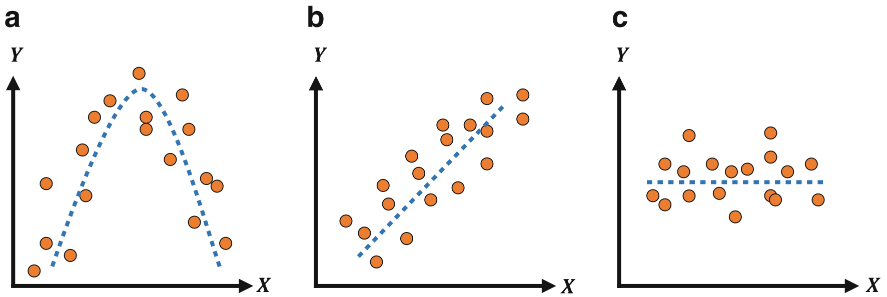


Fig. 4.1 Difference between association and correlation. (**a**) Association between *Y* and *X*, but no correlation. (**b**) Association and correlation between *Y* and *X*. (**c**) No correlation and no association between *Y* and *X*

illustrate those differences in the (X, Y) plane. In Fig. 4.1a, the graph reveals a relationship between *X* and *Y*. This relationship is non-linear meaning that the variables are associated but not linearly: there is no correlation. Figure 4.1b displays a linear relationship between *X* and *Y*. The variables are correlated and thereby associated. Last, Fig. 4.1c illustrates a situation where no association exists between the variables.

The statistical tools for the identification of relationships between variables depend on their nature. Categorical (or qualitative) variables describe a quality or characteristic of a data unit. Numerical (or quantitative) variables give an account of a numerically measured value. When the two variables that are compared are numerical, then testing for correlation is the relevant approach. It determines the degree to which two variables tend to move together linearly. The chi-square test is an inferential test that uses data from a sample to make conclusions about the relationship between two categorical variables. When one variable is numerical and the other is categorical, one usually tests differences between means (the analysis of variance or ANOVA extends the approach by analyzing the differences among several means).

When faced with more than two variables, it is also possible to provide a multidimensional representation of the problem using correlation-based methods (also known as factor analysis methods) such as principal component analysis and multiple correspondence analysis. Principal component analysis offers a way of identifying patterns in data when the variables are numerical. Multiple correspondence analysis is used on the other hand when the variables are categorical. Both methods offer an all-encompassing picture of the phenomena in play. The idea is to reduce the dimensionality of a data set by plotting all the observations on 2D graphs depending on how close the observations are with respect to their characteristics. Observations can then be divided into groups according to their proximity and these groups can serve to identify the beneficiaries of a particular intervention. For instance, the approach can be used to create a typology of jurisdictions according to their socio-economic characteristics to better explain differences in public spending. In education, one can identify profiles of students and bring to light the determinants of success and failure at school. The approach can be used in health to differentiate types of patients. Those typologies may in return be used to inform policy-makers about particular needs and possible interventions.

The remaining of the chapter is organized as follows. Section 4.2 explains how to compute a correlation coefficient and how to assess its statistical significance. Section 4.3 provides a description of the chi-square test of independence. Section 4.4 is about testing differences between means. Sections 4.5 and 4.6 develop the methodology of principal component analysis and multiple correspondence analysis, respectively.

4.2 Testing for Correlation

The covariance and correlation coefficients are used to assess a possible linear association between two numerical variables. They are simple to calculate and to interpret. For a population of size N, the covariance between two variables X and Y is given by:

$$\sigma_{X,Y} = \frac{1}{N}\sum\nolimits_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \text{ (population)}$$

where X_i and Y_i are the values of X and Y for the ith unit while μ_X and μ_Y denote the mean of the variables. The counterpart of the formula for a sample of size n is:

$$s_{x,y} = \frac{1}{n-1}\sum\nolimits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \text{ (sample)}$$

The sample covariance offers an average measure of the deviations from both means $\bar{x}$ and $\bar{y}$. Its sign shows the direction of the linear relationship between the variables. Figure 4.2 illustrates the approach. When most observations lie in the North-East and South-West quadrants, the sum $\sum(x_i - \bar{x})(y_i - \bar{y})$ contains positive

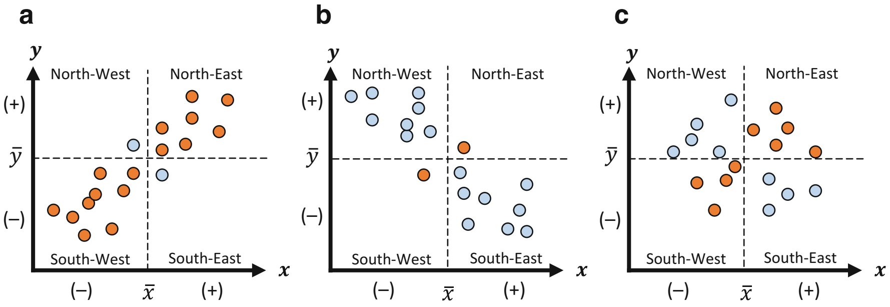


Fig. 4.2 Understanding covariance. (**a**) Positive linear relationship between x and y. (**b**) Negative linear relationship between x and y. (**c**) No relationship between x and y

values mainly (emphasized in orange) and the covariance is positive. When observations lie in the North-West and South-East quadrants, the sum is mostly made of negative values (in blue on the graph). The covariance is negative. Last, when the individuals are equally distributed among the quadrants, the covariance approaches zero.

The correlation coefficient, also referred to as Pearson product-moment correlation coefficient or Pearson's r, offers a normalized measure of the covariance. It is computed as:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \text{ (population)}$$

and

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} \text{ (sample)}$$

where σ_X and σ_Y denote the standard deviation of variables X and Y respectively, and s_x and s_y are their sample counterparts:

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2}; \sigma_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu_Y)^2};$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}; s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In Excel, one can use the *CORREL* function to compute the correlation coefficient.

The correlation coefficient provides a measure of linear association between two numerical variables. Its value ranges between -1 and +1. The sign of the correlation coefficient depends on whether the variables are positively or negatively related. If the coefficient is close to 1, the variables are said to be positively and linearly associated: as the value of one variable increases, the value of the other also tends to do so. If the coefficient is close to -1, the variables are said to be negatively and linearly associated: as the value of one variable increases, the value of the other tends to decrease.

The magnitude of the correlation coefficient determines the strength of the correlation. Although there is no convention, the following guidelines provide a proxy of how one usually describes the correlation coefficient:

Weak correlation: $0 \leq |r_{x,y}| < 0.4$;
Moderate correlation: $0.4 \leq |r_{x,y}| < 0.6$;
Strong correlation: $0.6 \leq |r_{x,y}| < 1$.

If the coefficient approaches 0, we generally conclude that there is no linear relationship between the variables. A correlation coefficient of −1 or +1 indicates a perfect linear relationship.

Graphically, the correlation coefficient is a measure of clustering around a line. In Fig. 4.3a for instance, there is a strong linear association between the two variables. The correlation coefficient is high and amounts to 0.96. This does not mean that 96% of the observations are concentrated. Assuming that Y is the dependent variable, the correlation coefficient measures instead the extent to which knowing the value of X helps to predict the value of Y. The squared correlation coefficient ${r_{x,y}}^2$ is the percentage of variation in Y that can be explained by knowing X, or vice-versa. In our example, we can say that $0.96^2 = 92\%$ of the variation in Y is accounted for by X. In Fig. 4.3b on the other hand, variable X moderately explains variable Y. The observations are more dispersed and the lower is the correlation coefficient.

A common mistake is also to think that the correlation coefficient measures the slope of the line around which the observations are clustered. This is not true. For instance, in Fig. 4.3c, the slope of the line is greater than the slope observed in Fig. 4.3a. Yet, the correlation coefficients are similar. Last, it should be stressed that the correlation coefficient measures linear relationships only. This means that any change in one variable must be associated with a constant proportional change in the other variable. In Fig. 4.3d, for instance, there is a clear association between the variables but the relationship is not linear. The correlation coefficient is found to be near zero.

The population correlation ρ is usually unknown. The sample statistic r can however be used to carry out tests of hypotheses. In hypothesis testing, one begins by making an assumption about a particular population parameter. This assumption is referred to as the null hypothesis H_0. Often, it assumes that the observed phenomenon is due to chance only. For instance, for the test for significance of correlation, this assumption is specified as:

$$H_0 : \rho_{X,Y} = 0 \text{ (the population correlation is zero)}$$

It may or may not be true. One also needs to define the alternative hypothesis denoted by H_1.

$$H_1 : \rho_{X,Y} \neq 0 \ \text{ (there is a real correlation)}$$

The alternative hypothesis states that the observed phenomenon is the result of some non-random cause. A statistical test is then implemented to determine whether there is enough evidence to reject the null hypothesis. If the correlation coefficient is found to be statistically different from zero, then we conclude that the correlation coefficient is significant.

Basically speaking, there are two ways of testing a hypothesis in statistics. The first approach relies on confidence interval estimation. A confidence interval gives an estimated range of values which is likely to include the population parameter ρ.

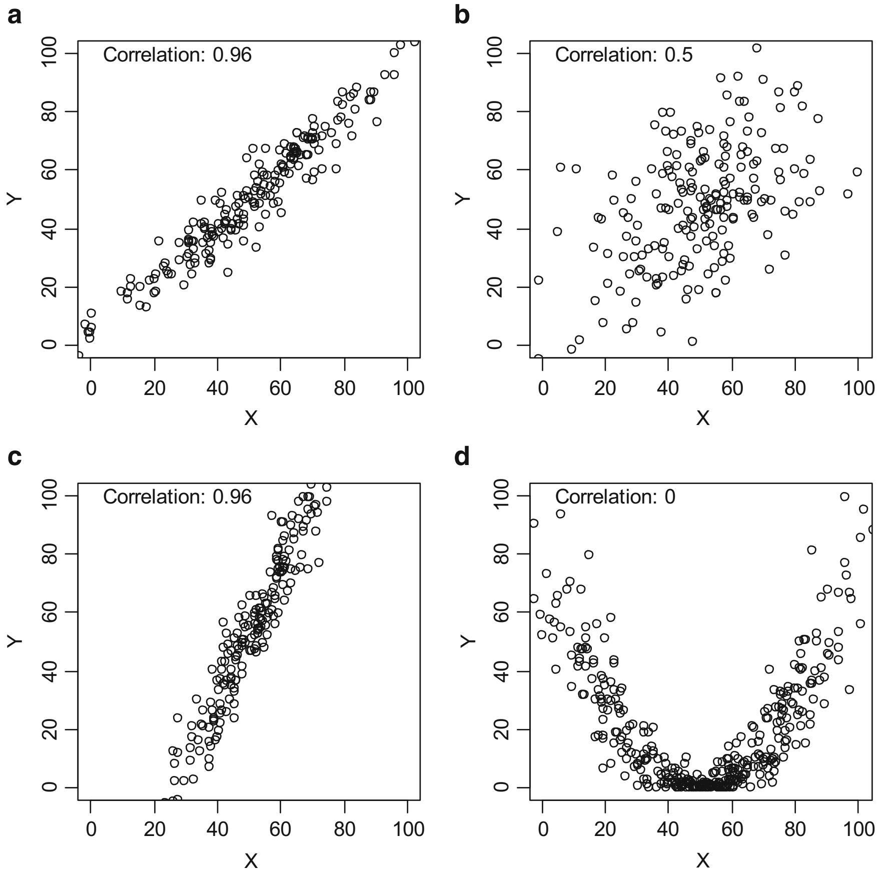


Fig. 4.3 Scatter plots and correlations. (**a**) Strong linear association. (**b**) Moderate linear association. (**c**) Different slope. (**d**) Non-linear association

Intervals are usually calculated at a 95% confidence level, so that 95% of the estimated intervals would include the true parameter. For instance, if the confidence interval of a given population parameter includes zero, then one would fail to reject the null hypothesis that the population parameter is null.

The second approach uses the correlation coefficient $r_{x,y}$ from the sample data to compute its standard error as:

$$se = \sqrt{\frac{1 - {r_{x,y}}^2}{n - 2}}$$

The value of the test statistic is given by the following *t*-value:

$$t^* = \frac{r_{x,y}}{se}$$

If the variables under examination are normally distributed, then this statistic follows a Student distribution with $n-2$ degrees of freedom. For a confidence level of 95%, we find the critical value for $n-2$ degrees of freedom and a significance level equal to 5%. The number of degrees of freedom is reduced by two because we need at least two observations to compute the correlation coefficient. The significance level is defined as 100% minus the confidence level. If $|t^{*}|$ is higher than the critical value, then the null hypothesis is rejected. On the other hand, if $|t^{*}|$ is lower than the critical value, we fail to reject the null hypothesis. In that case, there is not sufficient evidence to conclude that there is a significant linear relationship between the variables.

Note that the test is two-tailed (or two-sided). To achieve a significance level of 5%, the absolute value of the test statistic must be greater than or equal to a critical value defined as $t_{\alpha/2}(df)$, where α denotes the significance level, and df stands for the number of degrees of freedom. Under this framework, no assumption is made as regards the sign of the correlation coefficient. Deviations of the correlation coefficient are considered possible in either direction from zero. Figure 4.4a provides an illustration with 58 degrees of freedom. The upper limit of the region of acceptance is equal to the value for which the cumulative probability of the Student distribution is equal to one minus the significance level divided by 2, i.e. 0.975. The lower limit is defined as the value for which the cumulative probability of the Student distribution is equal to 5% divided by 2, i.e. 0.025. Overall, this yields a confidence level of 95%. Equivalently we say that the significance level (grey area in Fig. 4.4a) amounts to 5%. Since the Student distribution is symmetric, the upper limit and the lower limit are equal in absolute value. Therefore, in practice, one does not need to compute both values. By convention, the focus is on the upper limit only.

It is also possible to perform a one-tailed (one-sided) test. In our example, we would test for instance:

$$H_0 : \rho_{X,Y} \leq 0 \text{ (the population correlation is lower than or equal } to \text{ 0)}$$

$$H_1 : \rho_{X,Y} > 0 \text{ (the population correlation is strictly positive)}$$

Here, we are interesting in one side of the Student distribution only, as illustrated in Fig. 4.4b. In that case, we implicitly assume that the correlation coefficient cannot be negative. By construction, the critical value, now denoted $t_{\alpha}(n-2)$, is lower than previously. This makes it easier to reject the null hypothesis and to obtain statistical significance. One-tailed tests should however be used with caution. With a one-tailed test, one has to make an assumption about the direction of the relationship and completely disregard the possibility of a relationship in the other direction.

To illustrate the test of correlation, consider a sample of 60 firms. Table 4.1 provides information about their research and development intensity (R&D expenditures divided by value added), the number of patents assigned to each firm, whether those firms have received a subsidy from the government (0 if no subsidy and 1 otherwise) and the sector they belong to (3 for high-technology, 2 for

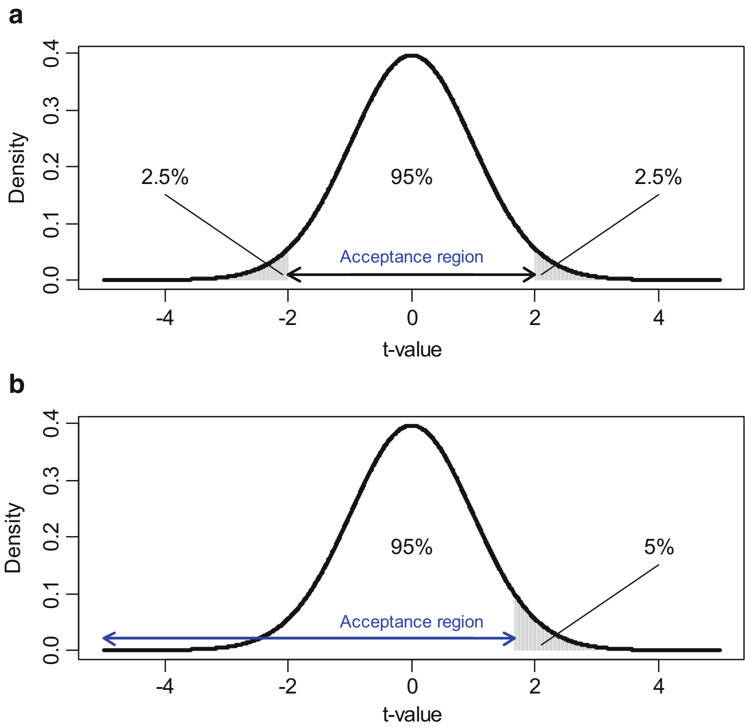


Fig. 4.4 The Student distribution ($df = 58$). (**a**) Two-tailed test. (**b**) One-tailed test

medium-technology and 1 for low-technology industry). Figure 4.5 provides the code to be used in R-CRAN if one wants to test whether the number of patents is correlated with R&D intensity. The database is uploaded using the command *read . table* and saved under the name D. A scatter plot is generated to better highlight the phenomenon in question (see Fig. 4.6). We can see that the observations are approximately concentrated around a fictitious line with positive slope.

The parameters of the test are specified as follows: n stands for the number of observations, r is the coefficient of correlation computed with the *cor* function, *se* denotes the standard error, and *tstar* is the test statistic. We find $t^* = 11.27$. The upper critical value ($t_{2.5\%}(58) = 2.0017$) is obtained using the command $qt(0.975, n-2)$. As can be deduced from the results, the test statistic (11.27) is greater than the critical value in absolute value, which means that we reject the null hypothesis. The population correlation is not zero. A similar conclusion can be reached with the *cor . test* command. This function directly provides information about the test statistic ($t = 11.2711$), the number of degrees of freedom ($df = 58$), the confidence interval of the population parameter [0.73, 0.89], and the sample estimates ($r_{x,y} = 0.83$). The p-value gives the level of significance for which one would be indifferent between rejecting and not rejecting H_0. In other words, if the p-value is

Table 4.1 Raw data for example 1

Firm	Intensity	Patents	Subsidy	Sector
1	0.23	42	1	3
2	0.19	31	0	3
3	0.20	20	0	3
4	0.19	33	1	3
5	0.18	30	1	3
6	0.18	43	0	3
7	0.17	37	1	3
8	0.23	27	1	3
9	0.11	7	0	3
10	0.16	21	1	3
11	0.14	13	0	3
12	0.20	20	1	3
13	0.21	42	1	3
14	0.18	35	1	3
15	0.14	13	0	3
16	0.21	26	1	2
17	0.11	24	0	2
18	0.10	1	0	2
19	0.13	33	0	2
20	0.17	49	1	2
21	0.08	3	0	2
22	0.11	11	0	2
23	0.14	28	0	2
24	0.11	3	1	2
25	0.09	4	0	2
26	0.17	32	1	2
27	0.08	2	0	2
28	0.13	22	0	2
29	0.08	2	0	2
30	0.12	16	0	2
31	0.07	1	0	2
32	0.16	21	1	2
33	0.18	39	1	2
34	0.08	2	0	2
35	0.16	25	1	2
36	0.06	0	0	2
37	0.12	19	0	2
38	0.20	58	1	2
39	0.10	7	0	2
40	0.04	0	0	1
41	0.09	13	1	1
42	0.01	2	0	1
43	0.04	13	0	1

(continued)

Table 4.1 (continued)

Firm	Intensity	Patents	Subsidy	Sector
44	0.04	11	0	1
45	0.07	2	0	1
46	0.07	5	0	1
47	0.00	0	0	1
48	0.02	0	0	1
49	0.08	2	1	1
50	0.02	2	0	1
51	0.01	4	0	1
52	0.05	0	0	1
53	0.05	6	0	1
54	0.02	0	0	1
55	0.02	4	0	1
56	0.07	19	0	1
57	0.00	0	1	1
58	0.08	12	0	1
59	0.08	14	0	1
60	0.01	0	0	1

less than the significance level $\alpha = 5\%$, the null hypothesis is rejected. On the other hand, if the p-value is greater the null hypothesis is not rejected.

Figure 4.5 also performs a one-tailed test. Using $qt(0.95, n-2)$, the critical value is found to be 1.67. Equivalently, one can include the entry "*greater*" in the *cor . test* function. As expected, the p-value is lower than 5%, which means that we do reject H_0: the population coefficient is not lower than nor equal to zero.

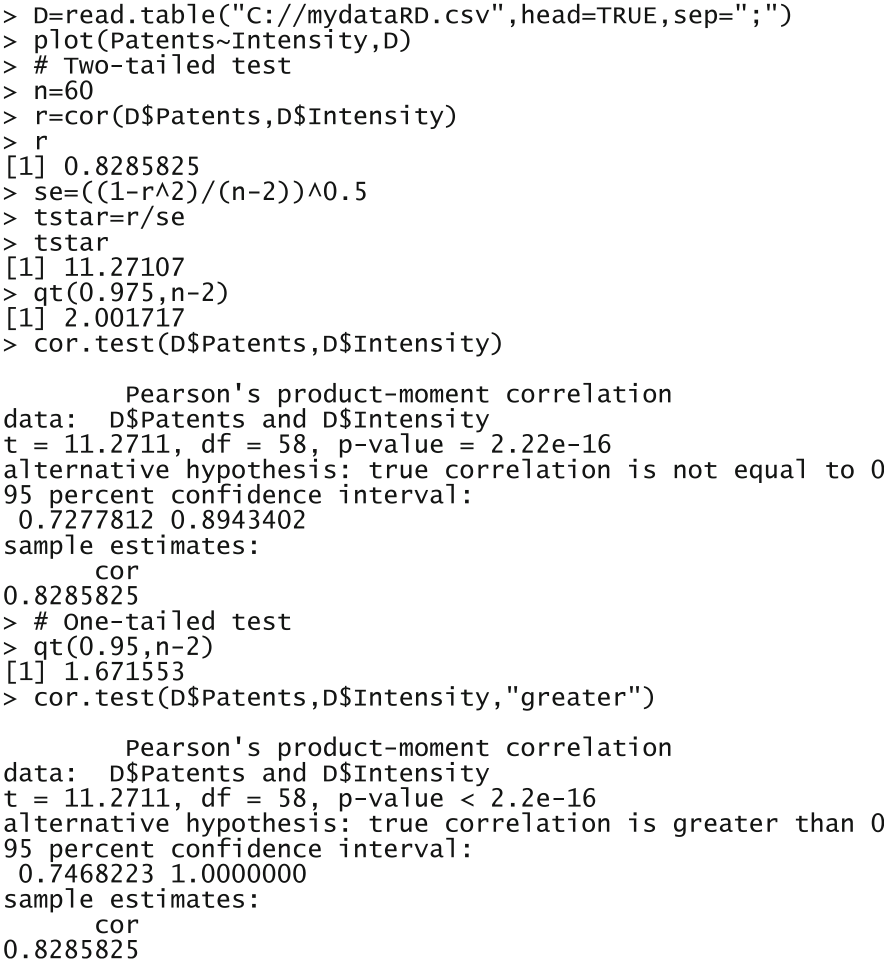
4.3 Chi-Square Test of Independence

This section explains how to statistically assess the strength and significance of a relationship between two categorical variables. The chi-square test of independence (or Pearson chi-square test) uses data from the sample to make conclusions about the relationship between categorical variables in the population. The alternative hypothesis suggests that the variables are associated but does not state that the relationship is necessarily causal:

$$H_0 : \text{The variables are not associated}$$

$$H_1 : \text{The variables are associated}$$

As for any statistical test, one has to compute a test statistic, namely the chi-square test statistic (denoted χ^2 hereafter), and examine whether this statistic is lower or higher than a critical value. The first step is to create a two-way or



```
> D=read.table("C://mydataRD.csv",head=TRUE,sep=";")
> plot(Patents~Intensity,D)
> # Two-tailed test
> n=60
> r=cor(D$Patents,D$Intensity)
> r
[1] 0.8285825
> se=((1-r^2)/(n-2))^0.5
> tstar=r/se
> tstar
[1] 11.27107
> qt(0.975,n-2)
[1] 2.001717
> cor.test(D$Patents,D$Intensity)

        Pearson's product-moment correlation
data:  D$Patents and D$Intensity
t = 11.2711, df = 58, p-value = 2.22e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7277812 0.8943402
sample estimates:
      cor
0.8285825
> # One-tailed test
> qt(0.95,n-2)
[1] 1.671553
> cor.test(D$Patents,D$Intensity,"greater")

        Pearson's product-moment correlation
data:  D$Patents and D$Intensity
t = 11.2711, df = 58, p-value < 2.2e-16
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.7468223 1.0000000
sample estimates:
      cor
0.8285825
```

Fig. 4.5 Testing for correlation in R-CRAN

contingency table. For this purpose, one needs to count how many observations are in each combination of row and column categories. By convention, when one variable is thought to be the explanatory variable, it is used to define the rows. The other is used to define the columns.

The second step is to compute the χ^2 statistic. The computation is based on a comparison between the observed frequencies, expressed in the two-way table, and the expected frequencies that would be observed under the null hypothesis. To illustrate, let us consider the data from example 1. Figure 4.7a shows the observed frequencies of variables *Sector* and *Subsidy*. From the marginal (total) row we know that the total share of firms that did not receive a subsidy is 40/60. Similarly, from the marginal column, we know that the probability that a firm belongs to sector 1 is 21/60. If the variables *Sector* and *Subsidy* were independent, the joint probabilities

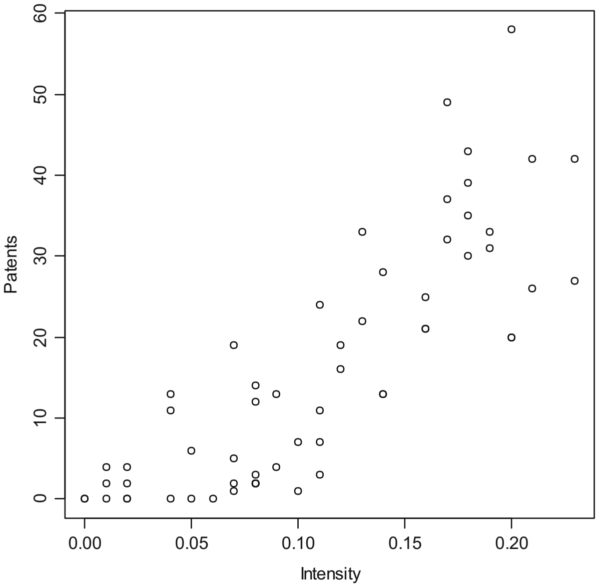
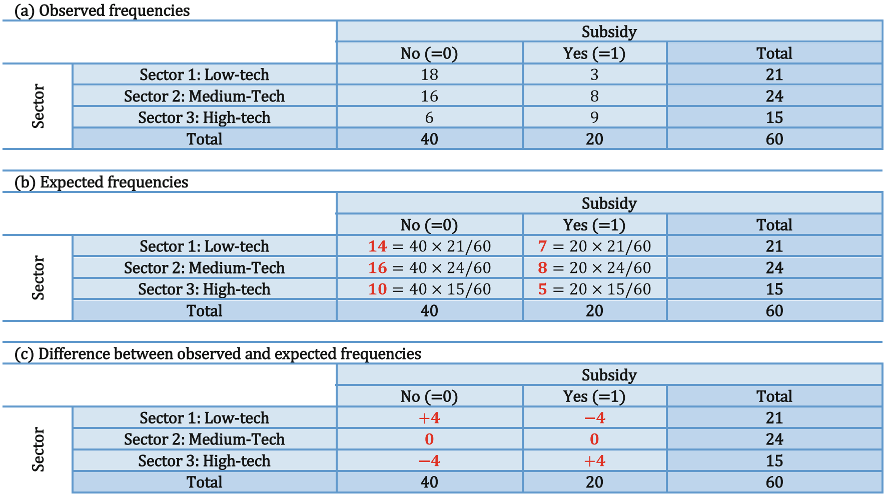


Fig. 4.6 Relationship between R&D intensity and patent claims



(a) Observed frequencies

		Subsidy		
		No (=0)	Yes (=1)	Total
Sector	Sector 1: Low-tech	18	3	21
	Sector 2: Medium-Tech	16	8	24
	Sector 3: High-tech	6	9	15
	Total	40	20	60

(b) Expected frequencies

		Subsidy		
		No (=0)	Yes (=1)	Total
Sector	Sector 1: Low-tech	14 = 40 × 21/60	7 = 20 × 21/60	21
	Sector 2: Medium-Tech	16 = 40 × 24/60	8 = 20 × 24/60	24
	Sector 3: High-tech	10 = 40 × 15/60	5 = 20 × 15/60	15
	Total	40	20	60

(c) Difference between observed and expected frequencies

		Subsidy		
		No (=0)	Yes (=1)	Total
Sector	Sector 1: Low-tech	+4	−4	21
	Sector 2: Medium-Tech	0	0	24
	Sector 3: High-tech	−4	+4	15
	Total	40	20	60

Fig. 4.7 Observed and expected frequencies: example 1

would equal the product of their marginal probabilities. Thus, under the null hypothesis, the probability that a firm from sector 1 does not receive a subsidy is:

$$\Pr\{\text{Sector 1}, \text{No}\} = \frac{40}{60} \times \frac{21}{60}$$

We should therefore have $60 \times \Pr\{\text{Sector } 1, \text{No}\} = 14$ non recipients in sector 1 (instead of 18 as stressed in Fig. 4.7a). By applying the same reasoning to the other cells of Fig. 4.7a, we are able to construct a two-way table showing the expected frequencies for each cell, as shown in Fig. 4.7b.

Categorical variables are said to be associated when the chance to fall into a particular category for one variable depends upon the category one falls into for the other variable. If the variables were not related, a firm from sector 3 would have the same probability to receive a subsidy than a firm from sector 2 or sector 1. A comparison of the observed frequencies with the expected frequencies can tell us whether this is true or not. For instance, we can see from Fig. 4.7a and b that the number of firms from sector 3 that received a subsidy is 9 while it should have been 5 under the null hypothesis. This evidences an association between the variables. The chi-square test extends the approach by comparing all the observed frequencies with the expected frequencies.

The chi-square statistic is computed as follows. We need to subtract each expected frequency from the observed frequency in each cell to get a difference table (observed minus expected) as shown in Fig. 4.7c. Those values points out how far from H_0 we are. Now the question remains whether they are significantly large. The test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency in cell i, E_i is the expected frequency in cell i and C is the total number of cells. The critical value is obtained from a chi-square distribution. The number of degrees of freedom is specified as:

$$df = (\text{number of row categories} - 1) \times (\text{number of column categories} - 1)$$

The chi-square test is a one-tailed test. If the test statistic is higher than the critical value $\chi^2_\alpha(df)$, with $\alpha = 5\%$, we reject the null hypothesis H_0 of independence. Moreover, the test is considered as valid if the minimum expected frequency is greater than 1 and if at least 80% of the expected frequencies are equal to or greater than 5. If this does not hold true, then one must merge some of the categories together.

Coming back to example 1 (see Fig. 4.7c), the test statistic is computed as:

$$\chi^2 = \underbrace{\frac{(4)^2}{14} + \frac{(0)^2}{16} + \frac{(-4)^2}{10}}_{\text{First column}} + \underbrace{\frac{(-4)^2}{7} + \frac{(0)^2}{8} + \frac{(4)^2}{5}}_{\text{Second column}} \approx 8.23$$

The number of degrees of freedom is:

$$df = (3-1) \times (2-1) = 2$$

	F(x)=Pr{X<x}									
df	**0.5%**	**1.0%**	**2.5%**	**5.0%**	**10.0%**	**50.0%**	**90.0%**	**95.0%**	**97.5%**	**99.0%**
1	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635
2	0.010	0.020	0.051	0.103	0.211	1.386	4.605	**5.991**	7.378	9.210
3	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345
4	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277
5	0.412	0.554	0.831	1.145	1.610	4.351	9.236	11.070	12.833	15.086
6	0.676	0.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812
7	0.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475
8	1.344	1.646	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090
9	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666
10	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209
11	2.603	3.053	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725
12	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217
13	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688
14	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141
15	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578
16	5.142	5.812	6.908	7.962	9.312	15.338	23.542	26.296	28.845	32.000
17	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409
18	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805
19	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191
20	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566
21	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932
22	8.643	9.542	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289
23	9.260	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638
24	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980
25	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314
26	11.160	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642
27	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963
28	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278
29	13.121	14.256	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588
30	13.787	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892
31	14.458	15.655	17.539	19.281	21.434	30.336	41.422	44.985	48.232	52.191
32	15.134	16.362	18.291	20.072	22.271	31.336	42.585	46.194	49.480	53.486
33	15.815	17.074	19.047	20.867	23.110	32.336	43.745	47.400	50.725	54.776
34	16.501	17.789	19.806	21.664	23.952	33.336	44.903	48.602	51.966	56.061
35	17.192	18.509	20.569	22.465	24.797	34.336	46.059	49.802	53.203	57.342

Fig. 4.8 Cumulative probabilities of the chi-square distribution

Figure 4.8 provides the cumulative probabilities of the chi-square distribution. Those values have been generated in Excel using the function *CHIINV*. For a confidence level of 95% and a number of degrees of freedom equal to 2, we find:

$$\chi^2_{5\%}(2) = 5.991$$

The test statistic is greater than the critical value. We thereby reject the null hypothesis of independence. The plus or minus signs observed in Fig. 4.7c give us the direction of the effects. The way each cell contributes to the chi-square statistic can also guide the interpretation. The larger is the difference between the observed and expected frequencies, the greater is the test statistic. In our example, it can be seen that the probability that a firm from sector 1 receives a subsidy is lower than if chance alone was operating, while a firm from sector 3 is actually more likely to receive a subsidy.

Function *chisq.test* in R-CRAN can be used to perform the chi-square test of independence. Figure 4.9 provides an example. First, the data are uploaded using

```
> D=read.table("C://mydataRD.csv",head=TRUE,sep=";")
> mytable=table(D$Sector,D$Subsidy)
> mytable
     0  1
  1 18  3
  2 16  8
  3  6  9
> chisq.test(mytable)

        Pearson's Chi-squared test

data:  mytable
X-squared = 8.2286, df = 2, p-value = 0.01634

> chisq.test(mytable)$expected
     0 1
  1 14 7
  2 16 8
  3 10 5

> library(lsr)
> cramersV(mytable)
[1] 0.370328
```

Fig. 4.9 The chi-square test using R-CRAN

the *read.table* function. The command *table* is then used to generate a two-way table which is used along with *chisq.test* to perform the test. As can be seen, the value of the chi-square statistic is significant (the *p*-value is lower than 5%). By including $*expected* in the command, we are also able to examine the frequencies that are expected under the null hypothesis.

The correlation coefficient cannot be used to measure the association between two categorical variables. Under this framework, an alternative exists which meets this purpose: Cramér's *V*. It is a statistic measuring the strength of dependency between two categorical variables via a value ranging from 0 to 1:

$$V = \sqrt{\frac{\chi^2}{nt}}$$

where n is the sample size, and t represents the minimum dimension minus 1:

$$\mathrm{t} = \min\{\text{ number of row categories} - 1, \text{number of column categories} - 1\}$$

The closer V is to 0, the smaller the association between the categorical variables. On the other hand, V being close to 1 is an indication of a strong association (but not necessarily of statistical significance). In our numerical application we have $t = 1$ and $n = 60$. Cramér's V is computed as:

$$V = \sqrt{\frac{8.23}{60 \times 1}} \approx 0.37$$

The same result is found in R-CRAN using the *cramersV* command from the package *lsr* (see Fig. 4.9). In general, two-way tables which have a larger value of V can be considered to have a stronger relationship between the variables.

4.4 Tests of Difference Between Means

This section explains how to statistically demonstrate that there is a significant difference between two or more groups when one variable is categorical and the other numerical.

A two-sample t-test, or independent samples t-test, is used when one wants to test the difference between two population means:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

The null hypothesis states that there is no difference between means μ_1 and μ_2 while the alternative hypothesis states that there is a difference. The sample sizes for the two groups may or may not be equal. The test statistic is defined as a t-score:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1 and s_2 denote the (sample) standard deviation of group 1 and group 2, respectively. The critical value is obtained from a Student distribution. The number of degrees of freedom is given by:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

The test is usually defined as a two-tailed test. If the test statistic t^* is higher in absolute value than the critical value $t_{\alpha/2}(df)$, we reject the null hypothesis.

Imagine for instance that we would like to assert in example 1 whether there is a difference in patent claims between the firms that did not receive a subsidy (group 1) and those who did (group 2). Table 4.1 can be transformed to better highlight the differences between the groups, as shown in Table 4.2. We have:

$$n_1 = 40; n_2 = 20; \bar{x}_1 = 9.90; \bar{x}_2 = 27.75; s_1 = 10.78; s_2 = 15.41$$

Table 4.2 Difference between recipients and non-recipients: example 1

	Group 1: non-recipients		Group 2: recipients	
	Firm	Patents	Firm	Patents
	2	31	1	42
	3	20	4	33
	6	43	5	30
	9	7	7	37
	11	13	8	27
	15	13	10	21
	17	24	12	20
	18	1	13	42
	19	33	14	35
	21	3	16	26
	22	11	20	49
	23	28	24	3
	25	4	26	32
	27	2	32	21
	28	22	33	39
	29	2	35	25
	30	16	38	58
	31	1	41	13
	34	2	49	2
	36	0	57	0
	37	19		
	39	7		
	40	0		
	42	2		
	43	13		
	44	11		
	45	2		
	46	5		
	47	0		
	48	0		
	50	2		
	51	4		
	52	0		
	53	6		
	54	0		
	55	4		
	56	19		
	58	12		
	59	14		
	60	0		
Counts		**40**		**20**
Mean		**9.90**		**27.75**
Standard dev.		**10.78**		**15.41**

```
> qt(0.975,29)
[1] 2.04523

> D=read.table("C://mydataRD.csv",head=TRUE,sep=";")
> t.test(Patents~Subsidy,D)

        Welch Two Sample t-test
data:  Patents by Subsidy
t = -4.6425, df = 28.596, p-value = 7.047e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.718505  -9.981495
sample estimates:
mean in group 0 mean in group 1
           9.90           27.75
```

Fig. 4.10 Two-sample t-test in R-CRAN: example 1

Using this information, we calculate the test statistic as follows:

$$t^* = \frac{9.90 - 27.75}{\sqrt{\frac{(10.78)^2}{40} + \frac{(15.41)^2}{20}}} \approx -4.64$$

The number of degrees of freedom is:

$$df = \frac{\left(\frac{(10.78)^2}{40} + \frac{(15.41)^2}{20}\right)^2}{\frac{\left((10.78)^2/40\right)^2}{40-1} + \frac{\left((15.41)^2/20\right)^2}{20-1}} \approx 28.60$$

The critical value is obtained from the Student table for a 95% confidence level and approximately 29 degrees of freedom. We actually have $t_{2.5\%}(29)=2.045$ (see Fig. 4.10 where the function *qt* is used to compute that value). This critical value is lower than $|-4.64|$. We thereby conclude that there is a significant difference in patent claims between the recipients and the non-recipients.

In Fig. 4.10, the function $t\,.\,test$ from R-CRAN performs the two-sample t-test. The p-value is lower than 5% which means that we do reject H_0. As can be seen, the software yields a confidence interval for the difference between means $\mu_1 - \mu_2$. As expected, zero does not belong to that confidence interval. Excel also offers a way to perform this test with the "Analysis ToolPak" add-in (which can be found in the Excel options). The command to be used is "t-test: Two-Sample Assuming Unequal Variances".

At this stage, we need to bring attention to two points. First, the test concludes that the observed difference between the recipients and the non-recipients is not due to chance alone. It does not, however, point out a causal relationship between being a recipient and having a high R&D productivity. It has been shown in the previous section that the subsidies were not randomly assigned, the firms of sector 3 being more likely to receive a subsidy. This points out a selection bias. To put it simply, the observed difference between the groups could be due to sectoral differences as well. The path analysis diagram would be:

$$\begin{tikzcd}
Subsidy \arrow[r] & Patents \\
Sector \arrow[u] \arrow[ur] &
\end{tikzcd}$$

Firms that received a subsidy are also those who belong to sector 3 (high-technology sector) and have a higher number of patent claims. Only the use of more advanced techniques such as quasi-experimental or experimental techniques can allow us to distinguish between the two potential effects of subsidy and sector. Second, note also that the *t*-test is based on the condition that the population generating the data is normally distributed or close enough to normal. The test may thus perform poorly if this assumption does not hold true. Histograms can be helpful in this matter to identify the shape of the distribution, assuming that the data are representative of the population.

The two-sample *t*-test can be extended to more than two groups. The approach, also known as one-way analysis of variance or ANOVA, determines whether any of the group means are significantly different from each other:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$$

$$H_1 : \text{Not all } \mu \text{ are equal}$$

where K denotes the total number of groups. The numerical variable is assumed to be normally distributed in each group k. The test also relies on the assumption of equal variances across groups. If the ANOVA test yields a significant result, we reject the null hypothesis. This means that at least two means are significantly different from one another. The test does not however explicit which groups are significantly different from the others.

ANOVA is based on a comparison of two different sources of variation: the between-group variability and the within-group variability. The between-group variability is the difference between the mean values of each group and the mean of the whole sample (or grand mean). If the group means are very similar, then this variability is low. The within-group variability on the other hand measures the variance in each group. Assume for instance that we have three groups and only two observations in each of these groups:

$$\underbrace{10\ 10}_{\text{Group 1}}\quad \underbrace{20\ 20}_{\text{Group 2}}\quad \underbrace{30\ 30}_{\text{Group 3}}$$

In that case, there is no variability within the groups (values are the same in each group) but there is some variability between the groups (the means are different). Assume now that the observations are distributed as follows:

$$\underbrace{5\ 35}_{\text{Group 1}}\quad \underbrace{15\ 25}_{\text{Group 2}}\quad \underbrace{10\ 30}_{\text{Group 3}}$$

In that case, variability is observed within the groups (there is heterogeneity among the observations in each group), but there is no difference between the group means, which are equal to 20 in the present case.

Formally, the test statistic is based on a ratio of the between-group variability to the within-group variability. We have:

$$F = \frac{\text{Average variability between groups}}{\text{Average variability within groups}}$$

If the between-group variability is much higher than the within-group variability, then it provides support for the alternative hypothesis. First, for each group k of size n_k, one must compute the sample mean $\bar{x}_k$. One also needs to compute the grand mean $\bar{x}_{GM}$ for the whole set of observations:

$$\bar{x}_{GM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The grand mean is defined as the total of all the data values x_i divided by the total sample size n.

The estimation of the between-group variability is termed between-mean square (MS_{between} hereafter) and is defined as the average variation of the group means around the grand mean:

$$MS_{\text{between}} = \frac{\text{Sum of squares between}}{\text{Degrees of freedom}} = \frac{\sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x}_{GM})^2}{K-1}$$

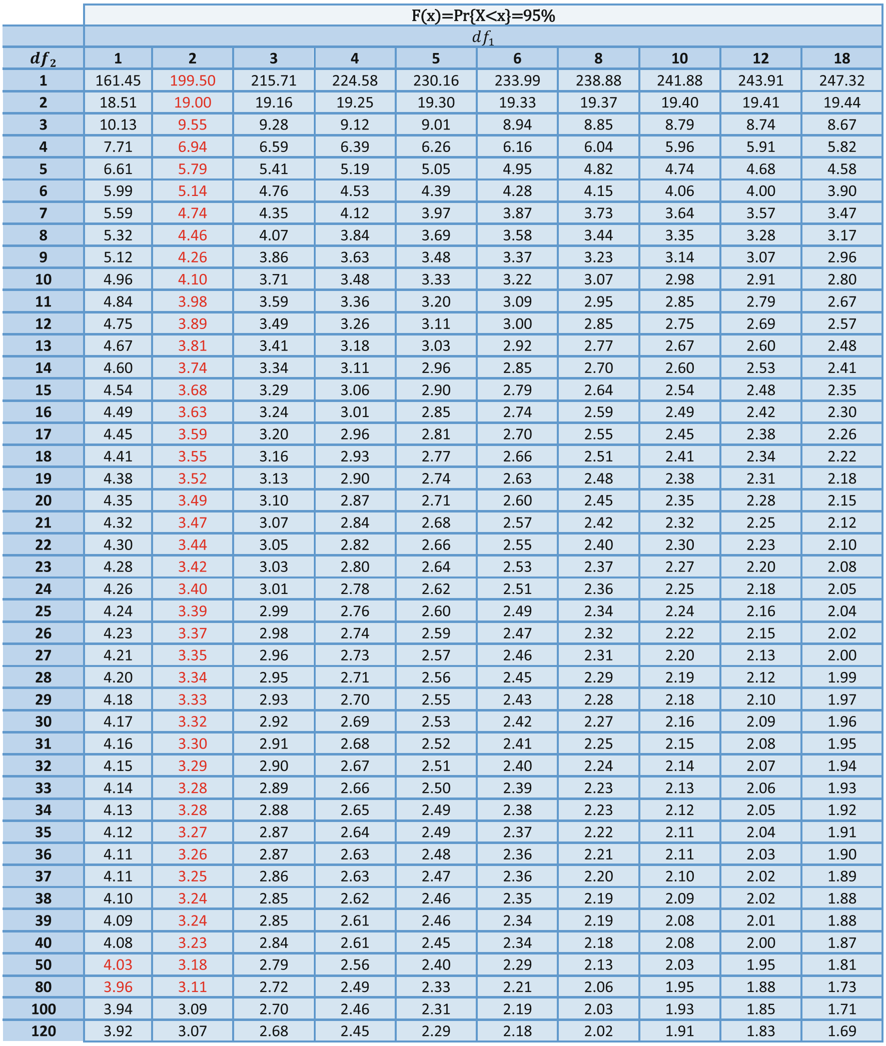
The estimation of the within-group variability is called within-mean square (MS_{within}). It is based on the sample variance observed in each group, denoted s_1^2, s_2^2, …, s_K^2, respectively. We have:

$$MS_{\text{within}} = \frac{\text{Sum of squares within}}{\text{Degrees of freedom}} = \frac{\sum_{k=1}^{K} (n_k - 1)s_k^2}{n-K}$$

It is the average variation of observations within each group around the group mean. Finally, the test statistic is defined as:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

It follows the Fisher-Snedecor distribution with $(K-1, n-K)$ degrees of freedom. It is a one-tailed test. Figure 4.11 provides a set of critical values for a 95% confidence level. This table has been constructed using the *FINV* function available



	F(x)=Pr{X<x}=95%									
	df_1									
df_2	1	2	3	4	5	6	8	10	12	18
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	241.88	243.91	247.32
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40	19.41	19.44
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.74	8.67
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.91	5.82
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.68	4.58
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	4.00	3.90
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.57	3.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.28	3.17
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.07	2.96
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.91	2.80
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.85	2.79	2.67
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.69	2.57
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.67	2.60	2.48
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.53	2.41
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.48	2.35
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.42	2.30
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.45	2.38	2.26
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.34	2.22
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.38	2.31	2.18
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.28	2.15
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.32	2.25	2.12
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.30	2.23	2.10
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.27	2.20	2.08
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25	2.18	2.05
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.24	2.16	2.04
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.22	2.15	2.02
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.20	2.13	2.00
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.19	2.12	1.99
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.18	2.10	1.97
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.09	1.96
31	4.16	3.30	2.91	2.68	2.52	2.41	2.25	2.15	2.08	1.95
32	4.15	3.29	2.90	2.67	2.51	2.40	2.24	2.14	2.07	1.94
33	4.14	3.28	2.89	2.66	2.50	2.39	2.23	2.13	2.06	1.93
34	4.13	3.28	2.88	2.65	2.49	2.38	2.23	2.12	2.05	1.92
35	4.12	3.27	2.87	2.64	2.49	2.37	2.22	2.11	2.04	1.91
36	4.11	3.26	2.87	2.63	2.48	2.36	2.21	2.11	2.03	1.90
37	4.11	3.25	2.86	2.63	2.47	2.36	2.20	2.10	2.02	1.89
38	4.10	3.24	2.85	2.62	2.46	2.35	2.19	2.09	2.02	1.88
39	4.09	3.24	2.85	2.61	2.46	2.34	2.19	2.08	2.01	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	2.00	1.87
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	2.03	1.95	1.81
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.95	1.88	1.73
100	3.94	3.09	2.70	2.46	2.31	2.19	2.03	1.93	1.85	1.71
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.91	1.83	1.69

Fig. 4.11 The Fisher-Snedecor distribution for a 5% risk level

in Excel. The decision will be to reject the null hypothesis if the test statistic is greater than the critical value $F_{5\%}(df_1, df_2)$.

To illustrate the method, assume that we would like to test whether there is a difference in patent claims between the industrial sectors of example 1. Table 4.3 reorganizes the raw data of Table 4.1 so that the groups and their characteristics are easily compared. The grand mean amounts to 15.85. We have:

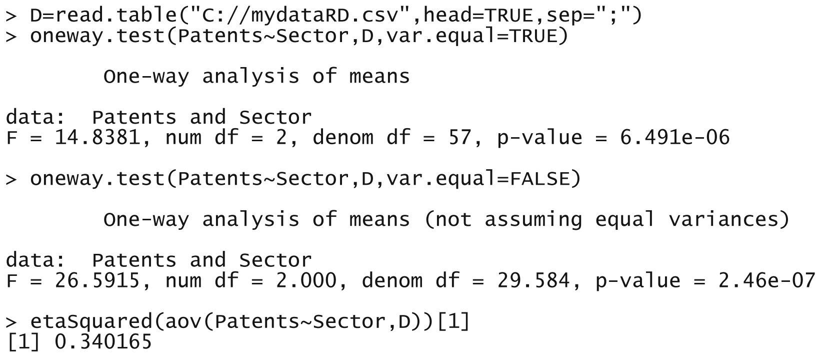
Table 4.3 Difference between recipients and non-recipients: example 1

	Group 1: sector 1		Group 2: sector 2		Group 3: sector 3	
	Firm	Patents	Firm	Patents	Firm	Patents
	40	0	16	26	1	42
	41	13	17	24	2	31
	42	2	18	1	3	20
	43	13	19	33	4	33
	44	11	20	49	5	30
	45	2	21	3	6	43
	46	5	22	11	7	37
	47	0	23	28	8	27
	48	0	24	3	9	7
	49	2	25	4	10	21
	50	2	26	32	11	13
	51	4	27	2	12	20
	52	0	28	22	13	42
	53	6	29	2	14	35
	54	0	30	16	15	13
	55	4	31	1		
	56	19	32	21		
	57	0	33	39		
	58	12	34	2		
	59	14	35	25		
	60	0	36	0		
			37	19		
			38	58		
			39	7		
Counts		**21**		**24**		**15**
Mean		**5.19**		**17.83**		**27.60**
Standard dev.		**5.93**		**16.40**		**11.50**
Grand mean	**15.85**					
Grand variance	**225.32**					

$$MS_{\text{between}} = \frac{21(5.19-15.85)^2 + 24(17.83-15.85)^2 + 15(27.60-15.85)^2}{3-1}$$

$$= \frac{4551.37}{2} = 2275.69$$

As for the within-group variability, we have:



```
> D=read.table("C://mydataRD.csv",head=TRUE,sep=";")
> oneway.test(Patents~Sector,D,var.equal=TRUE)

        One-way analysis of means

data:  Patents and Sector
F = 14.8381, num df = 2, denom df = 57, p-value = 6.491e-06

> oneway.test(Patents~Sector,D,var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  Patents and Sector
F = 26.5915, num df = 2.000, denom df = 29.584, p-value = 2.46e-07

> etaSquared(aov(Patents~Sector,D))[1]
[1] 0.340165
```

Fig. 4.12 One way ANOVA test in R-CRAN: example 1

$$MS_{\text{within}} = \frac{(21-1)\times(5.93)^2 + (24-1)\times(16.40)^2 + (15-1)\times(11.50)^2}{60-3}$$
$$= \frac{8740.89}{57} = 153.35$$

Finally, we get:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \approx 14.84$$

From Fig. 4.11, the critical value is found to be $F_{5\%}(2,57) \approx 3.1$, which is lower than the test statistic. The decision is to reject the null hypothesis. At least one of the group means is different. Figure 4.12 performs the test in R-CRAN. One can also relax the assumption of equal variances by setting $var\,.\,equal = FALSE$. In that case, the test produces the result of what is termed Welch's ANOVA. It is a form of one-way ANOVA that does not assume equal variances. The decision to reject the null hypothesis is confirmed.

Last, it is possible to measure the extent of the difference between means by computing the square correlation coefficient or eta-squared:

$$\eta^2 = \frac{\text{Sum of squares between}}{\text{Sum of squares total}} = \frac{\sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x}_{GM})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

It ranges from 0% to 100% and measures the extent to which the differences observed between groups (between-group variability) contribute to the total variance. In our example, the grand variance is displayed at the bottom of Table 4.3:

$$\text{Grand variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 225.32$$

We therefore have:

$$\eta^2 = \frac{21(5.19 - 15.85)^2 + 24(17.83 - 15.85)^2 + 15(27.60 - 15.85)^2}{\text{Grand variance} \times (\text{n} - 1)}$$

Replacing n with 60, we find 0.34 approximately. This means that sector membership accounts approximately for one third of the observed variance in patent claims. Figure 4.12 provides the coding to be used in R-CRAN to compute that value.

4.5 Principal Component Analysis

When faced with more than two numerical variables, it is very difficult or even impossible to provide a graphic representation of the data using standard plots. Principal components analysis overcomes this issue by projecting the whole set of observations on a 2-dimension map. The axes of the graph represent a combination of the variables so as to plot the observations with a minimum loss of information. The approach relies on maximizing the variance in the data, in order to highlight the heterogeneity of the observations. Each individual and each variable can be represented on this map, which provides an all-encompassing picture of the relationships in play.

Given the mathematical difficulties in creating the map manually, the approach requires a specific statistical environment, e.g., XLSTAT or the package FactoMineR in R-CRAN. Basically speaking, the purpose of the method is to reduce the dimensionality (number of variables) of the data set while retaining as much of the variability in the data as possible. Figure 4.13 illustrates the methodology with two variables. The raw data is initially represented through a scatter plot on the (X, Y) mapping. The first step is to search for the axis C_1 that best accounts for the majority of the variability in the data (see Fig. 4.13b). This axis is usually referred to as first principal component or first dimension. Once the first component is found, a second axis C_2 is constructed in a similar manner, subject to the constraint that C_2 is orthogonal to C_1. The second principal component accounts for the majority of the remaining variability and will not be correlated to C_1 (Fig. 4.13c). The approach continues until one gets as many components as there are variables. By construction, the contribution of C_1 to the total variance will be very high, while the contribution of C_2 will be lower. Note also that the origin of the graphic is placed at the center of gravity of the data, which is defined as the point $(\bar{X}, \bar{Y})$ in the interior of the data with mean coordinates (Fig. 4.13c). One can then decide to project the observations. In Fig. 4.13d, for instance, the observations are projected on the first component only, which results in a single line summarizing

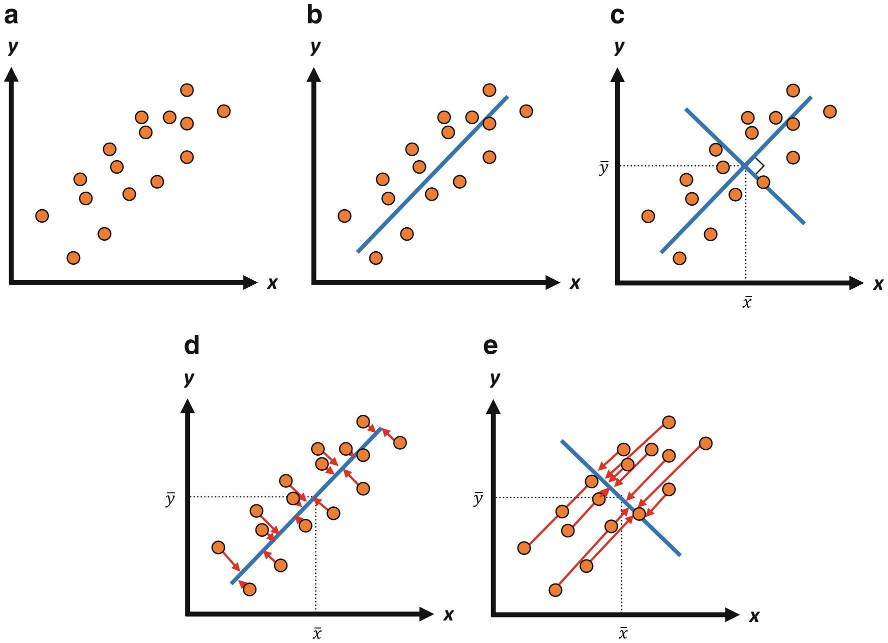


Fig. 4.13 The principal component analysis methodology. (**a**) Raw data. (**b**) Construction of the first component. (**c**) Construction of the second component. (**d**) Projection on the first component. (**e**) Projection on the second component

the information. Figure 4.13e uses the second component but provides less information about the variability in the data.

Figure 4.13 provides the reader with a simple illustration of how the method works. There is actually no need to perform a PCA in this particular example. There are two variables only and a simple scatter plot reveals the whole distribution pattern. In practice, one uses PCA when the number of variables to be projected is larger than two, in which case there are as many components as there are variables. The orthogonal projection is made on a plane defined by two of the resulting components (instead of a single straight line as illustrated in Fig. 4.13d and e). The first two components, C_1 and C_2, are generally used in this respect since, by construction, they offer more information about the heterogeneity of the observations. The remaining of the section illustrates this approach.

Mathematically speaking, the method consists in computing the mean value $\bar{x}_k$ for each variable x_k, $k = 1\ldots K$, and by centering the data. For all individual i one needs to subtract off the mean:

$$x'_{ik} = x_{ik} - \bar{x}_k$$

Each centered variable x'_{ik} has a mean of zero. The second step consists in calculating the covariance matrix which contains information about the variance and covariance of the centered variables:

$$V = \begin{bmatrix} s^2_{x'_1} & s_{x'_1,x'_2} & \cdots & s_{x'_1,x'_K} \\ s_{x'_1,x'_2} & s^2_{x'_2} & \cdots & s_{x'_2,x'_K} \\ \vdots & \vdots & \cdots & \vdots \\ s_{x'_1,x'_K} & s_{x'_2,x'_K} & \cdots & s^2_{x'_K} \end{bmatrix}$$

The variance appear along the diagonal and covariance appear in the off-diagonal elements. Entries are symmetric with respect to the main diagonal. Based on this covariance matrix, a software can calculate eigenvalues and eigenvectors.

The eigenvectors are the directions in which the data vary, i.e. the principal components. The eigenvalues on the other hand provide information about how much variance there is in those directions. One usually lists the eigenvalues in order from largest to smallest. The eigenvector with the highest eigenvalue is the first component, and so on. By construction, the sum of those values is equal to the number of variables. By dividing each eigenvalue by the number of variables we obtain a measure of how much variance each component extracts:

$$\begin{array}{c}\text{Contribution of each component} \\ \text{in the total variance}\end{array} = \frac{\text{Eigenvalue of the component}}{\text{Number of variables}}$$

In practice, it is common to use the Kaiser criterion which suggests to exclude any dimension with an eigenvalue lower than 1.

The principal component analysis approach is best exemplified in a public policy context. Imagine that we have (fictitious) data about a set of 60 districts that have been conferred responsibility for several welfare programs (e.g., protection of single mothers and children, social assistance for the disabled, aid to the elderly, and social welfare for the unemployed). Using information from Table 4.4, one would like to examine the link that may exist between socio-demographic data and per capita social expenditures. Variable *Social_Exp* denotes the level of expenditures per head in each district; *Income* is the mean taxable income; *Unemprate* represents the unemployment rate; *Shareof*60 is the share of people aged 60 and over; *Population* measures the number of inhabitants, *Density* is the population density per square kilometer. Variables *N_family*, *N_disabled*, *N_elder* and *N_benefits* represent the number of families, disabled, elder and unemployed who receive social assistance, respectively. For simplicity of exposition, we assume that the number of recipients in each district depends on eligibility criteria defined by a central government. In contrast, the amount of aid that each individual receives is within the discretion of the districts.

Principal component analysis can be used to provide a two-dimensional representation of the data that captures most of its variance. The method can also be used

Table 4.4 Data for example 2

Name	Social_ Exp	Income	Unemprate	Shareof60	Population	Density	N_families	N_disabled	N_elders	N_benefits
Allegan	184	6145	0.09	0.3	172580	28	255	847	2205	1720
Almedia	200	7260	0.14	0.21	1276453	128	2667	2882	7007	25302
Anangu	174	6463	0.18	0.22	884890	145	2016	2578	4879	29536
Asbury	141	6813	0.1	0.19	857295	127	1617	2948	6054	8617
Balnarring	168	6591	0.11	0.28	232252	34	404	928	1645	2112
Bartolo	165	7052	0.12	0.22	551498	90	1359	1780	1786	8253
Blackdale	171	7204	0.11	0.18	1085203	146	1522	3555	6158	13090
Bluewall	169	6674	0.08	0.23	250917	50	486	793	1419	1934
Bluford	195	6928	0.11	0.26	325077	35	890	695	2728	4136
Bridgemere	161	7200	0.11	0.25	314616	50	524	956	1835	3894
Brightwick	195	6681	0.12	0.22	731657	153	1105	3448	5206	10211
Brocton	148	5951	0.09	0.25	208285	42	171	699	1992	1591
Bulgandry	134	6893	0.14	0.19	1124006	165	1616	3288	5366	18194
Burlington	147	7564	0.1	0.2	616372	91	1150	1782	2122	7548
Campo	195	6275	0.1	0.3	159657	31	204	548	1526	2083
Coldwood	182	6254	0.13	0.27	304968	57	425	1052	2078	5677
Crystalash	202	5862	0.06	0.27	73507	14	44	272	1025	599
Dorwall	148	6296	0.12	0.2	729551	102	1493	1872	3788	9021
Eastsage	193	6233	0.11	0.24	481349	81	806	1921	5717	4778
Fayville	143	7338	0.1	0.18	565935	69	1240	1867	2788	8468
Fogview	158	6266	0.1	0.23	196378	32	447	951	1364	2267
Glassmallow	180	6263	0.07	0.22	283893	55	667	1358	2221	1798
Golconda	157	6595	0.11	0.19	715062	136	1672	1922	3084	12328
Goldbay	187	6147	0.1	0.22	192797	31	523	589	862	2655

Goldcrest	144	6648	0.12	0.24	639772	94	1037	2147	6539	7100
Gunyarra	221	6690	0.12	0.28	226776	33	614	1254	2651	3602
Hedgeview	226	6196	0.16	0.17	2554942	445	9117	6666	11607	62219
Hempstead	150	7222	0.12	0.16	765111	131	1671	2143	2839	9435
Holtsville	215	6245	0.1	0.24	292912	48	941	1092	1999	4083
Iceport	183	5699	0.16	0.19	1440495	216	4308	4649	8364	32302
Janlea	153	6990	0.1	0.22	602853	76	859	2053	4518	8460
Kakoma	179	7070	0.11	0.25	594245	78	720	1922	5817	10232
Lakesite	227	6580	0.12	0.27	222764	50	529	992	2923	3366
Lallat	193	6204	0.17	0.28	389051	95	1127	1934	2360	6617
Leavittsburg	144	8030	0.11	0.18	1566215	482	1710	2147	6539	7100
Maeystown	154	6104	0.1	0.22	229150	43	404	603	1294	2443
Makowata	187	6600	0.11	0.25	547472	64	796	2552	2957	6006
Malchi	189	6685	0.12	0.22	527687	85	965	1725	1827	6974
Mallowford	144	7248	0.1	0.2	369897	61	645	939	1620	3779
Marblehaven	112	6957	0.08	0.17	625854	143	719	1286	2330	5800
Matlacha	221	6760	0.15	0.19	1239784	197	2955	4616	8559	25457
Mentone	166	7971	0.09	0.14	1182543	200	2245	3834	4910	10397
Montgomery	141	10313	0.08	0.15	1353723	593	2003	3191	3031	10976
Murilla	136	6396	0.11	0.24	344574	57	619	1364	3420	3718
Nippering	161	6340	0.14	0.2	554595	90	1376	2002	4852	10179
Parkersburg	195	6364	0.12	0.27	342629	60	599	1351	2829	5315
Pepeekeo	196	5951	0.12	0.26	205519	55	207	822	1760	3818
Pepin	158	7228	0.16	0.26	891598	149	836	2417	5181	21929
Pomaria	171	6406	0.13	0.21	496948	139	806	1145	3405	12177
Pottersville	142	6445	0.1	0.24	535322	80	745	2332	3484	4315
Rockhollow	132	6591	0.12	0.23	397941	57	757	1191	2178	5997

(continued)

Table 4.4 (continued)

Name	Social_ Exp	Income	Unemprate	Shareof60	Population	Density	N_families	N_disabled	N_elders	N_benefits
Rosesea	196	7085	0.09	0.26	353941	64	564	1438	2928	5287
Saugus	129	6425	0.11	0.22	381791	65	506	1167	2078	4272
Sedley	212	7006	0.11	0.24	332773	45	906	1601	2461	4056
Stubbo	154	6824	0.11	0.19	137234	225	217	253	536	2020
Waggoner	166	9012	0.08	0.14	1133653	628	2261	2570	3119	13270
Warburto	204	11788	0.09	0.18	1419110	8063	2463	5207	6604	21113
Watonga	204	6802	0.15	0.14	1386023	5873	3373	3971	5491	39292
Westerwood	178	9106	0.11	0.17	1225473	5002	1877	4869	4905	21673
Wildefalcon	147	8040	0.1	0.13	1100782	883	1547	2577	2972	15356

```
# First step: choice of the set of variables
> D=read.table("C://mydataPCA.csv",head=TRUE,sep=";")
> library(FactoMineR)
> myPCA1=PCA(D[,c(6,8,9,10,11)])
> names(myPCA1)
[1] "eig"  "var"  "ind"  "svd"  "call"
> myPCA1$eig
      eigenvalue        % of variance     cumulative % of variance
comp 1 4.39943055       87.988611         87.98861
comp 2 0.27273892       5.454778          93.44339
comp 3 0.16321975       3.264395          96.70778
comp 4 0.09503159       1.900632          98.60842
comp 5 0.06957920       1.391584          100.00000
```

Fig. 4.14 Initial principal component analysis: program in R-CRAN

to identify groups or clusters of observations. Figures 4.14 and 4.15 provide the code to be implemented in R-CRAN.

The first step consists in choosing the set of variables to analyze**.** In Fig. 4.14, the database is uploaded using the *read . table* command. The command *library* is used to load the package *FactoMineR*. Then the *PCA* command from this package produces the preliminary principal component analysis using columns 6, 8, 9, 10, 11 of Table 4.4. This choice of variables is purely illustrative. Yet, one needs to make sure that the analysis is sufficiently informative given the variables used. Basically speaking, observations must be sufficiently dispersed. We can see from Fig. 4.16 that this condition does not hold true. Observations are rather concentrated around the first dimension (horizontal axis). Figure 4.17 helps visualizing the variables in the same plane, made of the first two components. The map draws what is termed a "correlation circle". When two variables are close to each other and far from the center, they are positively and linearly associated (the correlation coefficient is close to 1). If they are orthogonal, the correlation coefficient is close to 0 (the variables are not linearly associated). If they are on the opposite side of the center, then they are negatively correlated (the correlation coefficient is close to −1). When the variables are close to the center and far from the circle, interpreting the correlations is hazardous.

For instance, it can be observed from Fig. 4.17 that the variables are correlated with the first dimension only (they all point at the same direction). The reason of this phenomenon lies in the fact that the first dimension relates to the population size (column 6, Table 4.4), which is the dominant explanatory variable in this analysis, as evidenced by using the command *myPCA*1$*eig* which displays the set of eigenvalues (see Fig. 4.14). The higher is the number of inhabitants, the higher is the number of recipients by construction. To overcome this issue, one simply needs to divide the number of recipients by the population size. It is also important to understand at this stage that one usually does not choose randomly the variables to examine. Instead, the study should rely on a well-documented theoretical background.

The second step is the examination of the eigenvalues. Using the command *names* we get a description of the different outputs available with the *PCA* command. Only component 1 has an eigenvalue larger than 1, which means that this component is the main and overpowering dimension in play, according to the Kaiser criterion. This again shows evidence of a spurious relationship among the variables. We should therefore adjust the set of variables that is analyzed. This is done in Fig. 4.15 where

```
#Second and third steps: analysis of eigenvalues and explanatory power
> D$S_families=D$N_families/D$Population
> D$S_disabled=D$N_disabled/D$Population
> D$S_elders=D$N_elders/D$Population
> D$S_benefits=D$N_benefits/D$Population
> row.names(D)=substr(D[,1],1,5)
> myPCA2=PCA(D[,c(2,3,4,5,7,12,13,14,15)])
> myPCA2$eig
       eigenvalue       % of variance     cumulative % of variance
comp 1  2.9562542       32.847269         32.84727
comp 2  2.3489020       26.098911         58.94618
comp 3  1.4854989       16.505543         75.45172
comp 4  0.8740063       9.711181          85.16290
comp 5  0.4477400       4.974889          90.13779
comp 6  0.3040148       3.377943          93.51574
comp 7  0.2308340       2.564822          96.08056
comp 8  0.2160636       2.400706          98.48126
comp 9  0.1366863       1.518736          100.00000
#Fourth step: contributions of variables and correlations
> myPCA2$var$contrib
                 Dim.1       Dim.2        Dim.3     Dim.4        Dim.5
Social_Exp  9.32863945 10.58491717 1.697051e+01  1.628554 32.443697039
Income     17.33881541  0.02655803 2.192668e+01  0.344663  0.016179257
Unemprate   0.83844919 29.15205973 7.640736e+00  7.437152  8.230985791
Shareof60  24.96108358  1.80322115 8.002155e-02  5.589126  0.000723749
Density     6.57133797  4.35222902 3.510059e+01  6.955903  1.888799810
S_families  0.73813298 18.43997477 1.596589e-04 56.006325  0.016786846
S_disabled 16.85286579  0.08219843 1.454621e+01  2.337157 51.511571569
S_elders   23.29672725  2.46583045 2.650640e+00  4.684241  5.751884516
S_benefits  0.07394838 33.09301125 1.084460e+00 15.016878  0.139371423
> myPCA2$var$cor
                 Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
Social_Exp  0.52514598  0.49862744  0.50209236 -0.11930492 -0.38113436
Income     -0.71594655 -0.02497643  0.57071931  0.05488512 -0.00851123
Unemprate   0.15743789  0.82749822 -0.33690212  0.25495328  0.19197243
Shareof60   0.85901868 -0.20580548  0.03447781  0.22101881 -0.00180014
Density    -0.44075555  0.31973363  0.72209334  0.24656647  0.09196147
S_families  0.14771962  0.65813139 -0.00154004 -0.69964191  0.00866957
S_disabled  0.70584244 -0.04394042  0.46484815 -0.14292272  0.48024774
S_elders    0.82988583 -0.24066562  0.19843191  0.20233775 -0.16047893
S_benefits  0.04675577  0.88165889 -0.12692379  0.36228229 -0.02498042
> #Fifth step: cluster analysis
> mytypo=HCPC(myPCA2)
> library(nlme)
> gsummary(mytypo$data.clust,groups=mytypo$data.clust$clust)
Social_Exp   Income Unemprate Shareof60   Density  S_families
1   155.0370 7447.481 0.1040741 0.1914815 659.03704 0.001777269
2   189.1000 6535.800 0.1540000 0.2070000 747.70000 0.002368114
3   184.7826 6457.783 0.1052174 0.2556522  56.08696 0.001769439
   S_disabled    S_elders S_benefits clust
1 0.002865900 0.004536798 0.01238117     1
2 0.003119170 0.005969523 0.02333198     2
3 0.004061012 0.009037466 0.01240340     3
> mytypo$desc.var$quanti.var
                Eta2      P-value
Unemprate  0.5947855 6.592072e-12
S_elders   0.5893868 9.612514e-12
S_benefits 0.5777873 2.126386e-11
Shareof60  0.5119378 1.322946e-09
S_disabled 0.4096918 2.990510e-07
Social_Exp 0.3281485 1.194783e-05
Income     0.2228392 7.578112e-04
S_families 0.1368827 1.506570e-02
> oneway.test(Density~clust,mytypo$data.clust,var.equal=TRUE)
        One-way analysis of means
data:  Density and clust
F = 1.4639, num df = 2, denom df = 57, p-value = 0.2399
```

Fig. 4.15 Main principal component analysis: program in R-CRAN

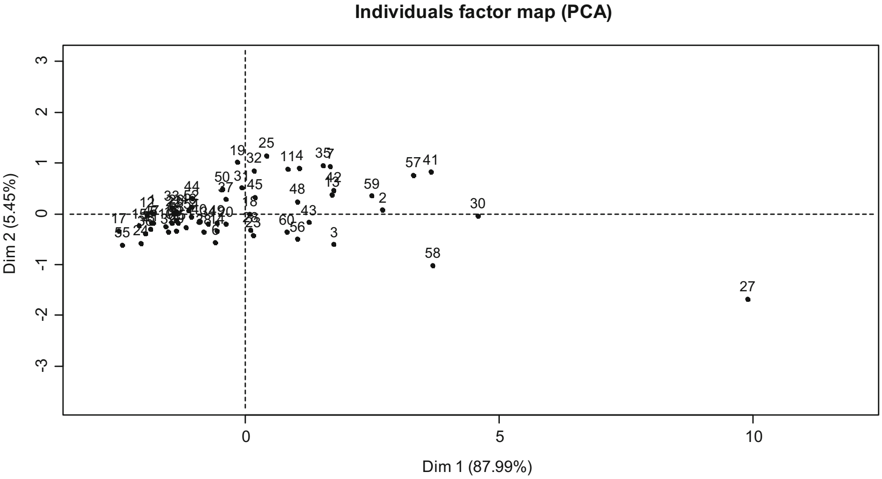


Fig. 4.16 Preliminary principal component analysis: set of observations

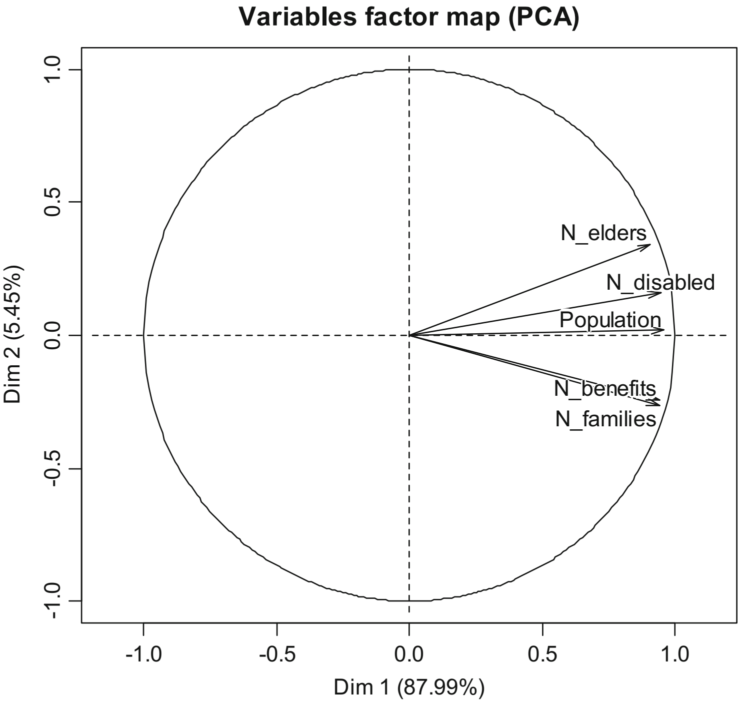


Fig. 4.17 Preliminary principal component analysis: set of variables

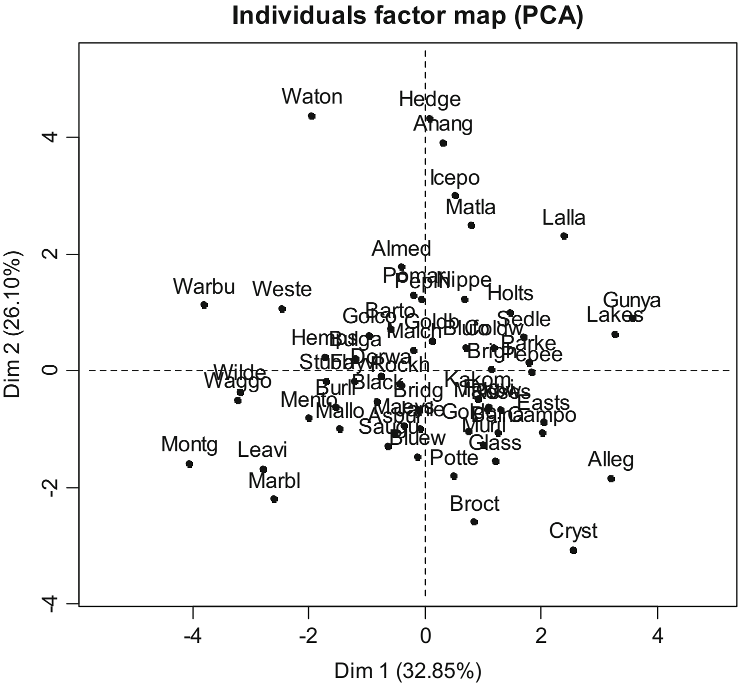


Fig. 4.18 Main principal component analysis: set of observations

data about the recipients are now expressed in percentage of the population ($D\$S\_families, D\$S\_disabled, D\$S\_elders, D\$S\_benefits$). Those new variables complete the database with additional columns by order of appearance (columns 12, 13, 14 and 15, respectively). Command $row\,.\,names(D)$ allows the name of the districts to appear now on the individuals factor map. Using the command $substr(D[, 1], 1, 5)$ we choose those names from database D using the first column (column *name* in Table 4.4) and extracting only the first five letters of those names. Figures 4.18 and 4.19 then provide the results using columns 2, 3, 4, 5, and 7 of Table 4.4 as well as the new set of variables 12, 13, 14 and 15. It can be seen from Fig. 4.18 that the observations are more dispersed than previously. From the new eigenvalues (Fig. 4.15), we can see that three dimensions have now an eigenvalue larger than one.

The third step investigates the explanatory power of the analysis. If we decide to focus the analysis on the first two components, one needs to ensure that C_1 and C_2 explain a sufficient amount of the variability in the data. As a rule of thumb, any analysis explaining less than 20% of the variance should be excluded. From Fig. 4.19, the first two components explain $26.10\,\% + 32.85\,\% = 58.95\%$ of the total variance. Note also that when one divides the eigenvalues values of Fig. 4.15 by the number of variables (here 9), we obtained the percentage of variance explained by the dimensions. For instance, for the first component we have: $32.85\% = 2.9562542/9$. The cumulative variance is then calculated by adding each variance to the running total of variances.

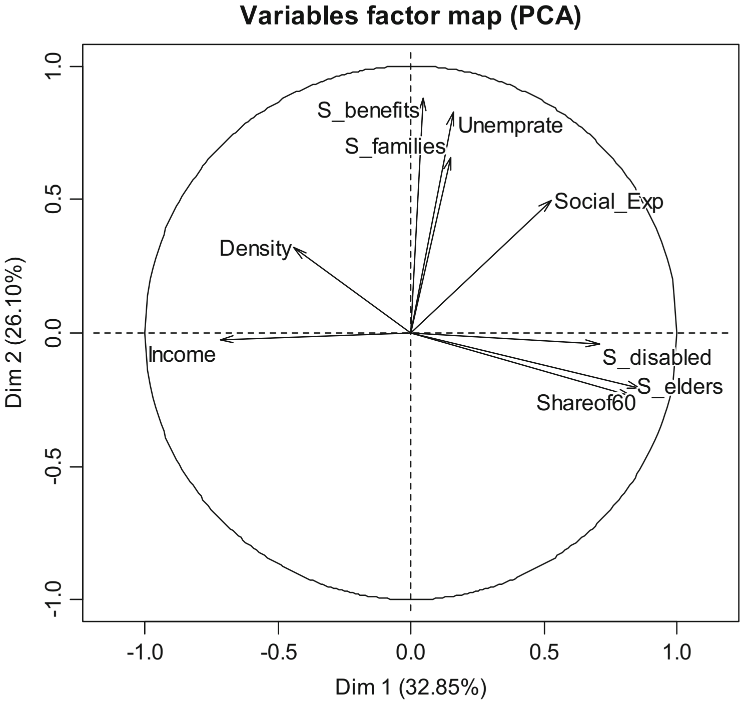
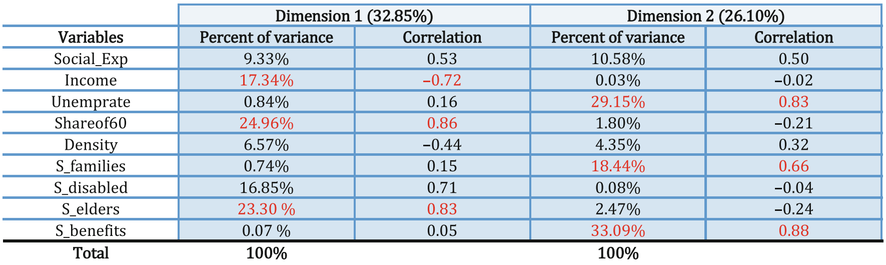


Fig. 4.19 Main principal component analysis: set of variables

The fourth step decomposes the variability in each principal component into contributions due to each variable. Those contributions can then be used together with the variables factor map to give a meaning to the axes. In Fig. 4.15, the command *myPCA*2$*var*$*contrib* displays the contributions. It can be seen that the variables that contribute most to the first component are *Shareof*60 (24.96%), *S_elders* (23.30%), and I*ncome* (17.34%). For the second dimension we have *S_benefits* (33.09%), *Unemprate* (29.15%) and *S_families* (18.44%). The command *myPCA*2$*var*$*cor* then displays the correlation coefficients between each component and each variable. Those correlations are related to how close the variables are to the axes of Fig. 4.19. In practice, it is common to provide information about correlations and variance as in Fig. 4.20. The first component is strongly structured by an opposition between a high share of elderly people on the East quadrant and high incomes on the West quadrant. The second component is explained mainly by high unemployment rates and a high share of unemployed people and families who benefit from social assistance (North quadrant).

The fifth step regroups the observations into clusters that are internally homogeneous (observations in each group must be as close as possible) but heterogeneous from one group to the other (the distance between each group must be sufficiently large). While there is no convention in this respect, the number of groups usually goes from two to eight. They can be constructed manually using information from the variables and individuals factor maps. For instance, in Fig. 4.19, it seems that



	Dimension 1 (32.85%)		Dimension 2 (26.10%)	
Variables	Percent of variance	Correlation	Percent of variance	Correlation
Social_Exp	9.33%	0.53	10.58%	0.50
Income	17.34%	−0.72	0.03%	−0.02
Unemprate	0.84%	0.16	29.15%	0.83
Shareof60	24.96%	0.86	1.80%	−0.21
Density	6.57%	−0.44	4.35%	0.32
S_families	0.74%	0.15	18.44%	0.66
S_disabled	16.85%	0.71	0.08%	−0.04
S_elders	23.30 %	0.83	2.47%	−0.24
S_benefits	0.07 %	0.05	33.09%	0.88
Total	**100%**		**100%**	

Fig. 4.20 Summary of components' structure

the disparities among the districts result from an opposition between the wealthier (West quadrant) and the poorer (East quadrant). The North quadrant highlights another group: those with a high unemployment rate. In R-CRAN, it is also possible to perform what is termed an agglomerative hierarchical clustering. Using the *HCPC* command, the software automatically constructs the groups in order to maximize the significance of differences between clusters. The method is an ANOVA-based approach (also known as Ward's method). Focusing on distance measurements, it merges the observations according to their proximity until one finally maximizes the ANOVA *F*-statistic. In R-CRAN, one simply needs to choose the number of clusters by clicking at the (or any other) level suggested by the *HCPC* command. In our example, the command suggests a number of three groups, as shown in Fig. 4.21.

Once the cluster typology has been created, it can be characterized with words and cross-analyzed with the variables used for the study (and possibly additional variables when those are available). One advantage of the *HCPC* command is that it reproduces the original data (renamed *data . clust*) with a supplementary column (called *clust*) containing the Ward partition. This supplementary column is simply a dummy variable specifying the cluster (e.g., cluster 1, 2 or 3) each unit belongs to. Using the command *gsummary* from the package *nlme*, it is then possible to provide summary statistics for each group, as shown in Fig. 4.15. The first entry in the *gsummary* command corresponds to the new database *mytypo*\$*data . clust*, and the second entry *mytypo*\$*data . clust*\$*clust* specifies the partition to be used. We thereby obtain the group means for each variable. Hence, it is possible to offer a description of the clusters. For instance, it can be seen that cluster 1 is characterized by a relatively low share of elderly people (19% on average), a low unemployment rate (10%) but a high income per capita (\$7447). Cluster 2 is defined on the other hand by a high unemployment rate (15%), a high population density (747 inhabitants per km^2) and a high proportion of families and unemployed people who benefit from social assistance. Last, cluster 3 is described by a high share of elderly people (25%), a low population density (56 inhabitants per km^2), and a high proportion of recipient elderly and disabled people.

Once those five steps have been implemented, the analysis must end with an answer to the research question. Table 4.5 displays the results of traditional ANOVA tests using the command *mytypo*\$*desc . var*\$*quanti . var* (Fig. 4.15). The

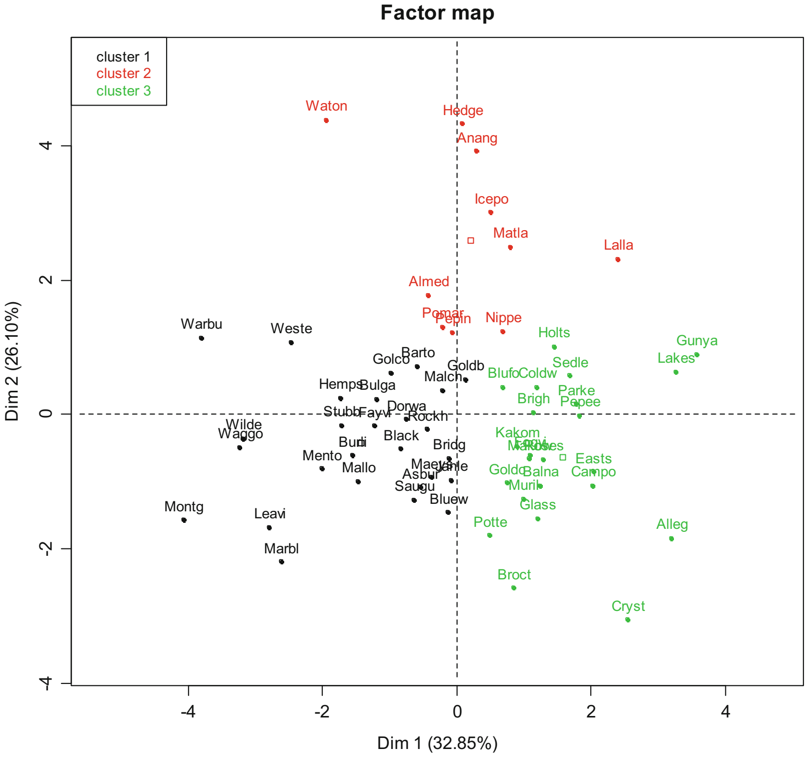


Fig. 4.21 Main principal component analysis: hierarchical clustering

first column *Eta*2 stands for the square correlation coefficient (see Sect. 4.4) while the second column offers the *p*-value of the ANOVA tests. Only the variables that yield a *p*-value lower than 5% are displayed. The observed differences between clusters are significant for all variables but *Density* (not displayed: the ANOVA test implemented using the *oneway . test* command confirms it). Thus it seems that the partition presented in Fig. 4.21 yields significant results. Three profiles of districts exist depending on their socio-demographics. Now the question remains whether there is a relationship between those profiles and the observed level of public spending. From Fig. 4.19, we can see that variable *Social_Exp* is pointing at the North-East quadrant where clusters 2 and 3 partly stand. As shown with the *gsummary* (Fig. 4.15), those two clusters are those that spend more on average ($189 for cluster 2 and $184 for cluster 3 against $155 for cluster 1). According to the ANOVA test (*p*-value=1.19e-05) those differences are also significant. This shows evidence of a relationship between the demand structure and public spending.

Table 4.5 Summary of the clustering's structure

Variables	Cluster 1 (mean value)	Cluster 2 (mean value)	Cluster 3 (mean value)	ANOVA test (significant difference)
Social_Exp	155.037	189.100	184.783	Yes
Income	7447.481	6535.800	6457.783	Yes
Unemprate	0.104	0.154	0.105	Yes
Shareof60	0.191	0.207	0.256	Yes
Density	659.037	747.700	56.087	No
S_families	0.0018	0.0024	0.0018	Yes
S_disabled	0.0029	0.0031	0.0041	Yes
S_elders	0.0045	0.0060	0.0090	Yes
S_benefits	0.0124	0.0233	0.0124	Yes

4.6 Multiple Correspondence Analysis

The approach of multiple correspondence analysis is very close to that of principal component analysis in that the method produces a 2D map of the data where each observation and each variable is represented. In its simplest form, when the observations are described by two categorical variables, the approach consists in producing a two-way table and performing what is called a correspondence analysis. In a more general manner, the approach can be applied to a larger set of categorical variables, in which case the method is termed multiple correspondence analysis.

Principal component analysis and multiple correspondence analysis are similar in that both procedures aim to provide a simple illustration of the phenomena at stake. There are slight differences however. First, multiple correspondence analysis is used to analyze a set of categorical variables. Each variable thus comprises several categories (e.g., "male" or "female") and each of those categories will be represented. The approach can accommodate numerical variables as long as they are recoded into classes. For instance, a variable such as "age" can be recoded into an ordinal variable including several levels, e.g., "less than 20", "between 21 and 30", "between 31 and 40", and so on. Second, the variables factor map does not provide a correlation circle. Two categories are said to be associated when they are close to each other on the multiple correspondence analysis factor map. Last, since all the categories will be represented on the map, one needs to choose accurately the set of categories to be drawn. When the number of categories is too high, the map can become unreadable.

Despite the previous differences, the method is quite similar to what has been presented in Sect. 4.5. First, one needs to examine how the observations are distributed on the individuals factor map. If one wants to understand the structure of the data, those observations must be sufficiently dispersed. Second, one should examine the eigenvalues that are associated with the analysis. The total number of components and, thereby, the total number of eigenvalues, depends now both on the number of variables K and the number of levels L_k in each variable k:

$$\text{Number of components or eigenvalues} = \sum_{k=1}^{K} L_k - K$$

The mean of the eigenvalues is equal to K. Under the multiple correspondence analysis framework, Kaiser's rule consists in excluding components that have an eigenvalue lower than $1/K$. Third, one needs to ensure that the first two components (C_1 and C_2) explain a sufficient amount of the variability in the data, i.e. at least 20% of the total variance. It is also possible to analyze other components (e.g., component C_3) if their eigenvalue is larger than $1/K$. Fourth, the contributions of the variables to the components serve to give a meaning to the map. The position of the categories on the map also helps to determine whether they yield a positive or a negative contribution. Fifth, a typology is created by clustering the observations according to their position on the map. Sixth, the partition must be defined with words and cross-analyzed with the other variables. Last, the typology should be used to answer the research question.

We illustrate the method using fictitious data about a health survey. The aim is to assess the relationships between a set of socio-demographic characteristics and three chronic diseases A, B and C (e.g., respectively diabetes, heart disease, and respiratory disease). First, using the individuals factor map, we will try to establish different profiles of individuals. Are some individuals similar with respect to their characteristics? Can we oppose a group of individuals to another one? Using the variables factor map, we will also examine how the different variables are associated. For instance, are some diseases more likely to appear in one group than in another? Are the differences between the group means statistically significant? The data set is provided in Table 4.6. It consists of $K = 7$ variables and $\sum L_k = 19$ categories: *Gender* (two levels: male or female); *Occupation* (five levels: unemployed, unskilled worker, skilled worker, manager or professional); *Education* (three levels: primary, secondary or higher education); *Residence* (three levels: rural, semi urban, or urban); *Disease A*, *Disease B* and *Disease C* (two levels: yes or no).

The R-CRAN program is provided in Fig. 4.22. First the data is uploaded with the *read.table* command and renamed as D. A multiple correspondence analysis is then implemented with the *MCA* function available with the package *FactoMineR*. Using the command *names* we get a description of the different outputs available with the *MCA* command. For instance, the command *myMCA$eig* yields the eigenvalues. According to the Kaiser criterion, those values must be higher than the inverse of the number of variables. This is for instance true for the first two dimensions which are larger than $1/K = 0.14$. Together, those dimensions account for 33.92% of the total variance. Figure 4.23 shows that the observations are sufficiently dispersed in the resulting system. The analysis can therefore pursue with an analysis of the variables.

Table 4.6 Data for example 3

Individual	Gender	Occupation	Education	Residence	Disease A	Disease B	Disease C
1	Male	Unemployed	Primary	Rural	Disease A:yes	Disease B:no	Disease C:no
2	Male	Manager	Higher	Urban	Disease A:no	Disease B:yes	Disease C:no
3	Male	Unskilled worker	Primary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
4	Male	Unemployed	Primary	Urban	Disease A:yes	Disease B:no	Disease C:no
5	Male	Skilled worker	Higher	Urban	Disease A:no	Disease B:yes	Disease C:no
6	Male	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
7	Male	Skilled worker	Higher	Urban	Disease A:no	Disease B:no	Disease C:no
8	Male	Unskilled worker	Primary	Urban	Disease A:no	Disease B:no	Disease C:no
9	Male	Skilled worker	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
10	Male	Unemployed	Secondary	Urban	Disease A:no	Disease B:no	Disease C:yes
11	Male	Unemployed	Primary	Urban	Disease A:no	Disease B:yes	Disease C:no
12	Male	Manager	Primary	Semi-urban	Disease A:yes	Disease B:no	Disease C:no
13	Male	Skilled worker	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
14	Male	Unskilled worker	Primary	Rural	Disease A:yes	Disease B:no	Disease C:no
15	Male	Professional	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
16	Male	Manager	Higher	Rural	Disease A:no	Disease B:no	Disease C:no
17	Male	Skilled worker	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
18	Male	Unskilled worker	Primary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
19	Male	Skilled worker	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
20	Male	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
21	Male	Unskilled worker	Primary	Rural	Disease A:yes	Disease B:no	Disease C:yes
22	Male	Unskilled worker	Primary	Urban	Disease A:no	Disease B:no	Disease C:no
23	Male	Manager	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes
24	Male	Professional	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
25	Male	Manager	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes

26	Male	Manager	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
27	Male	Manager	Higher	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
28	Male	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
29	Male	Skilled worker	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes
30	Male	Manager	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
31	Female	Skilled worker	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
32	Female	Manager	Secondary	Semi-urban	Disease A:yes	Disease B:no	Disease C:no
33	Female	Skilled worker	Secondary	Urban	Disease A:no	Disease B:no	Disease C:yes
34	Female	Unskilled worker	Primary	Urban	Disease A:no	Disease B:no	Disease C:no
35	Female	Unemployed	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no
36	Female	Professional	Higher	Urban	Disease A:yes	Disease B:no	Disease C:no
37	Female	Manager	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
38	Female	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
39	Female	Unemployed	Primary	Rural	Disease A:no	Disease B:no	Disease C:no
40	Female	Manager	Higher	Rural	Disease A:no	Disease B:yes	Disease C:no
41	Female	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
42	Female	Skilled worker	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
43	Female	Skilled worker	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
44	Female	Unskilled worker	Primary	Semi-urban	Disease A:yes	Disease B:no	Disease C:no
45	Female	Skilled worker	Secondary	Rural	Disease A:no	Disease B:yes	Disease C:no
46	Female	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
47	Female	Skilled worker	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
48	Female	Unemployed	Primary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
49	Female	Skilled worker	Higher	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
50	Female	Unemployed	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
51	Female	Unemployed	Higher	Semi-urban	Disease A:yes	Disease B:no	Disease C:no
52	Female	Manager	Secondary	Rural	Disease A:no	Disease B:no	Disease C:no

(continued)

Table 4.6 (continued)

Individual	Gender	Occupation	Education	Residence	Disease A	Disease B	Disease C
53	Female	Professional	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes
54	Female	Skilled worker	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:yes
55	Female	Unemployed	Primary	Semi-urban	Disease A:no	Disease B:yes	Disease C:no
56	Female	Skilled worker	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes
57	Female	Professional	Higher	Rural	Disease A:no	Disease B:yes	Disease C:no
58	Female	Manager	Secondary	Semi-urban	Disease A:no	Disease B:no	Disease C:no
59	Female	Manager	Higher	Rural	Disease A:no	Disease B:no	Disease C:yes
60	Female	Unemployed	Primary	Urban	Disease A:no	Disease B:no	Disease C:yes

```
> D=read.table("C://mydataMCA.csv",head=TRUE,sep=";")
> library(FactoMineR)
> myMCA=MCA(D[,2:8])
> names(myMCA)
[1] "eig"  "call" "ind"  "var"  "svd"

> myMCA$eig
       eigenvalue     % of variance        cumulative % of variance
dim 1  0.33261974       19.402818                   19.40282
dim 2  0.24887359       14.517626                   33.92044
dim 3  0.19200928       11.200541                   45.12099
dim 4  0.16385826       9.558399                    54.67938
dim 5  0.15646295       9.127005                    63.80639
dim 6  0.13835487       8.070701                    71.87709
dim 7  0.12844755       7.492774                    79.36986
dim 8  0.10818008       6.310505                    85.68037
dim 9  0.09632747       5.619102                    91.29947
dim 10 0.08672598       5.059016                    96.35849
dim 11 0.04098758       2.390942                    98.74943
dim 12 0.02143835       1.250570                    100.00000

> myMCA$var$contrib
                        Dim 1     Dim 2     Dim 3     Dim 4     Dim 5
Female               1.422887  0.007335 12.714660  5.881834  2.995213
Male                 1.422887  0.007335 12.714660  5.881834  2.995213
Manager              2.365098 11.071600  0.045309  0.067513 12.501410
Professional         1.261130  2.229042  7.068973 38.011546  1.022983
Skilled worker       6.939383  3.582375  0.000928 13.749688  2.055799
Unemployed           5.644497  0.783730 14.222770  1.407241 20.917692
Unskilled worker 18.138696  0.133935  4.108135  2.050654  6.759118
Higher               7.515689  8.440127  8.969409  2.983678  0.285050
Primary             24.786450  1.459062  0.195716  0.101187  0.015580
Secondary            2.262882 15.486619  6.851930  2.120100  0.410044
Rural                0.430353  2.350919  5.896896  1.617353  2.563523
Semi-urban           0.783467  0.060078  9.452323  0.271180 15.008541
Urban                4.820344  5.740759  1.958002  6.068803 13.638779
Disease A:no         2.140559  0.015447  0.043012  2.838967  0.792631
Disease A:yes       12.129834  0.087534  0.243734 16.087479  4.491574
Disease B:no         1.748599  8.041709  0.115662  0.071470  0.585609
Disease B:yes        5.745398 26.422757  0.380034  0.234829  1.924142
Disease C:no         0.125190  3.989230  4.255057  0.157149  3.127178
Disease C:yes        0.316657 10.090406 10.762791  0.397495  7.909921

> mytypo=HCPC(myMCA)
> names(mytypo)
[1] "data.clust" "desc.var"   "desc.axes"  "call"      "desc.ind"
> mytypo$desc.var$test.chi2
                p.value df
Education  9.226797e-15  4
Disease.B  1.776817e-10  2
Occupation 5.740610e-08  8
Disease.A  1.530932e-04  2
Disease.C  2.446689e-03  2
> mytable=table(mytypo$data.clust$Residence,mytypo$data.clust$clust)
> chisq.test(mytable)

        Pearson's Chi-squared test
data:  table(mytypo$data.clust$Residence, mytypo$data.clust$clust)
X-squared = 8.966, df = 4, p-value = 0.06196
Warning message:
In chisq.test(table(mytypo$data.clust$Residence,
mytypo$data.clust$clust)) : Chi-squared approximation may be incorrect
```

Fig. 4.22 Multiple correspondence analysis in R-CRAN: example 3

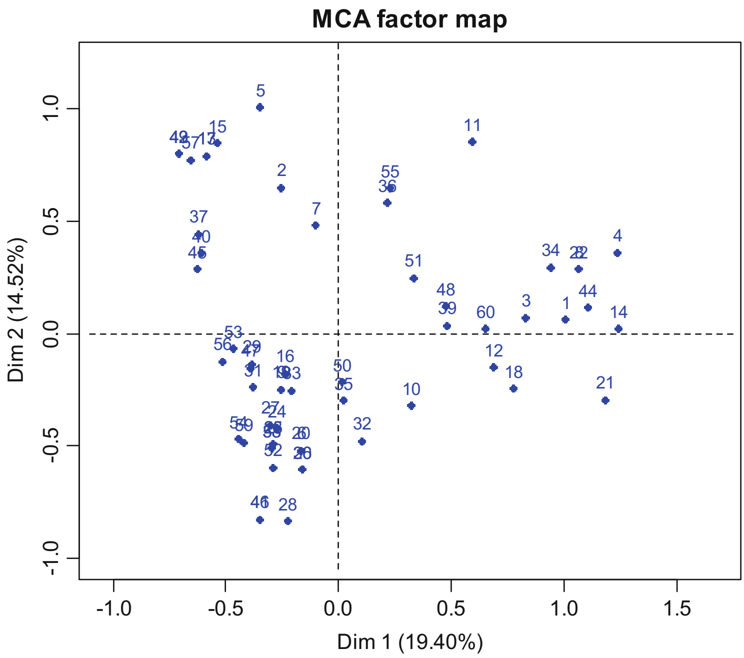


Fig. 4.23 Individuals factor map: example 3

In Fig. 4.24, the variables are presented with the purpose of highlighting the main associations. For instance, it can be seen that *Education* and *Occupation* are associated as well as *Disease . C* and *Residence*. Variable *Gender* appears at the origin of the map, suggesting that this variable is unlikely to contribute to the axes. Figure 4.25 provides further information by displaying the whole set of categories. The command *myMCA*\$*var*\$*contrib* provides the contributions of those categories to each dimension. The set of variables that contributes most to the first component includes *Primary* (24.78%) and *Unskilled . worker* (18.13%). For the second dimension, the main contribution is attributed to *Disease . B* : *yes* (26.42%).

Using Fig. 4.25, we can also visualize how each category is related to the others. For instance, "unemployed" and "unskilled workers" are associated with "primary education" and "disease *A*". Professionals and skilled workers appear along with "higher education" and tend towards "disease *B*". Those whose place of residency is located in a rural area are more likely to get disease *C*. In other words, a typology with three groups seems to be apparent. A cluster analysis can further confirm this statement.

Using the *HCPC* command, we obtain the partition of Fig. 4.26. Three groups are underlined. The program in Fig. 4.22 offers to test the differences observed between the mean groups using traditional chi-square tests. The command *mytypo*\$*desc . var*\$*test . chi*2 offers the *p*-values for the most significant variables

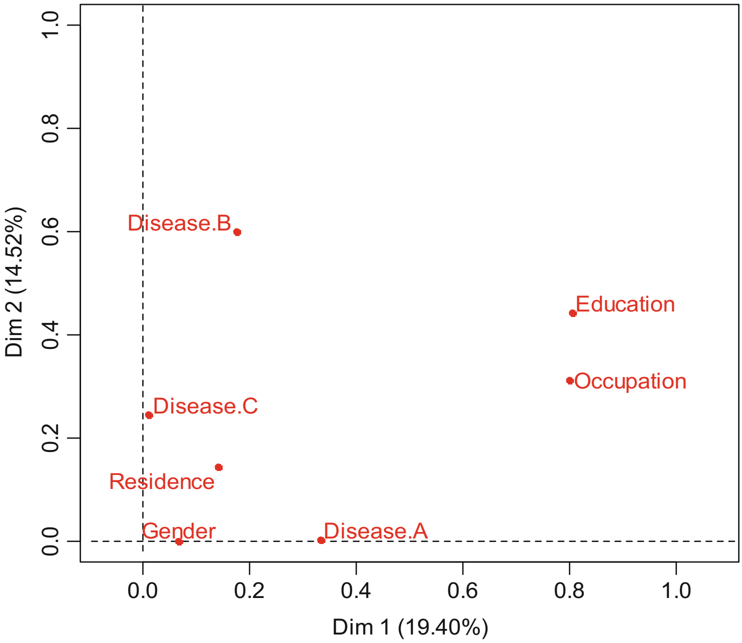


Fig. 4.24 Variables factor map: example 3

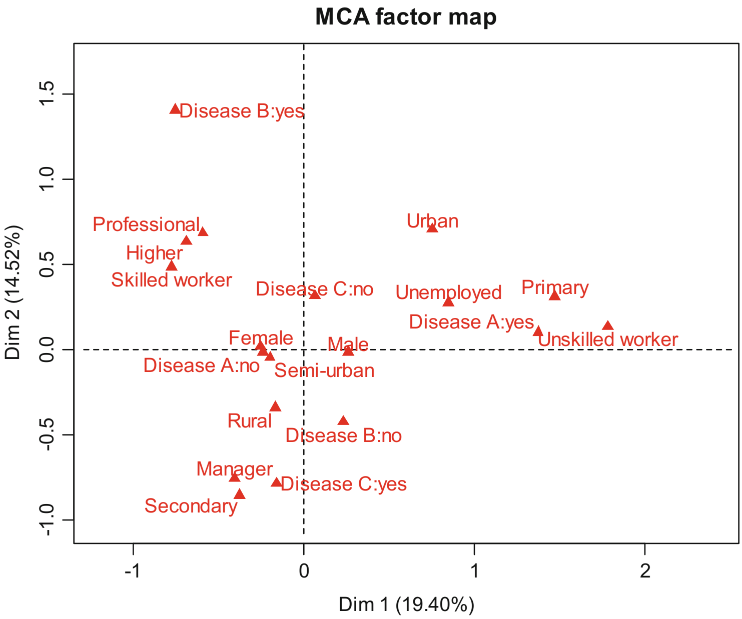


Fig. 4.25 Categories factor map: example 3

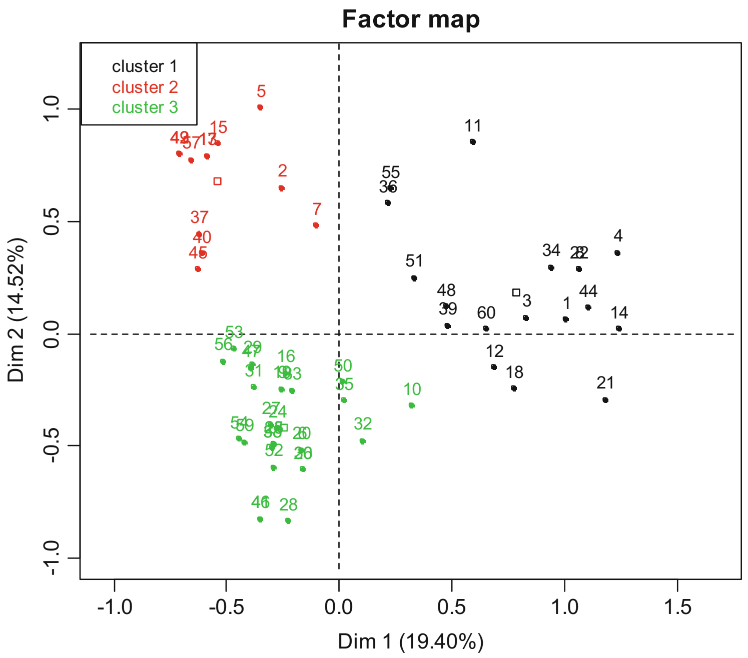


Fig. 4.26 Hierarchical clustering: example 3

(whose p-value is lower than 5%). The tests confirm that an association exists between the resulting partition and the following variables: *Education*, *Disease . B*, *Occupation*, *Disease . A* and *Disease . C*. As for variable *Residence*, since this variable does not appear in the results, the chi-square test must be completed manually as in Sect. 4.3. Using the *chisq . test* command, we find that the p-value amounts to 0.061 and is not far from 5%.

The analysis can go further by examining the different relationships highlighted by the multiple correspondence analysis. Using chi-square tests, we can for instance test if there is an association between *Education* and *Disease . A* or *Disease . B*. We could also test for a relationship between *Residence* and *Disease . C*. Two-way tables would ideally complete the analysis by providing a description of the distributions among the different categories. Last, the results could serve different policy purposes. For instance, a more cost-efficient screening process of a given disease could be implemented in relation with what has been observed in terms of socio-demographics. Differentiated prevention policies could also be set up according to the different areas at risk.

Bibliographical Guideline

The conception of the correlation coefficient is attributed to Sir Francis Galton, a cousin of Charles Darwin. In a study dating from 1877, he examined how the size of a sweet pea depends on the size of the parent seed (Galton 1877). Karl Pearson then developed a more rigorous treatment of the mathematics of the coefficient. Pearson (1900) also introduced what became known as the chi-square test. A rudimentary form of principal component analysis can be found in Galton (1889), Pearson (1901), and MacDonell (1902), while correspondence analysis was first discussed in Pearson (1906).

Several textbooks can familiarize the reader with the various concepts and techniques presented in this chapter. Rosenthal and Rosenthal (2011) offer an introduction to descriptive statistics (e.g., the correlation coefficient), inferential statistics (the chi-square and ANOVA tests) and data interpretation. Lang and Secic (2006) offer detailed guidelines for reporting and interpreting statistical relationships in biomedical science. Last, Giudici (2005) and Tufféry (2011) describe applied data mining methods such as principal component analysis and multiple correspondence analysis with the purpose of exploring and modeling large databases.

References

Galton, F. (1877). Typical laws of heredity. *Nature, 15*, 492–495.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Giudici, P. (2005). *Applied data mining: Statistical methods for business and industry*. New York: Wiley.

Lang, T. A., & Secic, M. (2006). *How to report statistics in medicine: Annotated guidelines for authors, editors, and reviewers*. Philadelphia, PA: ACP.

MacDonell, W. R. (1902). On criminal anthropometry and the identification of criminals. *Biometrika, 1*, 177–227.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series, 5*, 157–175.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series, 6*, 559–572.

Pearson, K. (1906). On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. *Biometrika, 5*, 176–178.

Rosenthal, G., & Rosenthal, J. A. (2011). *Statistics and data interpretation for social work*. New York: Springer.

Tufféry, S. (2011). *Data mining and statistics for decision making*. Wiley.

Econometric Analysis 5

5.1 Understanding the Basic Regression Model

Econometrics is as statistical tool that investigates cause-and-effect relationships with the aim of testing a theory or hypothesis, quantifying it, and providing indications about the evolution of some outcome of interest. The approach is particularly relevant in the public sector to provide a better understanding of the context, or to study the effect of a particular policy intervention. For example, one may use econometrics to examine the heterogeneity of local public spending using data about socio-demographics. Patient behavior or drug efficiency can be predicted based on the patients' characteristics (e.g., age, smoker type, income or gender). Failure at school can be analyzed based on revision time or lecture attendance.

The econometric methodology relies on mathematical models which are supposed to offer a simplified but accurate representation of the process under examination. Broadly speaking, a model consists in one or several equations that the evaluator wishes to estimate. Those equations include the variables that are thought to be relevant in explaining the phenomenon in question. A dependent variable y (explained or endogenous variable) is expressed as a function of several independent variables $x_1, x_2, \ldots, x_K$ (explanatory or exogenous variables, or regressors). We test:

$$y = f(x_1, x_2, \ldots, x_K)$$

This type of analysis is usually referred to as "regression analysis". Once the model is estimated on a sample, the parameters of the model serve to test whether the independent variables (the x's), have a significant impact on the outcome of interest (y). It is also possible to use the estimated function to analyze possible scenarios, to forecast the future or to guide policy formulation.

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_5

In its simplest form, the econometric approach sets a dependent variable as a function of a single independent variable. This type of analysis, also known as simple linear regression, defines a population regression function as follows:

$$y_i = \alpha_1 + \alpha_2 x_i + \epsilon_i \text{ (population)}$$

where i denotes the units under evaluation (e.g., individuals, patients, students, facilities, cities, etc.), α_1 and α_2 are the parameters to be estimated and ϵ_i is what is called an error term or random disturbance. For any unit i, the observed value of y_i is the sum of two components, a deterministic part $(\alpha_1 + \alpha_2 x_i)$ and a stochastic part (ϵ_i) which is meant to capture all the factors the model omits. We have the following path analysis diagram:

$$\begin{tikzcd} x \arrow[r] & y \\ & \epsilon \arrow[u] \end{tikzcd}$$

The stochastic error term ϵ is present in the equation because (1) other variables could also affect the dependent variable, (2) measurement errors are possible, (3) the linear functional form could be inaccurate (e.g., nonlinear relationships could be in play), (4) unpredictable or purely random variations of the dependent variable can never be ruled out.

The objective of an econometric study is to provide a numerical value to both the deterministic part and the stochastic part using data from a sample. The method of ordinary least squares (OLS) is the most common approach in this respect. Roughly speaking, it estimates the coefficients α_1 and α_2 so as to minimize the stochastic part. We thereby obtain a function that is as close as possible to the observed data points. Figure 5.1 provides an illustration. The aim is to reduce the differences between the observed outcomes (y) and the responses predicted by the linear approximation of the data (the deterministic part). In Fig. 5.1a, variable x and variable y are linearly associated. The correlation coefficient could serve as a measure of this association. It will not, however, provide an estimate of the equation of the line around which the observations are clustered. In theory, many models could be estimated (see Fig. 5.1b). The OLS method offers a solution by minimizing the distances between the regression line and the data points, as shown in Fig. 5.1c.

Formally, assume that we have gathered information about n units $i = 1 \ldots n$ as well as the value they take with respect to x and y. The optimization problem consists in searching for α_1 and α_2 so as to minimize the sum of squared errors:

$$\min_{\{\alpha_1, \alpha_2\}} \sum_{i=1}^{n} (\epsilon_i)^2$$

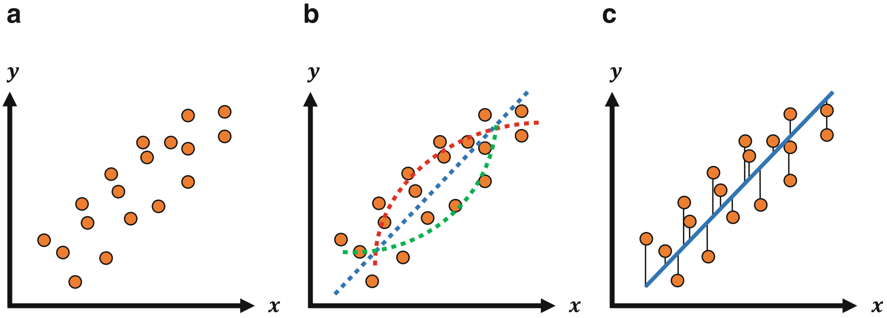


Fig. 5.1 The OLS method. (**a**) Linear association between x and y, (**b**) Estimation of the relationship, (**c**) The OLS method

By using squared residuals, we focus on "positive distances" and avoid positive and negative residuals canceling each other out. Once estimated, the sample counterpart of the model (the sample regression function) is written as:

$$\widehat{y}_i = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_{1i} \quad \text{(sample)}$$

which is the equation of the regression line. The sample regression function can also be expressed as:

$$y_i = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_{1i} + \widehat{\epsilon}_i \quad \text{(sample)}$$

where $\widehat{\epsilon}_i = y_i - \widehat{y}_i$ denotes the difference between the observed value y_i and the value $\widehat{y}_i$ predicted by the model.

As can be noticed, the notations differ depending on whether we are dealing with the sample function or the population function. Table 5.1 illustrates those differences. We put a "hat" over the parameters to indicate that they correspond to a sample estimator of the population parameter. By convention, the terms $\widehat{\epsilon}_i$ $(i = 1\ldots n)$ are termed residuals and, by construction, their sum $\sum_{i=1}^{n} \widehat{\epsilon}_i$ and their mean are equal to zero. They represent the vertical distances between each data point and the corresponding point on the regression line. The term fitted value is used to denote each point $(\widehat{y}_i)$ on that line.

Figure 5.2 provides an illustration. For each value x_i, the regression line yields the fitted value $\widehat{y}_i$ that is associated. By construction, this line passes through the center of gravity of the data $(\bar{x}, \bar{y})$. Coefficient $\widehat{\alpha}_1$ represents the constant (or intercept) while $\widehat{\alpha}_2$ represents the slope. In economic terms, $\widehat{\alpha}_1$ yields the amount variable y would reach on average if x was equal to zero. The slope $\widehat{\alpha}_2$ corresponds to the additional change in y we would observe on average if x was increasing by one unit. We indicate "on average" because the relationship between y and x is inexact as not all the data points lie on the regression line, as shown in Fig. 5.2. For all i, the distance between the line $(\widehat{y}_i)$ and the observed value (y_i) is the residual $\widehat{\epsilon}_i$. The lower are the residuals, the better the model fits the data.

Table 5.1 Econometrics vocabulary

Population regression function		Sample regression function	
y_i	Observed value of the dependent variable	$\widehat{y}_i$	Fitted value, predicted value, estimator of y_i
x_i	Independent variable, regressor	x_i	Independent variable, regressor
α_1	Parameter, regression coefficient	$\widehat{\alpha}_1$	Estimated coefficient, estimator of α_1
α_2	Parameter, regression coefficient	$\widehat{\alpha}_2$	Estimated coefficient, estimator of α_2
ϵ_i	Error term, errors, disturbances	$\widehat{\epsilon}_i$	Residual term, residuals

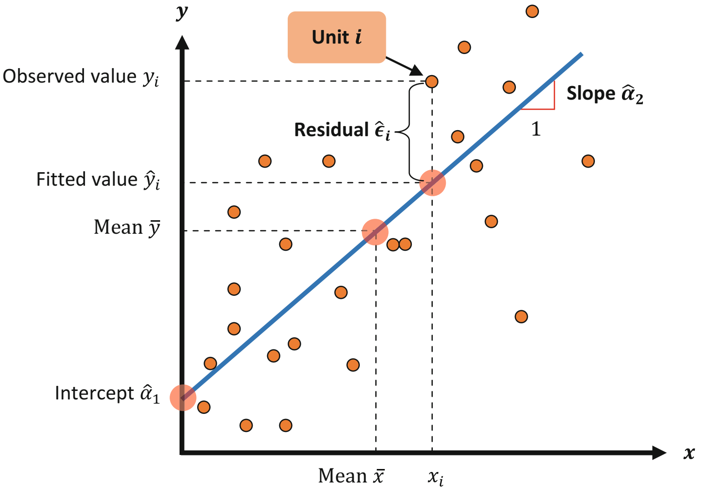

y
Unit i
Observed value y_i
Slope $\widehat{\alpha}_2$
Residual $\widehat{\epsilon}_i$
1
Fitted value $\widehat{y}_i$
Mean $\bar{y}$
Intercept $\widehat{\alpha}_1$
x
Mean $\bar{x}$
x_i

Fig. 5.2 Fitted and observed values

The variability of the dependent variable y can be defined as the sum of two components, the explained sum of squares (ESS) and the residual sum of squares (RSS), which relate to the deterministic part and stochastic part of the model, respectively:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{TSS}}=\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{\text{ESS}}+\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{RSS}}$$

The term TSS denotes the total sum of squares and is a proxy for the total variance in the data. It tells us how much variation there is in the dependent variable. The explained sum of squares measures the amount of variation in the dependent variable that is explained by the model. The residual sum of squares on the other hand measures the variation in the dependent variable that is not explained by the model. The latter can also be written as $\text{RSS} = \sum_{i=1}^{n} (\widehat{\epsilon}_i)^2$. The lower is RSS or the higher is ESS, the better is the explanatory power of the model.

Using the previous equation, it is possible to describe how well the model fits the observations (the goodness of fit) via the coefficient of determination, denoted R^2. It is computed as the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

It yields the proportion of variation, or percentage of the total variation in the dependent variable, that is explained by the regression model. It lies between 0 and 100%. This coefficient can be readily interpreted only for regressions with intercept. It is possible to get a negative value when a constant is not included in the regression equation. Last, when there is a single independent variable, R^2 is also the square of the sample correlation coefficient relating the dependent and the independent variable:

$$R^2 = \left(r_{x,y}\right)^2 \text{ (simple linear regression)}$$

In that case, R^2 measures the strength of the linear association between the variables.

To illustrate the methodology, assume that we have data about 60 (fictitious) districts that have been conferred responsibility for several welfare programs, providing assistance to single mothers and children, disabled, elderly, and the unemployed. Table 5.2 provides the dataset. We would like to highlight the determinants of per capita social expenditures, denoted *Social_Exp* hereafter. Among the set of independent variables, we have *Income* (the mean taxable income); *Unemprate* (the unemployment rate); *Shareof*60 (the share of people aged 60 and over); *Population* (number of inhabitants), *Density* (population density per km2). Variables *N_family*, *N_disabled*, *N_elder* and *N_benefits* represent the number of families, disabled, elder and unemployed who receive social assistance, respectively. Figure 5.3 performs a simple linear regression in R-CRAN. A similar methodology can be employed in Excel: one simply needs to click anywhere in a scatter plot. Then, on the Layout tab in the Analysis group, one clicks Trendline, then chooses "Linear" to calculate the least squares fit for a line.

Figure 5.3 starts with the *read.table* command which is used to upload the database in R-CRAN using the path *C://mydataOLS.csv*, which denotes the location of the file. The file format is .csv, with ";" as a separator. This format can be easily created with Excel. For computational convenience, the database is renamed *D*. Command *head* returns the first part of the data. This command is used to check that the data have been correctly uploaded. The first analysis consists in assessing graphically whether there is an association between social spending per head and the rate of unemployment. The command *plot* is used to this purpose and

Table 5.2 Data for example 1

Name	Social_ Exp	Income	Unemprate	Shareof60	Population	Density	N_families	N_disabled	N_elders	N_benefits
Allegan	184	6145	0.09	0.3	172,580	28	255	847	2205	1720
Almedia	200	7260	0.14	0.21	127,6453	128	2667	2882	7007	25,302
Anangu	174	6463	0.18	0.22	884,890	145	2016	2578	4879	29,536
Asbury	141	6813	0.1	0.19	857,295	127	1617	2948	6054	8617
Balnarring	168	6591	0.11	0.28	232,252	34	404	928	1645	2112
Bartolo	165	7052	0.12	0.22	551,498	90	1359	1780	1786	8253
Blackdale	171	7204	0.11	0.18	1,085,203	146	1522	3555	6158	13,090
Bluewall	169	6674	0.08	0.23	250,917	50	486	793	1419	1934
Bluford	195	6928	0.11	0.26	325,077	35	890	695	2728	4136
Bridgemere	161	7200	0.11	0.25	314,616	50	524	956	1835	3894
Brightwick	195	6681	0.12	0.22	731,657	153	1105	3448	5206	10,211
Brocton	148	5951	0.09	0.25	208,285	42	171	699	1992	1591
Bulgandry	134	6893	0.14	0.19	1,124,006	165	1616	3288	5366	18,194
Burlington	147	7564	0.1	0.2	616,372	91	1150	1782	2122	7548
Campo	195	6275	0.1	0.3	159,657	31	204	548	1526	2083
Coldwood	182	6254	0.13	0.27	304,968	57	425	1052	2078	5677
Crystalash	202	5862	0.06	0.27	73,507	14	44	272	1025	599
Dorwall	148	6296	0.12	0.2	729,551	102	1493	1872	3788	9021
Eastsage	193	6233	0.11	0.24	481,349	81	806	1921	5717	4778
Fayville	143	7338	0.1	0.18	565,935	69	1240	1867	2788	8468
Fogview	158	6266	0.1	0.23	196,378	32	447	951	1364	2267
Glassmallow	180	6263	0.07	0.22	283,893	55	667	1358	2221	1798
Golconda	157	6595	0.11	0.19	715,062	136	1672	1922	3084	12,328
Goldbay	187	6147	0.1	0.22	192,797	31	523	589	862	2655

Goldcrest	144	6648	0.12	0.24	639,772	94	1037	2147	6539	7100
Gunyarra	221	6690	0.12	0.28	226,776	33	614	1254	2651	3602
Hedgeview	226	6196	0.16	0.17	2,554,942	445	9117	6666	11,607	62,219
Hempstead	150	7222	0.12	0.16	765,111	131	1671	2143	2839	9435
Holtsville	215	6245	0.1	0.24	292,912	48	941	1092	1999	4083
Iceport	183	5699	0.16	0.19	1,440,495	216	4308	4649	8364	32,302
Janlea	153	6990	0.1	0.22	602,853	76	859	2053	4518	8460
Kakoma	179	7070	0.11	0.25	594,245	78	720	1922	5817	10,232
Lakesite	227	6580	0.12	0.27	222,764	50	529	992	2923	3366
Lallat	193	6204	0.17	0.28	389,051	95	1127	1934	2360	6617
Leavittsburg	144	8030	0.11	0.18	1,566,215	482	1710	2147	6539	7100
Maeystown	154	6104	0.1	0.22	229,150	43	404	603	1294	2443
Makowata	187	6600	0.11	0.25	547,472	64	796	2552	2957	6006
Malchi	189	6685	0.12	0.22	527,687	85	965	1725	1827	6974
Mallowford	144	7248	0.1	0.2	369,897	61	645	939	1620	3779
Marblehaven	112	6957	0.08	0.17	625,854	143	719	1286	2330	5800
Matlacha	221	6760	0.15	0.19	1,239,784	197	2955	4616	8559	25,457
Mentone	166	7971	0.09	0.14	1,182,543	200	2245	3834	4910	10,397
Montgomery	141	10,313	0.08	0.15	1,353,723	593	2003	3191	3031	10,976
Murilla	136	6396	0.11	0.24	344,574	57	619	1364	3420	3718
Nippering	161	6340	0.14	0.2	554,595	90	1376	2002	4852	10,179
Parkersburg	195	6364	0.12	0.27	342,629	60	599	1351	2829	5315
Pepeekeo	196	5951	0.12	0.26	205,519	55	207	822	1760	3818
Pepin	158	7228	0.16	0.26	891,598	149	836	2417	5181	21,929
Pomaria	171	6406	0.13	0.21	496,948	139	806	1145	3405	12,177
Pottersville	142	6445	0.1	0.24	535,322	80	745	2332	3484	4315
Rockhollow	132	6591	0.12	0.23	397,941	57	757	1191	2178	5997

(continued)

Table 5.2 (continued)

Name	Social_ Exp	Income	Unemprate	Shareof60	Population	Density	N_families	N_disabled	N_elders	N_benefits
Rosesea	196	7085	0.09	0.26	353,941	64	564	1438	2928	5287
Saugus	129	6425	0.11	0.22	381,791	65	506	1167	2078	4272
Sedley	212	7006	0.11	0.24	332,773	45	906	1601	2461	4056
Stubbo	154	6824	0.11	0.19	137,234	225	217	253	536	2020
Waggoner	166	9012	0.08	0.14	1,133,653	628	2261	2570	3119	13,270
Warburto	204	11,788	0.09	0.18	1,419,110	8063	2463	5207	6604	21,113
Watonga	204	6802	0.15	0.14	1,386,023	5873	3373	3971	5491	39,292
Westerwood	178	9106	0.11	0.17	1,225,473	5002	1877	4869	4905	21,673
Wildefalcon	147	8040	0.1	0.13	1,100,782	883	1547	2577	2972	15,356

```
> D=read.table("C://mydataOLS.csv",head=TRUE,sep=";")
> head(D)
        Name Social_Exp Income Unemprate Shareof60 Population
1    Allegan        184   6145      0.09      0.30     172580
2    Almedia        200   7260      0.14      0.21    1276453
3     Anangu        174   6463      0.18      0.22     884890
4     Asbury        141   6813      0.10      0.19     857295
5 Balnarring        168   6591      0.11      0.28     232252
6    Bartolo        165   7052      0.12      0.22     551498
  Density N_families N_disabled N_elders N_benefits
1      28        255        847     2205       1720
2     128       2667       2882     7007      25302
3     145       2016       2578     4879      29536
4     127       1617       2948     6054       8617
5      34        404        928     1645       2112
6      90       1359       1780     1786       8253
> plot(D$Social_Exp~D$Unemprate)
> lm(D$Social_Exp~D$Unemprate)

Call:
lm(formula = D$Social_Exp ~ D$Unemprate)

Coefficients:
(Intercept)  D$Unemprate
      141.7        269.5
> abline(lm(D$Social_Exp~D$Unemprate))
> mean(D$Social_Exp)
[1] 172.1167
> mean(D$Unemprate)
[1] 0.1128333
> points(mean(D$Social_Exp)~mean(D$Unemprate),col="red",pch=19)
> cor(D$Social_Exp,D$Unemprate)
[1] 0.2375224
> cor(D$Social_Exp,D$Unemprate)^2
[1] 0.05641687
```

Fig. 5.3 Simple linear regression with R-CRAN: example 1

offers the scatter plot of Fig. 5.4. The command *lm* fits a linear model on those observations. We estimate the following relationship:

$$Social\_Exp = 141.7 + 269.5 \times Unemprate$$

In most cases, the intercept (141.7) is not worth interpreting because it illustrates an extreme situation (here a zero unemployment rate) that is unlikely to occur and for which no data point is actually available for estimation purpose. The slope on the other hand is worth the examination. In the present case, it means that an increase in the unemployment rate by one percentage point (0.01) yields on average an increase by $2.7 in per capita social expenditures.

Using the *abline* function, the analysis goes further by plotting the regression line in the (x, y) system (see Figs. 5.3 and 5.4). The mean vector (*mean* (*D*$*Social_Exp*) ~ *mean*(*D*$*Unemprate*)) is also drawn on the graph. As can be seen, the regression line does pass through the center of gravity of the scatter

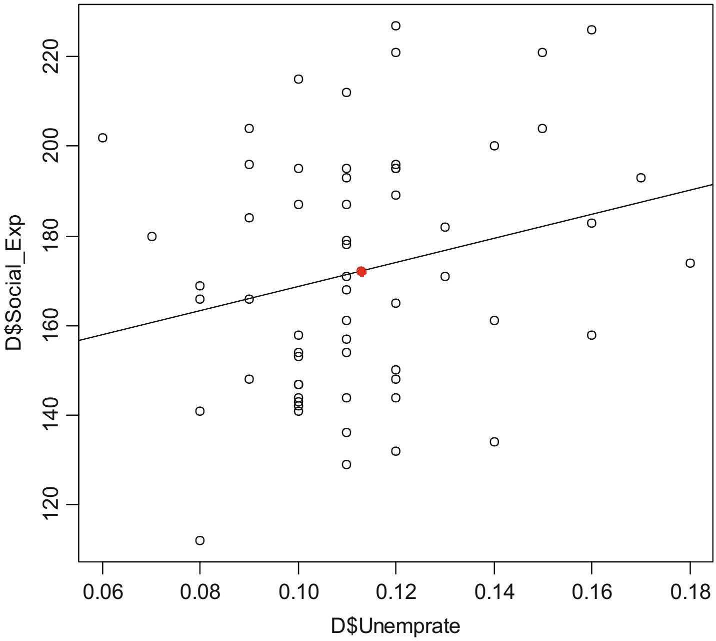


Fig. 5.4 Unemployment rate and social expenditures: example 1

plot. Last, the correlation coefficient, as well as the squared correlation coefficient, are computed. It can be observed that there is a weak correlation between the two variables of interest ($\rho = 0.23$). The coefficient of determination amounts to $R^2 = 0.23^2 = 0.056$. The model is thus found to explain 5% only of the variation in per capita social expenditures. This result is not surprising. In practice, it is rare that a single variable model explains most of the variation in the dependent variable.

As we will see later, it is possible to extend the previous simple regression model to a multiple regression model by including additional explanatory factors in the equation. In that context, existing theories and common sense are used to motivate the choice of the variables. The expected sign of the model's parameters should be defined *ex ante* using those theoretical arguments. One then needs to decide the data sample upon which the model will be estimated. Various types of data can be used in this respect. Time series data gives information about one single unit i over several periods of time. In that case, the subscript t is used as a replacement of i to index the observations. Conversely, cross-sectional data focuses on one single period of time and provide information about a set of individuals that are indexed by i. Panel data offers a combination of those two dimensions by considering a set of individuals i over several time periods t.

One should distinguish the econometric approach, which is based on observational data, from a randomized controlled experiment, which attempts to isolate the effects of a treatment through randomization. Understanding those differences is important, especially when the research study is supposed to guide the formulation of public policies. An econometric analysis does not attempt to manipulate or influence the environment. Data are only observed, collected and interpreted. In an experiment on the other hand, we compare a treatment group with a control group that has similar characteristics on average except for the fact of receiving the treatment (a comparison of the mean outcome in each group allows the average treatment effect to be quantified). The problem with observational data is that we do not control for all those characteristics that may also influence the outcome of interest. The analysis becomes multidimensional. The use of econometrics is an attempt to control for all those extra factors. By estimating a coefficient for each variable, each effect is supposed to be isolated. In other words, the econometric approach is based on a *ceteris paribus* reasoning (all else being equal). Each estimated coefficient represents the impact of a variable, holding constant the effects of the other independent variables.

While appealing, the econometric approach must be implemented with care. Many biases may result from model misspecifications. For instance, the true functional form of the relationships under examination is usually unknown. If we use the wrong form, we could reach misleading conclusions about the effects of the independent variables. Moreover, it is possible that an omitted variable has a link with both the dependent variable and one or more of the independent variables (spurious relationship). In such circumstances, we could erroneously conclude that a variable influences another while it actually does not. The dependent variable may also be part of a system of simultaneous equations, which may result in biases if the model does not account for that simultaneity. Last, a regression analysis, even perfectly implemented, cannot prove causality unless more sophisticated techniques are employed (e.g., Granger causality test or quasi-experimental techniques).

The outline of the chapter is as follows. Section 5.2 generalizes the simple linear model to a multiple setting which can be used to predict the unknown value of a dependent variable from the known value of two or more variables. Section 5.3 describes the assumptions underlying the method of ordinary least squares. Section 5.4 is about the choice of variables and Sect. 5.5 is concerned with the form those variables take in the regression equation. Section 5.6 explains how to deal with biases. Section 5.7 is about model selection and how to interpret regression analysis results. Last, Sect. 5.8 extends the approach to the case where the dependent variable is binary, i.e. takes on values 0 or 1.

5.2 Multiple Regression Analysis

Multiple regression analysis is an extension of the simple linear regression model. The approach is used when one wants to regress the dependent variable on two or more independent variables. The aim is not to estimate the equation of a line but, instead, to estimate a multidimensional linear equation. Formally, the multiple regression equation of y on $x_2, \ldots, x_K$ is given by:

$$y_i = \alpha_1 + \alpha_2 \mathrm{x}_{2\mathrm{i}} + \alpha_3 x_{3\mathrm{i}} + \ldots + \alpha_K x_{Ki} + \epsilon_i \text{ (population)}$$

The terms are analogous to those of simple linear regression: ϵ_i stands for the error term, α_1 is the intercept, and each coefficient α_k, $k > 1$, represents the impact on average of a one unit increase in x_k on the dependent variable y, holding constant the other independent variables. Under this setting, the sample regression function is written as:

$$\widehat{y}_i = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_{2\mathrm{i}} + \widehat{\alpha}_3 x_{3\mathrm{i}} + \ldots + \widehat{\alpha}_K x_{Ki} \text{ (sample)}$$

or

$$y_i = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_{2\mathrm{i}} + \widehat{\alpha}_3 x_{3\mathrm{i}} + \ldots + \widehat{\alpha}_K x_{Ki} + \widehat{\epsilon}_i \text{ (sample)}$$

where $\widehat{y}_i$ stands for the fitted values, $\widehat{\epsilon}_i$ is the residual term and the $\widehat{\alpha}$'s are the estimated coefficients.

Equivalently, the model can be written in matrix form:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_i \\ \vdots \\ \widehat{y}_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{21} & \ldots & x_{K1} \\ 1 & x_{22} & \ldots & x_{K2} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{2i} & \ldots & x_{Ki} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{2n} & \ldots & x_{Kn} \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_K \end{bmatrix}, \widehat{\boldsymbol{\alpha}} = \begin{bmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \\ \widehat{\alpha}_3 \\ \vdots \\ \widehat{\alpha}_K \end{bmatrix},$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{bmatrix}, \widehat{\boldsymbol{\epsilon}} = \begin{bmatrix} \widehat{\epsilon}_1 \\ \widehat{\epsilon}_2 \\ \vdots \\ \widehat{\epsilon}_i \\ \vdots \\ \widehat{\epsilon}_n \end{bmatrix}$$

Note that the unit vector $x_1 = 1, \ldots, 1$ is included among the independent variables so that the weighting coefficient α_1 on that term represents the regression constant. Under this setting, the previous population regression function becomes:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \text{ (population)}$$

The sample counterpart of the equation is specified as:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\alpha}} \text{ or } \mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\epsilon}} \text{ (sample)}$$

The OLS estimator is given by $\widehat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. For simplicity of exposition, we will not detail how to solve the optimization problem. As already stressed in Sect. 5.1, the OLS procedure consists in finding the coefficients that minimize the stochastic part of the model, i.e. the squared residuals.

Once the regression equation has been estimated, it is crucial to assess the goodness of fit of the model. One issue is that the R^2 automatically increases when extra explanatory variables are added to the model (even if they are not relevant in explaining the phenomenon in question). Adding new variables causes the stochastic part as measured by RSS to become smaller. To overcome this issue, one usually prefers to examine what is termed the adjusted R^2:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-K}\right)\left(1 - R^2\right) = R^2 - \left(1 - R^2\right)\left(\frac{K-1}{n-K}\right)$$

The closer the coefficient is to 100%, the better the model fits the data. The key difference between R^2 and $\bar{R}^2$ is that the adjusted R^2 is always lower than its unadjusted version. It does not automatically increases when new independent variables are included. For this reason, it offers a useful criterion for measuring the explanatory power of a model.

The adjusted R^2 tells you how well the model predicts the dependent variable. It is a measure of the strength of the model. To go further, it is possible to use a F-test of overall significance with null and alternative hypotheses as follows:

$$H_0 : \alpha_2 = \ldots = \alpha_K = 0 \text{ (all parameters taken jointly are not significant)}$$

$$H_1 : H_0 \text{ is not true (not all parameters are simultaneously zero)}$$

The test statistic is based on the (unadjusted) R^2:

$$F^* = \frac{R^2}{1 - R^2} \times \frac{n-K}{K-1}$$

The F-test is used to assess whether the strength of the model is due to some non-random cause. Under the null hypothesis, statistic F^* follows a Fisher distribution with $K-1$ and $n-K$ degrees of freedoms. The critical value is denoted $F_\alpha(K-1, n-K)$, where α is the significance level, usually 5%. If $F^* > F_{5\%}(K-1, n-K)$ then the null hypothesis is rejected: the independent variables are jointly significant. Note that a larger R^2 leads to a higher value of F^*, which means that when R^2 increases notably, everything else being equal, then there should be stronger evidence that at least some of the coefficients are non-zero.

A related question is whether the independent variables significantly influence the dependent variable. Statistically, this is equivalent to testing the null hypothesis that the estimated coefficients are zero:

$$H_0 : \alpha_k = 0 \text{ (the effect of } x_k \text{ is not statistically significant)}$$

$$H_1 : \alpha_k \neq 0 \text{ (} x_k \text{ has a statistically significant impact on } y\text{)}$$

The test statistic is defined as:

$$t_k^* = \frac{\widehat{\alpha}_k}{se(\widehat{\alpha}_k)}$$

where $se(\widehat{\alpha}_k)$ denotes the standard error of the estimated coefficient. Under the null hypothesis, the test statistic follows the Student distribution with $n - K$ degrees of freedom. The critical value is denoted $t_{\alpha/2}(n - K)$ where α is the significance level, often 5%. The testing strategy is to reject the null hypothesis if the test statistic in absolute value is higher than the critical value. If this is to be the case, one usually concludes that "variable $x_k$significantly influences the dependent variable while controlling for other independent explanatory variables". For $n - K > 100$ and a 5% significance level, the critical value converges to 1.96. In that case, the strategy is to reject the null hypothesis if:

$$|t_k^*| > 1.96$$

Note that the test is two-tailed because the alternative hypothesis allows for both negative and positive values of $\widehat{\alpha}_k$.

The standard error $se(\widehat{\alpha}_k)$ of the estimated coefficients is provided in the regression output of most statistical packages. Mathematically, one defines the sample variance of the OLS estimator via the following covariance matrix:

$$\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} se(\widehat{\alpha}_1)^2 & \sigma_{\widehat{\alpha_1},\widehat{\alpha_2}} & \dots & \sigma_{\widehat{\alpha_1},\widehat{\alpha_K}} \\ \sigma_{\widehat{\alpha_1},\widehat{\alpha_2}} & se(\widehat{\alpha}_2)^2 & \dots & \sigma_{\widehat{\alpha_2},\widehat{\alpha_K}} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{\widehat{\alpha_1},\widehat{\alpha_K}} & \sigma_{\widehat{\alpha_2},\widehat{\alpha_K}} & \dots & se(\widehat{\alpha}_K)^2 \end{bmatrix}$$

This matrix holds the squared standard errors in the diagonal elements and the covariances $\sigma_{\widehat{\alpha_k},\widehat{\alpha_l}}$ $(k \neq l)$ in the off-diagonal elements. What matters for statistical testing is the diagonal of the matrix. The matrix is obtained from the product of $\widehat{\sigma}^2$ with $(\mathbf{X}^{'}\mathbf{X})^{-1}$. The scalar $\widehat{\sigma}^2$ stands for the sample variance of the residuals. It is an estimate of the unobserved variance of the error term. We have:

$$\widehat{\sigma}^2 = \frac{1}{n - K}\sum_{i=1}^{n}\widehat{\epsilon}_i^2$$

where $n - K$ stands for the number of degrees of freedom in the model. We can see that the elements of the covariance matrix decrease with the variance of the error term, which means that t_k^* increases as $\widehat{\sigma}^2$ decreases, $k = 1 \dots K$. Intuitively, the lower is the stochastic part, the higher the chances that the estimator $\widehat{\boldsymbol{\alpha}}$ reflects the true value of $\boldsymbol{\alpha}$.

Let us now illustrate the approach using the dataset from example 1. Figure 5.5 performs a multiple regression analysis using *Social_Exp* as a dependent variable and *Unemprate*, *Income* and *Shareof*60 as dependent variables. The choice of those variables is purely illustrative. The command *lm* fits a linear model on those

```
> D=read.table("C://mydataOLS.csv",head=TRUE,sep=";")
> myreg=lm(Social_Exp~Unemprate+Income+Shareof60,D)
> summary(myreg)

Call:
lm(formula = Social_Exp ~ Unemprate + Income + Shareof60, data = D)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.888 -18.333  -2.069  17.757  55.717 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.495e+01  4.848e+01   0.927  0.35781   
Unemprate   3.126e+02  1.434e+02   2.180  0.03349 * 
Income      4.727e-03  3.918e-03   1.207  0.23267   
Shareof60   2.707e+02  9.303e+01   2.910  0.00517 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.36 on 56 degrees of freedom
Multiple R-squared: 0.1823,	Adjusted R-squared: 0.1385 
F-statistic: 4.161 on 3 and 56 DF,  p-value: 0.009892

> qt(0.975,56)
[1] 2.003241
> qf(0.975,3,56)
[1] 3.35945
> confint(myreg,"Shareof60")
             2.5 %   97.5 %
Shareof60 84.36819 457.0714
> confint(myreg,"Income")
              2.5 %     97.5 %
Income -0.003121093 0.01257488
```

Fig. 5.5 Multiple linear regression with R-CRAN: example 1

observations. A name is given to that regression (*myreg*). The *summary* function then provides a detailed presentation of this *lm* object. By order of appearance, the software provides summary statistics about the residuals (the min, max, and the quartiles), the estimated coefficients (their value, their standard error, their t-statistic, and their p-value), the sample error of the residuals, the number of degrees of freedom, the R^2 and $\bar{R}^2$ coefficients, and the result of the F-test of overall significance. Using the *qt* and *qf* functions, the program offers the critical values $t_{\alpha/2}(n-K)=t_{2.5\%}(56)$ and $F_{\alpha}(\mathrm{K}-1,n-K)=F_{5\%}(3,56)$ that will be used for hypothesis testing. Last, the program ends with the command *confint* which estimates a confidence interval for the coefficient of *Shareof*60 and *Income*. Each of these outputs is further detailed below.

First, it is essential to assess the appropriateness of the model by analyzing the residuals. Summary statistics can be very informative in this respect. For instance, the minimum and maximum values can be used to detect a sample peculiarity. Any observation with a large residual, i.e. whose actual value (y_i) is unusually far away from its fitted value ($\widehat{y}_i$), is suspicious. It may indicate an outlier in y, a measurement error or any other problem. The first and third quartiles (1Q and 3Q) can serve as a proxy for the overall quality of the model. As a rough indication, any value that is close to or higher than the mean value of the dependent variable indicates high residuals. For instance, in Fig. 5.5, the minimum and the maximum amount to –43.888 and +55.717, respectively. The first and third quartiles on the other hand are –18.333 and +17.757. Given the mean of the dependent variable ($172.1167 from Fig. 5.3) those values are coherent. Last, the median serves as an indication of the asymmetry of the distribution. A large median in absolute value suggests that the residuals are not distributed equally around their mean, which by construction amounts to zero.

The p-values ($Pr(>|t|)$) of Fig. 5.5 help determining the significance of the coefficients (i.e. whether they are or not significantly different from zero). Those values stand for the significance level $\alpha \in [0, 1]$ at which we are indifferent between rejecting or accepting the null hypothesis given the sample data under examination. A small p-value (typically less than 5%) indicates strong evidence against the null hypothesis. A large p-value (greater than 5%) indicates that one fails to reject the null hypothesis. As a rough and quick indication, one can also look at the asterisks "*", which point out the level of significance for each coefficient. In Fig. 5.5, at a 5% significance level, *Unemprate* and *Shareof*60 both yield a significant impact on the dependent variables (their p-value is lower than 5%), while *Income* does not yield a significant impact. Likewise, we can look at the t-values and compare them to the relevant critical value, here $t_{2.5\%}(56) = 2.003$. As can be seen, those t-values are higher than the critical value for *Unemprate* and *Shareof*60, but lower for *Income*.

The t-values from Fig. 5.5 are obtained from the values of the estimates and their standard errors. For instance, for *Unemprate*, we have:

$$t^*_{Unemprate} = \frac{\widehat{\alpha}_{Unemprate}}{se\left(\widehat{\alpha}_{Unemprate}\right)} = \frac{312.6}{143.4} \approx 2.18$$

Confidence intervals offer an alternative method for assessing the significance of a given parameter. A margin of error is computed as follows:

$$e(\widehat{\alpha}_k) = t_{\alpha/2}(n - K) \times se(\widehat{\alpha}_k)$$

The confidence interval is then given by $\widehat{\alpha}_k \pm e$. If zero belongs to that interval, then we reject H_0 at the significance level α. To illustrate, let us consider the impact of *Shareof*60. The standard error amounts to 93.03. At a 5% significance level, the margin of error is computed as 2.003 × 93.03 ≈ 186.3. The confidence interval of the coefficient is thus 270.7 ± 186.3 = [84.4,457]. In Fig. 5.5, one reaches the same conclusion using the command *confint*(*myreg*, "*Shareof*60"). As already stressed, the coefficient is significantly different from zero. Conversely, the same calculation for *Income* brings a confidence interval that includes 0, confirming the non-significant impact of that variable on social expenditures.

The sign of the estimated coefficient indicates the direction of the relationship. Once significance has been assessed, it is important that the sign of the estimates are consistent with what is expected. In practice, only the coefficients significantly different from 0 are examined (here, *Unemprate* and *Shareof*60). For instance in Fig. 5.5, we can see that *Unemprate* and *Shareof*60 yield a positive impact on per capita social expenditures. A one percentage point increase in the unemployment rate and in the share of people aged over 60 yields on average an increase in per capita social spending of \$3.1 and \$2.7, respectively. Whether those signs are in accordance with the existing literature would require further analysis as it depends on the context (the sample data at hand) and the motivations behind the study (what is the theoretical background?).

The last step when performing a regression analysis consists in checking the overall strength of the regression model. In Fig. 5.5, the adjusted R^2 amounts to 13.85%, which means that the model explains 13.85% of the variation in the dependent variable. This value can be obtained from the unadjusted R^2:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-K}\right)\left(1 - R^2\right) = 1 - \left(\frac{60-1}{60-4}\right)(1 - 18.23\%) \approx 13.85\%$$

The F-statistic is derived as follows:

$$F^* = \frac{R^2}{1-R^2} \times \frac{n-K}{K-1} = \frac{18.23\%}{1 - 18.23\%} \times \frac{60-4}{4-1} \approx 4.16$$

This statistic is larger than the critical value $F_{\alpha/2}(3, 56) = 3.359$ which indicates overall significance of the model: not all coefficients are zero. Equivalently, the p-value of the F-test (0.009892) is lower than 5%.

Note that there is no specific rule as regards the minimum value the adjusted R^2 is supposed to reach. Depending on the field of analysis, the coefficient can reach very different values that can be lower than 30% in some cases and larger than 60% in others. Moreover, as we shall see later, a very large R^2 (higher than 90%) can be suspicious as well. It may be a sign of a spurious relationship between the variables (which appear to be significant because each of them is related to a third one). On the other hand, a low $\bar{R}^2$ is problematic when the purpose of the model is to ground predictions about the determinants of the dependent variable.

5.3 Assumptions Underlying the Method of OLS

The method of ordinary least squares is based on a set of assumptions which are required not only to estimate accurate coefficients, but also for statistical inference, i.e. for testing whether the model parameters are non-zero. The main assumptions are described below:

Linearity As already stated, the relationship between the dependent variable and the independent variables is linear in the parameters. This assumption, however, is not as restrictive as it looks. It says linear in the parameters, not the variables. The variables can be modified at convenience using logarithmic ($\ln x_{ik}$) or polynomial forms (x_{ik}^2, x_{ik}^3, etc.). Moreover, the same independent variable can be included several times in the same regression model using different transformations. This introduces a lot of flexibility in the estimation procedure. As long as the equation remains in an additive form, the model can be estimated by OLS.

Independent Variables Are Exogenous All the independent variables are uncorrelated with the error term:

$$\mathrm{Cov}(x_{ki}, \epsilon_i) = 0 \; \forall k$$

When a dependent variable is correlated with the error term, it means that the variable is associated with factors that the model omits. In that case, the variable is said to be endogenous. The method of instrumental variables must be used to carry out inference. If the variables are not correlated with the error terms, then the variables are said to be exogenous. In that context, the method of least squares is valid.

No Collinearity Between the Independent Variables The independent variables must be linearly independent. They should not be correlated (collinearity) nor expressed as a linear combination of the other independent variables (multicollinearity). If one fails to meet this assumption, it becomes difficult or even impossible to disentangle the effects of the independent variables. A way to detect and thereby avoid this problem is to run a correlation matrix of all independent variables. Correlations among independent variables that are larger than, say 0.6, can be considered as problematic. Another situation where multicollinearity occurs is when two or more independent variables show a deterministic linear relationship (perfect multicollinearity), for instance when both the share of males and females are introduced in the right-hand side of the regression model. In that case, the share of males is perfectly correlated with the share of females and most statistical packages will either not run or drop one or more variables (e.g., by including NA's in the regression output). Whether it is perfect or imperfect, the usual solution to multicollinearity is to omit the offending variables.

Equal Variance of Errors When the error terms are distributed with constant variance, they are said to be homoscedastic:

$$\mathrm{Var}(\epsilon_i) = \sigma$$

In the simple linear regression model, this means that the variance of the observations along the line of best fit remains similar along the line. This assumption is labeled homoscedasticity. When the error terms are distributed with unequal variance, we use instead the term heteroscedasticity. In that case, it becomes difficult to assess the true standard error of the coefficients. Confidence intervals may be too wide or too narrow which could invalidate statistical tests of significance.

Independence of Errors The size of each error term is not influenced by the size of other error terms:

$$\mathrm{Cov}\left(\epsilon_i, \epsilon_j\right) = 0, \quad i \neq j$$

This assumption is very likely to be violated with time series data because each error term at time t is likely to be correlated with the error at time $t-1$. If this is to be the case, the error terms are said to be auto-correlated. Autocorrelation problems (aka serial correlation) generally occur when the model is misspecified and does not fully assess the evolution of the dependent variable through time. The coefficient estimators as well as the standard errors can be biased.

Normality of Errors This assumption is stated as:

$$\epsilon_i \sim N(0, \sigma^2)$$

where the symbol ~ means "distributed as", N denotes the normal distribution, 0 stands for the mean of residuals, and σ^2 is the (unobserved) variance of the error term. Simply put, it means that the errors terms are close to zero on average, but could possibly reach negative or positive values. This assumption is crucial. Non-normally distributed errors can distort the significance tests (t- and F-tests).

As we will see in Sects. 5.4–5.7, assessing whether the classical assumptions of the OLS method hold true is an important step of regression analysis. In particular, since residuals offer an estimate of the unobservable statistical errors, examining their distribution is essential. Any atypical pattern in the way they are distributed can be considered as a sign of misspecification. The graphs in Fig. 5.6 provide a schematization. First, Figure 5.6 plots the residuals against the fitted value. This type of analysis can be used to evaluate how well the model fits the data or meets the assumptions underlying the OLS method. For instance, in Fig. 5.6a, no particular pattern is observed. The residuals display a Gaussian (normal) white noise with mean 0 and constant variance. In Fig. 5.6b, the residuals exhibit a nonlinear pattern. The linear regression model is not adapted to the data. Figure 5.6c illustrates a situation where two points are far away from the majority of the data, meaning that those two observations are not well predicted by the model. The distribution of residuals around their mean (i.e. zero) is thus asymmetric, which indicates a violation of the normality assumption. Figure 5.6d, e illustrates a situation of heteroscedasticity. The residual plots indicate that the variance of the residuals is increasing or decreasing with the fitted values. Last, Fig. 5.6f plots the residuals at year t as a function of the residuals at year $t-1$, revealing a strong positive correlation in the residuals.

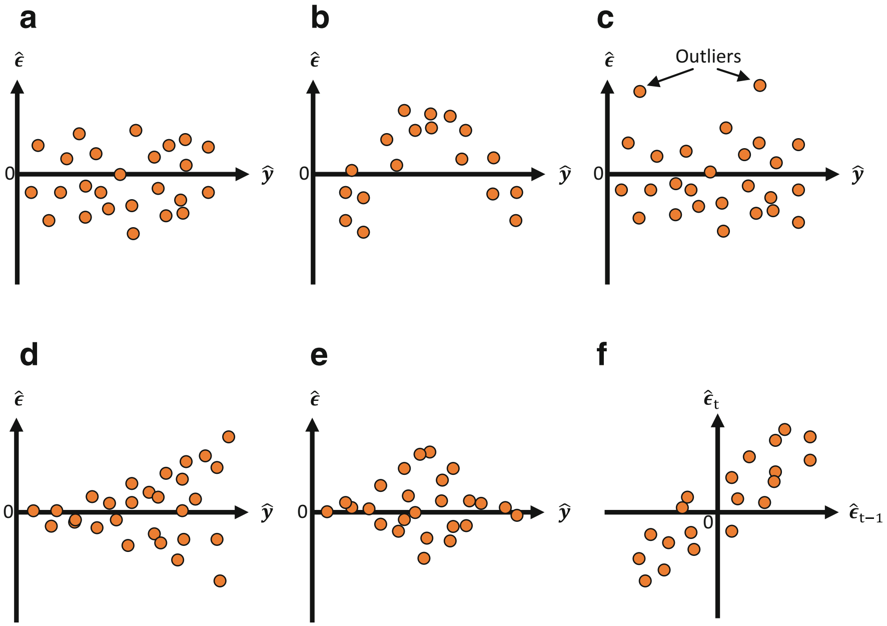


Fig. 5.6 Graphic representation of regression assumptions. (**a**) No pattern in residuals, (**b**) nonlinearity, (**c**) non-normality, (**d**) heteroscedasticity, (**e**) heteroscedasticity, (**f**) Autocorrelation

5.4 Choice of Relevant Variables

An important step in conducting an econometric analysis is to characterize the problem, specify the objectives of the study, and select the variables to be analyzed. This step, also termed specification, is essential as it will determine the validity of the regression analysis. A badly conceived specification may yield the wrong statistical inferences. The choice of a specification also impacts the way information is collected. Therefore, this step should not be taken lightly. In this respect, there are no comprehensive rules. The choice of variables mainly depends upon the understanding of the context. Yet, the following guidelines may help specify an accurate regression model and avoid estimation biases.

The main motivation for choosing one specification among others is theory. If theory says that a given variable should depend on some other variable, then this relationship should be examined in priority. For instance, it has been argued in economics that several factors such as price and income affect the demand for a good or service. For most goods (termed "normal goods"), there must be a positive relationship between a consumer's income and the demand for the good and there must be a negative relationship between the price of the product and the quantity consumers are willing to buy (the so-called Law of Demand). In that context, the usual econometric specification for modeling the demand for a good specifies the quantity demanded by the consumers as a function of their income and the price of the good itself.

The second motivation for selecting a set of variables is empirical evidence. Nowadays, the literature is vast and rich. Many empirical studies have been carried out in all fields. Provided that they come from a reliable information source (e.g., government agency, peer-reviewed journal), those studies should serve as a reference for any subsequent analysis. Those who have conducted those studies have been confronted with similar problems, e.g., sampling procedure, data collection and choice of variables. Any information about the ways they have addressed them may save a lot of time and effort. The empirical literature can also serve as a benchmark with respect to the expected sign and level of the estimated coefficients. Any impact that has not been supported by previous studies should always be considered with caution.

In many occasions, a reason for choosing one variable instead of another is data availability. While convenient at first, a specification based only the availability of data can be difficult to motivate. Make sure that this will not jeopardize the final quality of the research study. Note also that a distinction exists among the explanatory variables depending on whether they are considered as variables of interest or control variables:

$$y = f\underbrace{(\text{Variables of interest}, \text{Control variables})}_{\text{Independent variables}}$$

Variables of interest are concerned with the main relationships under examination, usually theory-based. For instance, in the context of public policies, it may be any variable over which policy-makers have direct control (e.g., whether the units have been selected for an intervention) or any variable which is the primary focus of the evaluator (e.g., the main determinants of the phenomenon under study). A control variable on the other hand is included in the model as it may improve the explanatory power of the model and, to some extent, reduce estimation biases.

Each estimated coefficient represents the impact of a variable, holding constant the effects of the other independent variables. If an important variable is omitted, then its impact is not kept constant for the estimation of the other coefficients, which may result in an omitted variable bias. Omitting a relevant variable (e.g., income or price in a demand equation) not only makes it impossible to estimate the impact of that variable, but can also create bias in the estimated coefficient of the other variables. Yet, this does not imply that the model should contain all available variables. While it does not cause bias (the coefficient still provide an accurate estimation of the population parameters), including an irrelevant variable that does not truly affect the dependent variable may distort the significance tests. Standard errors become larger and the hypothesis that the variables yield a significant impact is more likely to be rejected. This is why theory should always be used to motivate a particular specification in the first place.

The choice of a particular specification should also be based on common sense. Several ground rules can be offered in this respect. What follows provides several examples using simulated data and R-CRAN as a statistical environment (*lm* command). For simplicity of exposition, only the related codes are provided.

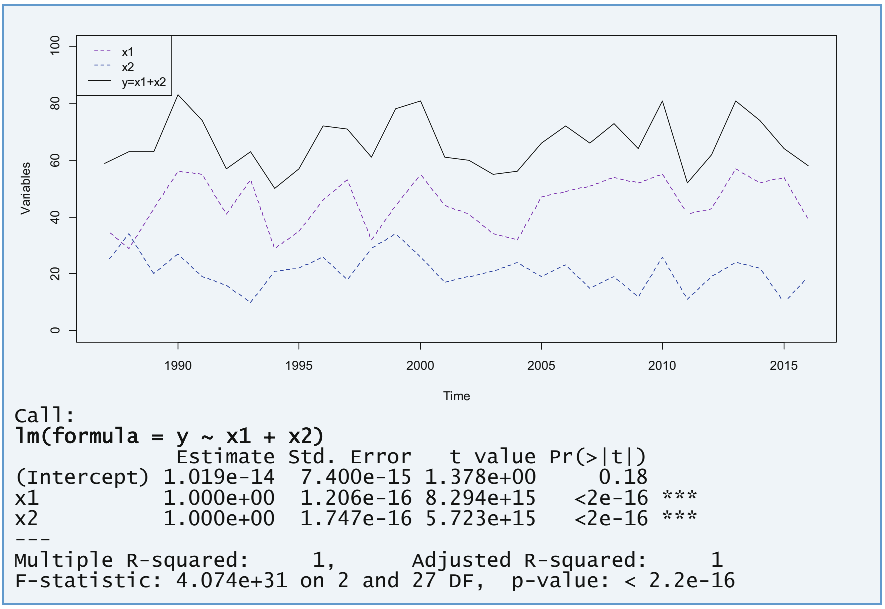
Perfect Fit Theoretical models are usually of two types: identity relationships and behavioral relationships. Identity relationships correspond to accounting equations which link perfectly the variables together. An example of such an identity is the GDP being equal to the sum of its component parts:

$$Y = C + I + G + X - M$$

where Y denotes gross domestic product, C is private consumption, I is private investment, G is government consumption, X is exports, and M is imports. Under this setting, econometrics is not required to estimate the equation because the relationship already is common knowledge. Behavioral relationships on the other hand are unknown *ex ante* and require theory and advanced statistical techniques to be fully assessed. The demand for a good offers an example. The econometric specification is based on a model of consumer behavior which attempts to identify the factors that influence the choices that are made by consumers.

While useless in most cases, estimating an identity relationship using OLS is still possible. In Fig. 5.7, the dependent variable $y = x_1 + x_2$ is the sum of two independent variables, x_1 and x_2. As can be seen, the regression output offers accurate results. The R^2 amounts to 100% and the independent variables yield a significant impact (the t-values are extremely high). Another situation where perfect fit ($R^2 = 1$) is attained is when there are as many parameters to be estimated as observations. In that case, the number of degrees of freedom is insufficient to provide an accurate result. The model is said to be saturated. For instance, if one estimates the equation of a line using two observations only, then the R^2 (or equivalently the correlation coefficient in that case) is equal to 1. Be advised that there is no convention with respect to sample size requirement for econometric use. As a rule of thumb, we may for instance rely on Green's rule which states that the sample size should be of at least $50 + 8K$ where K is the number of parameters to be estimated. In practice, it also depends on data availability. A minimum of 30 data points is usually considered as "safe".

Multicollinearity One of the classical assumptions of the method of least squares is that the independent variables should be independent of one another. If this assumption does not hold true (multicollinearity), the calculation and interpretation of the estimated coefficients are affected. The analysis may erroneously conclude that some relevant variables yield no significant impact. Consider for instance Fig. 5.8 where x_1 is correlated with both y and x_2. When y is regressed on x_1 and x_2, variable x_1 does not yield a significant impact. The reason is that x_1 and x_2 are linearly associated. It is as if x_2 was explaining the whole variation in y by itself. On the other hand, when y is regressed on x_1 alone, the impact is found to be significant.



```
Call:
lm(formula = y ~ x1 + x2)
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.019e-14  7.400e-15 1.378e+00     0.18
x1          1.000e+00  1.206e-16 8.294e+15   <2e-16 ***
x2          1.000e+00  1.747e-16 5.723e+15   <2e-16 ***
---
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 4.074e+31 on 2 and 27 DF,  p-value: < 2.2e-16
```

Fig. 5.7 Perfect relationship between y, x_1 and x_2

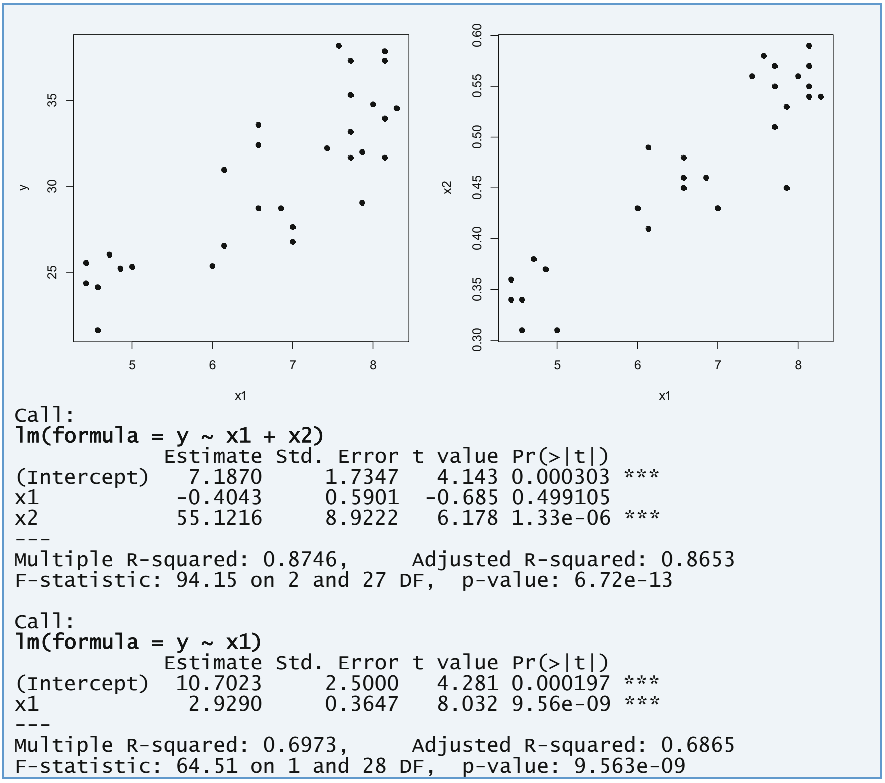
As already stressed in Sect. 5.3, the usual solution to avoid any misinterpretation is to rule out the offending variables.

Time Trends When faced with time series data, the dependent variable may grow or decline automatically over time. An omitted variable bias will occur if this evolution is not accurately taken into account. A commonly accepted way to deal with this problem is to include a linear trend in the regression equation:

$$y_t = \alpha_1 + \alpha_2 x_{2t} + \cdots + \alpha_K x_{Kt} + \beta trend_t + \epsilon_t$$

where *trend* is a sequential numbering of the time periods $(1, 2, \ldots, n)$ usually beginning with a value of 1. Coefficient β represents the increment at which the dependent variable changes, on average, in each time period. If β is positive, then the dependent variable increases by β from one period to the other. If it is negative, then the dependent variable decreases over time. Note that the dependent variable can also be expressed in logarithm so that the slope coefficient becomes a direct estimate of the percentage growth rate per period (see next section for a discussion about the different functional forms).

In practice, a graphical analysis guides the choice whether to include a time trend or not. Consider for example Fig. 5.9. It can be seen that variables y and x do not experience the same progression. While y shows a growing tendency through time, x is stationary. By stationary, we mean that x is not trending upwards or downwards. Explaining y solely on the basis of x cannot yield fine results in terms of goodness of

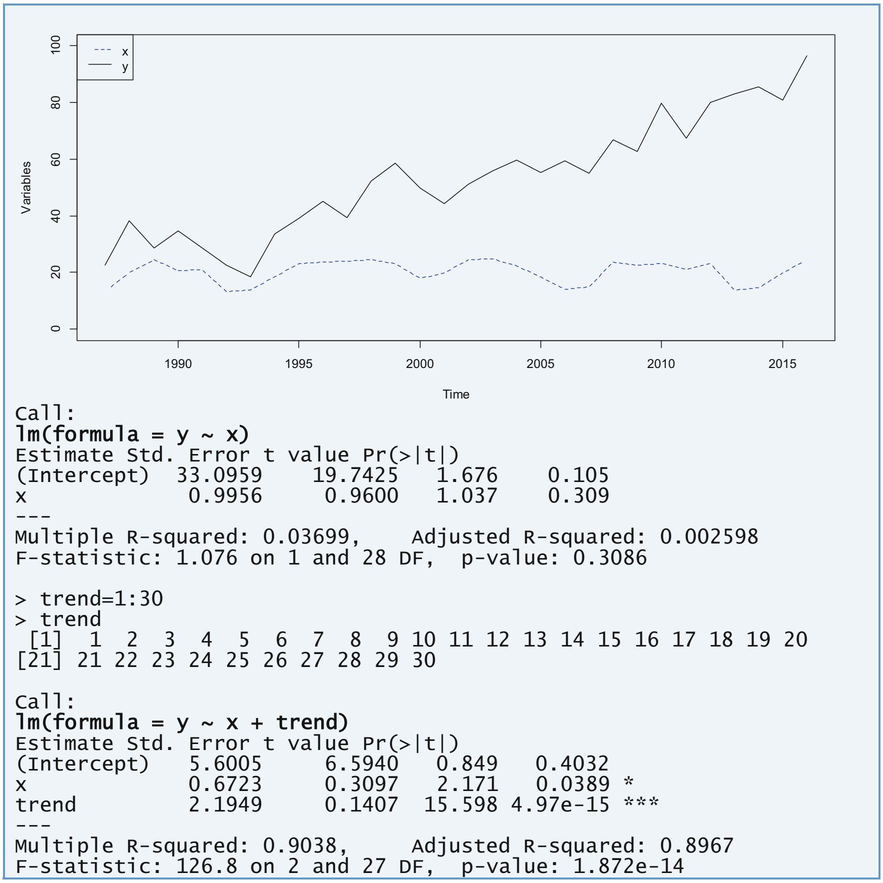


```
Call:
lm(formula = y ~ x1 + x2)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1870     1.7347   4.143 0.000303 ***
x1           -0.4043     0.5901  -0.685 0.499105
x2           55.1216     8.9222   6.178 1.33e-06 ***
---
Multiple R-squared: 0.8746,     Adjusted R-squared: 0.8653
F-statistic: 94.15 on 2 and 27 DF,  p-value: 6.72e-13

Call:
lm(formula = y ~ x1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.7023     2.5000   4.281 0.000197 ***
x1            2.9290     0.3647   8.032 9.56e-09 ***
---
Multiple R-squared: 0.6973,     Adjusted R-squared: 0.6865
F-statistic: 64.51 on 1 and 28 DF,  p-value: 9.563e-09
```

Fig. 5.8 Collinearity between x_1 and x_2

fit as pointed out by the quite low $\bar{R}^2$ (0.0025). This has two important consequences. First, the growth tendency of y is not assessed by the model. Second, the coefficient of x is found to be non-significant. The results improve radically when a trend variable is included ($trend = 1:30$ where 30 is the total number of time periods). In that case, both the impact of x and *trend* appear to be significant. The $\bar{R}^2$ amounts to 89.67%. On average, the dependent variable increases by 2.19 units each year.

Spurious Relationship A spurious relationship is observed when two variables are independent from each other and yet are found to be significantly associated due to the influence of a third, unobserved variable. This phenomenon may distort considerably what comes out of the data. A typical example is when one ignores a common trend in time series data. If a dependent variable and an independent variable are both nonstationary, they will automatically appear as correlated because of the influence of time. A way to solve this issue is to make the variables stationary by specifying them in first-difference. A similar situation occurs when a



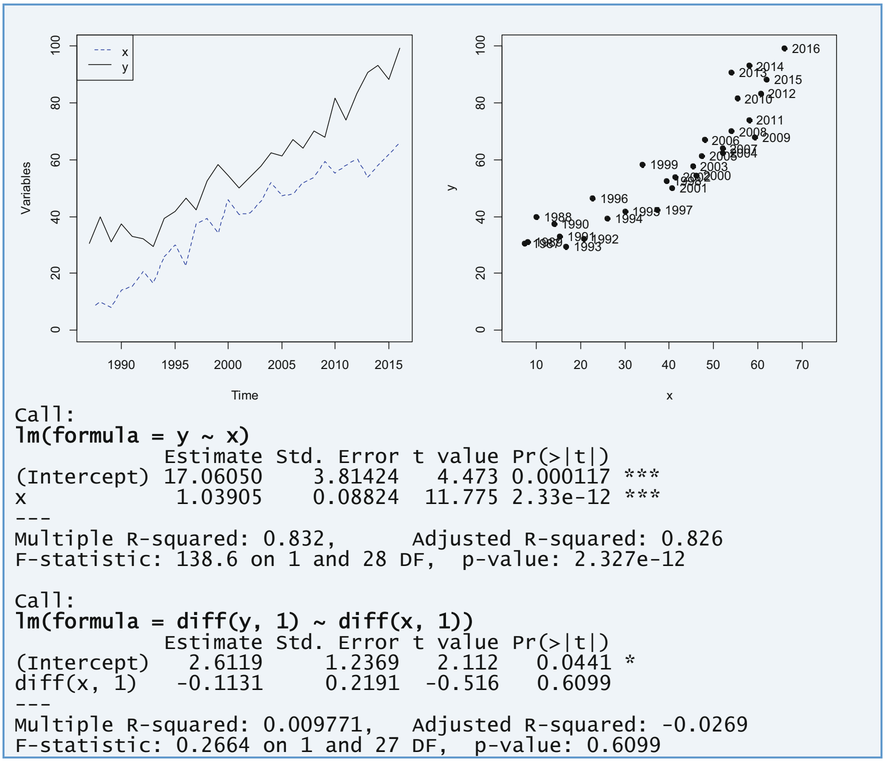
```
Call:
lm(formula = y ~ x)
Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.0959    19.7425   1.676    0.105
x              0.9956     0.9600   1.037    0.309
---
Multiple R-squared: 0.03699,    Adjusted R-squared: 0.002598
F-statistic: 1.076 on 1 and 28 DF,  p-value: 0.3086

> trend=1:30
> trend
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
[21] 21 22 23 24 25 26 27 28 29 30

Call:
lm(formula = y ~ x + trend)
Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.6005     6.5940   0.849   0.4032
x              0.6723     0.3097   2.171   0.0389 *
trend          2.1949     0.1407  15.598 4.97e-15 ***
---
Multiple R-squared: 0.9038,     Adjusted R-squared: 0.8967
F-statistic: 126.8 on 2 and 27 DF,  p-value: 1.872e-14
```

Fig. 5.9 Projecting time trends

trend in the data is attributed to inflation. A solution in that case is to deflate the series, i.e. to specify them in real terms.

Figure 5.10 offers an example. Both variables x and y grow over the years. When y is regressed on x, the analysis erroneously concludes to a significant relationship among the variables. We face a spurious regression. The scatter plot of y on x points out the problem. Smallest values of x and y are observed at the beginning of the observation period (before 1990) and largest values appear after 2010. The analysis is invalid because both variables are correlated with a third variable: time. The solution is to use first-differences. Rather than regressing y_t on x_t, we regress $\Delta y_t = y_t - y_{t-1}$ on $\Delta x_t = x_t - x_{t-1}$. The model becomes:

$$\Delta y_t = \beta_1 + \beta_2 \Delta x_{\mathrm{t}} + \epsilon_t$$



```
Call:
lm(formula = y ~ x)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.06050    3.81424   4.473 0.000117 ***
x            1.03905    0.08824  11.775 2.33e-12 ***
---
Multiple R-squared: 0.832,      Adjusted R-squared: 0.826
F-statistic: 138.6 on 1 and 28 DF,  p-value: 2.327e-12

Call:
lm(formula = diff(y, 1) ~ diff(x, 1))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6119     1.2369   2.112   0.0441 *
diff(x, 1)   -0.1131     0.2191  -0.516   0.6099
---
Multiple R-squared: 0.009771,   Adjusted R-squared: -0.0269
F-statistic: 0.2664 on 1 and 27 DF,  p-value: 0.6099
```

Fig. 5.10 Spurious regression

In such a model, when a constant β_1 is included, it estimates the coefficient of the trend in the dependent variable. On average, and at each period of time, the dependent variable increases by β_1 units. In Fig. 5.10, applying this method and using function *diff*, we find that y is not influenced by x anymore and that y automatically increases each year by 2.61. Note that the unadjusted R^2 is now negative, meaning actually that a model based on x only fits the data really poorly.

Spurious relationships are not limited to time series. Similar problems may occur when faced with cross-sectional data, especially when one does not control for size effects (the number of inhabitants in a set of jurisdictions, size of hospitals, countries by GDP). Consider for example a small city with one primary school, no railway station and a very low traffic. A comparison of this city with a larger city with shopping centers, multiple car parks and traffic schemes would be misleading. One would for instance erroneously conclude that school expenditures and traffic congestion are linked. The larger is the population, the higher are the number of cars and the number of pupils at school. Similarly, comparing two hospitals of different size may yield ambiguous results. Larger hospitals will face higher cost just because the numbers of patients is larger. Last, comparing countries based on their GDP is not right because this measure is expected to increase with population.

The obvious solution to those problems is to express the variables in per capita terms, e.g., expenditures per inhabitant or per patient. Another way to control for those size effects is to divide the variables of the study by a variable that indirectly accounts for the size of the units under examination. For instance, instead of expressing public expenditures in per capita terms we may consider government spending as a percentage of gross domestic product.

Structural breaks appear when there is an unexpected shift in a time series due for instance to some institutional change or new regulation. The usual and easiest approach to account for that phenomenon is to include a binary variable, aka dummy variable, in the regression equation. Such a variable switches from zero to one at the date of the breakpoint. Consider the following simple linear regression model:

$$y_t = \alpha_1 + \alpha_2 x_t + \epsilon_t$$

Suppose we suspect that the structure of the model has changed at a given point in time $t = t^*$. We would create a dummy d that takes value 0 for $t < t^*$ and 1 for $t \geq t^*$. Then two possibilities arise. First, the structural change may affect the intercept coefficient only, in which case the dummy variable is introduced as an additive term:

$$y_t = \alpha_1 + \alpha_2 x_t + \alpha_3 d_t + \epsilon_t$$

Here α_3 denotes the effect of the structural break on the constant term. On average, when $t < t^*$, the dummy equals zero and the intercept coefficient amounts to α_1, while for $t \geq t^*$ the dummy equals one and the constant is $\alpha_1 + \alpha_3$. Second, it is possible that the structural change affects both the constant and the slope coefficient in which case the dummy variable is also introduced as a multiplicative term, either as:

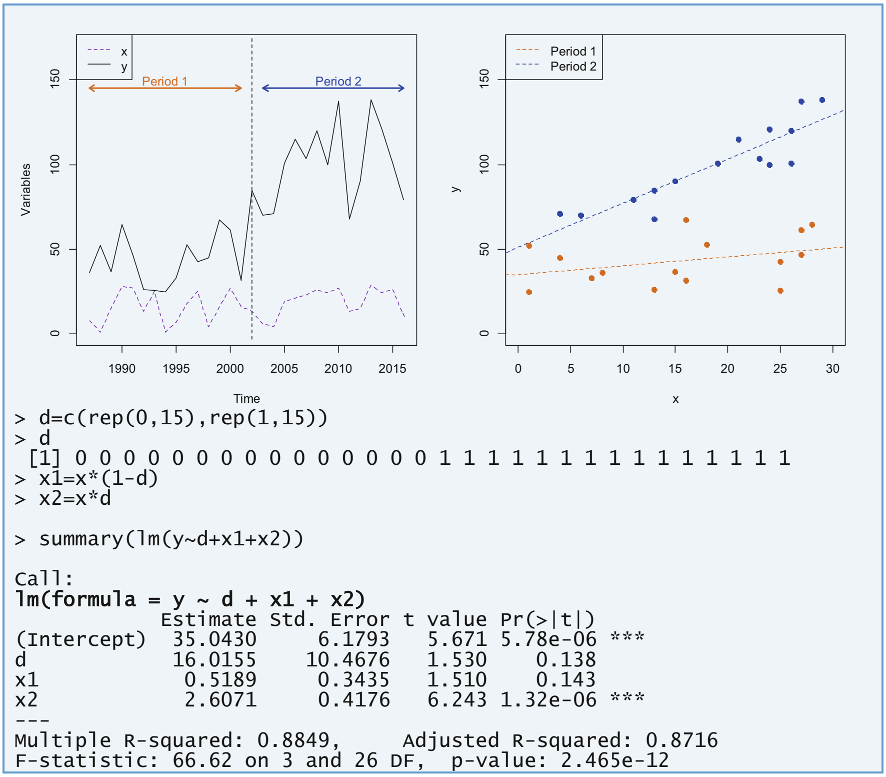
$$y_t = \alpha_1 + \alpha_2 x_t + \alpha_3 d_t + \alpha_4 d_t \times x_t + \epsilon_t$$

or:

$$y_t = \beta_1 + \beta_2 d_t + \beta_3 (1 - d_t) \times x_t + \beta_4 (d_t) \times x_t + \epsilon_t$$

On average, when $t < t^*$, the slope amounts to β_3, while for $t \geq t^*$ it is β_4. Not only does the dummy variable account for the change in the constant but also for the change in the slope coefficient.

Figure 5.11 provides an illustration. It can be seen that the evolution of y is unstable through time. Before 2002 (period 1), the values of y are smaller and less dispersed. After 2002 (period 2), y reaches higher and more dispersed values. To account for a possible structural change a dummy d is created with "zeros" for the first 15 periods, and "ones" for the last 15 periods. Variable x is then split into two elements: $x1 = x^*(1 - d)$ and $x2 = x^* d$. Using these variables, the regression output yields a significant result for the slope coefficient only for the second period. In



```
> d=c(rep(0,15),rep(1,15))
> d
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
> x1=x*(1-d)
> x2=x*d

> summary(lm(y~d+x1+x2))

Call:
lm(formula = y ~ d + x1 + x2)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.0430     6.1793   5.671 5.78e-06 ***
d            16.0155    10.4676   1.530    0.138
x1            0.5189     0.3435   1.510    0.143
x2            2.6071     0.4176   6.243 1.32e-06 ***
---
Multiple R-squared: 0.8849,     Adjusted R-squared: 0.8716
F-statistic: 66.62 on 3 and 26 DF,  p-value: 2.465e-12
```

Fig. 5.11 Structural breakpoint

period 1, the impact is found to be non significant while in period 2, an increase in x generates on average an increase in y of 2.60 units. Those relationships are displayed on the scatter plot of y on x. In blue are depicted the observations for the second period. Visually, the slope of the regression line is higher for this period. In the first period, represented in orange, the slope is closer to zero.

5.5 Functional Forms of Regression Models

Another difficulty in regression analysis is the choice of the functional form that best fit the data. Many functional forms are possible. They include for instance the double-log model, semi-log models, and polynomial models. In the double-log model, also known as Log-Log, all variables are expressed in natural logarithm. For instance, under the framework of the simple linear regression, we have:

$$\ln y_i = \alpha_1 + \alpha_2 \ln x_i + \epsilon_i$$

where ln y_i and ln x_i stand for the dependent variable and the independent variable, respectively. One asset of the double-log model is that the slope coefficient measures the elasticity of y with respect to x. We have on average:

$$\alpha_2 = \frac{dy/y}{dx/x}$$

By definition, an elasticity is the percentage change in y for a one percentage increase in x (if we set $dx/x = 1$ percent, we have $dy/y = \alpha_2$ percent). For instance, if the estimated coefficient is 1.2 that means that a 1% increase in x will generate on average a 1.2% increase in y. Graphically, the slope α_2 determines the shape of the regression curve. If $0 < \alpha_2 < 1$, the impact of the independent variable is positive and becomes smaller as its value increases (Fig. 5.12a). If $\alpha_2 > 1$, the impact of the independent variable is positive and becomes larger as its value increases (Fig. 5.12b). If $\alpha_2 = 1$, the impact of the independent variable is positive and constant (Fig. 5.12c). If $\alpha_2 < 0$, the impact of the dependent variable is negative (Fig. 5.12d).

It should be reminded that a distinction exists between a "percent change" and a "percentage point change". While the former denotes the proportion of increase or decrease in a given variable, the latter indicates whether a rate goes up or down and by how many points. For instance, in Fig. 5.5, variables were not expressed in natural log. We found that a one percentage point increase in the unemployment rate was generating on average an increase of $3.1 in per capita social spending. Here, by "percentage point increase" we mean the unemployment rate plus one percent. If the variable was expressed in logarithm, we would be referring to a "percent change", i.e. the unemployment rate times $(1 + 1\,\%\,)$.

Semi-log models, also referred to as Log-Lin and Lin-Log models, express one variable in natural logarithm and the other in a linear form. In the Log-Lin model, the natural logarithm of the dependent variable is taken, but not that of the independent variable:

$$\ln y_i = \alpha_1 + \alpha_2 x_i + \epsilon_i$$

The slope coefficient yields the percentage change in the dependent variable induced on average by a one unit increase in the independent variable:

$$\alpha_2 = \frac{dy/y}{dx}$$

If we set $dx = 1$, we have $dy/y = \alpha_2$. If $\alpha_2 > 0$, the impact of the independent variable is positive and becomes larger as its value increases (Fig. 5.12e). If $\alpha_2 < 0$, the impact of the dependent variable is negative and becomes smaller (Fig. 5.12f). In the Lin-Log model, only the independent variable is expressed in logarithm:

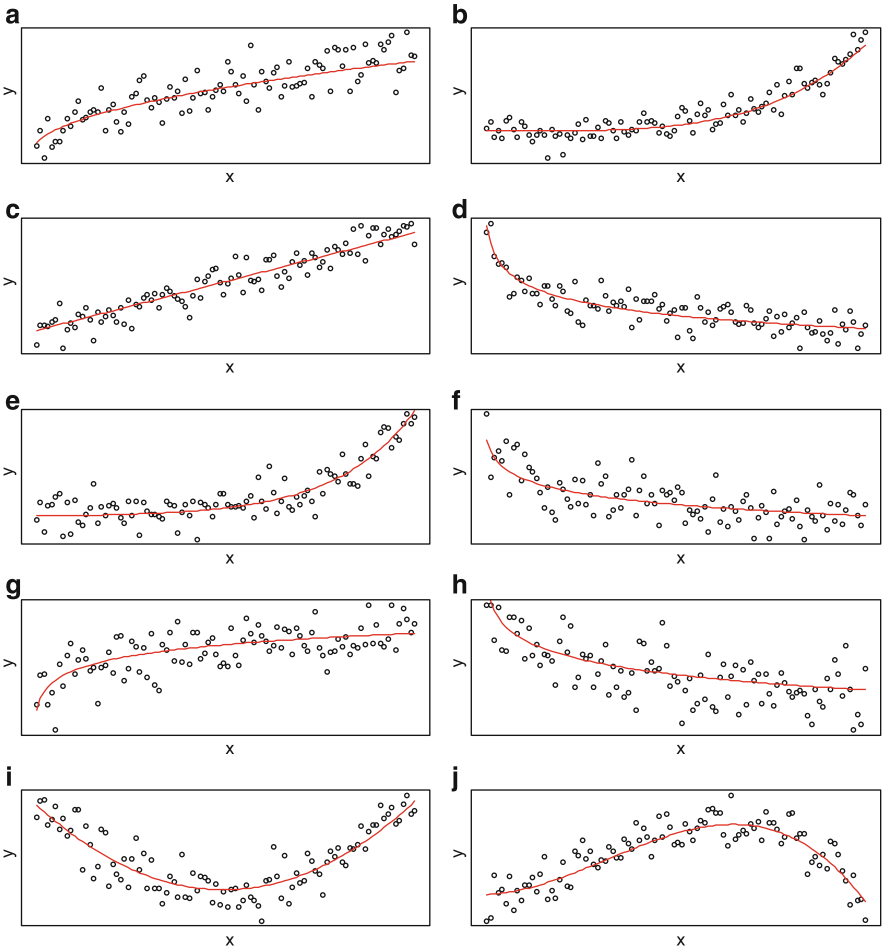


Fig. 5.12 Functional forms. (**a**) Double-log model: 0 < α2 < 1, (**b**) double-log model: α2 > 1, (**c**) double-log model: α2 = 1, (**d**) double-log model: α2 < 0, (**e**) log-lin model: α2 > 0, (**f**) log-lin model: α2 < 0, (**g**) lin-log model: α2 > 0, (**h**) lin-log model: α2 < 0, (**i**) quadratic model: α1, α3 > 0, α2 < 0, (**j**) cubic model: α1, α2, α3 > 0, α4 < 0

$$y_i = \alpha_1 + \alpha_2 \ln x_i + \epsilon_i$$

We have on average:

$$\alpha_2 = \frac{dy}{dx/x}$$

If we set $dx/x = 1\%$, we get $dy = \alpha_2/100$. A one percent increase in the independent variable yields a $\alpha_2/100$ change in the dependent variable. If $\alpha_2 > 0$, the impact of the independent variable is positive and becomes smaller as its value increases

(Fig. 5.12g). If $\alpha_2 < 0$, the impact of the independent variable is negative and becomes smaller (Fig. 5.12h).

It is also possible to choose a polynomial functional form that allows the relationship between x and y to be modeled in a more complex way. For instance, adding a squared term x_i^2 in the equation is useful when one wants to fit a U-shaped distribution, as illustrated in Fig. 5.12i. Formally, we write:

$$y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + \epsilon_i$$

This equation is usually referred to as a polynomial of degree two or quadratic. The model can also be extended by adding a cube term (cubic model), as shown in Fig. 5.12j:

$$y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2 + \alpha_4 x_i^3 + \epsilon_i$$

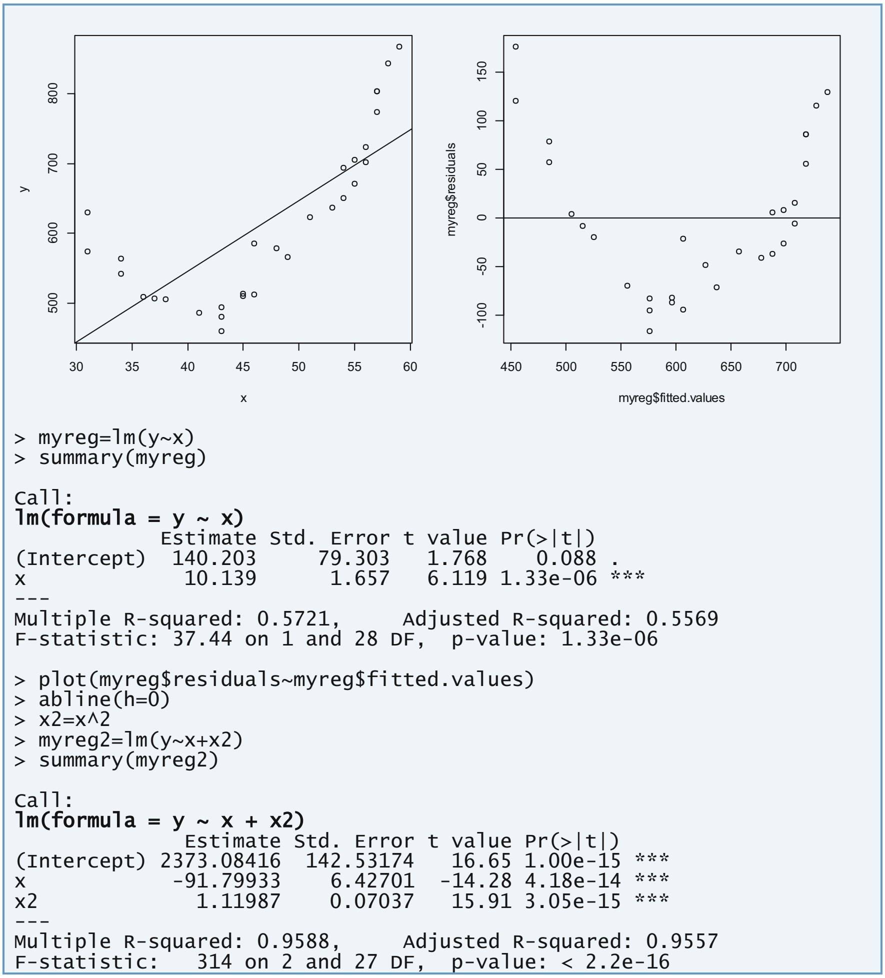
While more general, the polynomial transformation entails a loss of simplicity in the interpretation of regression coefficients.

In simple linear regression analysis (one independent variable only), a simple way to choose the best functional form is through the examination of a scatter plot, as shown in Fig. 5.12. The task is more difficult when more than one independent variable is included. In that case, we may rely on residual plots showing the residuals as a function of the fitted values. Any non-random pattern can be an indication that the regression function is not linear. An alternative is to rely on the adjusted R^2 of the models in competition. For instance, we can express the dependent variable linearly and compare the simple linear model with the Lin-Log or quadratic model. Or we can express the dependent variable in logarithm and compare the double-log model with the Log-Lin or any other functional form where ln$y$ is the dependent variable. When two models are compared, the best model is the one with the highest $\bar{R}^2$.

5.6 Detection and Correction of Estimation Biases

An important step that should not be neglected in econometrics is the examination of residual terms. Residual plots can be very useful in detecting for instance non-linearity, non-normal residuals, heteroscedasticity or autocorrelation. Those problems may themselves be generated by a model misspecification. More formally, several tests exist to diagnose a specific pattern in the way residuals are distributed: the Jarque-Bera test of normality, the Breusch-Pagan test for heteroscedasticity, the Durbin-Watson test for autocorrelation. This section offers a description of those procedures after testing for non-linearity.

Non Linearity If the linear functional form is not appropriate to fit the data, then the estimated coefficients as well as the fitted values will be biased. For instance, as shown in Fig. 5.13, the first scatter plot shows a nonlinear pattern between x and y.

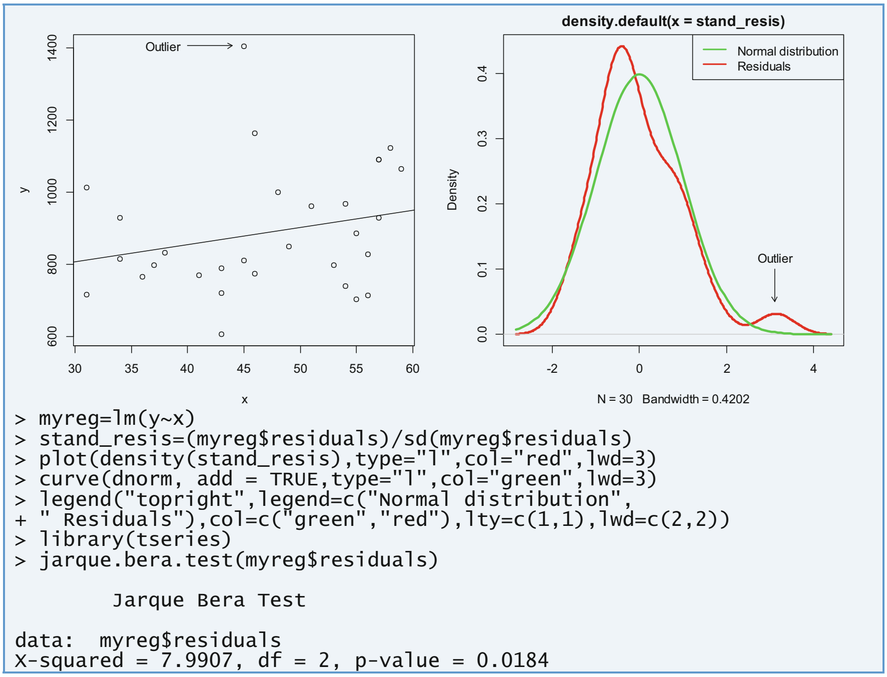


```
> myreg=lm(y~x)
> summary(myreg)

Call:
lm(formula = y ~ x)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  140.203     79.303   1.768    0.088 .
x             10.139      1.657   6.119 1.33e-06 ***
---
Multiple R-squared: 0.5721,     Adjusted R-squared: 0.5569
F-statistic: 37.44 on 1 and 28 DF,  p-value: 1.33e-06

> plot(myreg$residuals~myreg$fitted.values)
> abline(h=0)
> x2=x^2
> myreg2=lm(y~x+x2)
> summary(myreg2)

Call:
lm(formula = y ~ x + x2)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2373.08416  142.53174   16.65 1.00e-15 ***
x            -91.79933    6.42701  -14.28 4.18e-14 ***
x2             1.11987    0.07037   15.91 3.05e-15 ***
---
Multiple R-squared: 0.9588,     Adjusted R-squared: 0.9557
F-statistic:   314 on 2 and 27 DF,  p-value: < 2.2e-16
```

Fig. 5.13 Violation of the linearity assumption

Using a linear model is inappropriate. Plotting the residuals ($\hat{\epsilon}$) of that linear regression as a function of the fitted values ($\hat{y}$) illuminates the problem. The command *abline*($h = 0$) is used to add a horizontal line to the plot. Data points above that line are underestimated while those below that line are overestimated. The closer a data point is to the line, the better the model estimates that observation. As expected, the residual plot points out a non-linear pattern. As can be seen, a possible solution to that problem is to use a quadratic form (x and x^2), which improves the goodness of fit as measured with the adjusted R^2.

Non-normal Residuals Violations of normality affect the results of the *t*-tests of statistical significance. Non-normality occurs for instance when outliers are present



```
> myreg=lm(y~x)
> stand_resis=(myreg$residuals)/sd(myreg$residuals)
> plot(density(stand_resis),type="l",col="red",lwd=3)
> curve(dnorm, add = TRUE,type="l",col="green",lwd=3)
> legend("topright",legend=c("Normal distribution",
+ " Residuals"),col=c("green","red"),lty=c(1,1),lwd=c(2,2))
> library(tseries)
> jarque.bera.test(myreg$residuals)

        Jarque Bera Test

data:  myreg$residuals
X-squared = 7.9907, df = 2, p-value = 0.0184
```

Fig. 5.14 Violation of the normality assumption

in the data (residual terms are large for some of the observations) or when the functional form is inappropriate. The distribution of residuals can be asymmetric (skewness coefficient different from 0), light-tailed or heavy-tailed (kurtosis coefficient different from 3). Visually, the easiest way to assess the normality of residuals is to draw a probability density function for the residuals and to compare it to the normal distribution. Figure 5.14 offers an illustration. The scatter plot of *y* on *x* points out an outlier. Once the regression line has been fitted using the *lm* function, the values of the residuals (*myreg*$*residuals*) are standardized, i.e. divided by their (sample) standard deviation (we do no need to subtract the mean as the sum of residuals is zero). The probability density function of the residuals is then drawn using the *plot* function. The density of the standard normal distribution is added to the graph using the *curve* command. As can be observed, the shape of the distribution is far from being normal on the right tail of the distribution.

The Jarque-Bera test offers a formal way of detecting non-normal residuals. The test looks jointly at the skewness and kurtosis of the residuals distribution via the following statistic:

$$JB = (n - K)\left[\frac{g_1^2}{6} + \frac{(g_2 - 3)^2}{24}\right]$$

where n is the sample size, K is the number of regressors, g_1 is the skewness coefficient, g_2 is the kurtosis coefficient. For normally distributed residuals, $g_1 = 0$ and $g_2 = 3$. The test hypotheses are specified as follows:

$$H_0 : \text{The residuals are normally distributed } (JB = 0)$$

$$H_1 : \text{The residuals are not normally distributed } (JB > 0)$$

Under the null hypothesis, the JB statistic follows a chi-square distribution with 2 degrees of freedom. If the p-value of the test is lower than 5%, one rejects the hypothesis that the residuals are normally distributed. If the null hypothesis fails, one can then check whether the violation of normality is due to the presence of outliers, i.e. anomalous values in the data. One can also try to transform the dependent variable or the explanatory variables as described in the previous section until one reaches normality. For instance, in Fig. 5.14, the command *jarque . bera . test* from the package *tseries* has been used to assess the normality of the residuals. It can be seen that the residuals are not normally distributed: the p-value amounts to 0.0184 (and is lower than 5%) and the null hypothesis is rejected.

Heteroscedasticity This problem frequently appears in cross-sectional data. When the units under examination are heterogeneous with respect to their size, large units (e.g., highly populated cities or countries, large hospitals) may exhibit high variance in the dependent variable. Smallest units on the other hand are more likely to be similar. This may generate a problem of heteroscedasticity where the residual variance is not constant across fitted values. If the residual plot exhibits such a pattern, then heteroscedasticity is said to be present and the estimation procedure needs to be modified accordingly. Figure 5.15 provides an example. In the scatter plot of y on x, observations are more dispersed for high values of x. This pattern is also observed in the residual plot where the variance is found to be not constant. As a consequence, the results of the t- tests cannot be trusted.

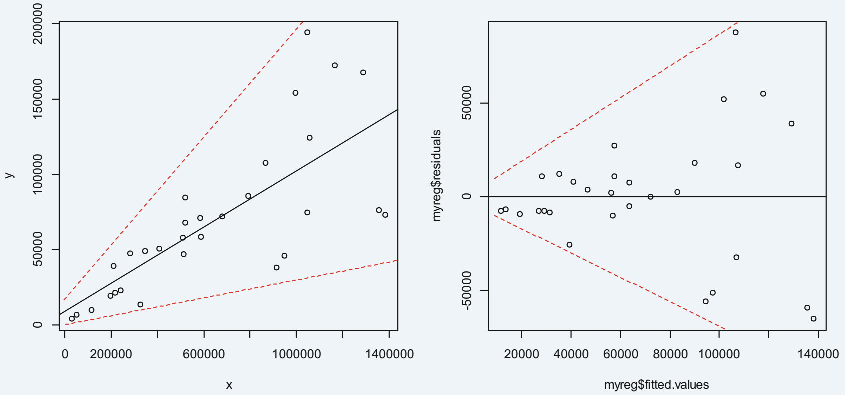
The Breusch-Pagan test can be implemented to detect heteroscedasticity. It tests whether the estimated variance of the residuals is dependent on the values of the regressors:

$$\widehat{\epsilon^2} = \beta_1 + \beta_2 x_{2\mathrm{i}} + \beta_3 x_{3\mathrm{i}} + \ldots + \beta_K x_{Ki} + u_i$$

where u_i is the error term of this regression. The test hypotheses are:

$$H_0 : \text{Homoscedasticity}$$

$$H_1 : \text{Heteroscedasticity}$$



```
> myreg=lm(y~x)
> summary(myreg)

Call:
lm(formula = y ~ x)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.607e+03  1.176e+04   0.732     0.47
x           9.380e-02  1.560e-02   6.013 1.77e-06 ***
---
Multiple R-squared: 0.5636,     Adjusted R-squared: 0.548
F-statistic: 36.16 on 1 and 28 DF,  p-value: 1.766e-06

> plot(myreg$residuals~myreg$fitted.values)
> abline(h=0)

> library(lmtest)
> bptest(myreg)
        studentized Breusch-Pagan test
data:  myreg
BP = 13.3175, df = 1, p-value = 0.0002629

> # Results with a corrected covariance matrix:
> library(sandwich)
> coeftest(myreg,vcov=vcovHC(myreg,type = "HC"))

t test of coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 8.6068e+03 7.6180e+03  1.1298    0.2681
x           9.3802e-02 1.8908e-02  4.9610 3.085e-05 ***
```

Fig. 5.15 Violation of the homoscedasticity assumption

If an F-test confirms that the independent variables are jointly significant then the null hypothesis of homoscedasticity is rejected. In some cases, expressing the dependent variable in logarithm or in per capita terms may solve the problem. In other cases, White's heteroscedasticity-corrected covariance matrix can be used to make inference:

$$HCE = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\widehat{\mathbf{\Omega}}\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1}$$

with $\widehat{\mathbf{\Omega}} = \text{diag}\left(\widehat{\epsilon}_1^2, \widehat{\epsilon}_2^2, \ldots, \widehat{\epsilon}_n^2\right)$ and where *HCE* stands for heteroscedasticity-consistent estimator. Statistical packages provide this matrix very easily. Note that using *HCE* yields a better estimate of the standard errors. It does not solve however what may have caused the problem in the first place. For instance, in Fig. 5.15, the Breusch-Pagan test is carried out with the command *bptest* from the package *lmtest*. The *p*-value of the test is lower than 5% which means that we reject the hypothesis of homoscedasticity. The command *coeftest* then recalculates the *t*-value using the corrected covariance matrix. The entry *HC* specifies White's estimator.

Autocorrelation This problem appears in time series data when the series have a common trend, or when the estimated relationship has a non-linear shape. A way to diagnose this problem is to plot the residuals at t as a function of the residuals at $t-1$. The Durbin-Watson test can also be implemented to detect the presence of autocorrelation. The test hypotheses are the following:

$$H_0 : \text{No autocorrelation}$$

$$H_1 : \text{Autocorrelation}$$

The test statistic is defined as:

$$DW = \frac{\sum_{t=2}^{n}\left(\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1}\right)^2}{\sum_{t=1}^{n}\left(\widehat{\epsilon}_t\right)^2}$$

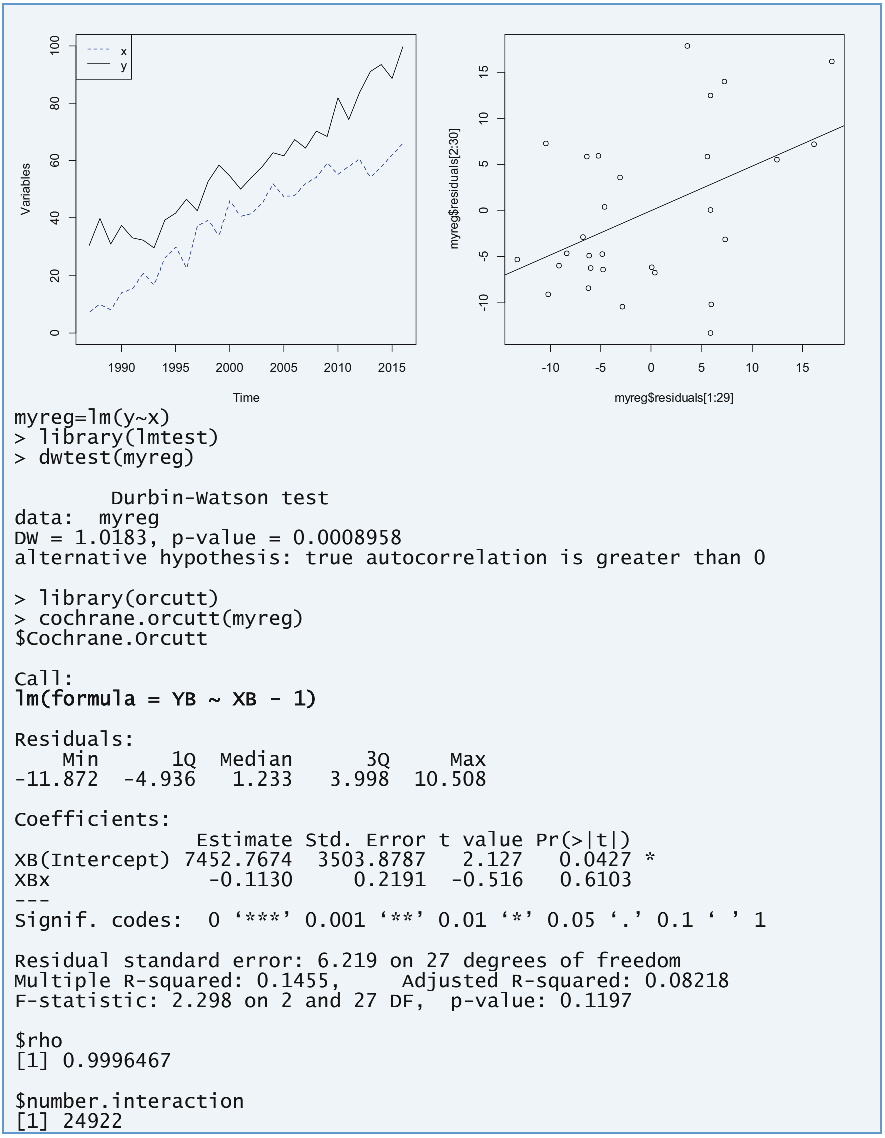
This statistic lies between 0 and 4. At a 5% significance level, it is compared to lower and upper critical values. If *DW* is close to 0, there is evidence of positive serial correlation. If *DW* is close to 4, there is evidence of negative serial correlation. If autocorrelation is detected, then the model needs to be modified by using for instance first-differences or non-linear transformation of the variables. The problem can also be resolved with the Cochrane–Orcutt method. The procedure is iterative and usually available in any statistical package. It consists in estimating:

$$\widehat{\epsilon}_t = \rho\widehat{\epsilon}_{t-1} + u_t$$

Using $\widehat{\rho}$ we transform the regression model by taking a quasi-difference:

$$y_t - \widehat{\rho}y_{t-1} = \alpha_1 + \alpha_2(x_{2t} - \widehat{\rho}x_{2t-1}) + \ldots + \alpha_\mathrm{K}(x_{Kt} - \widehat{\rho}x_{Kt-1}) + \epsilon_i$$

The procedure is then reiterated until no important variation in the estimated value of $\widehat{\rho}$ is observed.



```
myreg=lm(y~x)
> library(lmtest)
> dwtest(myreg)

        Durbin-Watson test
data:  myreg
DW = 1.0183, p-value = 0.0008958
alternative hypothesis: true autocorrelation is greater than 0

> library(orcutt)
> cochrane.orcutt(myreg)
$Cochrane.Orcutt

Call:
lm(formula = YB ~ XB - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.872  -4.936   1.233   3.998  10.508

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
XB(Intercept) 7452.7674  3503.8787   2.127   0.0427 *
XBx             -0.1130     0.2191  -0.516   0.6103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.219 on 27 degrees of freedom
Multiple R-squared: 0.1455,     Adjusted R-squared: 0.08218
F-statistic: 2.298 on 2 and 27 DF,  p-value: 0.1197

$rho
[1] 0.9996467

$number.interaction
[1] 24922
```

Fig. 5.16 Test of autocorrelation

Figure 5.16 offers an illustration. Both the dependent variable y and the independent variable x are non-stationary. For this reason, regressing y on x yields a problem of autocorrelation. The scatter plot of the residuals at year t (observations 2–30) against the residuals at year $t - 1$ (observations 1–29) exhibits a pattern. The Durbin-Watson test (command *dwtest*) confirms the presence of autocorrelation. The p-value is lower than 5%. To solve this issue, the iterative procedure of

Cochrane–Orcutt can be used (command *cochrane . orcutt* from the package *orcutt*). Using this approach, the dependent variable does not yield a significant impact on y. The slope parameter $\widehat{\rho}$ is found to be 0.99 and the software needs 24,922 iterations to reach a stable result.

5.7 Model Selection and Analysis of Regression Results

It is not an easy task to select the final model that will be included in a policy report. Quite often, the method of selection relies on trial and error. Once a model is estimated, the regression output must be analyzed meticulously. If the estimated coefficients are not significant or in the opposite direction to what is expected, one should question the choice of variables and the manner in which they are measured and expressed. What happens if one of the variables is excluded? What if one additional control variable is included? It is also important to check that the coefficients do not change significantly when extra variables are included or excluded. Instability in the estimated coefficients is often a sign of multicollinearity. The examination of a correlation matrix can be very helpful in this respect, in order to detecting those potential problems. The adjusted R^2 and the F-test of overall significance can also be used to assess the quality of the model. Does the goodness of fit improve extensively when an extra variable is added to the equation? Last but not least, once the final model is established, it is important to check that the residuals satisfy the classical assumptions of the OLS method. If not, then it means that a few problems remain with regard to the chosen specification.

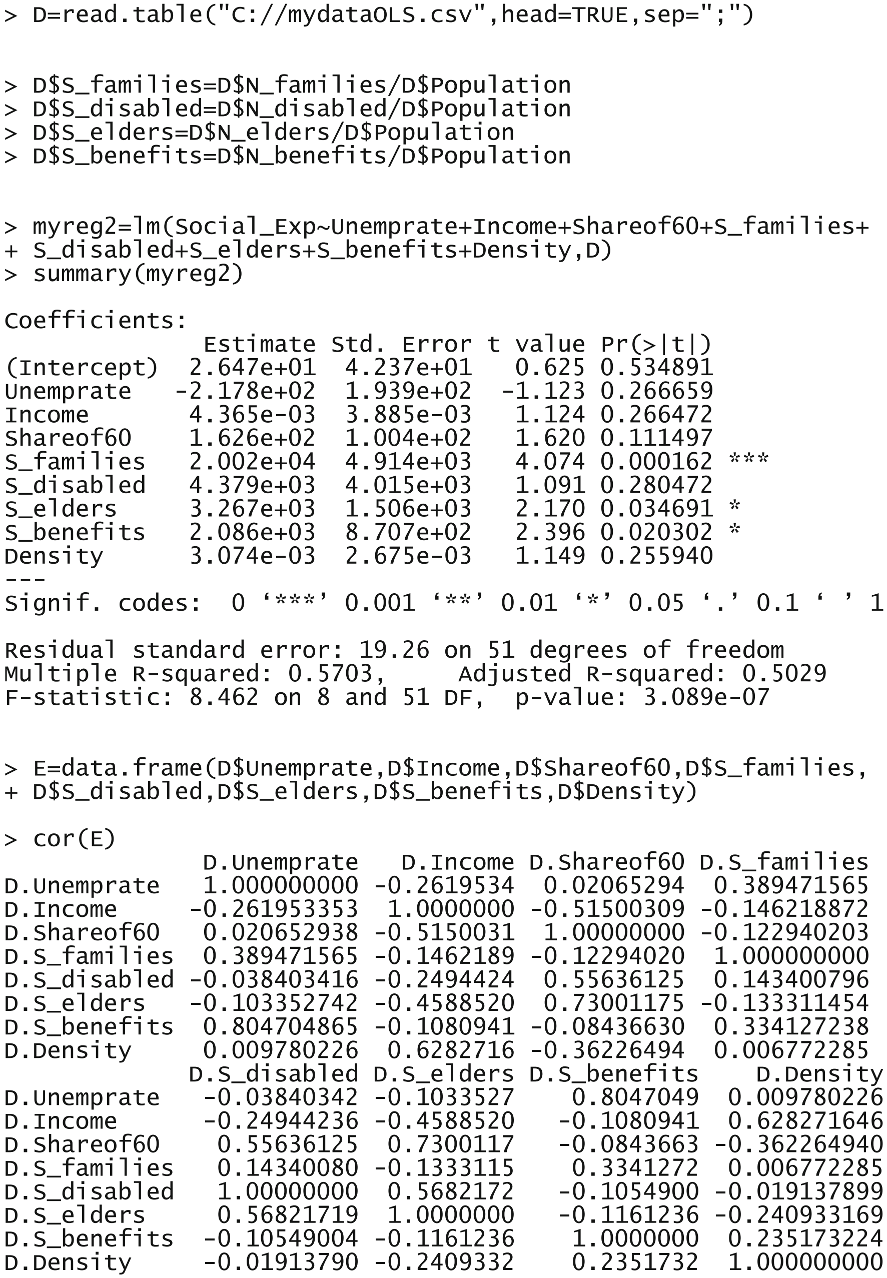
Consider for instance example 1. In Fig. 5.5, we have made the choice of including three variables only (as a matter of simplicity, we do not discuss the theory behind the model):

Model 1

$$Social\_Exp = \alpha_0 + \alpha_1 Unemprate + \alpha_2 Income + \alpha_3 Shareof60 + \epsilon_i$$

The adjusted R^2 of the model amounts to 13.85%. The regression equation can actually be modified so as to improve the goodness of fit. For instance, so far we did not use information about variables *N_family*, *N_disabled*, *N_elder* and *N_benefits* which represent the number of families, disabled, elder and unemployed who receive social assistance, respectively. The population density could also be included in the model. Does the goodness of fit increase when those variables are included?

To avoid spurious regression or any other misspecification, we cannot express the variables in level. All the variables must be expressed in per capita terms. This is done in Fig. 5.17 where data about social beneficiaries are now expressed in percentage of the population and renamed *S_families*, *S_disabled*, *S_elders*, and *S_benefits*. Those new variables complete the database with additional columns. A new model is estimated:



```
> D=read.table("C://mydataOLS.csv",head=TRUE,sep=";")


> D$S_families=D$N_families/D$Population
> D$S_disabled=D$N_disabled/D$Population
> D$S_elders=D$N_elders/D$Population
> D$S_benefits=D$N_benefits/D$Population


> myreg2=lm(Social_Exp~Unemprate+Income+Shareof60+S_families+
+ S_disabled+S_elders+S_benefits+Density,D)
> summary(myreg2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.647e+01  4.237e+01   0.625 0.534891
Unemprate   -2.178e+02  1.939e+02  -1.123 0.266659
Income       4.365e-03  3.885e-03   1.124 0.266472
Shareof60    1.626e+02  1.004e+02   1.620 0.111497
S_families   2.002e+04  4.914e+03   4.074 0.000162 ***
S_disabled   4.379e+03  4.015e+03   1.091 0.280472
S_elders     3.267e+03  1.506e+03   2.170 0.034691 *
S_benefits   2.086e+03  8.707e+02   2.396 0.020302 *
Density      3.074e-03  2.675e-03   1.149 0.255940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.26 on 51 degrees of freedom
Multiple R-squared: 0.5703,     Adjusted R-squared: 0.5029
F-statistic: 8.462 on 8 and 51 DF,  p-value: 3.089e-07


> E=data.frame(D$Unemprate,D$Income,D$Shareof60,D$S_families,
+ D$S_disabled,D$S_elders,D$S_benefits,D$Density)

> cor(E)
              D.Unemprate   D.Income D.Shareof60 D.S_families
D.Unemprate   1.000000000 -0.2619534  0.02065294  0.389471565
D.Income     -0.261953353  1.0000000 -0.51500309 -0.146218872
D.Shareof60   0.020652938 -0.5150031  1.00000000 -0.122940203
D.S_families  0.389471565 -0.1462189 -0.12294020  1.000000000
D.S_disabled -0.038403416 -0.2494424  0.55636125  0.143400796
D.S_elders   -0.103352742 -0.4588520  0.73001175 -0.133311454
D.S_benefits  0.804704865 -0.1080941 -0.08436630  0.334127238
D.Density     0.009780226  0.6282716 -0.36226494  0.006772285
             D.S_disabled D.S_elders D.S_benefits    D.Density
D.Unemprate   -0.03840342 -0.1033527    0.8047049  0.009780226
D.Income      -0.24944236 -0.4588520   -0.1080941  0.628271646
D.Shareof60    0.55636125  0.7300117   -0.0843663 -0.362264940
D.S_families   0.14340080 -0.1333115    0.3341272  0.006772285
D.S_disabled   1.00000000  0.5682172   -0.1054900 -0.019137899
D.S_elders     0.56821719  1.0000000   -0.1161236 -0.240933169
D.S_benefits  -0.10549004 -0.1161236    1.0000000  0.235173224
D.Density     -0.01913790 -0.2409332    0.2351732  1.000000000
```

Fig. 5.17 Model selection using R-CRAN: example 1 (part 1)

Model 2

$$Social\_Exp = \alpha_0 + \alpha_1 Unemprate + \alpha_2 Income + \alpha_3 Shareof60 + \alpha_4 S\_families$$

$$+\alpha_5 S\_disabled + \alpha_6 S\_elders + \alpha_7 S\_benefits + \alpha_8 Density + \epsilon_i$$

The goodness increases from 13.85% (Fig. 5.5) to 50.29% (Fig. 5.17). Yet, we should be careful because we do not want to introduce variables that are too much correlated with each other. For instance, *Unemprate* and *Shareof*60 do not show a significant impact anymore. This is a sign of multicollinearity which is confirmed by the computation of a correlation matrix (see function *cor* in Fig. 5.17). *Unemprate* is strongly correlated with *S_benefits* ($r = 0.80$), while *Shareof*60 is strongly correlated with *S_elders* ($r = 0.73$). Moreover, variable *Income* seems to be correlated with the population density ($r = 0.63$).

From the previous discussion, we exclude *Shareof*60, *Unemprate* and *Density* from the regression equation. Figure 5.18 estimates the following model:

Model 3

$$Social\_Exp = \alpha_0 + \alpha_1 Income + \alpha_2 S\_families + \alpha_3 S\_disabled$$

$$+\alpha_4 S\_elders + \alpha_5 S\_benefits + \epsilon_i$$

The model yields an adjusted R^2 equal to 48.88%. At this stage, we may decide to implement a Box-Cox test to determine the form of the dependent variable. The Box-Cox procedure consists in finding the optimal value of a parameter λ that yields the final form of the dependent variable:

$$\text{Endogenous variable} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ \ln(y) \text{if } \lambda = 0 \end{cases}$$

If $\lambda = 0.5$ we can use the square root of y in the regression analysis. If $\lambda = 1$ we may use y, or $\ln(y)$ if the parameter approaches 0. Figure 5.18 implements the procedure using the command *boxcox* from the package *MASS*. It generates the graph of Fig. 5.19. The horizontal line indicates a 95% confidence interval about the maximum observed value of λ. As can be seen, the procedure yields a parameter λ close to one, which means that we can keep the dependent variable in a linear form.

The question remains about the form of the independent variables. We may try to express the independent variables in logarithm:

Model 4

$$Social\_Exp = \alpha_0 + \alpha_1 \ln(Income) + \alpha_2 \ln(S\_families) + \alpha_3 \ln(S\_\ disabled)$$

$$+\alpha_4 \ln(S\_elders) + \alpha_5 \ln(S\_benefits) + \epsilon_i$$

```
> myreg3=lm(Social_Exp~Income+S_families+S_disabled+S_elders+
+ S_benefits,D)
> summary(myreg3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.548e+01  2.893e+01   0.535 0.594688
Income      6.706e-03  2.890e-03   2.320 0.024132 *
S_families  1.744e+04  4.807e+03   3.629 0.000633 ***
S_disabled  7.054e+03  3.720e+03   1.896 0.063289 .
S_elders    4.658e+03  1.323e+03   3.521 0.000883 ***
S_benefits  1.657e+03  5.087e+02   3.258 0.001944 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.53 on 54 degrees of freedom
Multiple R-squared: 0.5321,     Adjusted R-squared: 0.4888
F-statistic: 12.28 on 5 and 54 DF,  p-value: 5.704e-08

> library(MASS)
> boxcox(myreg3)

> myreg4=lm(Social_Exp~log(Income)+log(S_families)+log(S_disabled)+
+ log(S_elders)+log(S_benefits),D)
> summary(myreg4)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      263.460    217.680   1.210  0.23143
log(Income)       47.575     26.830   1.773  0.08183 .
log(S_families)   24.130      9.309   2.592  0.01225 *
log(S_disabled)   18.641     13.190   1.413  0.16331
log(S_elders)     30.186     10.251   2.945  0.00476 **
log(S_benefits)   22.552      8.077   2.792  0.00722 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.39 on 54 degrees of freedom
Multiple R-squared: 0.4391,     Adjusted R-squared: 0.3871
F-statistic: 8.453 on 5 and 54 DF,  p-value: 5.846e-06

> plot(myreg3$residuals~myreg3$fitted.values)
> abline(h=0)

> stand_resis=(myreg3$residuals)/sd(myreg3$residuals)
> plot(density(stand_resis),type="l",col="red",lwd=3,ylim=c(0,0.5))
> curve(dnorm, add = TRUE,type="l",col="green",lwd=3)
> legend("topright",legend=c("Normal distribution",
+ " Residuals"),col=c("green","red"),lty=c(1,1),lwd=c(2,2))

> library(tseries)
> jarque.bera.test(myreg3$residuals)

        Jarque Bera Test
data:  myreg3$residuals
X-squared = 1.5011, df = 2, p-value = 0.4721

> library(lmtest)
> bptest(myreg3)

        studentized Breusch-Pagan test

data:  myreg3
BP = 4.1453, df = 5, p-value = 0.5287
```

Fig. 5.18 Model selection using R-CRAN: example 1 (part 2)

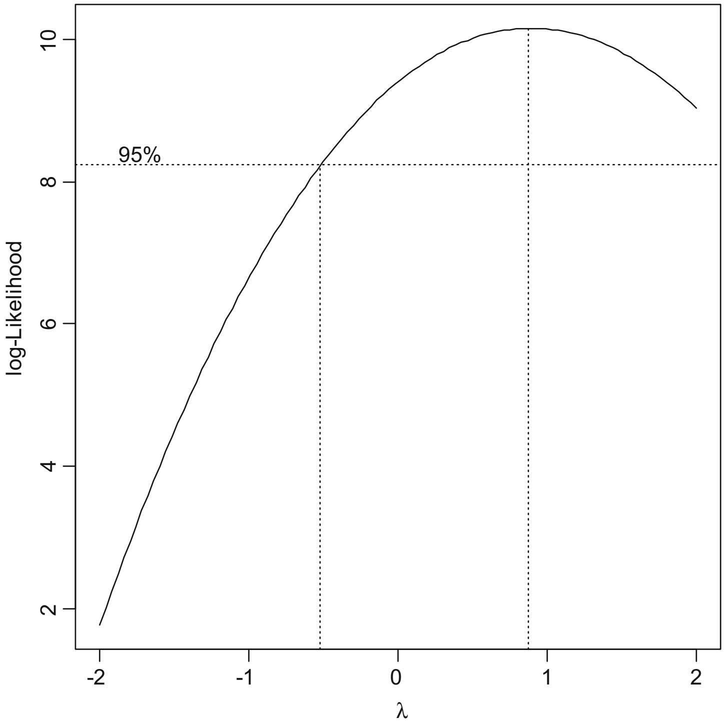


Fig. 5.19 Parameter of the Box-Cox power transformation

As can be seen from Fig. 5.18, the goodness of fit decreases when the independent variables are expressed in natural log (see *myreg*4). Overall, the best model seems to be Model 3.

The last step consists in examining the distribution of residuals. Figure 5.18 offers an example using the third model (*myreg*3). As already stated in Sect. 5.6, the scatter plot of residuals versus fitted values is very useful when conducting a residual analysis. The graph is used to detect non-linearity, non-normality (e.g., outliers) and heteroscedasticity (unequal error variances). Using information about *myreg*3, Fig. 5.20 is created with the command *plot (myreg3*$*residuals*~ *myreg3*$*fitted.values)*. The residuals (displayed on the y axis) are expressed as a function of fitted values (displayed on the x axis). The command *abline*($h = 0$) draws an horizontal line at $y = 0$. The residuals fit no particular pattern: the residuals are clustered around the horizontal axis, which suggests that the relationship is indeed linear. The distribution is symmetric around that axis and does not exhibit outliers, which supports the assumption of normality. The variance of residuals is not a function of the fitted values, meaning that the residuals are homoscedastic.

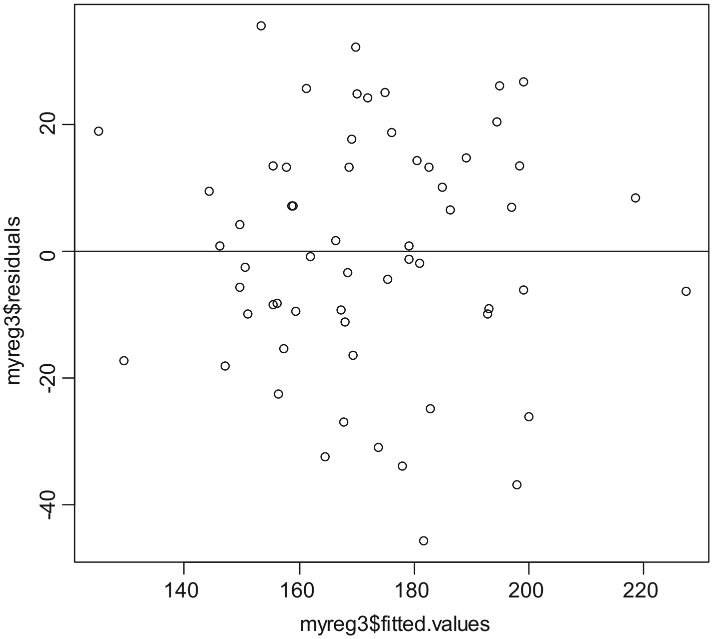


Fig. 5.20 Residuals versus fits plot

The normality of residuals is further confirmed in Fig. 5.21. To create this graph, the values of the residuals (*myreg*3*$residuals*) are standardized (*stand_resis*). Function *curve* displays the density of the standard normal distribution (see Fig. 5.18). Both density curves have a similar pattern. The tests also confirm that the residuals are normal (*jarque.bera.test*) and homoscedastic (*bptest*), the respective *p*-values being much above the 5% threshold.

Table 5.3 displays the previous results in a more polished form. This is the standard way to present regression outputs. For each variable, the numbers that are not in brackets are the regression coefficients. The sign of the coefficient indicates the direction of the relationship between the dependent variable and the independent variable. For instance, a negative relationship indicates that the dependent variable increases as the independent variable decreases. An asterisk is included whenever this impact is found to be significant. Three asterisk means that the *p*-value is lower than 0.1%, two asterisks indicates a 1% significant level, and one asterisk stands for a 5% significance level. Values in brackets indicate *t*-values. The higher are those values, the lower are the *p*-values associated with the *t*-tests. Note that “being significant” does not mean that the effect is significantly large. It means only that the observed effect is not due to chance. Only the regression coefficients measure the magnitude of the effects.

What do the regression results tell us? According to the result of Model 3, the mean income and the shares of social beneficiaries all have a statistically significant relationship with per capita social expenditures. For instance, an increase in income

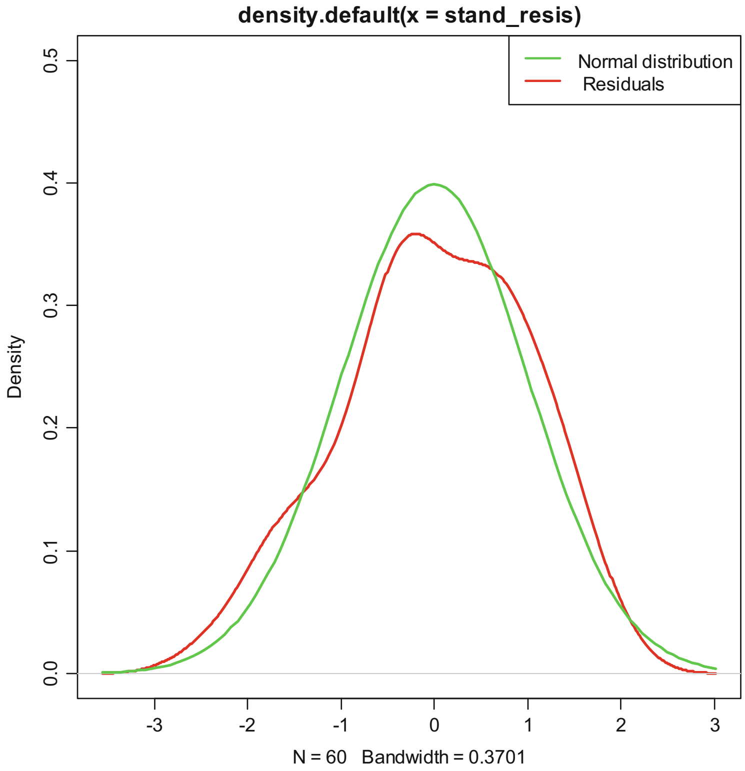


Fig. 5.21 Probability density function of residuals

by $100 would generate an increase in per capita social expenditures by $6.7 (see column (3) of Table 5.3). Assessing whether those effects are consistent with the existing literature would require further expertise as each model is context-dependent.

5.8 Models for Binary Outcomes

So far, the dependent variable has always been numeric which allows the OLS method to be used without worries. There some situations, however, where the dependent variable is categorical, which makes the OLS approach irrelevant. Examples include the analysis of individual choice and responses to survey questions, e.g., whether or not to buy a good, or to accept a particular policy. To illustrate, let the dependent variable y_i be a dichotomous random variable which takes the values 0 and 1 or equivalently “No/Yes”. The probability of observing either $y_i = 1$ or $y_i = 0$ is determined by a set of factors $x_{2i}, \ldots, x_{Ki}$ and a set of parameters $\alpha_1, \ldots, \alpha_K$. Since the dependent variable is dichotomous, the usual

Table 5.3 Estimation results (OLS): example 1

	Social_Exp			
	(1)	(2)	(3)	(4)
Intercept	49.95 (0.927)	26.47 (0.625)	15.48 (0.535)	263.46 (1.210)
Unemprate	312.6* (2.180)	−217.8 (−1.123)		
Income	0.0047 (1.207)	0.0043 (1.124)	0.0067* (2.320)	47.575 (1.773)
Shareof60	270.7** (2.910)	162.6 (1.620)		
S_families		20,020*** (4.074)	17,440*** (3.629)	24.130* (2.592)
S_disabled		4379 (1.091)	7054 (1.896)	18.641 (1.413)
S_elders		3267* (2.170)	4658*** (3.521)	30.186** (2.945)
S_benefits		2086* (2.396)	1657** (3.258)	22.552** (2.792)
Density		0.0030 (1.149)		
Adj. R^2	0.1385	0.5029	0.4888	0.3871
N	60	60	60	60
F-stat	4.161***	8.462***	12.28***	8.453***

***, **, and * indicate a significance level of 0.1%, 1% and 5%, respectively. *t-values* are in brackets

linear approach with ordinary least squares is not suitable for estimating this relationship. Consider for instance Fig. 5.22 where y_i is explained by one single dependent variable x. The individual observations (displayed in orange) do not follow a linear pattern. Individuals who face extreme values of x are more likely to say either "yes" or "no". In contrast, for intermediate values of x, they are more likely to hesitate. Two families of models can be used to estimate such a relationship: logit and probit models. Both models are forms of generalized linear models (GLMs) and adopt the maximum likelihood estimation (LME) method. The procedure requires specific distribution functions and estimates the parameters so that the probability of observing the sample outcome is as high as possible.

Formally, the probability that $y_i = 1$ is defined by a link function F which connects the dependent variables to the outcome:

$$\Pr(y_i = 1) = F(\alpha_1 + \alpha_2 x_{2\mathrm{i}} + \ldots + \alpha_K x_{Ki})$$

F is a cumulative distribution function which is expressed either as a logistic distribution function in the logit model or as a normal distribution function in the

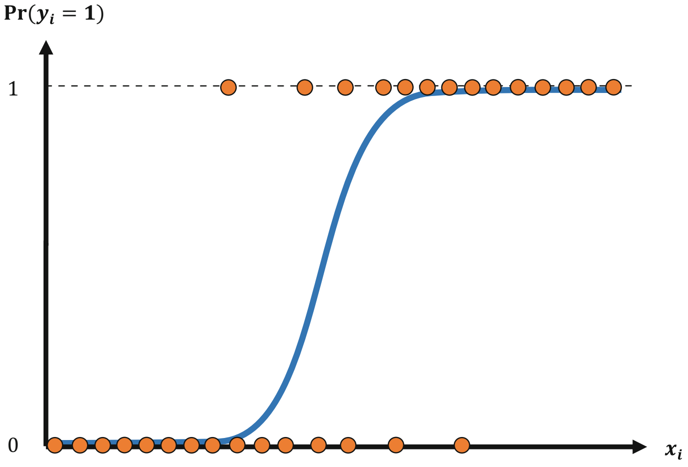


Fig. 5.22 A qualitative response regression model

probit model. Generally speaking, if a variable Z has mean μ_Z and variance σ_Z^2, those distributions are expressed as follows:

$$\text{Logit model} : F(z) = \frac{1}{1 - e^{-z}}$$

$$\text{Probit model} : F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sigma_Z \sqrt{2\pi}} e^{-\frac{(t-\mu_Z)^2}{2\sigma_Z^2}} dt$$

By replacing z with $\alpha_1 + \alpha_2 \mathrm{x}_{2\mathrm{i}} + \ldots + \alpha_K x_{Ki}$, the models can be expressed as follows:

$$\text{Logit model} : \ln \frac{y_i}{1 - y_i} = \alpha_1 + \alpha_2 \mathrm{x}_{2\mathrm{i}} + \ldots + \alpha_K x_{Ki}$$

$$\text{Probit model} : \Phi(y_i)^{-1} = \alpha_1 + \alpha_2 \mathrm{x}_{2\mathrm{i}} + \ldots + \alpha_K x_{Ki}$$

As with the method of least squares, the value of y_i depends on a set of independent variables and unknown parameters. Based on those specifications, it is possible to compute a log-likelihood function ln$L$ that is maximized for estimating the unknown parameters. Empirically, the logistic and normal cumulative functions do not differ much. The estimates obtained using the logit and probit models are often very close. For this reason, there is no exact rule as to which model should be used. Because of its simplicity, the logit model is often preferred.

An important feature of logit and probit models is that the parameters of the model, unlike the linear regression model, do not represent marginal effects. Those

Table 5.4 Data for example 2

Patient	S	Age	Smoker
1	0	63	1
2	0	63	1
3	0	56	0
4	0	61	1
5	0	64	0
6	0	61	1
7	0	64	1
8	0	61	1
9	0	69	0
10	0	61	1
11	0	67	0
12	0	49	0
13	0	57	0
14	0	59	0
15	0	69	0
16	0	61	1
17	0	63	0
18	0	69	0
19	0	59	0
20	0	64	0
21	0	64	0
22	0	41	0
23	0	61	0
24	0	54	1
25	0	64	1
26	0	61	0
27	0	68	0
28	0	40	1
29	0	65	0
30	0	69	0
31	0	64	0
32	0	64	0
33	1	61	1
34	1	73	0
35	1	65	1
36	1	79	0
37	1	56	1
38	1	69	1
39	1	62	1
40	1	73	1
41	1	93	1
42	1	61	1
43	1	62	1

(continued)

Table 5.4 (continued)

Patient	S	Age	Smoker
44	1	69	1
45	1	63	0
46	1	87	0
47	1	64	0
48	1	71	1
49	1	59	1
50	1	68	0
51	1	68	0
52	1	68	1
53	1	67	1
54	1	72	1
55	1	59	1
56	1	87	1
57	1	81	1
58	1	76	0
59	1	68	1
60	1	73	0

effects depend on the values of the dependent variable. To overcome this issue, one usually evaluates the marginal impacts at the sample means of the data, i.e. by setting the independent variables at their mean. Statistical packages, such as R-CRAN, usually make the task easier by providing those values automatically.

Assume for instance that we would like to assess whether the characteristics of a set of patients affect their probability to undergo a particular treatment, e.g., strategy 0 ($y_i = 0$) or strategy 1 ($y_i = 1$). The data is cross-sectional and presented in Table 5.4. It consists in $n = 60$ patients of different age (*Age*) who used to smoke or not (*Smoker*). All patients underwent treatment, either with strategy 0 ($S = 0$) or with strategy 1 ($S = 1$). The codes used in R-CRAN are presented in Fig. 5.23. The logit and probit models are estimated using the *glm* command. Both models specifies S as a function of *Age* and *Smoker*. Option *family* specifies the link function to be used in the model.

The command *summary* displays the results. The coefficients cannot be directly interpreted because they do not represent marginal effects. Yet, we can already see that there is a significant positive relationship between the probability of receiving treatment $S = 0$ and both *Age* and *Smoker*. Option $x = TRUE$ in the *glm* command indicates whether the exogenous variables should be saved for subsequent analysis. This option is required if one wants to obtain the marginal effects at the sample means, using the package *erer*. The regression output can be interpreted using the *maBina* command: being one year older increases the probability of undergoing strategy $S = 1$ by 6.5–6.7%. Being a smoker increases this probability by 52.5–53.1%.

```
> D=read.table("C://mydataBINARY.csv",head=TRUE,sep=";")
> mylogit=glm(S~Age+Smoker,D,family=binomial(link="logit"),x=TRUE)
> summary(mylogit)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50494  -0.72031  -0.08169   0.60866   2.01953

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.88718    5.77310  -3.272  0.00107 **
Age           0.26964    0.08532   3.160  0.00158 **
Smoker        2.37366    0.79753   2.976  0.00292 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.911  on 59  degrees of freedom
Residual deviance: 53.441  on 57  degrees of freedom
AIC: 59.441
Number of Fisher Scoring iterations: 6

> myprobit = glm(S~Age+Smoker,D, family=binomial(link="probit")
+ ,x=TRUE)
> summary(myprobit)

Call:
glm(formula = S ~ Age + Smoker, family = binomial(link = "probit"),
    data = D, x = TRUE)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50933  -0.73487  -0.02507   0.59835   2.02072

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.37050    3.20069  -3.553 0.000382 ***
Age           0.16259    0.04752   3.422 0.000623 ***
Smoker        1.43203    0.45052   3.179 0.001480 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.911  on 59  degrees of freedom
Residual deviance: 53.080  on 57  degrees of freedom
AIC: 59.08

Number of Fisher Scoring iterations: 7
> # Marginal effects:
> library(erer)
> maBina(mylogit)$out
            effect error t.value p.value
(Intercept) -4.701 1.440  -3.264   0.002
Age          0.067 0.021   3.140   0.003
Smoker       0.531 0.143   3.719   0.000
> maBina(myprobit)$out
            effect error t.value p.value
(Intercept) -4.528 1.276  -3.549   0.001
Age          0.065 0.019   3.409   0.001
Smoker       0.525 0.139   3.778   0.000

> logLik(mylogit)
'log Lik.' -26.72026 (df=3)
> logLik(myprobit)
'log Lik.' -26.53975 (df=3)
```

Fig. 5.23 Estimation of logit and probit model using R-CRAN: example 2

Note that the output of the *glm* command differs slightly from that of the *lm* command. The term "Fisher scoring iterations" relates to the way the model is estimated, which is based on successive trial and error until the log-likelihood function is maximized. The number of iterations corresponds to the number of times the software implements that process. The null deviance is a measure of the lack of fit of the model when the equation includes only the intercept. In that case, only one parameter is estimated and the number of degrees of freedom is $n - K = 60 - 1 = 59$. Likewise, the residual variance is a measure of the lack of fit, but when the model includes the independent variables. The number of degrees of freedom is 57 because three parameters are being estimated. The larger is the difference between the null deviance and the residual deviance, the better is the explanatory power of the model. Last, when two models are compared, the best model is the one with the lowest Akaike Information Criterion (*AIC*) value.

The computation of the residual deviance and the *AIC* is based on the value of the maximized log-likelihood function. We have:

$$\text{Residual Deviance} = -2\ln\left(L^*\right)$$

$$AIC = 2K - 2\ln\left(L^*\right)$$

where L^* is the maximum value of the likelihood function for the model. Consider for instance Fig. 5.23. The value of $\ln L^*$ is obtained with the *logLik* command. For the logit model, the residual deviance and the AIC are computed as follows:

$$\text{Residual Deviance} = -2 \times (-26.72026\ ) = 53.44052$$

$$AIC = 2 \times 3 - 2 \times (-26.72026\ ) = 59.44052$$

Those statistics are very useful when comparing two models.

Bibliographical Guideline

The term "regression" comes from genetics and has been popularized by Galton (1886), an English polymath, cousin of Darwin, who was interested in the link between the characteristics of children and those of their parents. The method of least squares originates in Pearson (1894). The analytical foundation of maximum likelihood estimations as well as further developments in econometrics have been introduced by Fisher (1922). The first known empirical studies to use multiple regressions in economics are those of Benini (1907), where a demand function for coffee is estimated using data from Italy, and Moore (1914), about economic cycles. The term "econometrics" is attributed to Ragnar Frisch, a Norwegian economist and the co-winner of the first Nobel Prize in economics in 1969 (with Jan Tinbergen). He was one of the founders of the Econometric Society and editor of Econometrica, a peer-reviewed academic journal of economics, for over twenty years. For further historical aspects the reader may consult Eatwell et al. (1990).

Since the seminal works of Pearson (1894) and Moore (1914), the econometric approach has gone through a number of methodological changes and developments. Far from being exhaustive, this chapter only provides the basics of regression techniques. To go further, the reader may rely on Gujarati and Porter (2009), Verbeek (2012) and Greene (2011). Those textbooks review the basic empirical techniques as well as the theoretical foundation of the econometric methods. Note that the rule of thumb for minimum sample size presented in Sect. 5.4 is derived from Green (1991).

It should be stressed that the econometric methods have long been decried in science. We may for instance quote Frisch's (1934) own criticism of the approach, which is nowadays still relevant. "*The data will frequently obey many more relations than those which the statistician happens to think of when he makes a particular regression study [. . .] As a matter of fact I believe that a substantial part of the regression and correlation analyses which have been made on economic data in recent years is nonsense for this reason*." To overcome this issue, many empirical analyses now rely on more complex techniques, such as quasi-experimental methods.

References

Benini, R. (1907). Sull'uso delle formole empiriche a nell'economia applicata. *Giornale degli economisti, 2nd series, 35*, 1053–1063.

Eatwell, J., Milgate, M., & Newman, P. (1990). *Econometrics*. Springer.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, 222*, 309–368.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute of Economics.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246–263.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499–510.

Greene, W. H. (2011). *Econometric analysis* (7th ed.). Hoboken, NJ: Pearson.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics*. McGraw-Hill Irwin.

Moore, H. L. (1914). *Economic cycles: Their law and cause*. New York: Macmillan Press.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London Series A, 185*, 71–110.

Verbeek, M. (2012). *A guide to modern econometrics* (4th ed.). Chichester: Wiley.

Estimation of Welfare Changes

6

6.1 Valuing the Consequences of a Project

Public projects have consequences on individual lives. This rather obvious statement nevertheless puts pressure on policy-makers to bring in relevant measures of how those consequences are perceived by their citizens. The estimation of welfare changes is thus a crucial step in the evaluation process, but also a tricky task.

Formally, let u_i denote the satisfaction or utility levels of agents $i = 1 \ldots n$. If the agents are able to express the level of satisfaction they get from being in one state (e.g., consumption bundle, health state) then their individual welfare is observable and those measurements can be used for policy-making purposes. For instance, assume that we would like to assess the change in welfare resulting from a particular public intervention. Let denote u_i^0 the utility that each agent i derives from the status quo and u_i^1 the utility they derive from the intervention. The main issue for the evaluator is to get a measure of the utility differences $u_i^1 - u_i^0$ for all agents $i = 1 \ldots n$. In this respect, one may rely on two different conceptualizations. First, we may assume that utility levels can be directly measured on a cardinal scale, in which case welfare is expressed in units of utility; or one may rely on the concept of consumer surplus, in which case welfare is measured in monetary units.

Utility refers to the benefit or satisfaction an agent derives from using a good or service. Surplus on the other hand is defined as the difference between the amount of money an agent is willing to pay for consuming a particular good or service and the price she or he actually pays. Although surplus and utility are different concepts, they are closely related: agents will derive extra satisfaction if the price they pay for a good decreases, first, because they will be able to consume more of that good and, second, because they will be able to use the extra money to purchase other goods.

To illustrate, consider an agent who is willing to pay \$4 for one unit of a good, \$7 for two units, \$9 for three units and \$10 for four units. The inverse demand curve (price as a function of quantity) can be schematized as follows:

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_6

$1			
$1	$1		
$1	$1	$1	
$1	$1	$1	$1
1 unit	2 units	3 units	4 units

The law of diminishing marginal utility says that as an agent uses more and more of a good, each additional unit yields less satisfaction. This is also reflected in the willingness to pay. Now assume that the actual price of the good is $3. At this price the agent will buy two units of the good:

$1			
$1	**$1**		
$1	**$1**	$1	
$1	**$1**	$1	$1
1 unit	**2 units**	3 units	4 units

The surplus amounts to $1 and is computed as the difference between the willingness to pay for those two units (here, $7) and the actual cost ($2 \times \$3 = \$6$).

Graphically, the surplus corresponds to the area (displayed in green in the diagram) below the inverse demand curve and above total spending (displayed in red). Should a reduction in price by $2 be observed, the surplus would increase by $5:

$1			
$1	**$1**		
$1	**$1**	**$1**	
$1	**$1**	**$1**	**$1**
1 unit	**2 units**	**3 units**	**4 units**

The agent derives additional satisfaction first because he or she will be able to buy two extra units of the good and, second, because he or she can buy additional goods with the extra $2 saved. Change in surplus is thus an approximation of differences in utility, expressed in monetary values.

Choosing whether and how to measure utility directly or through consumer surplus mainly depends on the context of the analysis. Broadly speaking, there are three ways of eliciting individual preferences with regard to a public project. The first two imply a monetization of consequences while the third does not. Let us briefly introduce them.

The first set of methods consists of stated preferences techniques whereby individuals declare what their perceptions are of the project and its consequences. A first and quite popular method is contingent valuation. A sample of individuals is picked from the population targeted by the project. They are asked how much they would be willing to pay for positive consequences (or receive in compensation for negative ones). The econometric treatment of their answers generates an average individual willingness to pay. Stated preference methods also include choice modeling techniques. They are quite suited to projects that have multiple characteristics that cannot be reduced to a single attribute. Respondents are then asked to evaluate multi-dimensional policy options. In this chapter, we focus on the discrete choice experiment method, as it is more and more often used to elicit preferences (the bibliographical guideline provides references on other choice modeling methods).

The second set of methods comprehends revealed preferences techniques. Instead of directly asking people what their perception of policy options is, preferences are inferred from what is observed on existing markets. For instance, the hedonic pricing method estimates the implicit price of non-market goods, e.g., proximity of a school or air quality, from their impact on real estate market prices. The method rests on the idea that the choice of a living place does not only reflect the preference for the property itself, but also integrates the environmental attributes. Regression analysis is then used to disentangle the different factors influencing market prices. Alternatively, the travel cost method uses information about the monetary and opportunity costs borne by individuals to reach a recreational site, and relates it to the demand for the site. The econometric analysis of that relation then estimates the individual marginal willingness to pay for the amenities of the site.

Finally, health technology assessment has developed an ambitious framework for evaluating individual perceptions of the health states they are in or may fall into. Measures of health-related quality of life allow to shift from objective statements of health states (for instance the walking condition after a hip surgery) to its subjective appreciation by individual who face it or are asked to act as if they were facing it. Quality adjusted life-years (or QALYs) aggregate two dimensions of health programs consequences, the quantity of life and the utility associated with each period lived under a particular health state. Contrary to revealed or stated preferences, the valuation of health-related quality of life does not involve any monetization of the consequences of a health program on individual welfare.

The outline of the chapter is as follows. Sections 6.2 and 6.3 develop two stated preferences methods: contingent valuation and discrete choice experiment. Sections 6.4 and 6.5 are about revealed preferences methods: hedonic pricing and travel cost method. For each section, we take time to demonstrate how statistical confidence intervals can be defined and applied. The use of confidence intervals, rather than point estimates, will reveal to the decision-maker the precision of the analysis. Finally, Sect. 6.6 proposes an introduction to valuation techniques for health outcomes.

6.2 Contingent Valuation

The contingent valuation method directly asks a sample of individuals from a target population how much they would be willing to pay or accept in compensation for gains or losses of non-market goods and services. The questionnaire is divided in two parts, a descriptive part and a set of questions. Since the survey is based on hypothetical scenarios, the descriptive part must contain a relevant set of information to approximate real-life situations. The consequences of the project are usually described on a yearly basis over a given time horizon. A typical scenario describes the type of good affected by the project, how the latter affects the quantity or quality of that good, whether it generates external effects, its location and duration. The scenario also explains the source of financing (users, taxpayers), the method of payment (fare, taxes), and its frequency (on-site, on a yearly basis). Finally, the status quo should be evoked: what if the project is not implemented? The description may be accompanied by support materials, such as charts or pictures, with care however as they sometimes induce biases due to framing effects.

The descriptive part is followed by a question about the willingness to pay of the respondent. The elicited value is thus contingent to the proposed scenario. The question can take the form of an open-ended question such as:

“*What is the most you would be willing to pay for . . . ?*”

Here the respondent has total freedom as regards the answer, which may be a problem if the respondent has no prior experience from using the good or insufficient knowledge about the satisfaction from using it. The approach may result in many missing values if the respondent refrains from answering. To overcome this problem, a closed-ended question can be employed with an appropriate set of amounts, or payment cards, among which respondents have to choose:

“*What is the most you would be willing to pay for . . .*”

□\$5□\$10□\$15□\$20□\$25□\$30□\$35□\$40□\$45□\$50

The amounts should be sufficiently different so that respondents do not hesitate too much between them.

In both cases, open- or closed-ended questions, the statistical treatment consists in computing the sample mean of declared individual willingness to pay. Let n represent the sample size and let w_i denote the willingness to pay of respondent i, $i = 1 \dots n$. The average individual willingness to pay ($AWTP$) is given by:

$$AWTP = \frac{\sum_{i=1}^{n} w_i}{n}$$

The confidence interval is defined as:

$$[AWTP - t \times se, AWTP + t \times se]$$

The estimated standard error of the mean $se = s/\sqrt{n}$ is computed from the sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (w_i - AWTP)^2}{n-1}}$$

Last, parameter t is obtained from a Student distribution table for $n - 1$ degrees of freedom and a 5% risk probability. In practice, for large samples ($n > 200$), the t-distribution converges on the normal distribution and the t-statistic is close to 1.96. In contrast, when n is small, the tails of the t-distribution decrease more slowly than the tails of the normal distribution, thus yielding higher confidence intervals.

Despite their simplicity, open- and closed-ended questions have been progressively less used in practice. Both types of questions may indeed induce individuals to choose an amount close to the price they may have experienced in similar (or seemingly so) situations. The bias in that case is tremendous as the elicited amount would approximate the price or cost of the good, and not the satisfaction derived from using it. The net welfare associated with the project is measured as the surplus, that is, the willingness to pay minus the price paid for the good. Should the survey elicit prices instead of willingness to pay, the surplus would approach zero, and the evaluator would be likely to reject the project for that reason.

To obtain adequate answers from the respondent, many studies prefer to employ the dichotomous choice approach advocated by the National Oceanic and Atmospheric Administration, also known as the NOAA method. Respondents are initially divided into subsamples each associated with different bid values. For instance, those bid values can be determined in a pre-survey, through an open-ended questionnaire. Then the survey asks respondents whether they would be willing to support the project considering the bid value they face:

$$\text{``}\textit{Assume that you have to pay } \$25 \textit{ for the project} \ldots$$

$$\textit{Would you be in favor of its implementation? } \square \textit{yes} \square \textit{ no}\text{''}$$

If the respondent is in favor of the project, it means that his/her willingness to pay is higher than the bid value, and lower otherwise. The approach thus does not directly provide a willingness to pay, but yields instead a lower bound if the answer is "yes" and an upper bound if the answer is "no".

With the NOAA method, the computation of the average willingness to pay implies the use of logit or probit econometric models. Let b_i denote the bid value faced by individual i and y_i the response. We set $y_i = 0$ if the answer is "no" and $y_i = 1$ if the answer is "yes". The probability of observing a positive response is determined by the bid value b_i and the individual willingness to pay w_i:

$$\Pr(y_i = 1) = \Pr(b_i < w_i)$$

The higher the bid value, the lower the probability that individual i accepts the project. Since the endogenous variable is dichotomous, the usual linear approach with ordinary least squares is not suitable for estimating this relationship. As illustrated in Fig. 6.1 where individual observations are represented in orange, the

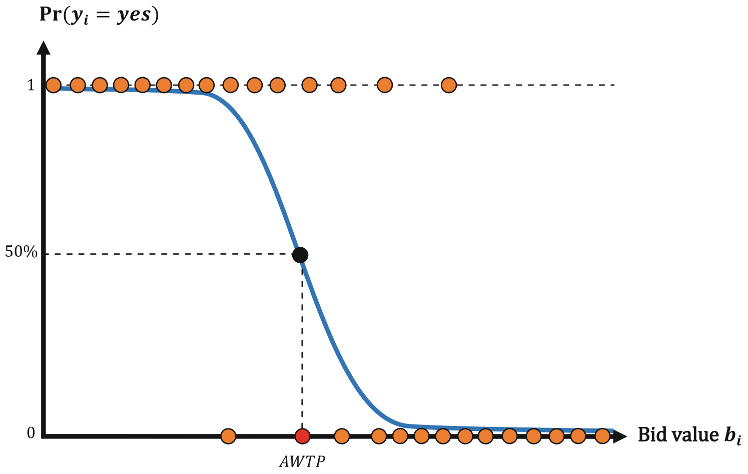


Fig. 6.1 The dichotomous choice method: illustration

model is nonlinear. Individuals who face extreme bid values are more likely to say either "yes" or "no". In contrast, for intermediate bid values, they are more likely to hesitate between supporting and rejecting the project.

The dichotomous choice approach consists in estimating the point of equal opportunity, i.e. the amount of the bid for which the probability of saying "yes" or "no" is 50%. In Fig. 6.1, the blue curve represents the fitted values estimated via a logit or a probit model. The inflection point, depicted in black, yields the average willingness to pay, i.e. what we aim to measure.

To illustrate the contingent valuation method, let us consider a very simple example made of $n = 20$ observations. The dataset is provided in Table 6.1. The first column results from the anonymization of individual respondents by random indexing. The second column gives the individual willingness to pay obtained with a closed-ended question. From the second column, the average willingness to pay can be computed using the *AVERAGE* function in Excel, which yields $AWTP = 22.80$. The confidence interval is defined as:

$$\left[22.80 - 2.09 \times \frac{7.78}{\sqrt{20}}, 22.80 + 2.09 \times \frac{7.78}{\sqrt{20}}\right] \approx 22.80 \pm 3.64$$

where $t = 2.09$ can be obtained using the *TINV* Excel function for a 5% risk probability and $n - 1 = 19$ degrees of freedom. Sample standard deviation $s = 7.78$ is obtained using the Excel function *STDEV*. Assume now that the target population is $N = 300{,}000$. The total willingness to pay (*TWTP*) of that population would amount to:

$$TWTP \approx (22.80 \times 300{,}000) \pm (3.64 \times 300{,}000) \approx 6{,}840{,}000 \pm 1{,}092{,}000$$

Table 6.1 Contingent valuation: example 1

Respondent	Closed-ended question	Dichotomous question		Socio-economic characteristics	
$i = 1\ldots20$	w_i	*Bid value* b_i	y_i	Income	Gender
1	13	5	1	1750	0
2	24	10	1	1750	0
3	27	15	1	2750	0
4	19	20	0	750	1
5	33	25	1	2750	0
6	39	30	1	1750	0
7	16	35	0	750	1
8	19	40	0	1250	1
9	18	45	0	750	1
10	21	50	0	750	0
11	18	5	1	1750	1
12	13	10	1	1750	1
13	32	15	1	2750	1
14	17	20	0	1750	0
15	31	25	1	2750	1
16	11	30	0	750	0
17	22	35	0	1250	0
18	33	40	0	1250	0
19	27	45	0	3250	0
20	23	50	0	1250	1
Mean	22.80			1675	0.45

These values can be included directly in a cost-benefit appraisal.

The third and fourth columns of Table 6.1 retrace the responses obtained with the dichotomous choice method. The bid value ranges from 5 to 50. To estimate the average willingness to pay, a logit model can be employed:

$$\ln\left(\frac{\Pr(y_i = 1)}{\Pr(y_i = 0)}\right) = \alpha_0 + \alpha_1 b_i$$

Solving for the value of b_i that yields $\Pr(y_i = 1) = \Pr(y_i = 0) = 1/2$, the left-hand side of the equation amounts to $\ln(1) = 0$. Thus, we have:

$$AWTP = -\frac{\alpha_0}{\alpha_1}$$

The average willingness to pay is simply given by the ratio of the estimated coefficients. Obtaining the confidence interval of this ratio is however not straightforward. The use of either the delta method or the bootstrap method is necessary. The delta method amounts to a linear approximation of the relationship between the

```
> D=read.table("C:\\data-CV.csv",head=TRUE, sep=";")
> reg=glm(y~Bid.value,family=binomial,data=D)
> summary(reg)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.95089    2.22855   2.222   0.0263 *
Bid.value   -0.19832    0.08403  -2.360   0.0183 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> AWTP=(-reg$coef[1]/reg$coef[2])
> AWTP
  24.96457
> plot(y~Bid.value,D)
> curve(predict(reg,data.frame(Bid.value=x),type="resp"),add=TRUE)

> library(car)
> deltaMethod(reg,"-Intercept/Bid.value")
                      Estimate       SE
-Intercept/Bid.value 24.96457 3.578421

> myfunction=function(data,i){data2=data[i,]
+ reg=glm(y~Bid.value,family=binomial,data2)
+ -reg$coef[1]/reg$coef[2]}

> library(boot)
> myboot=boot(D,myfunction,R=10000)
> quantile(myboot$t,c(.025, .975))
    2.5%    97.5%
17.47853 32.45224
```

Fig. 6.2 Contingent valuation with R-CRAN: example 1

coefficients and then computes the variance for large sample inference. The bootstrap method uses instead the sample as a surrogate population, and artificially creates a large number of subsamples (known as bootstrap samples) for the purpose of approximating the sampling distribution. The bootstrap subsamples are generally of the same size as the initial sample, and created with replacement. For large samples, the delta and bootstrap methods coincide asymptotically.

At this point, using Excel is rather difficult. Figure 6.2 provides the codes to be used in R-CRAN and the corresponding outcomes. The final name of the dataset is D. The $read.table$ command is used to upload the data file. Primary data is stored in disc C: in the form of an Excel comma separated values file (with suffix .csv), as is confirmed by the $sep = ";"$ command. The $head = TRUE$ command identifies column names. The glm function is used to regress y on the variable $Bid.value$. As can be seen from the estimation results (only outcomes directly used in the analysis are displayed), we obtain a significant and negative impact: the higher the bid value, the lower the probability is to accept the project. The average willingness to pay is obtained by dividing the intercept by the second coefficient, i.e. $AWTP = -(4.95089/-0.19832) = 24.96$. The $plot$ and $curve$ commands allow to check this value by drawing both the observations and the fitted values on a same graph (Fig. 6.3). As can be observed, the inflection point is indeed around 25.

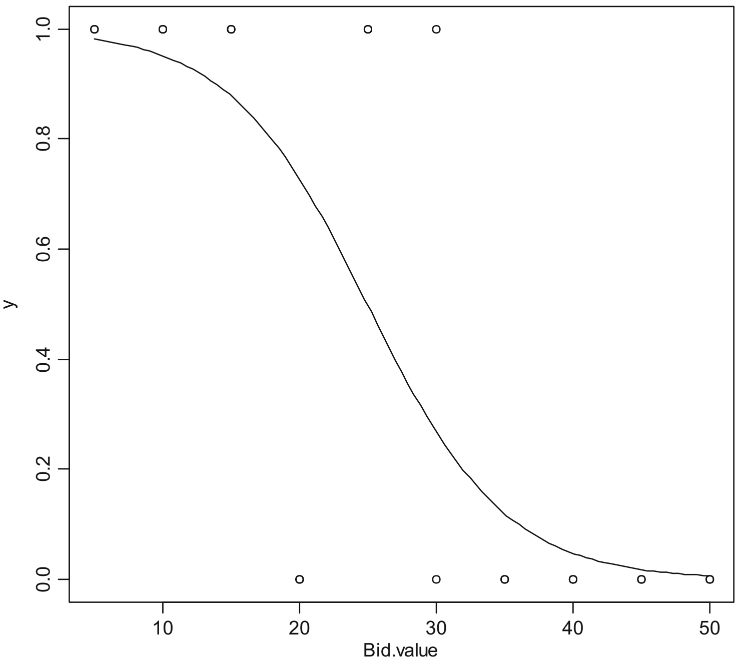


Fig. 6.3 The dichotomous choice method: example 1

The function *deltaMethod* available with the package *car* can be used to approximate the standard error *se* of the average willingness to pay. Using information from Figure 6.2, the confidence interval amounts to:

$$[AWTP - 1.96 \times 3.578, AWTP + 1.96 \times 3.578] = 24.96 \pm 7.01$$

In our case, given the small number of observations, bootstrapping is however more appropriate. First, we need to specify the function that we would like to estimate. In Fig. 6.2, *myfunction* is defined by two entries: a database and a random index for the bootstrap sample. In this function, a subsample denoted *data*2 is created and is used for estimating the relationship between $Bid.value$ and y. The average willingness to pay is defined as the ratio of the estimated coefficients ($-reg\$coef[1]/reg\$coef[2]$). Once the function is created, the *boot* command (from the *boot* package) is used to compute the confidence intervals. The *boot* command uses both the original dataset D and the function *myfunction* to generate randomly $R = 10{,}000$ bootstrap samples. The computation is based on the quantile method where *myboot*$t denotes the vector of bootstrap statistics, i.e. the vector containing the estimated values of *AWTP* for each bootstrap sample. The bootstrapping generates here the confidence interval [17.48, 32.45]. Due to the random nature of the *boot* process, each bootstrapping is likely to yield slightly different results. Finally, due to the small size of our numerical example, the bootstrap procedure may generate R-CRAN error messages that do not affect the results nor the methodology.

The contingent valuation survey should finally end with questions about socio-demographic characteristics, such as the age and gender of the respondent, his/her income or level of education. Behavioral questions (what do you do? where? when?), opinion questions (what do you think about?) and motives questions (why?) can also provide useful information in order to interpret the willingness to pay. More importantly, the socio-demographic variables (gathered in vector $CONTROLS_i$) can be used ex post to correct sample representativeness. In the context of closed-ended questions, we may estimate for instance:

$$w_i = f(CONTROLS_i)$$

Such an equation, sometimes referred to as a valuation function, may have different functional forms *f*. For instance, the endogenous variable does not necessarily need to be a linear function of its arguments.

The valuation function relates the respondent's answer to his/her socioeconomic characteristics. Consider for instance the last two columns of Table 6.1. Variables *Gender* (coded as 0 for female and 1 for male) and *Income* can be used to predict the individual willingness to pay. Let us assume a linear valuation function:

$$w_i = \alpha_0 + \alpha_1 Income_i + \alpha_2\, Gender_i$$

Once the coefficients α_0, α_1 and α_2 are determined, one can use the estimation results and the true population's characteristics (denoted *MeanIncome* and *MeanGender* afterwards) to predict the average willingness to pay:

$$AWTP = \alpha_0 + \alpha_1 MeanIncome + \alpha_2\, MeanGender$$

Figure 6.4 provides the corresponding coding. The first step consists in estimating the model via OLS using the *lm* command. As can be seen from the regression summary, only *Income* significantly affects willingness to pay. The greater is this variable, the higher is the willingness to pay. In contrast, *Gender* has no significant impact.

Assume now that the sample is not representative of the target population and that both income and share of women are overvalued in our sample. This would mean that the mean sample value of the willingness to pay ($AWTP = 22.88$) does not provide a good estimation of the population's average willingness to pay. Suppose for instance that the mean income and the share of men in the population amounts to 1500 (instead of 1675 in Table 6.1) and 51% (instead of 45%), respectively. In Fig. 6.4, a new database, denoted *E*, is constructed with these values. The *predict* function uses this new database and the coefficients obtained in the previous regression (*reg*) to predict the average willingness to pay. The confidence interval is $[8.20, 33.48]$. As expected, the average willingness to pay

```
> D=read.table("C:\\data-CV.csv",head=TRUE, sep=";")
> reg=lm(w~Income+Gender,D)
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.102043   3.800270   2.921  0.00952 **
Income       0.006511   0.001666   3.909  0.00113 **
Gender      -0.049938   2.746047  -0.018  0.98570
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Income=1500
> Gender=0.51
> E=data.frame(Income, Gender)
> predict(reg,E,interval="prediction")
      fit      lwr      upr
1 20.8436 8.204579 33.48262
```

Fig. 6.4 Valuation function with R-CRAN: example 1

(*fit* = 20.84) is lower with population characteristics than with sample characteristics.

A similar approach can be used under the framework of dichotomous questions (the NOOA method) by directly including population's characteristics in the logit estimation:

$$\ln\left(\frac{\Pr\{y_i=1\}}{\Pr\{y_i=0\}}\right)=\alpha_0+\alpha_1 b_i+\alpha_2 Income+\alpha_3\ Gender$$

Focusing again on the point of equal opportunity and using the population's characteristics yields:

$$AWTP=-\frac{\alpha_0+\alpha_2 MeanIncome+\alpha_3\ MeanGender}{\alpha_1}$$

The method to compute the confidence interval is similar to what has been done previously, i.e. bootstrapping, except that now the mean income (1500), the share of male in the population (0.51), and their coefficients are included in *myfunction*. The code in R-CRAN is detailed in Fig. 6.5. The confidence interval in that case is $[10.97, 31.51]$.

As it may result in several biases, the contingent valuation method has been intensively criticized. In particular, the way in which the scenario is presented to the respondents may strongly influence their responses. Since the method does not involve real cash transactions, the respondents may also overstate their true preferences. There are however several ways to deal with these biases. For instance, the questions can also be rephrased by asking the respondents to report their willingness to pay to avoid the loss of the good, their willingness to abstain from an improvement in the quality of the good, or their willingness to accept a worsening in the quality of the good. Moreover, one or several follow-up questions can be included to improve the precision of the dichotomous choice method. A

```
> D=read.table("C:\\data-CV.csv",head=TRUE, sep=";")
> myfunction=function(data,i){
+ data2=data[i,]
+ reg=glm(y~Bid.value+Income+Gender,family=binomial,data2)
+ -(reg$coef[1]+reg$coef[3]*1500+reg$coef[4]*0.51)/reg$coef[2]}

> library(boot)
> myboot=boot(D,myfunction,R=10000)
> quantile(myboot$t, c(.025, .975))
    2.5%    97.5%
10.97516 31.51289
```

Fig. 6.5 The NOAA method with R-CRAN: example 1

higher bid value is proposed to those in favor of the project, and a lower amount for those against it. This bidding game has however been also criticized as being subject to a starting point bias. The initial bid provides a reference point for the undecided respondent who may regard the proposed amount as an approximation of the true cost of the project. Despites its drawbacks, contingent valuation still remains the simplest approach to elicit the willingness to pay of individuals.

6.3 Discrete Choice Experiment

Choice modeling methods, among which discrete choice experiment, involve the construction of a hypothetical market and, as such, resembles a market research survey. The approach is typically used when several projects with multiple characteristics are evaluated. Contrary to contingent valuation, the purpose is not directly to elicit an individual willingness to pay, but instead to ask respondents to state a preference over a set of public goods or services using multiple scenarios. The questionnaire generally starts with a detailed description of the context such as why the survey is carried out, what are the study area, the status quo, and management issues. It is also reminded that the objective of the survey is to determine the citizens' preference and that the results will be used to design future public policies. The survey then provides the respondents with information about the policy options, within the time framework relevant to the decision context. Last, respondents are asked to evaluate those policy options.

In contrast to what is practiced in contingent valuation, the questionnaire also often incorporates the cost and payment vehicle of the strategy for the users themselves. Each respondent thus mentally associates the utility derived from each option with the expense he or she will bear. Socio-demographic questions can also be included to provide a better understanding of the results or to correct a potential representativeness bias.

Formally, each option S_j, $j = 1 \ldots J$ combines attributes a_k, $k = 1 \ldots K$ by assigning them levels a_{kj}. Each attribute a_k has a number of levels l_k. These levels must comprehend the most relevant aspects of the decision problem but should also be kept to a tractable number. Concision and comprehensiveness may not be so easy to conciliate. Systematic literature reviews and focus groups are useful in this matter. In practice, the number of attributes ranges from three to ten. The levels can be expressed in qualitative or quantitative terms and their number must also be kept small, usually from two to six.

To illustrate the discrete choice approach, let us consider a public health project dealing with the management of a chronic disease (e.g., diabetes or chronic heart failure) that requires intermittent but regular care. Assume that the standard management program is hospital-based. As an alternative, decision-makers may wish to consider home-based patient's care. Eliciting patients' preferences for one solution or the other is of paramount significance since they will influence adherence to the management program. The discrete choice experiment makes it possible by identifying the factors influencing patients' choice of a care pathway. An illustrative choice set in the context of chronic disease management is given in Table 6.2. It shows a choice card comparing two policy options that each provide different combinations of attribute levels. Patient's copayment consists of the share of total cost not covered by social or private insurance. Then follow attributes describing care procedures as well as access to information sessions (e.g., with dieticians and other people with the same disease). For instance, level "Same nurse" means that if the respondent chooses option 2, he or she will always get healthcare from the same person. The choice card should evidence a trade-off between the various attributes of care delivery. The final row of Table 6.2 is for decision between the two options.

The economic foundations of discrete choice experiment lie in random utility theory, in which utility has a deterministic and a probabilistic component. They are also rooted in Lancaster's theory of consumer behavior, by which goods and services take their value from their characteristics (you do not buy a car but rather its color, image, engine, available space, etc.). The utility function is thus multi-attribute. Let $u_i(S_j)$ denote the utility derived by individual i from option S_j. The level of satisfaction is assumed to depend linearly on attributes $a_{1j}, \ldots, a_{kj}, \ldots a_{Kj}$ that characterize option S_j:

$$u_i(S_j) = \beta_0 + \beta_1 a_{1j} + \ldots + \beta_k a_{kj} + \ldots + \beta_K a_{Kj}$$

Let y_i represent the choice of individual i among the J options of a given choice card. The probability that i will choose option j is:

Table 6.2 Choice card: example 2

Attributes a_k	Option 1 Hospital-based care	Option 2 Home-based care
Patient co-payment	\$160	\$50
How often you see the nurse	On a daily basis	On a weekly basis
Continuity of contact	Different nurse	Same nurse
Emergency care	Hospital emergency services	Standard emergency services
Access to group education class	No	Yes
Which option would you prefer?	□	□

$$\Pr\{y_i = j\} = \Pr\{\max(u_i(S_1), \ldots, u_i(S_J)) = u_i(S_j)\}$$

The choice of option j thus depends on the characteristics of the competing options, which leads to a conditional logit model. By estimating the model, we obtain a value for each coefficient $\beta_0, \beta_1, \ldots, \beta_K$.

Assume now that a_1 is the cost attribute. To obtain the marginal willingness to pay for attribute a_k ($k \neq 1$) one needs to compute the marginal rate of substitution of a_1 for a_k ($MRS_{a_1 \to a_k}$). By definition, this is the amount of money the individual is willing to give up in exchange for one extra unit of a_k while maintaining the same level of utility. For example, assume that attributes $a_3 \ldots a_K$ are held constant. With respect to a_2, we have $du = \beta_1 da_1 + \beta_2 da_2 = 0$, which yields:

$$MRS_{a_1 \to a_2} = \frac{da_1}{da_2} = -\frac{\beta_2}{\beta_1}$$

This ratio represents the implicit price of attribute a_2. More generally, we have:

$$MRS_{a_1 \to a_{k \neq 1}} = \frac{da_1}{da_k} = -\frac{\beta_k}{\beta_1}, k = 2 \ldots K$$

The choice experiment provides estimates of those marginal rates of substitution and allows to compare the relative satisfaction derived from the various attributes of the project.

We now move on to the practical implementation of a choice experiment. The first step in the implementation process is to design a tractable choice framework. The number of possible scenarios may increase exponentially with the number of attributes and their levels, generating high costs of survey administration as well as a heavy cognitive burden on respondents who would be faced with long and complex interviews. The set of all possible combinations of the levels of the attributes is labeled a full factorial design. For instance, assume that there are five attributes a_k with four levels each ($l_k = 4$ for $k = 1 \ldots 5$). This yields $\prod_{k=1}^{K=5} l_k = 4^5 = 1024$ hypothetical options. Thus, a major issue with respect to choice modeling is the conception of an experimental design that will balance informational content (a number of options sufficient enough to provide reliable data) and tractability of the interview process and data management.

In practice, it is convenient to use an algorithm that generates a fractional design, namely a subset of the full factorial design. An orthogonal fractional design is such that the attributes are uncorrelated. Consider for instance a full factorial design for three attributes $\{a_1, a_2, a_3\}$ with three levels $\{1, 2, 3\}$ each ($l_k = 3$ for $k = 1 \ldots 3$). This yields $\prod_{k=1}^{K=3} l_k = 3^3 = 27$ possible options:

$$(1,1,1); (2,1,1); (3,1,1); (1,2,1); (1,3,1); \ldots$$

If one had to select a subset of three options, it would be useless to choose $(1,1,1)$; $(2,2,2)$; $(3,3,3)$ as it would be impossible to isolate the effect of each attribute. An orthogonal fractional design aims instead to avoid multicollinearity between attributes, so as to minimize the loss of estimation power.

Let us now consider a more detailed illustration. Suppose that we would like to estimate the economic value of a natural park (example 3). The attributes and levels chosen to describe the options are presented in Table 6.3. For each attribute, levels are arranged in increasing order, quantitatively or qualitatively. For instance, \$15 represents the third highest level of additional annual tax for the creation of the park, out of four possible payments. Similarly for qualitative attributes, “camping inside the park, in unorganized campsite” describes the second level of the camping attribute. The full factorial design for Table 6.3 is thus $\{\{1,2,3,4\},\{1,2,3\},\{1,2\},\{1,2,3\}\}$ and it generates $\prod_{k=1}^{K=4} l_k = 4\times 3\times 2\times 3 = 72$ possible alternatives, denoted $S_1 \ldots S_{72}$ hereafter.

To reduce the cost of survey administration, assume that the respondents are allocated to four interview groups (usually referred to as blocks), each group facing four choice cards that each display two competing options. We can use R-CRAN, and the package *AlgDesign*, to optimize the choice set. In Fig. 6.6 the command *gen.factorial* generates the full factorial design. The vector $c(4,3,2,3)$ represents the number of levels l_k for each attribute a_k, i.e. four levels for the tax, three for amenities, two for facilities, and three for the camping attribute. The command *optBlock* is used to generate the fractional design. This command generates randomly, but with a minimum loss of information, the options that will form the 16 choice cards. The command *optBlock* assigns the options so as to minimize D-error, a usual criterion in the design of optimal choice experiments. The entry “~.” specifies that all variables from the design are to be used linearly. The option *center = FALSE* states that the levels are not to be centered. As can be seen, 16 choice cards are generated (termed “block” in R-CRAN although this term will be used afterwards to denote each group of individuals). In what follows, as a matter of simplicity, choice cards $B1$ to $B4$ will be allocated to the first group/block of individuals, choice cards $B5$ to $B8$ will be allocated to the second block,

Table 6.3 Attributes and levels in discrete choice modeling: example 3

Attributes a_k	Levels	No.
Additional tax per year	\$5	1
	\$10	2
	\$15	3
	\$20	4
Amenities	Basic (toilets)	1
	Medium (toilets and picnic area)	2
	High (toilets, picnic area and exercise station)	3
Recreational facilities	Basic (jogging)	1
	Medium (jogging and children playground)	2
Camping	Not inside the park	1
	Inside the park, in unorganized campsite	2
	Inside the park, in organized campsite	3

```
> library(AlgDesign)
> myfulldesign=gen.factorial(levels=c(4,3,2,3),
+ center=FALSE,varNames=c("a1","a2","a3","a4"))
> myfulldesign
   a1 a2 a3 a4
1   1  1  1  1
2   2  1  1  1
3   3  1  1  1
...
70  2  3  2  3
71  3  3  2  3
72  4  3  2  3

> myblock=optBlock(~.,myfulldesign,rep(2,16))
> myblock$Blocks
$B1               $B5               $B9               $B13
   a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4
16  4  1  2  1    2   2  1  1  1    1   1  1  1  1    21  1  3  2  1
17  1  2  2  1    72  4  3  2  3    23  3  3  2  1    57  1  3  1  3

$B2               $B6               $B10              $B14
   a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4
20  4  2  2  1    4   4  1  1  1    52  4  1  1  3    50  2  1  1  3
56  4  2  1  3    64  4  1  2  3    61  1  1  2  3    69  1  3  2  3

$B3               $B7               $B11              $B15
   a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4
3   3  1  1  1    9   1  3  1  1    24  4  3  2  1    13  1  1  2  1
62  2  1  2  3    49  1  1  1  3    63  3  1  2  3    60  4  3  1  3

$B4               $B8               $B12              $B16
   a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4       a1 a2 a3 a4
12  4  3  1  1    5   1  2  1  1    25  1  1  1  2    14  2  1  2  1
65  1  2  2  3    68  4  2  2  3    45  1  3  2  2    59  3  3  1  3
```

Fig. 6.6 Fractional design with R-CRAN: example 3

and so on. Remember that the fractional design is optimized but also generated randomly so that different sets of cards are likely to be generated each time the program is run.

Following a usual but not systematic practice, the status quo or "do-nothing" option is included in each choice card as a reference point. This anchoring of choices to the current situation allows respondents to conceive the change that could be brought in by the project under its various forms. Another advantage of including the status quo is that respondents do not feel like the public decision-maker imposes a new policy at their expense, with the only choice of variants of an inflicted project. The attributes of the status quo are coded with 0 values. The final number of options in each choice card is thus three. An example of choice card is provided in Table 6.4, where options S_{12} and S_{65} as well as the status quo are confronted. This choice card corresponds in Fig. 6.6 to the fourth choice set $B4$ of the first group of respondents.

The data structure associated with our experimental design is presented in Table 6.5. For the sake of simplicity, only twelve hypothetical individuals (variable *ind*) were interviewed, i.e. three individuals per block (variable *block*), each facing four choice cards (*card*). The choice variable is denoted *y*. The variables *tax*, *fac*, *amen* and *camp* stand for the tax, facilities, amenities and camping attributes,

Table 6.4 A choice card: example 3

Attributes a_k	Option 1: S_{12} $a_{1,12}=4$ $a_{2,12}=3$ $a_{3,12}=1$ $a_{4,12}=1$	Option 2: S_{65} $a_{1,69}=1$ $a_{2,69}=2$ $a_{3,69}=2$ $a_{4,69}=3$	Option3: status quo
Additional tax per year	$20	$5	$0
Amenities	High (toilets, picnic area and exercise station)	Medium (toilets and picnic area)	No amenities
Recreational facilities	Basic (jogging)	Medium (jogging and children playground)	No recreational facilities
Camping	Not inside the park	Inside the park, in organized campsite	No camping site
Which option would you prefer?	□	□	□

respectively. When one looks at the columns, each attribute appears three times as there are three options in each choice card (i.e. options 1, 2 or 3). Furthermore, when one reads the rows, each choice card appears three times as there are three individuals per block. For example, choice card $B4$ of Table 6.4 has been displayed in Table 6.5 (block 1). When faced with this choice card, respondents 1 to 3 have to compare option 1 (i.e S_{12}) versus option 2 (i.e. S_{65}) and option 3 (the status quo). Their answer may be different and will be recorded in column y either with 1, 2 or 3. Moving to the next block implies a new set of choice cards (e.g., numbered from $B5$ to $B8$ for block 2).

As can be seen from Table 6.5, the structure of the data is complex. There is one row for each choice situation and there are as many columns for the attributes as there are options. Table 6.6 provides a numerical example of such data. In addition to the choice experiment questions, data on the respondent's annual taxable income is collected (*income*). This format is known as a "wide" format in R-CRAN. In Fig. 6.7, to transform the data (initially stored on memory space C: in the program) in a suitable format for R-CRAN, we use the command $mlogit.data$ from the *mlogit* package. Each individual has responded to four choice cards. To take this panel dimension into account, one has to add the argument $id = "ind"$ in the formula so that the software differentiates each individual. The argument *varying* indexes the variables that are option-specific (i.e. from column five $tax1$ to column sixteen $camp3$). The command returns a dataset E in "long" format which has fewer columns than in the wide format, the caveat being that the value of income is repeated twelve times (three options by four choice cards). Two additional variables are created: *chid*, which represents the choice index (i.e. *card*), and *alt*, which denotes the index of each option in each choice set. The *head* command shows that respondent 1, whose income is 252,000, has chosen option 2 (the *TRUE* outcome

Table 6.5 Data structure for a discrete choice experiment

				Option 1				Option 2				Status quo			
Ind. ***i***	Block	Choice card	Choice *y*	a_1 tax1	a_2 amen1	a_3 fac1	a_4 camp1	a_1 tax2	a_2 amen2	a_3 fac2	a_4 camp2	a_1 tax3	a_2 amen3	a_3 fac3	a_4 camp3
i=1	1	B1										0	0	0	0
i=1	1	B2										0	0	0	0
i=1	1	B3										0	0	0	0
i=1	1	B4	?	20	3	1	1	5	2	2	3	0	0	0	0
i=2	1	B1										0	0	0	0
i=2	1	B2										0	0	0	0
i=2	1	B3										0	0	0	0
i=2	1	B4	?	20	3	1	1	5	2	2	3	0	0	0	0
i=3	1	B1										0	0	0	0
i=3	1	B2										0	0	0	0
i=3	1	B3										0	0	0	0
i=3	1	B4	?	20	3	1	1	5	2	2	3	0	0	0	0
i=4	2	B5										0	0	0	0
i=4	2	B6										0	0	0	0
i=4	2	B7										0	0	0	0
i=4	2	B8										0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i=12	4	B13										0	0	0	0
i=12	4	B14										0	0	0	0
i=12	4	B15										0	0	0	0
i=12	4	B16										0	0	0	0

Table 6.6 Data set for example 3

ind	block	card	y	tax1	amen1	fac1	camp1	tax2	amen2	fac2	camp2	tax3	amen3	fac3	camp3	income
1	1	1	2	4	1	2	1	1	2	2	1	0	0	0	0	252,000
1	1	2	1	4	2	2	1	4	2	1	3	0	0	0	0	252,000
1	1	3	2	3	1	1	1	2	1	2	3	0	0	0	0	252,000
1	1	4	2	4	3	1	1	1	2	2	3	0	0	0	0	252,000
2	1	1	2	4	1	2	1	1	2	2	1	3	0	0	0	180,000
2	1	2	1	4	2	2	1	4	2	1	3	0	0	0	0	180,000
2	1	3	2	3	1	1	1	2	1	2	3	0	0	0	0	180,000
2	1	4	2	4	3	1	1	1	2	2	3	0	0	0	0	180,000
3	1	1	1	4	1	2	1	1	2	2	1	0	0	0	0	396,000
3	1	2	1	4	2	2	1	4	2	1	3	0	0	0	0	396,000
3	1	3	2	3	1	1	1	2	1	2	3	0	0	0	0	396,000
3	1	4	2	4	3	1	1	1	2	2	3	0	0	0	0	396000
4	2	5	2	2	1	1	1	4	3	2	3	0	0	0	0	324,000
4	2	6	2	4	1	1	1	4	1	2	3	0	0	0	0	324,000
4	2	7	1	1	3	1	1	1	1	1	3	0	0	0	0	324,000
4	2	8	3	1	2	1	1	4	2	2	3	0	0	0	0	324,000
5	2	5	2	2	1	1	1	4	3	2	3	0	0	0	0	180,000
5	2	6	3	4	1	1	1	4	1	2	3	0	0	0	0	180,000
5	2	7	1	1	3	1	1	1	1	1	3	0	0	0	0	180,000
5	2	8	2	1	2	1	1	4	2	2	3	0	0	0	0	180,000
6	2	5	3	2	1	1	1	4	3	2	3	0	0	0	0	252,000
6	2	6	2	4	1	1	1	4	1	2	3	0	0	0	0	252,000
6	2	7	1	1	3	1	1	1	1	1	3	0	0	0	0	252,000
6	2	8	3	1	2	1	1	4	2	2	3	0	0	0	0	252,000
7	3	9	2	1	1	1	1	3	3	2	1	0	0	0	0	396,000

(continued)

Table 6.6 (continued)

ind	block	card	y	tax1	amen1	fac1	camp1	tax2	amen2	fac2	camp2	tax3	amen3	fac3	camp3	income
7	3	10	2	4	1	1	3	1	1	2	3	0	0	0	0	396,000
7	3	11	1	4	3	2	1	3	1	2	3	0	0	0	0	396,000
7	3	12	2	1	1	1	2	1	3	2	2	0	0	0	0	396,000
8	3	9	2	1	1	1	1	3	3	2	1	0	0	0	0	252,000
8	3	10	2	4	1	1	3	1	1	2	3	0	0	0	0	252,000
8	3	11	1	4	3	2	1	3	1	2	3	0	0	0	0	252,000
8	3	12	2	1	1	1	2	1	3	2	2	0	0	0	0	252,000
9	3	9	3	1	1	1	1	3	3	2	1	0	0	0	0	180,000
9	3	10	3	4	1	1	3	1	1	2	3	0	0	0	0	180,000
9	3	11	3	4	3	2	1	3	1	2	3	0	0	0	0	180,000
9	3	12	2	1	1	1	2	1	3	2	2	0	0	0	0	180,000
10	4	13	1	1	3	2	1	1	3	1	3	0	0	0	0	180,000
10	4	14	2	2	1	1	3	1	3	2	3	0	0	0	0	180,000
10	4	15	1	1	1	2	1	4	3	1	3	0	0	0	0	180,000
10	4	16	3	2	1	2	1	3	3	1	3	0	0	0	0	180,000
11	4	13	1	1	3	2	1	1	3	1	3	0	0	0	0	396,000
11	4	14	2	2	1	1	3	1	3	2	3	0	0	0	0	396,000
11	4	15	1	1	1	2	1	4	3	1	3	0	0	0	0	396,000
11	4	16	2	2	1	2	1	3	3	1	3	0	0	0	0	396,000
12	4	13	1	1	3	2	1	1	3	1	3	0	0	0	0	252,000
12	4	14	2	2	1	1	3	1	3	2	3	0	0	0	0	252,000
12	4	15	2	1	1	2	1	4	3	1	3	0	0	0	0	252,000
12	4	16	2	2	1	2	1	3	3	1	3	0	0	0	0	252,000

```
> D=read.table("C:\\data-DCE.csv",head=TRUE,sep=";")
> library(mlogit)
>E=mlogit.data(D,choice="y",shape="wide",sep="",varying=5:16,id="ind")
> head(E)
    ind block card     y income alt tax amen fac camp chid
1.1   1     1    1 FALSE 252000   1   4    1   2    1    1
1.2   1     1    1  TRUE 252000   2   1    2   2    1    1
1.3   1     1    1 FALSE 252000   3   0    0   0    0    1
2.1   1     1    2  TRUE 252000   1   4    2   2    1    2
2.2   1     1    2 FALSE 252000   2   4    2   1    3    2
2.3   1     1    2 FALSE 252000   3   0    0   0    0    2

> reg=mlogit(y~tax+amen+fac+camp,E)
> summary(reg)

Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
2:(intercept) -1.86881    1.21412 -1.5392 0.1237493
3:(intercept)  6.46877    2.31864  2.7899 0.0052724 **
tax           -0.62148    0.27144 -2.2896 0.0220456 *
amen           1.60850    0.47623  3.3776 0.0007312 ***
fac            2.75535    0.78786  3.4973 0.0004700 ***
camp           1.11510    0.68574  1.6261 0.1039206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -28.983
McFadden R^2:  0.39015
Likelihood ratio test : chisq = 37.084 (p.value = 1.731e-07)

> reg=mlogit(y~tax+amen+fac+camp|income,E)
> summary(reg)

Coefficients :
                 Estimate  Std. Error t-value  Pr(>|t|)
2:(intercept) -1.2426e+00  2.1087e+00 -0.5893 0.5556848
3:(intercept)  1.1596e+01  3.5676e+00  3.2503 0.0011527 **
tax           -6.9493e-01  2.8606e-01 -2.4293 0.0151265 *
amen           1.7778e+00  5.2646e-01  3.3769 0.0007331 ***
fac            2.8741e+00  8.2476e-01  3.4848 0.0004925 ***
camp           1.3201e+00  7.4636e-01  1.7688 0.0769321 .
2:income      -3.6687e-06  6.3218e-06 -0.5803 0.5616975
3:income      -1.8566e-05  9.3958e-06 -1.9760 0.0481598 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -25.911
McFadden R^2:  0.4548
Likelihood ratio test : chisq = 43.228 (p.value = 1.0513e-07)

> library(car)
> deltaMethod(reg,"-amen/tax")
          Estimate        SE
-amen/tax 2.558261 0.8897657
> deltaMethod(reg,"-fac/tax")
         Estimate       SE
-fac/tax 4.135866 1.644336
```

Fig. 6.7 Multinomial logit regressions with R-CRAN: example 3

for y when $alt = 2$) when faced with the first choice card ($chid = 1$) while rejecting option 1 as well as option 3 (the *FALSE* outcome for y for both $alt = 1$ and $alt = 3$).

In Fig. 6.7, the data is first analyzed with a conditional logit model using the command *mlogit*. The McFadden R^2 is similar to the usual coefficient of determination in conventional analysis. The likelihood ratio test is significant meaning that we reject the hypothesis that all coefficients except the intercept are jointly zero. The coefficients of $2:(intercept)$ and $3:(intercept)$ stand for option specific effects. The impact of the tax variable is significant and negative while amenities and facilities attributes yield a significant and positive coefficient. The camping attribute is not significant. For attributes with a significant impact on the probability of choice, we calculate the marginal rates of substitution:

$$MRS_{a_{tax}\rightarrow a_{amen}} = -\frac{\alpha_{amen}}{\alpha_{tax}} = \frac{1.60850}{0.62148} \approx 2.59$$

$$MRS_{a_{tax}\rightarrow a_{fac}} = -\frac{\alpha_{fac}}{\alpha_{tax}} = \frac{2.75535}{0.62148} \approx 4.43$$

The ratio 2.59 is the marginal rate of substitution of the amenities attribute in terms of dollars. It means that to increase by one level the quality of amenities, a respondent is willing to pay at most 2.59 dollars more. The satisfaction derived from recreational facilities is found to be higher, with an implicit price approximately equal to $4.43.

To improve the explanatory power of the model, a mixed logit model is estimated by including “| *income*” in the regression formula. The goodness of fit increases to 45%. Marginal rates of substitution become:

$$MRS_{a_{tax}\rightarrow a_{amen}} = -\frac{\alpha_{amen}}{\alpha_{tax}} = \frac{1.7778}{0.69493} \approx 2.56$$

$$MRS_{a_{tax}\rightarrow a_{fac}} = -\frac{\alpha_{fac}}{\alpha_{tax}} = \frac{2.8741}{0.69493} \approx 4.14$$

The implicit price of the amenities attribute remains stable while it decreases slightly for the facilities attribute. Furthermore, the coefficient on “$3:income$“ is negative and significant which indicates that the probability of choosing the status quo decreases with income, thus highlighting a likely income effect on the demand for that particular public good. Finally, the delta method is the simplest (if not perfectly accurate due to the small size of the sample) manner to compute the confidence intervals ($1.96 \times se$). The model yields:

$$MRS_{a_{tax}\rightarrow a_{amen}} \approx 2.56 \pm 1.96 \times se \approx 2.56 \pm 1.74$$

$$MRS_{a_{tax}\rightarrow a_{fac}} \approx 4.14 \pm 1.96 \times se \approx 4.14 \pm 3.22$$

Discrete choice experiments allow to estimate a multi-characteristic valuation function that can be used to deduce the attributes of the public project that do matter and the individual marginal willingness to pay for them. It may be less prone to the “yea-saying” bias than the contingent valuation method as it gives the opportunity for respondents to choose among alternatives. Yet, the method is costly and still subject to criticisms. In particular, the cognitive burden of choice experiment surveys may strongly influence the preferences that are stated by the respondents.

6.4 Hedonic Pricing

The hedonic pricing method uses the value of a surrogate good or service to approximate the implicit price of a non-market good. Most of its applications concern environmental quality or amenities and their impact on the real estate market. In particular, the method has been often used to provide a value to improved air or water quality, or conversely to reduced noise nuisance.

Hedonic pricing is based on a set of assumptions about individuals' behavior and their location decision. In particular, the real-estate market is assumed to be competitive with freedom of access and in equilibrium. Individuals are presumed to perceive the evaluated environmental attribute and integrate it as a dimension of their location decision. That decision is the result of an optimal choice given the characteristics of each property (including the environmental attribute) and the budget constraint of individuals. Under these conditions, the places where people choose to live in should not only reflect their preference for the property itself but also their preference for the environmental good. As a consequence, through a regression analysis, house prices could be used to estimate a value of their associated environmental amenities.

Figure 6.8 illustrates the setting. Each house is characterized by its location on a one-dimensional axis of air quality z. As there should be a higher demand for better environmental quality, the house unit price $p(z)$ is (here linearly) increasing with the distance from the source of pollution ($p^{'}(z)$ is strictly positive). In the top-panel of Fig. 6.8, lower rents are paid for homes in more polluted areas. Consider now a household living in house number 2. Its marginal willingness to pay is displayed at the bottom-panel of Fig. 6.8, and is assumed to be decreasing as stated in standard consumer theory. The curve stands for the marginal satisfaction the household derives from air quality. The marginal price $p^{'}(z)$ is also displayed (it is assumed to be constant, but this is not necessarily so). It represents what the household should pay for a new location with air quality improved by one unit, i.e. by "moving next door". As can be seen, location number 2 represents the household's optimal choice as moving next door would imply an additional price higher than the benefits in terms of improved environmental quality. Observations of marginal price and corresponding level of air quality thus give the equilibrium point of each household. This is the keystone of hedonic pricing. If we consider marginal changes in environmental quality, the marginal rent should provide a good proxy for the marginal willingness to pay for a change in location. For larger changes, however, the method would provide only an upper bound (as shown in Fig. 6.8). Hedonic pricing is thus suitable mostly for measuring the impact of small changes in environmental quality.

One of the difficulties of the approach lies in the choice of a functional form for the model. With hedonic pricing, the final aim is not to provide an estimation of

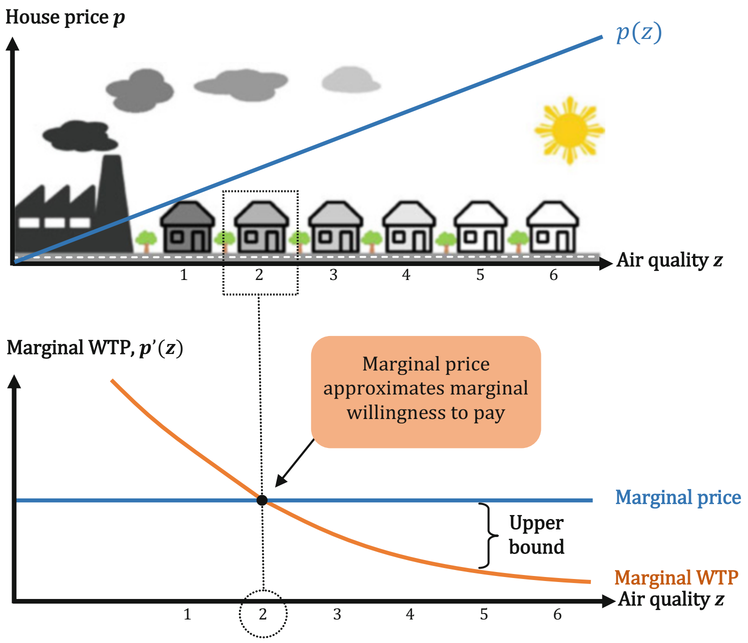


Fig. 6.8 Hedonic pricing: illustration

house price, but rather to approximate the marginal impact of the environmental attribute z on the endogenous variable (here, house price p). Consider for instance the linear approach:

$$p = CONTROLS + \beta_1 z$$

In this case, the coefficient β_1 gives directly the implicit price or hedonic price we are looking for ($\partial p/\partial z = \beta_1$). It denotes the approximate change in the price p for a unit change in the environmental attribute z. Now, should we consider instead a semi-log model, the conclusion would actually differ:

$$\ln p = CONTROLS + \beta_1 z$$

Coefficient β_1 would give the approximate percent change in the price p for a unit change in the environmental good ($\partial p/\partial z = \beta_1 e^{(CONTROLS+\beta_1 z)} = \beta_1 p$).

How can we select the best functional form? While comparing the goodness of fit (R^2) can be helpful with respect to the choice of exogenous variables, this is not the case anymore as regards the endogenous variable. Instead, one has to implement a Box-Cox test. The latter aims to find a monotonic transformation of data, using power functions, in order to improve the model fit. It consists in finding the optimal value of a parameter λ to make the data more normal distribution-like. Once the parameter is known, the endogenous variable is transformed as follows:

$$\text{Endogenous variable} = \begin{cases} \dfrac{p^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0 \\ \ln(p) \text{if } \lambda = 0 \end{cases}$$

For instance, if $\lambda = 0.5$ we can use the square root of p in the regression analysis. If $\lambda = 1$ we may use p, or $\ln(p)$ if the parameter approaches 0.

To illustrate the method let us consider a simple example, the estimation of the value of urban green space (parks and sport fields) in a given city. The data set is provided in Table 6.7 and contains thirty geo-coded apartment rents (on a year-basis), divided into five different districts of equivalent size. It provides information on one single intrinsic variable of the real estate, namely the surface area in square meters. In order to investigate how the presence of urban green spaces is capitalized in real estate prices, we also have information about the number of green spaces in each district (see last column).

Figure 6.9 provides the estimation results. While the first model ($reg1$) formulates the endogenous variable in logarithm, the second model ($reg2$) does not transform the endogenous variable. For the first regression, both the *surface* and *green . spaces* variables yield significant coefficients with the expected positive sign. An increase in surface by one square meter generates an increase in price of 0.7%. Similarly, an additional green space in a given district yields an increase in the renting price of 1.7%. Reasoning in terms of growth rate is however not convenient as the value cannot be used as such. Instead, we have to settle an initial value for the rent if we want to approximate the marginal willingness to pay for an additional green space. For instance, for the first housing unit, an additional green space in district 1 will yield:

$$\text{New rent} = \text{Old rent} \times (1 + 1.7384\%) = 5153 \times (1 + 1.7384\%) = 5242.57$$

Equivalently, this means that the marginal willingness to pay of this household is:

$$WTP = \text{Hedonic price} = \text{Old rent} \times 1.7384\% = 89.57$$

Should the public sector build a new park in district 1, then the willingness to pay of household 1 would be at most \$89.57 per year. In Fig. 6.9, we automated the results using the command $D\$hedonic\,.price = D\$rent * reg1\$coef[3]$ where $reg1\$coef$ [3] denotes the third coefficient of the regression output, i.e. 1.7384%.

The logarithmic form implies a different hedonic price for each household. As shown in Fig. 6.9, we may also provide a hedonic price for each district using the command *gsummary* available with the package *nlme*. The command allows to compute the mean ($FUN = mean$) of each variable in dataset D for each district ($groups = D\$district$). By construction, the higher is the average rent in a given district, the higher is the hedonic price. Moreover, using the average rent for the

Table 6.7 Data for hedonic pricing: example 4

Housing unit	District	Rent	Surface	Green spaces
1	1	5153	20	21
2	1	6484	30	21
3	1	4322	20	21
4	1	4308	26	21
5	1	5710	35	21
6	2	6899	60	11
7	2	8359	80	11
8	2	9469	90	11
9	2	8369	65	11
10	2	7234	70	11
11	3	16,705	79	24
12	3	10,955	35	24
13	3	14,833	78	24
14	3	9805	35	24
15	3	11,767	56	24
16	4	10,835	112	5
17	4	12,885	116	5
18	4	3638	32	5
19	4	12,510	116	5
20	4	4635	32	5
21	5	14,050	117	12
22	5	21,621	190	12
23	5	14,522	130	12
24	5	33,253	321	12
25	5	13,274	123	12
26	6	34,303	324	19
27	6	11,654	100	19
28	6	4675	23	19
29	6	3687	20	19
30	6	11,717	89	19

whole sample ($mean(D\$rent)*reg1\$coef[3]$), the model gives a mean hedonic price equal to $195.64.

The results of the linear model (*reg*2) are different as the hedonic price is now the same for each district and each household. The estimation results yield a hedonic price equal to $260.06. At this stage, it should not be necessary to point out how important the bias imposed by a misspecification in the functional form would be. The *boxcox* command from the package *MASS* may provide some insights as regards the functional form we should employ for the endogenous variable. Figure 6.10 shows that the value of λ approaches 1, giving in this case support to the linear model. Hence, should public sector authorities be willing to

```
> D=read.table("C://data-HPM.csv",head=TRUE,sep=";")
> reg1=lm(log(rent)~surface+green.spaces,D)
> summary(reg1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.269172   0.166585  49.639  < 2e-16 ***
surface      0.007038   0.000736   9.563 3.67e-10 ***
green.spaces 0.017384   0.008445   2.058   0.0493 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> D$hedonic.price=D$rent*reg1$coef[3]
> head(D$hedonic.price)
[1]  89.57821 112.71591  75.13235  74.88898  99.26093 119.93015

> library(nlme)
> gsummary(D,groups=D$district,FUN=mean)
  housing.unit district    rent surface green.spaces hedonic.price
1            3        1  5195.4    26.2           21      90.31528
2            8        2  8066.0    73.0           11     140.21693
3           13        3 12813.0    56.6           24     222.73735
4           18        4  8900.6    81.6            5     154.72536
5           23        5 19344.0   176.2           12     336.27030
6           28        6 13207.2   111.2           19     229.59000

> mean(D$rent)*reg1$coef[3]
green.spaces
    195.6425
> reg2=lm(rent~surface+green.spaces,D)
> summary(reg2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1405.191    895.729  -1.569    0.128
surface         99.146      3.957  25.054  < 2e-16 ***
green.spaces   260.062     45.408   5.727 4.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(MASS)
> boxcox(reg2)
```

Fig. 6.9 Hedonic pricing with R-CRAN: example 4

build a new park in the city, the cost per household should not exceed $260.06. The confidence interval is directly obtained from the regression results:

$$\left[MWTP - t_{\alpha/2} \times se, MWTP + t_{\alpha/2} \times se\right] \approx 260.062 \pm 93.16$$

Parameter $t_{\alpha/2} = 2.0518$ is obtained from a Student distribution table for $n - K$ degrees of freedom (where $n = 30$ is the number of housing units and $K = 3$ is the number of estimated coefficients) and a 5% significance level. The standard error ($se = 45.408$) comes directly from the estimation results for $reg2$ in Fig. 6.9.

With the recent availability of large geo-coded databases, hedonic pricing has been extensively used in the past years. However, it is suitable only for observable

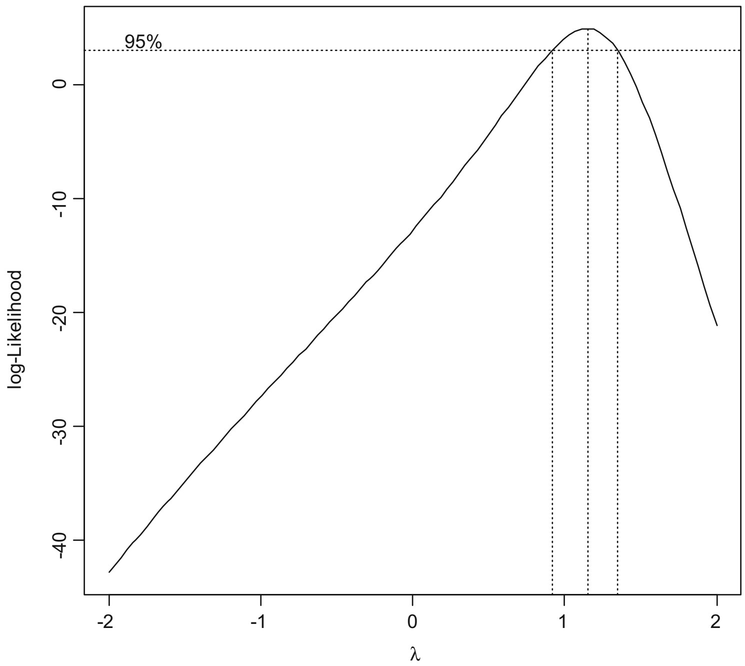


Fig. 6.10 Box-Cox test: example 4

attributes. For instance, while the taste and odor of water may be inoffensive, water turbidity may affect the housing price more significantly than the presence of harmful but invisible pollutants. Moreover, the need for a large database makes the method difficult to implement, especially as there may be multicollinearity problems if insufficient variance is observed between observations. For instance, the largest houses may be also those located far away from the source of pollution, in a wealthy neighborhood, with a low crime rate, which makes it impossible for the evaluator to extract the true effect of the environmental good. Last, real estate may be affected by external factors, like taxes or interest rates, making the results relatively complex to interpret, depending heavily on model specification, and requiring a high degree of econometric expertise. The fact that the hedonic price method is based on real choices is however an asset and, as such, it remains a key figure of monetary valuation methods.

6.5 Travel Cost Method

The travel cost method evaluates the value of recreational sites (e.g., forest, park, castle, beach, etc.) using information about the cost that people incur to visit the site in a given time period, usually a year. The basic tenet is to gather information about the distance between the recreational site and the starting point of travel, and to examine how this distance affects the cost and number of visits. The cost associated with the trip includes both (1) monetary expenses, such as gasoline, average wear and tear on car, the admission price and (2) the opportunity cost of time. The method consists in estimating a function that relates the demand for the site, measured by the number of visits, to the visit cost. Individual marginal willingness to pay is thus directly estimated as for any marketed good. There are two approaches in this respect, the zonal travel cost method and the individual travel cost method, depending on whether we deal with aggregate or individual data.

The zonal travel cost method groups respondents into zones of residence, so that visit costs increase with the average distance of each zone to the site. Data consist of information about visitors' geographical origin, e.g., their zip code or equivalent. Within each zone, travels costs to the recreational site are assumed identical. When the cost increases, we should observe a decrease in the visit rate, which is computed as the number of visits from a zone divided by the zone population. Based on these observations, the estimated demand function can be used to appraise the economic benefits resulting for instance from a reduction in the cost of access to the site, or an improvement of its recreational quality.

The individual travel cost method shares the same theoretical premises as the zonal travel cost method but uses individual records on the number of visits, distance and travel costs, plus a set of socio-economic characteristics, thereby allowing a more precise estimation of the demand curve. Whether we should use the individual or zonal travel cost method mainly depends on three factors: (1) the cost of survey administration as the individual travel cost method requires a larger set of personal information; (2) whether the site implies on average several visits per individual since the individual travel cost method requires sufficient variance among observations; (3) the zonal method requires that individuals in a zone be relatively similar in terms of travel costs: an accurate specification of zonal divisions is therefore of high importance. As for the last point, since higher distances do not necessarily imply higher travel time, both distance and time aspects have to be taken into account when specifying the zones.

To illustrate the zonal travel cost approach, let us consider the numerical example provided in Table 6.8. The area encompassing visitors' origins is divided into five zones, indexed from 1 to 5. For each zone, information on the number of visits and population size is collected for a given year. The visit rate is obtained by dividing the number of visits from each zone by the zone population. To compute the cost of one visit we need information on travel distance and travel time. Travel time is generally computed by assuming an average vehicle speed and using the round-trip distances from each zone. Coupled with data on average distance costs (e.g., a standard cost per km for operating an automobile, here \$0.8 per km) and time costs (usually the average hourly wage, which could also differ by zone, here \$0.5 per min) these variables yield the distance costs (column 7) and the time costs (column 8). If any, admission fees also have to be included to compute the visit cost.

Figure 6.11 illustrates the econometric approach. First, the two variables (*VisitCost* and *VisitRate*) are created using information from Table 6.8. The regression results obtained with the command *lm* yield the following demand function:

Table 6.8 The zonal travel cost method: example 5

Zones	Total visits	Population	Visit rate	Travel distance (km)	Travel time (min)	Distance cost $0.8/km	Time costs $0.5/min	Cost per visit
1	20,000	8000	2.50	5	5	$4	$2.5	$6.5
2	25,000	12,500	2.00	10	6	$8	$3	$11
3	50,000	31,250	1.60	15	7	$12	$3.5	$15.5
4	11,000	13,750	0.80	20	12	$16	$6	$22
5	17,000	42,500	0.40	25	19	$20	$9.5	$29.5

$$Demand \approx 3.04492 - 0.09378 \times VisitCost$$

The confidence interval can be obtained using the command *confint*. It is important to understand that the endogenous variable is the visit rate and, as such, the estimation results provide an estimated value of this item, denoted *Demand*. Using the previous equation and the real value of the cost (or equivalently the command $\$fitted.values$), we obtain the estimated demand for each zone:

$$Demand = (2.435, 2.013, 1.591, 0.981, 0.278\ )$$

The estimated demand function is displayed at Fig. 6.12. Function *plot* displays the observations while the *abline* command draws the regression line.

By definition, the average willingness to pay in a given zone is defined as the sum of two components: (1) the cost of visiting the site on average and (2) any excess amount which people in the zone would be willing to pay on average but do not have to pay, i.e. the surplus. The regression equation can be used to compute these two elements. For instance, for a visit cost of $22 (zone 4), the first component is defined as the area of the rectangle in Fig. 6.12 (area labeled “Cost”). We have:

$$Cost = VisitCost \times Demand$$

This area represents the (estimated) average spending incurred by the zonal population for visiting the recreational site. The second element relates to the net welfare the zonal population derives because it pays less than what it was willing to pay. In our case, since we are dealing with the equation of a demand curve (quantity as a function of price and not the other way around), the surplus is defined as the area of the triangle on the right hand side of Fig. 6.12 (area labeled “Surplus”). Integration can be used to calculate this amount.

In Fig. 6.11, using the command *function* we more formally specify the demand function we will analyze. The intercept and slope are respectively the estimated

```
> VisitRate=c(2.5,2,1.6,0.8,0.4)
> VisitCost=c(6.5,11,15.5,22,29.5)
> Population=c(8000,12500,31250,13750,42500)
> reg=lm(VisitRate~VisitCost)
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.044920   0.136280   22.34 0.000196 ***
VisitCost   -0.093782   0.007268  -12.90 0.001005 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(reg, level = 0.95)
                 2.5 %      97.5 %
(Intercept)  2.6112166  3.47862269
VisitCost   -0.1169135 -0.07065098

> reg$fitted.values
        1         2         3         4         5
2.4353352 2.0133151 1.5912951 0.9817106 0.2783439

> plot(VisitRate~VisitCost,xlim=c(0,35),ylim=c(0,2.5))
> abline(reg)
> abline(h=0,lty=3)

> Demand=function(x){as.numeric(reg$coef[1]+reg$coef[2]*x)}
> Demand(22)
[1] 0.9817106

> # Intercept on the x-axis (cost axis)
> maxcost=uniroot(Demand,c(0,35))$root
> maxcost
[1] 32.46798

> # Surplus per capita (for cost $22)
> S_22=integrate(Demand,22,maxcost)$value
> S_22
[1] 5.138265
> # Total surplus
> TS_22=S_22*Population[4]
> TS_22
[1] 70651.14
> # Total Cost
> TC_22=22*Demand(22)*Population[4]
> TC_22
[1] 296967.5

> # Surplus per capita (for cost $20)
> Demand(20)
[1] 1.169275
> S_20=integrate(Demand,20,maxcost)$value
> S_20
[1] 7.28925
> # Total surplus
> TS_20=S_20*Population[4]
> TS_20
[1] 100227.2
> # Total Cost
> TC_20=20*Demand(20)*Population[4]
> TC_20
[1] 321550.7

> # Net total surplus
> TS_20-TS_22
[1] 29576.05
```

Fig. 6.11 Zonal travel cost method with R-CRAN: example 5

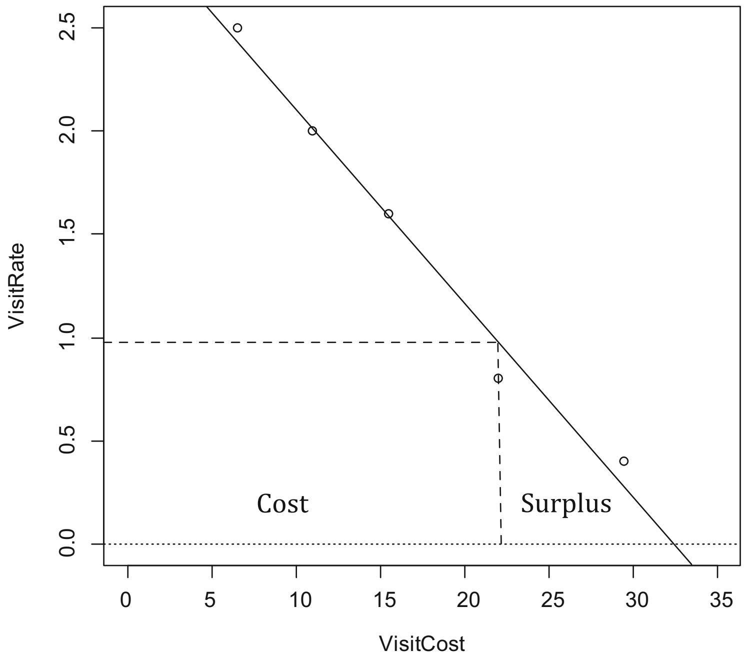


Fig. 6.12 Demand function: example 5

coefficients *reg*$*coef*[1] and *reg*$*coef*[2]. Consider zone 4. For a visit cost of $22, the estimated demand (visit rate) is 0.981:

$$Demand(22) \approx 3.04492 - 0.09378 \times 22 \approx 0.981$$

The corresponding surplus is the area under the demand curve at the bottom right of Fig. 6.12 from a visit cost of $22 to the x-axis intercept (obtained with the *uniroot* command). Since we are dealing with a rate (number of visits divided by the population size), this surplus represent the "average" surplus of the zone. Integrating the demand curve (using *integrate*) over those bounds yields a per capita surplus ($5.138) that is to be multiplied by the population of zone 4 in order to obtain a total surplus of $70,651. The estimated total cost they incur ($296,967) is obtained from the multiplication of the average cost (*VisitCost* × *Demand*(22)) with the population of zone 4 (*Population*[4]).

Consider now a new road project that could improve access to the recreational site by reducing the visit cost of zone 4 to $20. Demand increases to 1.169 (compared to the previous visit rate of 0.981 for a cost of $22):

$$Demand(20) \approx 3.04492 - 0.09378 \times 20 \approx 1.169$$

The per capita surplus is found to be \$7.289 (instead of \$5.138). Surplus for zone 4 is now \$100,227 for a total cost of \$321,550. The net surplus for zone 4 that could be generated by the new road is then obtained by comparing this surplus with the one obtained previously:

$$\textit{Net surplus} = \$100,227 - \$70,651 = \$29,576$$

Should the road project cost more than \$29,576 per year in total, then inhabitants from zone 4 would not get a net benefit from it. These results can of course be replicated to the other zones.

The travel cost method rests on the empirical techniques used by economists to estimate market good values. Given its simplicity and the low costs of survey administration, the travel cost method represents a very useful tool for monetizing non-market goods. Several assumptions are however made which may weaken the approach. In particular, people are assumed to react in a similar manner to distance costs, travel costs and admission fees. The measure of the opportunity cost of time is also subject to controversy, especially if people enjoy the travel time itself, which would yield an overestimation of the costs. Moreover, while hedonic pricing assumes that people choose their place of residency according to the different attributes of the competing locations, the travel cost method makes the opposite assumption. Should we relax the no-mobility hypothesis, the travel cost method would actually fail in estimating the demand for the recreational site. If individuals change their place of residency so as to live closer to the recreational site, the price of a trip would actually become endogenous. Those with the lowest travel costs would also be those with the highest preference for the site. Last, there exists a random-utility version of the method assuming that individuals will choose the recreational site they prefer out of all possible sites. This approach requires however information on all potential sites that a visitor might select, their quality characteristics, the travel costs to each site, thus making the method difficult to implement in practice.

6.6 Health-Related Quality of Life

One important issue when assessing the benefits of a health program is whether the valuation of outcomes should comprehend the viewpoint of the patient. By means of illustration, let us take the case of a public health program for cancer screening (e.g., colorectal cancer screening of men and women from 50 to 75 years old). Any decision concerning the allocation of healthcare resources involves demonstrating the performance of the competing strategies (e.g., no screening, standard screening, and innovative screening, more or less repeated). The decision-maker may for instance rely on quantitative measures such as survival rates. For example, a cohort and cost-effectiveness model may show that the most cost-effective strategy (e.g., innovative screening) yields, on average, a gain of five life-years. This alternative whose outcome is *a priori* most desirable would thereby be selected. Yet, from the patients' standpoint, the survival rate does not encompass all the dimensions of the health program. The strategies in competition may bring in additional benefits and harms (notwithstanding costs) that are important as well. Examples of harms are the side effects and painful compliance with the screening process. Another instance is the case of chronic diseases: patients may live a life in a state far worse than what they could define as perfect health. Those are attributes that can be accounted for by the evaluator.

With a single and quantitative measure of outcome (e.g., lives saved, life-years lived or events avoided), there is an explicit or underlying one-dimensional scale, which usually takes the form of an arithmetic scale (e.g., number of cancers avoided during the implementation of the program) or that of a time scale (e.g., total quantity of life years lived by a cohort over a given time horizon). Yet, in health, outcomes can also have qualitative attributes relating to quality of life. Life expectancy is one thing, quality adjusted survival is another and crucial one. This is why QALYs (for Quality Adjusted Life Years) have become a standard outcome measure in the cost-effectiveness analysis of health programs. The QALY indicator provides a single scale measure of both quantity and quality of life. Outcomes are thus valued through the elicitation of individual preferences with respect to the various health states individuals or patients may be confronted with, either "potentially" if the elicitation framing involves people from the general population, or "actually" if it directly considers patients under the assessed condition.

Figure 6.13 provides an illustration of an intervention program (the blue curve) that would replace a strategy of status quo (the orange curve). During the first phase, quality of life deteriorates, for instance because of a heavy treatment involving side-effects. The second phase sees an improvement, for instance associated with the recovery process. The last time sequence involves quantitative (decrease in mortality) and qualitative (decrease in morbidity) gains. As can be understood from this example, the first step in the valuation of the outcomes of a health program is to measure individual preferences for each health state at the time they occur. To do so, a multi-attribute utility function is used to model the preferences of patients.

Formally, patients are characterized by a vector of physical and psychological attributes meant to comprehend the aspects of individual well-being relevant to the assessed health intervention. A health state $H_j = (a_{j1}, a_{j2}, \ldots, a_{jk}, \ldots, a_{jK})$, $j = 1 \ldots J$, is described through K attributes where a_{jk} is the level of attribute k in state H_j, $k = 1 \ldots K$. For instance, if attribute k is the ability to walk and j is the health state associated with a hip surgery recovery phase, a_{jk} will describe the corresponding walking speed. Each attribute k has a number of levels l_k (e.g., five walking speeds, from very low to normal). Each health state is thus represented as a K-dimensional vector of attributes. The keystone of health-related quality of life methods is to measure the utility level associated with each of these multi-attribute health states.

Among the empirical methods for measuring preferences, the most frequent are the standard gamble, the time trade-off and more recently discrete choice experiments. Basically speaking, the respondent, a patient or an individual from the general population, is faced with a description of the health states that includes

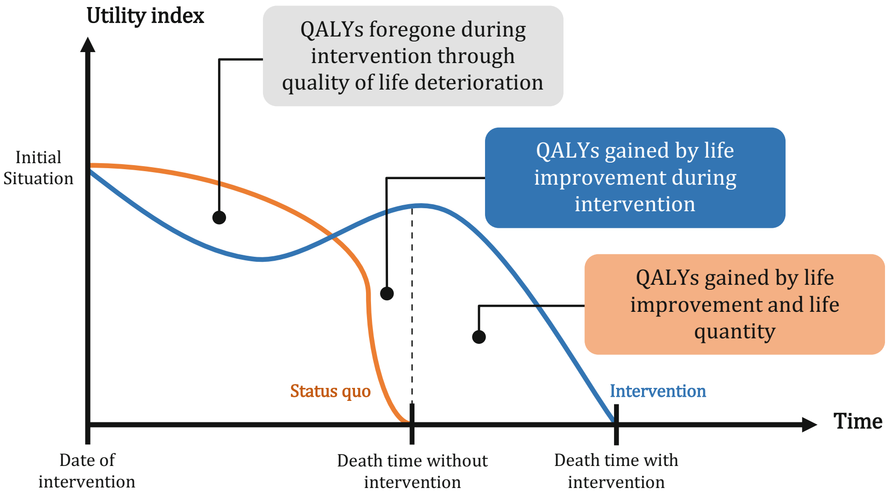

Utility index
QALYs foregone during intervention through quality of life deterioration
Initial Situation
QALYs gained by life improvement during intervention
QALYs gained by life improvement and life quantity
Status quo
Intervention
Time
Date of intervention
Death time without intervention
Death time with intervention

Fig. 6.13 QALYs gained from an intervention

clinical considerations (e.g., post-surgery status), side-effects (e.g., nauseas) and functional consequences (e.g., patient confined to bed). The interview technique is meant to elicit individual preferences with regard to those health states.

The standard gamble method rests on the Von Neumann and Morgenstern expected utility function, denoted EU hereafter. First, let $u = u(H_j)$ denote the utility a respondent derives when facing state H_j, $j = 1 \ldots J$. The higher is the utility u for a particular state, the higher is the preference for that state. For the evaluation of health state H_j, the respondent is confronted with the following choice frame: either choose the certain lottery $A = \{H_j; 1\}$, which amounts to remaining in H_j, or pick the risky lottery $B = \{H_+,H_-; p_j, 1 - p_j\}$ where he or she may end up in perfect health (H_+) with probability p_j, or face death (H_-) with probability $(1 - p_j)$. For simplicity of exposition, we assume that $u(H\_) < u(H_j) < u(H_+)$. The expected utility EU for lottery A and lottery B is respectively given by:

$$EU(A) = u(H_j)$$

$$EU(B) = p_j \times u(H_+) + (1 - p_j) \times u(H_-)$$

If we set $u(H\_) = 0$ and $u(H_+) = 1$ then $EU(B) = p_j$. Thus, at the point of indifference between the two lotteries, we get $u(H_j) = p_j$.

Figure 6.14 displays the corresponding decision tree. The square is a choice node, the circle is a chance node. For example, the respondent is questioned on state H_j and is asked "For which value of p_j are you indifferent between accepting the risky gamble and remaining with certainty in health state H_j?" or "From which value of p_j are you willing to accept the risky gamble?" If for a state H_1, the

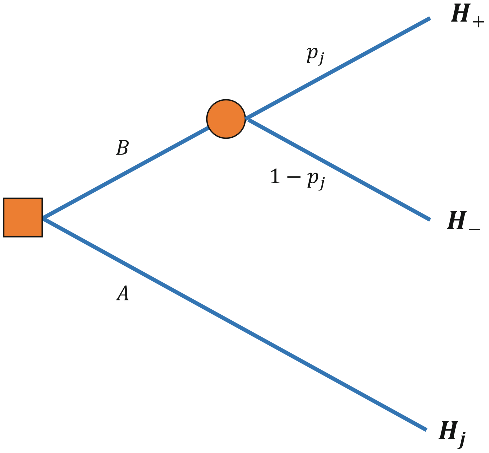


Fig. 6.14 A standard gamble

respondent declares $p_1 = 0.90 = u(H_1)$, it means that to give up the certainty of state H_1, the individual requires as much as a 90% probability that the risky gamble will be favorable. State S_1 is highly valued. If on the contrary, for a state H_2, the answer is $p_2 = 0.10 = u(H_2)$, then to give up the certainty of state H_2, the individual does not require more than a 10% probability that the risky gamble will be favorable. State H_2 is poorly valued.

An alternative preference measurement is the time trade-off method. The patient is confronted with two options. The first one is to live in state H_j for T years (the life expectancy in this health state) followed by death. The second option is to live in state H_+ for $t < T$ years, followed by death. The interviewer proposes to vary time t from T to 0 until the patient is indifferent between the two options. The utility associated with health state H_j is then $u(H_j) = t/T$. Figure 6.15 exemplifies the approach.

Finally, discrete choice experiment techniques can be used to elicit preferences for health states, as shown in Table 6.9 (for the attribute framework) and Table 6.11 (for an example). Attributes must be chosen carefully in order to comprehensively, precisely, and yet simply depict the condition under scrutiny.

In practice, if measuring preferences appears too costly and complex whenever a new health question is tackled, one may also rely on pre-existing health status classification systems. Among the many available systems, popular ones are for instance (1) the EQ-5D-5L from the EuroQoL Group, based on the time trade-off and discrete choice experiment methods with additive utility, or (2) the Health Utility Index based on standard gamble with multiplicative utility. For sake of simplicity, we focus here only on the EQ-5D-5L classification system.

The EQ-5D-5L questionnaire (fully available from the EuroQol Research Foundation) is widely used in clinical trials and recommended as well by a vast majority of national or international health technology assessment agencies. The questionnaire provides a generic patient-reported outcome. It has five dimensions or attributes which in turn have five qualitative levels of increasing degrees of severity

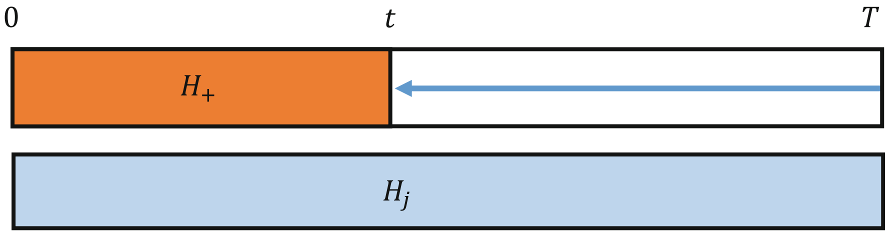


Fig. 6.15 Time trade-off

Table 6.9 Choice card for health state evaluation

Attributes	Health state S_1	Health state S_2
1	a_{11}	a_{21}
. . .	. . .	. . .
k	a_{1k}	a_{2k}
. . .	. . .	. . .
K	a_{1K}	a_{2K}
Which option is better?	□	□

(Table 6.10). Respondents are asked to indicate the extent to which they face problems on the following generic dimensions of health: mobility; self-care; usual activities (e.g., work, study, housework, and family or leisure activities); pain/discomfort; anxiety/depression. The level descriptors (no, slight, moderate, severe, extreme/unable) are respectively captured by values 1, 2, 3, 4, 5. A health state for instance characterized by a situation of severe problems in walking about, moderate problems with self-care, slight problems doing usual activities, extreme pain or discomfort, no anxiety or depression, is coded 43251.

The initial EQ-5D questionnaire contained only three levels of severity, which nevertheless generated 243 (3^5) health states. With five levels of severity, the number of health states reaches 3125 (5^5). Neither system could afford a full investigation of all situations through interviews. Consequently, and without going into details that may change over time and along ongoing methodological improvements, respondents are randomly selected from the general population of the country carrying the study, usually a sample from 400 to 1000 individuals. Around 100 health states are valued. A standard protocol is that respondents are asked to evaluate 10 EQ-5D-5L health states by time trade-off. Afterwards, they are instructed to carry discrete choice valuation of 10 pairs of EQ-5D-5L health states. Table 6.11 provides such an example with state 13345 and state 52454. Note that this discrete choice setting does not involve a monetary attribute.

Regression methods (e.g., multinomial regressions) are then used to estimate values for the whole set of health states. For instance, Table 6.12 shows the central estimates for England with 2016 Office of Health Economics data and uses them for the valuation of utility decrements associated with health states 13345 and 52454 (please note that not all statistical corrections are reported here, so that the results

Table 6.10 The EQ-5D-5L descriptive system

Attributes/Dimensions	Level	
Mobility		
	I have no problem in walking about	□
	I have slight problems in walking about	□
	I have moderate problems in walking about	□
	I have severe problems in walking about	□
	I am unable to walk about	□
Self-care		
	I have no problem washing or dressing myself	□
	I have slight problems washing or dressing myself	□
	I have moderate problems washing or dressing myself	□
	I have severe problems washing or dressing myself	□
	I am unable to wash or dress myself	□
Usual activities		
	I have no problem doing my usual activities	□
	I have slight problems doing my usual activities	□
	I have moderate problems doing my usual activities	□
	I have severe problems doing my usual activities	□
	I am unable to do my usual activities	□
Pain/discomfort		
	I have no pain or discomfort	□
	I have slight pain or discomfort	□
	I have moderate pain or discomfort	□
	I have severe pain or discomfort	□
	I have extreme pain or discomfort	□
Anxiety/depression		
	I am not anxious or depressed	□
	I am slightly anxious or depressed	□
	I am moderately anxious or depressed	□
	I am severely anxious or depressed	□
	I am extremely anxious or depressed	□

are approximations). The severity of the condition in health state 52454 is such that it gets a negative utility value (see last row of Table 6.12).

It should be stressed that the EQ-5D-5L questionnaire is conceived to be applicable for any health condition so that they can be used across different patient populations and diseases. In some cases, however, one must distinguish between condition-specific measures and generic measures. With the former, extra questionnaires must be used to assess specific clinical impacts (e.g., joint laxity in the hip surgery example).

Once each health state has been valued, one must proceed to the computation of QALYs. Formally, it is a measure that combines both length and quality of life dimensions into a single indicator. The approach consists in multiplying the utility

Table 6.11 Choice card for the EQ-5D-5L

	State A (13345)	State B (52454)
Mobility	I have no problem in walking about	I am unable to walk about
Self-care	I have moderate problems washing or dressing myself	I have slight problems washing or dressing myself
Usual activities	I have moderate problems doing my usual activities	I have severe problems doing my usual activities
Pain/discomfort	I have severe pain or discomfort	I have extreme pain or discomfort
Anxiety/depression	I am severely anxious or depressed	I am severely anxious or depressed
Which is better, state A or state B?	□	□

value $u(H_j)$ associated with each health state H_j by the number of years lived in that state. For example, 2 months lived in state H_j is computed as $2/12 \times u(H_j)$. QALYs can then be incorporated in a decision-analytic model together with medical costs to implement a cost effectiveness analysis (see the related chapter). For instance, the EQ-5D-5L classification system provides indicators of health-related quality of life that run from death (or even situations worse than death) to "perfect" health on a QALY scale. As time elapses after an intervention, patients go through various health states (e.g., hospitalization for hip surgery, then convalescence period followed by progressive return to normal walk) that generates quality-adjusted life-years lived. Those are compared with what would have come out of a life without that intervention (e.g., no trauma associated with hip surgery but recurrent difficulties to walk). The difference over the considered time horizon would breed quality-adjusted life-years gained. This difference in effectiveness can then be related to cost considerations.

Bibliographical Guideline

The economic foundations of stated preferences techniques can be traced to Bowen (1943) and Ciriacy-Wantrup (1947) who recognized the advantages of using public opinion surveys to value public goods. Since then, many developments and applications have emerged over a wide variety of themes such as recreational sites, air and water quality. Carson (2011) provides a comprehensive review of this large literature and its recent developments.

In particular, contingent valuation enjoyed a revival in 1993 after the publication of a report by a panel of experts, chaired by Nobel Prize laureates Kenneth Arrow and Robert Solow. Following the Exxon Valdez oil catastrophe that occurred in 1989, this panel was created under the auspices of the National Oceanic and Atmospheric Administration, a branch of the United States Department of Commerce, to appraise the validity of contingent valuation measures for natural

Table 6.12 Example of EQ-5D-5L value set and health state values

Utility decrements	Estimate	13345	52454
Constant	1	1	1
Mobility			
No	0	0	
Slight	−0.051		
Moderate	−0.063		
Severe	−0.212		
Unable	−0.275		−0.275
Self-care			
No	0		
Slight	−0.057		−0.057
Moderate	−0.076	−0.076	
Severe	−0.181		
Unable	−0.217		
Usual activities			
No	0		
Slight	−0.051		
Moderate	−0.067	−0.067	
Severe	−0.174		−0.174
Unable	−0.190		
Pain/discomfort			
No	0		
Slight	−0.060		
Moderate	−0.075		
Severe	−0.276	−0.276	
Unable	−0.341		−0.341
Anxiety/depression			
No	0		
Slight	−0.079		
Moderate	−0.104		
Severe	−0.296		−0.296
Unable	−0.301	−0.301	
Value for health state 13345	1−(0+0.076+0.067+0.276+0.301)=0.28		
Value for health state 52454	1−(0.275+0.057+0.174+0.341+0.296)=−0.143		

resource damage. The report advocated the use of carefully designed surveys and concluded that "*contingent valuation studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive-use values*" (NOAA 1993, p. 43).

Choice modeling and the discrete choice experiment methodology find their conceptual basis in Lancaster's (1966) theory of consumer demand which assumes that consumers' utility for goods depend on the characteristics those goods contain. The approach became popular in marketing after the development of the random utility model formulated as a conditional logit model by the Nobel Laureate McFadden (1973). Since then, discrete choice experiment has been intensively used in diverse fields, e.g., tourism, irrigation water, climate change, air quality, etc. A review of the recent literature is available in Carson and Czajkowski (2014). It is also commonly used in the health sector to elicit preferences for healthcare products and programs (Ryan et al. 2008).

The theoretical foundations of the hedonic pricing method are attributed to Lancaster's (1966) consumer theory and Rosen's (1974) model of market behavior for differentiated goods. One of the first hedonic pricing studies is that of Ridker and Henning (1967). Using cross-section data for the St. Louis metropolitan area in the USA, they provided empirical evidence that sulfate air pollution, housing characteristics, accessibility, and neighborhood characteristics affect property values. More specifically, they argued that the coefficient on the air pollution variable in the regression equation can be interpreted as the average willingness to pay for air quality improvements for all St. Louis households. The approach was criticized a few years later by Freeman (1971) who emphasized the fact that the population coefficient provides a correct measure of willingness to pay if and only if the air quality improvement is small. Since then, the method has been used for estimating the effect of different attributes in many fields (see for instance Chau et al. 2004).

The origin of the travel cost method is attributed to Hotelling (1947), who argued in a letter to a park service director that visitation rates should be inversely related to the distance travelled to reach a site. In this document, Hotelling considered concentric zones defined around the park so that the cost of travel to the park from all points in one of these zones is approximately constant. Hotelling's approach was first applied by Trice and Wood (1958) on data obtained from visitors in the Sierras. Clawson (1959) and Clawson and Knetsch (1966) then expanded the idea assuming that the experience of users from one location zone should provide an accurate measure of what people in other location zones would do if costs were the same. Nowadays, the method is still often referred to as the Clawson-Knetsch approach. In a context were recreational sites were becoming more and more attractive, the approach was particularly popular in the sixties when the United States Congress actually required that recreation be considered in program valuation for water projects (see e.g., Banzhaf 2010, for an historical perspective on nonmarket valuation and recreation demand).

The measure of health-related quality of life is a major challenge to health technology assessment and policy evaluation. Integrating clinical evidence and individual values (Huninck et al. 2007) requires a single measure of both quantity of life and quality of life. QALYs are meant to reach that objective and we have focused here on the EQ-5D-5L valuation protocol (Oppe et al. 2014). However, the approach does not provide a definite answer to the challenge. Vast areas of health policies cannot convincingly rely on QALYs, as evidenced in cancer (Garau et al. 2011) and pediatric (Ungar 2011) treatments. Furthermore, health-related quality of life does not fully comprehend all the complex dimensions of individual well-being.

In this respect, the ICECAP methodology is a promising avenue (Al Janabi et al. 2012).

References

Al Janabi, H., Flynn, T., & Coast, J. (2012). Development of a self-reported measure of capability well-being for adults: The ICECAP-A. *Quality of Life Research, 21*, 167–176.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). Report of the NOAA panel on contingent valuation. *Federal Register, 58*, 4602–4614.

Banzhaf, H. S. (2010). Consumer surplus with apology: A historical perspective on nonmarket valuation and recreation demand. *Annual Review of Resource Economics, 2*, 18.1–18.25.

Bowen, H. R. (1943). The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics, 58*, 27–48.

Carson, R. T. (2011). *Contingent valuation: A comprehensive bibliography and history*. Cheltenham: Edward Elgar.

Carson, R. T., & Czajkowski, L. (2014). The discrete choice experiment approach to environmental contingent valuation. In S. Hess & A. Daly (Eds.), *Handbook of choice modelling*. Cheltenham: Edward Elgar.

Chau, K. W., Yiu, C. Y., Wong, S. K., & Wai-Chung, L. L. (2004). Hedonic price modelling of environmental attributes: A review of the literature and a Hong Kong case study. In *Encyclopaedia of life support systems*. Oxford: UNESCO.

Ciriacy-Wantrup, S. V. (1947). Capital returns from soil-conservation practices. *Journal of Farm Economics, 29*, 1181–1196.

Clawson, M. (1959). *Method for measuring the demand for, and value of, outdoor recreation*. Washington, DC: Resources for the Future, *10*.

Clawson, M., & Knetsch, J. L. (1966). *Economics of outdoor recreation*. Baltimore: Johns Hopkins University Press.

Freeman, A. M. I. I. I. (1971). Air pollution and property values: A methodological comment. *Review of Economics and Statistics, 53*, 415–416.

Garau, M., Shah, K., Mason, A., Wang, Q., Towse, A., & Drummond, M. (2011). Using QALYs in cancer. *Pharmacoeconomics, 29*, 673–685.

Hotelling, H. (1947). *Letter of June 18, 1947, to Newton B. Drury. Included in the report the economics of public recreation: An economic study of the monetary evaluation of recreation in the National parks, 1949. Mimeographed*. Washington, DC: Land and Recreational Planning Division, National Park Service.

Huninck, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A., et al. (2007). *Decision making in health and medicine: Integrating evidence and values*. Cambridge: Cambridge University Press.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*, 132–157.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–135). New York: Wiley.

Oppe, M., Devlin, N., van Hout, B., Krabbe, P., & de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health, 17*, 445–453.

Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics, 49*, 246–257.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in perfect competition. *Journal of Political Economy, 82*, 34–55.

Ryan, M., Gerard, K., & Amaya-Amaya, M. (Eds.). (2008). *Using discrete choice experiments to value health and health care*. Heidelberg: Springer.

Trice, A. H., & Wood, S. E. (1958). Measurement of recreation benefits. *Land Economics, 34*, 195–207.

Ungar, W. (2011). Challenges in health state valuation in pediatric economic evaluation. *Pharmacoeconomics, 29*, 641–652.

Part II

Ex ante Evaluation

Financial Appraisal 7

7.1 Methodology of Financial Appraisal

Financial appraisal evaluates the financial attractiveness of an investment project by analyzing the timing and value of cash flows that result from its implementation. The approach is regularly carried out by companies in the private sector to assess whether investment strategies are commercially profitable. It is also in use in the public sector and particularly relevant for appraising large public works or public-private partnership projects. In the public sector, it is often a preliminary step to a more detailed and complex "economic appraisal" which, under the auspices of a cost benefit analysis (CBA), aims at assessing the impacts of the project on the well-being of the stakeholders. While a financial analysis examines the projected revenues with the aim of assessing whether they are sufficient to cover expenditures and/or to make the investment sufficiently profitable, an economic analysis goes further by examining the satisfaction derived from the consumption of public services. The approaches are thereby different, but also complementary, as a project that is financially interesting is not necessarily economically viable and vice versa.

The financial methodology should not be mistaken with what is done in accounting. Simply put, an accountant prepares financial statements such as revenue statements, balance sheets and cash flows, and makes sure that these records are compliant with law requirements. A financial manager on the other hand provides decision-makers with financial advices and support (through performance indicators and graphics) by weighing the costs and revenues of a certain course of actions and planning for the long-term. The emphasis is placed on decision-making with the aim of better allocating the budget. The analysis is often global in the sense that the focus is on the financial health of the entity as a whole, but can be project-specific (so-called project appraisal), to better adjudge and compare the attractiveness of competing investment strategies.

A crucial issue in project appraisal is the timing of cash flows. Figure 7.1 provides an illustration. First, a public project is characterized by a time horizon. It represents the maximum number of years for which cash flow forecasts are

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_7

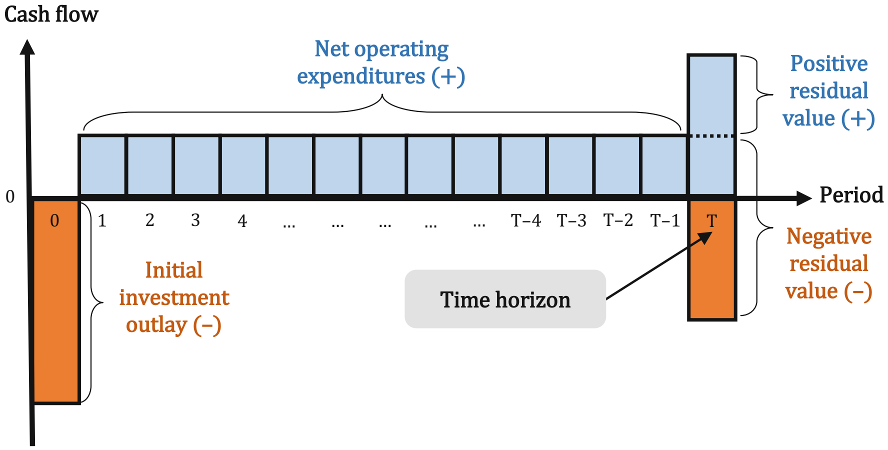

Cash flow
Net operating expenditures (+)
Positive residual value (+)
0
Period
0
1
2
3
4
...
T−4
T−3
T−2
T−1
T
Initial investment outlay (−)
Time horizon
Negative residual value (−)

Fig. 7.1 A typical cash flow stream

provided (usually the same as the CBA time horizon). Selecting the right time horizon is crucial. A time horizon that does not capture all future revenues and costs can make a project's return on investment seem better or worse than it is. This may in return affect policy recommendations. Based on survey results, the "Guide to cost-benefit analysis of investment project" prepared for the European Commission in 2014, advocates a time horizon of at least 20 years and at most 30 years for the majority of infrastructures; while for productive investments, it is about 10 years.

Second, an investment project is defined by a stream of cash flows whose evolution is irregular in time. Identifying and forecasting those cash flows constitutes the core of any financial appraisal. Cash flow forecasts are usually provided in spreadsheets which include information on:

1. **Investment costs:** initial outlay incurred at the beginning of the project's life (land, buildings, equipment, licenses, patents, other pre-production expenses, etc.), extraordinary maintenance, residual value;
2. **Operating costs and revenues:** ongoing running costs (labor, energy costs, maintenance costs, security costs, administration costs) and revenues (sales, user charges);
3. **Sources of financing**: equity, loans, bonds and other financial resources.

As illustrated in Fig. 7.1, a project is characterized by (1) large costs at the beginning of the projects' life, (2) recurrent but smaller inflows observed all along the project's life and (3) a residual value when the project ends. Operating costs and revenues emerge recurrently during the implementation of the project. Investment flows appear in the first periods mostly, i.e. during the construction process. An exception is when some items can be sold off (thus generating a positive residual value) or when an extraordinary maintenance generates a large outflow at a point in time (when an item breaks down or becomes obsolete). The residual value can sometimes be negative if the asset has to be removed from where it was used (e.g., nuclear waste placed in long-term storage).

The financial attractiveness of a public project is determined by its ability to be sustainable and profitable. To assess whether this is the case or not, a financial appraisal produce three types of analysis:

1. **Sustainability analysis**.
2. **Analysis of the profitability of the investment**.
3. **Analysis of the profitability of the capital**.

Financial sustainability is defined as the capacity of the project revenues to cover the costs. The analysis relates the sources of financing to the financial outflows in order to check whether the project risks of running out of money from one year to the other. Financial profitability is the ability of the project to achieve a satisfactory rate of return. Here, a distinction is made between the public entity that commissions the project and the partner investors that support the project through additional funding. As their objective functions differ, there are two definitions of profitability. First, the profitability of the investment is examined to make sure that there exists no better source of revenue for the public. Second, the return on capital can be evaluated, with the aim of assessing whether the project is commercially profitable for the partner investors.

Table 7.1 summarizes the differences between the different types of analysis. The “–“ and “+” signs indicate whether an item should be considered as an outflow or an inflow, respectively. As can be seen, the financial sustainability analysis takes into account all the possible cash flows in order to evaluate the difference between the money that goes in and the money that goes out. The approach enables the identification of financing shortfalls that may occur during the project. The

Table 7.1 The main elements of financial appraisal

	Financial sustainability	Financial profitability of the investment	Financial profitability of the capital
Total investments			
Total investments costs	–	–	
Operating revenues and costs			
Total operating costs	–	–	–
Total operating revenues	+	+	+
Sources of financing			
Loans	+		
Equity	+		–
Loans reimbursement	–		–

examination of cash flows is made on an annual basis. The approach is different when it comes to the profitability of a project. The main question is whether the revenues as a whole are worth the money invested. The analysis consists first in comparing the initial investment with the net operating revenues (profitability of the investment). In this case, profitability is examined regardless of loan repayments. This avoids double counting as the loan repayments also reflect the initial investment cost for which the money was borrowed. Second, the analysis compares the amount of capital with the commercial profit induced by the project (profitability of the capital). Here, external contributions (equity) appear with a negative sign as they are expenditures from the point of view of the partner investors. As they also determine the final profit, loan repayments have to be taken into account in this case (total investments are excluded).

A key feature of the profitability analysis is that it does not consider the cash flows on a year basis but, instead, totals them up in order to compare the global net revenue generated by the project with the initial investment or capital. To do so, the approach relies on discount factors, which weight the cash flows according to their position in time. The approach is based on the idea that a dollar today is worth more than a dollar tomorrow. The time value of money is a central concept in finance. There are several situations where dollars at different points in time are compared, for instance when money is borrowed (bank loans), when money is invested (saving accounts, bonds), or when one wants to evaluate an investment project (discounted cash flow analysis).

The remainder of the chapter is structured as follows. Section 7.2 introduces the concept of the value of time with a focus on interest rate effects. Section 7.3 explains how to assess the financial sustainability of a project through the examination of its sources of financing. Section 7.4 is about discounting and the profitability of the investment and capital. Section 7.5 discusses and presents alternative methods of ranking investments. Section 7.6 discusses the treatment of inflation in project appraisal. Last, Sect. 7.7 introduces the basics of a sensitivity analysis, in order to identify the variables that are critical in profitability analysis.

7.2 Time Value of Money

The time value of money is a concept underlying many financial techniques. The intuition behind it is that a dollar in hand today is worth more than a dollar received tomorrow. The reason is simple: a dollar today can be invested to earn interests tomorrow. The time value of money is thus related to the concept of opportunity cost. Trade-offs between current and future dollars depend on the rate of return or interest rate one can earn by investing.

There are two basic methods to account for the effects of interest accumulation. One is to compute the future value of an investment. This process by which cash flows are expressed in terms of their future value is called compounding. The second approach consists in removing the interest effect over time by computing the present value of a future payment. The approach is referred to as discounting.

Understanding the effects of interest rates is essential if one wants to apprehend the whys and wherefores of the financial approach. Basically speaking, the sum on which the interest is being paid is termed the principal. It is the amount of money borrowed or invested. The interest rate is the ratio of interest paid to the principal.

Interests can be classified as simple interest or compound interest. Compound interest refers to the situation in which interests are calculated using a base that changes over time. With simple interest on the other hand, the base on which interests are calculated is fixed. Assume for instance that one invests \$10,000 for 10 years and receives 5% per year in interest. The growth of the investment is depicted in Table 7.2. With simple interest, the amount earned does not change. One multiplies the principal with the 5% rate. The interest amounts to $\$10{,}000 \times 5\,\% = \$500$ each year. With compound interest, the principal accumulates. One earns \$500 at the end of the first year; the balance becomes \$10,500. The interest of the second year is computed as $5\,\% \times \$10{,}500 = \$525$. The process is then reiterated until one reaches the time horizon. As can be seen from Table 7.2, the total interest amounts to \$5000 with simple interest, and to \$6289 with compound interest. Compound interest works to the investor's advantage but against that of the borrower.

While simple interest is described in many textbooks, it has limited practical use. The reason is the time inconsistency generated by the process of calculating the interest. If one earns \$500 after one year, one should be able to re-invest that money in a similar manner to earn additional interest. This is why, in practice, compound interest is more commonly used in finance, to compute the interest charged for a loan, or to compute a present value. In the remaining of the chapter, when we will refer to interest, we will always have in mind a compound interest.

Formally, if an amount P_0 (the initial principal) is invested for one year at an interest rate equal to r then, at the end of the year, one earns $P_0(1+r)$. The second year, one earns $P_0(1+r) \times (1+r) = P_0(1+r)^2$, and so on. The value of an

Table 7.2 Simple versus compound interest: example 1

	5% simple interest		5% compound interest	
Year	Balance	Interest	Balance	Interest
0	\$10,000		\$10,000	
1	\$10,500	5%×\$10,000=\$500	\$10,500	5%×\$10,000=\$500
2	\$11,000	5%×\$10,000=\$500	\$11,025	5%×\$10,500=\$525
3	\$11,500	5%×\$10,000=\$500	\$11,576	5%×\$11,025=\$551
4	\$12,000	5%×\$10,000=\$500	\$12,155	5%×\$11,576=\$579
5	\$12,500	5%×\$10,000=\$500	\$12,763	5%×\$12,155=\$608
6	\$13,000	5%×\$10,000=\$500	\$13,401	5%×\$12,763=\$638
7	\$13,500	5%×\$10,000=\$500	\$14,071	5%×\$13,401=\$670
8	\$14,000	5%×\$10,000=\$500	\$14,775	5%×\$14,071=\$704
9	\$14,500	5%×\$10,000=\$500	\$15,513	5%×\$14,775=\$739
10	\$15,000	5%×\$10,000=\$500	\$16,289	5%×\$15,513=\$776
Total		**\$5000**		**\$6289**

investment at the end of a time horizon T over which interest is compounded is called the future value of an investment. It is defined as:

$$F_T = P_0(1+r)^T$$

Coming back to example 1 (Table 7.2), the principal is $P_0 = \$10{,}000$ and invested at 5% interest for 10 years. The future value of the investment is $F_{10} = \$10{,}000 \times (1+5\%)^{10} \approx \$16{,}289$. This sum of money represents the value the investment has after 10 years. It comprises both the interest ($6289) and the principal ($10,000).

Compound interest is also used to determine the present value of a future sum of money. The approach, also known as discounting, has the advantage to allow different investment strategies to be compared regardless of their time horizon. The present value of a future payment is computed as:

$$P_0 = \frac{F_T}{(1+r)^T}$$

In this setting, the interest rate is also termed a discount rate, and the ratio $1/(1+r)^T$ is referred to as a discount factor.

Consider for instance an investment strategy S_1 that yields a unique cash flow $F_{10} = \$16{,}289$ at year $T = 10$. At a 5% discount rate, the present value of the investment is computed as $P_0 = \$16{,}289/(1+5\%)^{10} = \$10{,}000$. This sum is lower than the future value because the money that is earned in the future is less valuable than the money that is earned today. Let us now examine an investment strategy S_2 that yields $F_{20} = \$20{,}000$ in 20 years at a similar rate. The present value of this lump sum is $P_0 = \$20{,}000/(1+5\%)^{20} = \$7538$. This amount is found to be lower than the one previously observed with strategy S_1. The sum earned ($20,000) does not compensate the large time horizon (20 years).

The present value of an annuity can be calculated in a similar manner. By definition, an annuity is a constant cash flow F that occurs at regular intervals for a fixed period of time ($F_t = F$ for all $t > 0$). The present value is determined by examining each periodic cash flow and discounting them back to the present:

$$P_0 = \frac{F}{(1+r)} + \frac{F}{(1+r)^2} + \ldots + \frac{F}{(1+r)^T}$$

Imagine for instance that one invests an amount at 5% interest so that one receives $10,000 per year for each of the next 3 years. To find the present value of this $10,000 3-year annuity, one can calculate the discount factor $1/(1+r)^t$ applying to each period t. Then, one multiplies each receipt by the discount factor, as shown in Table 7.3. The sum of the resulting figures yields the net present value of the annuity, which is found to be $27,232 in the present case.

Table 7.3 Present value of an annuity: example 2

Year	Amount received	Discount factor at 5%	Present value at 5%
1	\$10,000	1/1.05=0.9524	\$9524
2	\$10,000	$1/1.05^2 = 0.9070$	\$9070
3	\$10,000	$1/1.05^3 = 0.8638$	\$8638
Total			\$**27,232**

The computation of the present value of an annuity can use a factorized version of the formula. In mathematics, it has been shown that the sum of a geometric series of the form $ax + ax^2 + \ldots + ax^T$ is finite as long as x is strictly less than 1. We have:

$$ax + ax^2 + \ldots + ax^T = ax \frac{[1 - x^T]}{[1 - x]}$$

Replacing a by F and x by $1/(1 + r)$ in this expression yields:

$$P_0 = \frac{F}{(1 + r)} \times \frac{\left[1 - 1/(1 + r)^T\right]}{[1 - 1/(1 + r)]} = F \times \frac{1 - (1 + r)^{-T}}{r}$$

The expression $[1 - (1 + r)^{-T}]/r$ is termed an annuity factor. It represents the weight by which the periodic payment must be multiplied to obtain the present value of the annuity. For instance, in example 2 (see Table 7.3), the present value is computed as:

$$P_0 = 10{,}000 \times \frac{1 - (1 + 5\%)^{-3}}{5\%} \approx \$27{,}232$$

This sum represents the initial amount of money that has to be invested.

The process of paying off a loan (plus interest) is similar to that of an annuity. An amortized loan, by definition, is made of a series of regular, equal payments termed amortization. Assume for instance that one has borrowed \$100,000 from a bank at 5% interest and that one has agreed to pay off this loan by making equal payments on a year basis during 10 years. The amount \$100,000 represents the present value P_0 of the loan. Using the inverse form of the annuity factor, the periodic payment can be computed as:

$$F = P_0 \times \frac{r}{1 - (1 + r)^{-T}}$$

We thus have $F = \$100{,}000 \times (5\%)/(1 - (1 + 5\%)^{-10}) \approx \$12{,}950$. Multiplying this amount with the time horizon yields the maturity value of the loan: $F \times 10 \approx \$129{,}505$. This amount represents the total amount one must repay to the lender. It includes the principal (\$100,000) and the interest (\$29,505).

Year	Payment	Interest	Principal	Balance
0				$100,000
1	$12,950	$5,000	$7,950	$92,050
2	$12,950	$4,602	$8,348	$83,702
3	$12,950	$4,185	$8,765	$74,936
4	$12,950	$3,747	$9,204	$65,733
5	$12,950	$3,287	$9,664	$56,069
6	$12,950	$2,803	$10,147	$45,922
7	$12,950	$2,296	$10,654	$35,267
8	$12,950	$1,763	$11,187	$24,080
9	$12,950	$1,204	$11,746	$12,334
10	$12,950	$617	$12,334	$0
Total	$129,505	$29,505	$100,000	

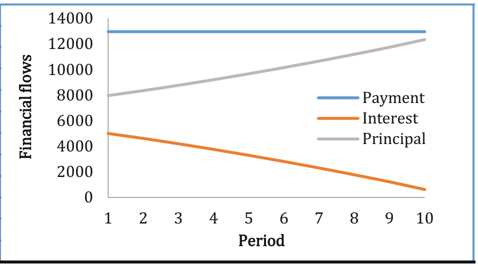


Fig. 7.2 Constructing an amortization schedule: example 3

The payment F that a borrower makes on an amortized loan partly pays off the principal (the original amount borrowed) and the interest (the fee the lender receives). In practice, it is common to provide the borrower with a table that shows how these portions vary through time. This list is called an amortization schedule. We illustrate the approach in Fig. 7.2. The numbers have been rounded off for simplicity of exposition. For the first period the interest (hereafter I_1) is computed by multiplying the 5% rate with the initial principal: $I_1 = P_0 \times r = \$100{,}000 \times 5\,\% = \$5000$. The difference obtained ($7950) between the payment $F = \$12{,}950$ and the interest $I_1 = \$5000$ yields a reduction in the remaining principal balance. We have $P_1 = \$100{,}000 - \$7950 = \$92{,}050$. In period 2, the interest is obtained from the multiplication of the 5% rate with the new balance: $I_2 = \$92{,}050 \times 5\%$. The approach is reiterated until one reaches the term of the loan. As can be seen from Fig. 7.2, at the beginning of the loan, large interest payments and small payments to the principal are made. As time goes on, the principal is reduced to the point that the payment covers mostly the principal. Because interests are computed on the current balance, they become progressively smaller as time increases.

Notice that annual interest can be compounded monthly, or quarterly. In this situation, one needs to take into account the number of periodic payments m per year in the previous formula:

$$F = P_0 \times \frac{r/m}{1 - (1 + r/m)^{-mT}}$$

The amortization schedule is defined similarly to what has been done previously. The relevant number of periodic payments becomes $m \times T$. First, the periodic payment F is computed using the formula above. Second, one obtains the interest portion I_1 by multiplying the principal P_0 with the rate r/m. Third, the principal portion is calculated as the difference between the periodic payment F and the interest I_1. The new principal balance is then determined from these expressions, and so on.

The concept of the time value of money should not be mistaken with that of depreciation. While the former is related to the return one may effectively earn from an investment, the latter is not associated with a real cash transaction. It is an accounting method that spreads the investment costs more equally over the lifespan of the project. The approach can be used for instance to claim special tax allowances. Most types of fixed assets such as buildings, machinery, vehicles, furniture, and equipment are depreciable. Certain intangible properties, such as patents, copyrights, and computer software are also depreciable. However, land value is usually not depreciable as it can be sold off at the end of the investment period. Many methods of depreciation exist. They include for instance:

1. **Straight-line method.** The asset's value is equally distributed over its estimated useful life:

   $\text{Depreciation per annum} = \frac{1}{\text{Useful life}}(\text{Cost} - \text{Residual value})$

   For example, an equipment costing $5000 with an estimated salvage value of $0 and an estimated life of 10 years, would be depreciated at the rate of $500 per year for each of the 10 years.
2. **Sum of the years' digits method.** It allocates a higher depreciation rate in the earlier years of the asset's useful life. Under this method, one needs first to calculate the sum of the years' digits. The level of depreciation at year t is then defined as:

   $\text{Depreciation at year } t = \frac{\text{Useful life} - (t-1)}{\text{Sum of the years digits}}(\text{Cost} - \text{Residual value})$

   If an equipment has a useful life of 5 years, the sum of the years' digits equals 1+2+3+4+5=15. The depreciation rate is computed as (5–0)/15=5/15 for the first year, (5–1)/15=4/15 the second year and so on. The depreciation rates should add up to 100%. The sum of the years' digits can also be computed as $n(n+1)/2$ where n denotes the useful life in years.

Depreciation techniques are based on the idea that one cannot deduct spending on fixed assets immediately as they have a long useful life. The approach has however no relation to the actual flow of money that takes place. Furthermore, depending on the method used, large differences may be observed between the true value and the book value of investment assets.

Spreadsheet software like Excel provide a large set of tools for those who want to compute items such as the present value or future value of an investment. Those tools are presented in Table 7.4. Some of them will be used in the remaining of this chapter to evaluate the sustainability and profitability of a project. For instance, loans from commercial or State banks allow public entities to leverage resources to finance a portion or all of a project's implementation costs. Assessing the cost of these loans is essential as it will determine the sustainability of the project. The concept of the time value of money is also central to project appraisal if one wants to assess the profitability of a project. In that case, all future cash flows must be estimated, discounted and compared to the initial sum of money engaged. In the remainder of the chapter, attention will be devoted to presenting these tools in the context they are used.

Table 7.4 Using spreadsheets with Excel

Context	Formula	Definition	Excel function
Lump sum	$F_T = P_0(1 + r)^T$	Future value of an investment	$FV(rate, nper, 0, pv)$ calculates the future value of an investment; *rate* is a constant interest rate; *nper* is the time horizon; *pv* is the lump sum. Entering a zero as the payment amount tells Excel there is no constant stream of payments.
	$P_0 = \frac{F_T}{(1+r)^T}$	Present value of an investment	$PV(rate, nper, 0, fv)$ calculates the present value of an investment; *rate* is a constant interest rate; *nper* is the time horizon; *fv* is the future value. Entering a zero as the payment amount tells Excel there is no constant stream of payments.
Annuity	$P_0 = F \times \frac{1-(1+r)^{-T}}{r}$	Present value of an annuity	$PV(rate, nper, pmt)$ calculates the present value of a an annuity; *rate* is a constant interest; *nper* is the time horizon; *pmt* is the payment made each period and cannot change over the life of the annuity.
Amortization schedule	$F = P_0 \times \frac{r}{1-(1+r)^{-T}}$	Payment	$PMT(rate, nper, pv)$ returns the payment amount for a loan based on an interest rate and a constant payment schedule; *rate* is the interest rate; *nper* is the time horizon; *pv* is the present value or principal of the loan.
	$P_t \times r$	Interest	$IPMT(rate, per, nper, pv)$ returns the interest payment for a given period for a loan based on periodic, constant payments; *rate* is a constant interest rate; *per* is the period for which the interest must calculated, *nper* is the time horizon; *pv* is the initial principal.
	$F - I_t$	Principal	$PPMT(rate, per, nper, pv)$ returns the payment on the principal for a given period for a loan based on periodic, constant payments; *rate* is a constant interest rate; *per* is the period for which the interest must calculated; *nper* is the time horizon; *pv* is the initial principal.

Source: Gathered and adapted from https://support.office.com

7.3 Cash Flows and Sustainability

A cash flow statement records the estimates of all cash receipts and expenditures that are expected to occur during a certain time period because of the project. It includes forecasts of financing, investment costs, operating costs and revenues. Only the cash flows induced by the project must be examined, in the sense that they are considered only if they induce a change compared to some status quo. For this reason, the analysis of cash flows is often referred to as “with versus without comparison” or “incremental cash flow analysis”.

Whether payments are for items already delivered or to be received in the future is of no concern to financial appraisal. What matters is the exact timing of cash proceeds and payments. They must represent real transactions, with a well-determined position in time. Similarly, a financial appraisal is not interested in whether inventories increase or decrease. Only cash outlays are to be recorded. Other accounting items should be ignored such as sunk costs (which have already been incurred and cannot be recovered) or depreciation (by which the costs of equipment are allocated a value over the duration of its useful life). Table 7.5 provides a list of typical cash flows as well as their position in time. The “–“ and “+” signs specifies whether those items represent an outflow or an inflow. These cash flows are usually net of VAT (tax on value added) if it is relevant to the context.

Investment costs include all costs related to the design and construction of the project (see Table 7.5a). A distinction is made between fixed assets, start-up costs and working capital. Fixed assets refer to investments needed to set up the project or to replace obsolete equipment (extraordinary maintenance). Examples comprise buildings, property, infrastructures, equipment, etc. Those goods are relatively durable and can be used repeatedly through the project’s life. Their purchase is usually concentrated in the first years of the project. A residual value can be included among the costs at the end of the project, generally as an inflow. However, the value can also be negative if the asset has to be removed from where it was used.

Start-up costs represent the inflows incurred to get the project started. They include the costs of preparatory studies, consulting services, training expenses, patents, research and development expenses.

Working capital refers to additional expenses that must be engaged to ensure that the project is implemented without delay (stocks, equipment spare parts). Only year-on-year increments in the level of working capital should be considered. These increments will be particularly large at the beginning of the project, when stocks and equipment spare parts must be built up. They will be lower and converging to zero as the project ends. At some point in time, no further investments in working capital will be recorded.

Operating costs are expenses associated with the operation, maintenance and administration of the project. They include raw materials, energy costs, labor, repairs and maintenance, insurance cost, quality control, and waste disposal costs. As shown in Table 7.5b, those items are regularly purchased. Costs such as interest and loan repayments are not included, because it would double count the capital cost for which money has been borrowed.

Operating revenues encompass the inflows obtained from selling a product or providing a service. The project’s output can be sold directly through user charges. Typical examples include the toll revenue from a road project, park entrance fees and annual park passes, admission fees for public swimming pools, fees levied on parents of public school, park rental rates, fees on gym and sport fields or equipment rentals. The project’s output can also be charged indirectly, via specific tax schemes (e.g., garbage disposal tax, fire protection or water taxes). Those user charges and

Table 7.5 A typical cash flow budget

		Year 0	Year 1	Year 2	. . .	Year t	. . .	Year T
(a) Total investments								
R1	Land	–			. . .		. . .	
R2	Buildings	–			. . .		. . .	
R3	Equipment	–			. . .		. . .	
R4	Extraordinary maintenance				. . .	–	. . .	
R5	Residual value				. . .		. . .	+
R6	**Total fixed assets: R1+. . .+R5**	**–**			. . .	**–**	. . .	**+**
R7	Licenses	–			. . .		. . .	
R8	Patents	–			. . .		. . .	
R9	Preparatory studies	–			. . .		. . .	
R10	**Total start-up costs: R7+R8 +R9**	**–**			. . .		. . .	
R11	Stocks		–	–	. . .		. . .	
R12	Equipment, spare parts		–	–	. . .		. . .	
R13	**Working capital: R11+R12**		**–**	**–**	. . .		. . .	
R14	**Total investment costs: R6 +R10+R13**	**–**	**–**	**–**	. . .	**–**	. . .	**–/+**
(b) Operating costs and revenues								
R1	Raw material		–	–	. . .	–	. . .	–
R2	Labor		–	–	. . .	–	. . .	–
R3	Electric power		–	–	. . .	–	. . .	–
R4	Maintenance		–	–	. . .	–	. . .	–
R5	Administrative costs		–	–	. . .	–	. . .	–
R6	Sales expenditures		–	–	. . .	–	. . .	–
R7	**Total operating costs: R1+. . .+R6**		**–**	**–**	. . .	**–**	. . .	**–**
R8	Sales		+	+	. . .	+	. . .	+
R9	Taxes		+	+	. . .	+	. . .	+
R10	**Total operating revenues: R8+R9**		**+**	**+**	. . .	**+**	. . .	**+**
R11	**Net operating revenues: R7+R10**		**–/+**	**–/+**	. . .	**–/+**	. . .	**–/+**
(c) Sources of financing								
R1	Loan	+			. . .		. . .	
R2	Private equity	+			. . .		. . .	
R3	Public contribution	+			. . .		. . .	
R4	**Total financial resources: R1+R2+R3**	**+**			. . .		. . .	
R5	Loans: principal		–	–	. . .	–	. . .	–
R6	Loans: interest		–	–	. . .	–	. . .	–
R7	**Loan reimbursement: R5+R6**		**–**	**–**	. . .	**–**	. . .	**–**
R8	**Net financing flows: R4+R7**	**+**	**–**	**–**	. . .	**–**	. . .	**–**

taxes usually do not accrue until the service in question is operational. For this reason, they are often devoted to ongoing facility operations only.

A project may also depend on extra sources of financing, for instance through equity (grantors, local investors, host government, bilateral or multilateral organizations), grants (supranational, federal, State or local funding) and debt (commercial lenders, State-owned banks, supranational investment banks, bonds). Financial appraisal usually records those flows in an extra table (see Table 7.5c). Equity contributions are made by external investors through share capital and other shareholder funds. Many infrastructure projects also benefit from State support in the form of equity participation and capital grants (transfers received from government or international organizations for the purpose of financing the acquisition of capital assets). In Table 7.5c, debt reimbursements are included and tied to an amortization schedule. Those payments have priority among the invested funds. Lenders can obtain any return or repayment before equity does.

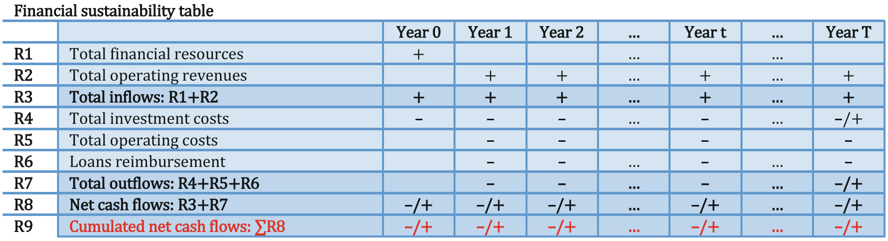
By definition, a project is said to be financially sustainable when it does not incur the risk of running out of cash in one of the examined periods. The sustainability analysis establishes this condition by summarizing all the cash flows in a single spreadsheet, as depicted in Fig. 7.3. Essentially, the analysis must relate all costs and debt repayment induced by the project to the different sources of financing. One needs first to compute the net cash flows at the point in time t they are observed:

$$\text{Net cash flows}_t = \text{Total inflows}_t - \text{Total outflows}_t$$

As shown in Fig. 7.3, the total inflows are defined by the total financial resources and operating revenues. The total outflows on the other hand are computed as the sum of investment costs, operating costs and loans reimbursement.

Second, once the net cash flows are computed, one must show that the project will have sufficient inflows to cover expenditures for investment and operation throughout the entire lifespan:

$$\sum_{k=1}^{t} \text{Net cash flows}_k \geq 0 \text{ for all } t = 1 \ldots T$$



Financial sustainability table

		Year 0	Year 1	Year 2	...	Year t	...	Year T
R1	Total financial resources	+			...		...	
R2	Total operating revenues		+	+	...	+	...	+
R3	**Total inflows: R1+R2**	+	+	+	...	+	...	+
R4	Total investment costs	–	–	–	...	–	...	–/+
R5	Total operating costs		–	–		–		–
R6	Loans reimbursement		–	–	...	–	...	–
R7	**Total outflows: R4+R5+R6**		–	–	...	–	...	–/+
R8	**Net cash flows: R3+R7**	–/+	–/+	–/+	...	–/+	...	–/+
R9	**Cumulated net cash flows: ∑R8**	–/+	–/+	–/+	...	–/+	...	–/+

Fig. 7.3 Verification of the financial sustainability

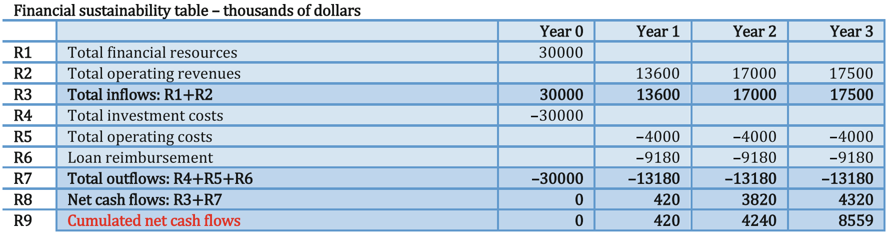
Table 7.6 Costs, revenues and resources: example 4

		Year 0	Year 1	Year 2	Year 3
Total Investments—thousands of dollars					
R1	Lands	–7000			
R2	Bridge infrastructure	–15,000			
R3	Equipment	–4000			
R4	Start–up costs	–1500			
R5	Road network	–2500			
R6	**Total investment costs: R1+R5**	**–30,000**	**0**	**0**	**0**
Operating revenues and costs—thousands of dollars					
R1	Raw materials		–2250	–2250	–2250
R2	Labor		–750	–750	–750
R3	Electric power		–300	–300	–300
R4	Maintenance		–450	–450	–450
R5	Administrative costs		–80	–80	–80
R6	Sales expenditures		–170	–170	–170
R7	**Total operating costs: R1+...+R6**	**0**	**–4000**	**–4000**	**–4000**
R8	Sales		13,600	17,000	17,500
R9	**Total operating revenues: R8**	**0**	**13600**	**17000**	**17500**
R10	**Net operating revenues: R7+R9**	**0**	**9600**	**13000**	**13500**
Sources of financing table—thousands of dollars					
R1	Loan	25,000			
R2	Private equity	2100			
R3	Public contribution	2900			
R4	**Total financial resources: R1+R2+R3**	**30,000**	**0**	**0**	**0**
R5	Loans: principal		–7930	–8327	–8743
R6	Loans: interest		–1250	–853	–437
R7	**Loan reimbursement: R5+R6**		**–9180**	**–9180**	**–9180**
R8	**Net financing flows: R4+R7**	**30,000**	**–9180**	**–9180**	**–9180**

Should this condition be fulfilled, it would ensure the availability of sufficient funds all along the project's life.

To illustrate the method, consider the construction of a bridge in a given jurisdiction. Table 7.6 provides a detailed presentation of the cash proceeds and payments. For simplicity of exposition, the time horizon is set to 3 years. In addition, we assume no residual value. The project involves an immediate outlay of $30 million (for the lands, infrastructure, equipment, the stocks and spare parts) and is followed by annual operating expenditures of $4 million (raw materials, labor, electric power, etc.). It generates annual revenues via a toll which amounts to $13.6 million for the first year, $17 million for the second year, and $17.5 million for the third year.

The bridge project is financed by a combination of debt and funds. The private and public sector contribute to the financing of the project for $2.1 million and $2.9 million, respectively. A loan is secured for an amount of $25 million. The current



Financial sustainability table – thousands of dollars

		Year 0	Year 1	Year 2	Year 3
R1	Total financial resources	30000			
R2	Total operating revenues		13600	17000	17500
R3	**Total inflows: R1+R2**	**30000**	**13600**	**17000**	**17500**
R4	Total investment costs	-30000			
R5	Total operating costs		-4000	-4000	-4000
R6	Loan reimbursement		-9180	-9180	-9180
R7	**Total outflows: R4+R5+R6**	**-30000**	**-13180**	**-13180**	**-13180**
R8	**Net cash flows: R3+R7**	**0**	**420**	**3820**	**4320**
R9	Cumulated net cash flows	**0**	**420**	**4240**	**8559**

Fig. 7.4 Sustainability analysis: example 4

interest rate is assumed to be 5% per annum. The formula provided in Sect. 2 can be used to compute the loan annuity:

$$F = \$25 \text{ million} \times \frac{5\%}{1-(1+5\%)^{-3}} \approx \$9.18 \text{ million}$$

The amortization schedule can be obtained using Excel formula such as *IPMT* and *PPMT* (see Table 7.4).

The sustainability analysis consists in assessing whether the revenues and other sources of financing are sufficient to cover the costs and the loan repayments. This comparison is made on a year basis. Figure 7.4 presents the net cash flows resulting from the bridge project (not discounted and before tax). At year 0, the project faces an inflow equal to 30 million dollars. This amount comprises the loan, the private equity and the public contribution. In the meanwhile, the investment amounts to 30 million dollars, which finally results in a net cash flow equal to zero. Financial sustainability is thus verified for the first period. In year 1, the project gets extra money from the toll (13,600), but the loan must be reimbursed (9180) and the usual operating expenses must be paid (4000). This yields a net cash flow equal to 420 thousand dollars. From row R9, we can see that financial sustainability is still verified as the net cumulated cash flow remains positive 0+420=420). Overall, the examination of row R9 yields the conclusion that financial sustainability is verified for all periods. The cumulated cash flows are always more than or equal to zero for all the years considered.

7.4 Profitability Analysis

The profitability analysis aims to compare the total revenues against the total costs observed over the whole projects' life. The question is not whether the project risks of running out of money from 1 year to another but, instead, whether the return on investment is sufficiently high. The approach makes use of discount factors to weight the cash flows according to their position in time. The approach is also referred to as discounted cash flow analysis. When investors commit funds to a project, they have an opportunity cost that derives from sacrificing a return on alternative investments. Discounted cash flow analysis makes it possible to take into account this implicit cost.

It has been shown in Sect. 7.2 that the interest rate can be used to convert a stream of cash flows in terms of present value. In a similar manner, the profitability analysis computes the financial net present value or *FNPV*. It is defined as the difference between the present value of cash inflows and the present value of cash outflows. Formally, let CF_t denote the net cash flows observed in period t. We have:

$$FNPV = CF_0 + \frac{CF_1}{(1+r)^1} + \ldots + \frac{CF_T}{(1+r)^T} = \sum_{t=0}^{T} \frac{CF_t}{(1+r)^t}$$

The weights $1/(1+r)^t$ by which the cash flows are multiplied are the discount factors and r is the financial discount rate. Those factors are lower than one and decrease as time increases. Cash flows are thereby considered of less importance if they occur at the end of the project's life. A positive financial net present value indicates that the projected earnings generated by the project exceed the costs, in terms of present dollars.

The choice of the discount rate is decisive. In theory, it represents the opportunity cost of funds, valued as the loss of return from an alternative investment. Assume for instance that we invest $50,000 today and receive $51,000 next year (strategy S_1). We can compare this return with that of an alternative strategy, e.g., a government bond at 5% interest (strategy S_2). To make this comparison possible, we can compute the net present value:

$$FNPV(S_1) = -\$50,000 + \frac{\$51,000}{(1+5\%)} \approx -\$1428$$

The net present value is negative. This means that strategy S_2 offers a higher return than strategy S_1. We reach a similar conclusion if, instead, we compare the future value of the investments: $50,000 invested at a 5% rate yields a return equal to $52,500. Hence, strategy S_2 remains the optimal choice even when values are expressed in future terms ($52,500 is greater than $51,000). Comparing present or future values is actually strictly equivalent as we have $(-\$52{,}500 + \$51{,}000)/(1+5\%) = -\$1428$.

In practice, the discount rate is recommended by government agencies such as the national Treasury, or supranational authorities such as the European Union. In most cases, it represents the safest alternative use. It is for instance approximated by the real return on government bonds, the long term real interest rate on commercial loans, or the return on a portfolio of securities in the international financial market. A positive financial net present value means that the project yields a return higher than these safer investment strategies.

An alternative to the net present value criterion is the financial internal rate of return or *FIRR*. It is the highest rate that the project can bear. Formally, it is defined as the value of r such that $FNPV(r) = 0$. Any strategy with a positive net present value will also have an internal rate of return that exceeds the discount rate. The term “internal” is used because this rate depends only on the cash flows generated by the investment and not on some other rate observed elsewhere. It cannot be determined by an algebraic formula. Three different techniques can be used to approximate its value:

1. **Trial and error method.** The approach consists in using different discount rates until one finds out which rate delivers a *FNPV* close to zero. One generally starts with a low discount rate and calculates the net present value. If the *FNPV* exceeds zero, the discount rate is increased. When it is negative, the discount rate is decreased.
2. **Linear interpolation.** The following heuristic can be used, although it may yield imprecise results. First, one needs to calculate two net present values using two different rates. Second, one uses the following formula to find the internal rate of return:
   $FIRR = r_1 - FNPV(r_1)\frac{(r_2 - r_1)}{FNPV(r_2) - FNPV(r_1)}$
   where r_1 and r_2 are two randomly chosen rates with $r_2 > r_1$. Depending on the value of r_1 and r_2, the computation can be more or less accurate.
3. **Spreadsheet software.** Excel can perform financial calculations. Relevant formulas are provided in Table 7.7.

   Assuming that all projects require the same amount of initial funds, the project with the highest rate of return is generally considered the best from the financial point of view.

To illustrate the approach, let us consider a simple example, as that provided in Table 7.8. The project involves an immediate outlay of –$30,000 with annual net operating income in each of 5 years of $7200. The initial outlay is timed for year

Table 7.7 Discounted cash flow analysis with Excel

Context	Formula	Definition	Excel function
Profitability of a project	$FNPV = \sum_{t=0}^{T}\frac{CF_t}{(1+r)^t}$	Net present value	$value0 + NPV(rate, value1, value2, \ldots)$ calculates the net present value; *value*0 represents the first cash flow (investment) and is excluded from the *NPV* formula because it occurs in period 0 and should not be discounted; *rate* is the discount rate and *value*1 , *value*2 , . . . is a series of future payments (range of cells containing the subsequent cash flows).
	r such that $FNPV(r) = 0$	Financial internal rate of return	$IRR(value0, value1, value2, \ldots)$ yields the internal rate of return for a series of cash flows (here *value*0, *value*1, *value*2), starting from the initial period. Values must contain at least one positive value and one negative value.

Source: Gathered and adapted from https://support.office.com

Table 7.8 Annual cash flows: example 5

Year	Net cash flow	Discount factor at 5%	Present value at 5%	Discount factor at 8%	Present value at 8%	Discount factor at 6%	Present value at 6%
0	–$30,000	1.00	–$30,000	1.00	–$30,000	1.00	–$30,000
1	$7200	0.95	$6857	0.93	$6667	0.94	$6792
2	$7200	0.91	$6531	0.86	$6173	0.89	$6408
3	$7200	0.86	$6220	0.79	$5716	0.84	$6045
4	$7200	0.82	$5923	0.74	$5292	0.79	$5703
5	$7200	0.78	$5641	0.68	$4900	0.75	$5380
FNPV			**$1172**		**–$1252**		**$329**

0, while net cash flows are spread equally from year 1 to 5. This is a typical pattern of cash flows for which the internal rate of return can be derived without computational difficulties. The net present value is the sum of the discounted annual net cash flows. First, assuming a discount rate of 5%, we can compute the discount factors for each period (third column). For this rate, the project is financially acceptable as the net present value is positive ($FNPV = \$1172$). Using trial and error, one can figure out the internal rate of return. A discount rate of 8% yields for instance a *FNPV* equal to –$1252. Using a lower discount rate of 6%, the *FNPV* starts converging to zero. We can also use a linear interpolation, for instance by using the 5% and 8% rates:

$$FIRR = 5\% - (\$1172)\frac{(8\% - 5\%)}{(-\$1252) - (\$1172)} \approx 6.4\%$$

A similar result can also be obtained using the IRR formula in Excel.

As future cash flows are more likely to be positive while the initial net cash flow CF_0 is generally negative (initial investment), the higher is the discount rate, the lower is the net present value. This result is illustrated in Fig. 7.5 where the investment strategy presented in Table 7.8 is evaluated using different discount rates. This is because future inflows are given less weight. This result however does not hold anymore if investment costs are observed all along the project's life, which is often the case with long-run public projects (planned or extraordinary maintenance, reinvestment, negative residual value, etc.). In that context, multiple *FIRR* can be found, which makes the approach useless. The *FNPV* on the other hand is always computable which renders it more suitable in many occasions.

The profitability of a project can be assessed from the point of view of the investment ($FNPV_I$ hereafter) or the capital ($FNPV_K$). Figure 7.6 describes the two types of analysis. The net present value of the investment ($FNPV_I$) assesses the ability of operating net revenues to sustain the investment costs, regardless financing and tax. As can be seen from Fig. 7.6a, the approach simply consists in

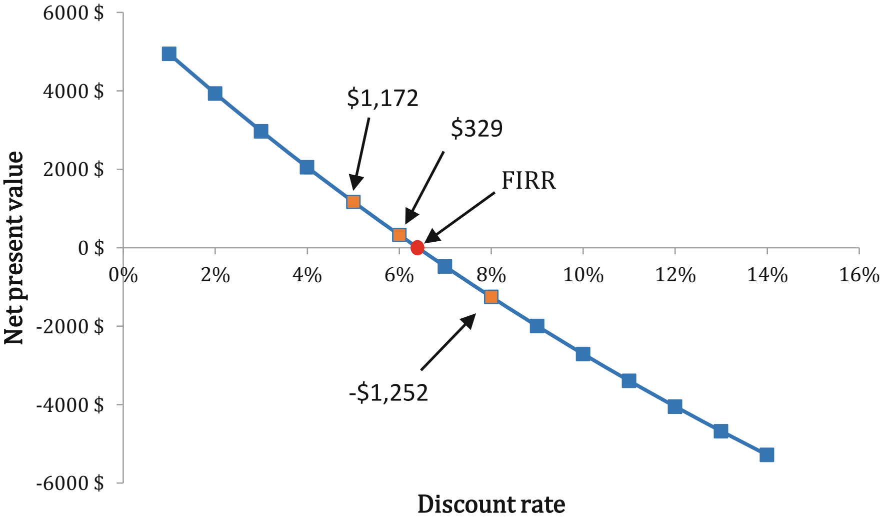
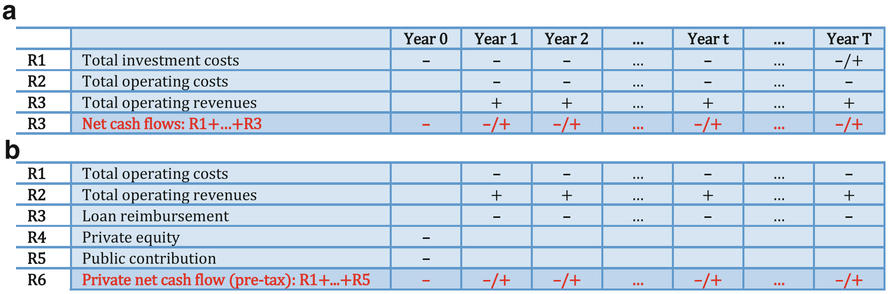


Fig. 7.5 Net present value and the discount rate: example 5


a

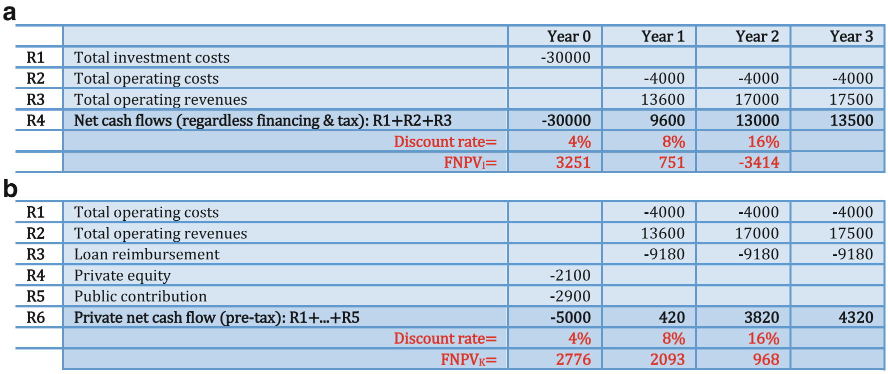
		Year 0	Year 1	Year 2	...	Year t	...	Year T
R1	Total investment costs	–	–	–	...	–	...	–/+
R2	Total operating costs		–	–	...	–	...	–
R3	Total operating revenues		+	+	...	+	...	+
R3	Net cash flows: R1+...+R3	–	–/+	–/+	...	–/+	...	–/+

b

R1	Total operating costs		–	–	...	–	...	–
R2	Total operating revenues		+	+	...	+	...	+
R3	Loan reimbursement		–	–	...	–	...	–
R4	Private equity	–						
R5	Public contribution	–						
R6	Private net cash flow (pre-tax): R1+...+R5	–	–/+	–/+	...	–/+	...	–/+

Fig. 7.6 Verification of the financial profitability. (**a**) Returns on investment—thousands of dollars. (**b**) Returns on capital—thousands of dollars

comparing the investment costs (row R1) with the net operating revenues (R2+R3). The approach is slightly more complex with respect to the profitability of the capital ($FNPV_K$). Here, the objective is to examine the project performance from the perspective of the external contributors, whether they are public or private. Following the European Commission methodology, the private net cash flow is defined before tax (see Fig. 7.6b). It is computed as the difference between the net operating revenues (R1+R2) minus the loan reimbursement (R3) and the external contributions (R4+R5). There are variations in this method, however, in that the profit may be also examined after tax. In that case, since depreciation is a deductible



a

		Year 0	Year 1	Year 2	Year 3
R1	Total investment costs	-30000			
R2	Total operating costs		-4000	-4000	-4000
R3	Total operating revenues		13600	17000	17500
R4	**Net cash flows (regardless financing & tax): R1+R2+R3**	**-30000**	**9600**	**13000**	**13500**
	Discount rate=	4%	8%	16%	
	$FNPV_I$=	3251	751	-3414	

b

R1	Total operating costs		-4000	-4000	-4000
R2	Total operating revenues		13600	17000	17500
R3	Loan reimbursement		-9180	-9180	-9180
R4	Private equity	-2100			
R5	Public contribution	-2900			
R6	**Private net cash flow (pre-tax): R1+...+R5**	**-5000**	**420**	**3820**	**4320**
	Discount rate=	4%	8%	16%	
	$FNPV_K$=	2776	2093	968	

Fig. 7.7 Profitability analysis: example 4. (**a**) Returns on investment—thousands of dollars. (**b**) Returns on capital—thousands of dollars

expense for profit tax purposes, the results are heavily influenced by the depreciation method.

Let us now consider again example 4. Figure 7.7 examines the profitability of the project using information from Table 7.6. The first part of the table compares the investment costs of $30 million with the net operating costs (9.6, 13 and 13.5 million dollars). There is no information about the capital and loan structure as the focus is on the profitability of the investment. The financial net present value is displayed for three discount rates: 4%, 8% and 16%, respectively. When the discount rate is 4%, we have:

$$FNPV_I(4\%) = -30,000 + \frac{9600}{1.04} + \frac{13,000}{1.04^2} + \frac{13,500}{1.04^3} \approx 3251 \text{ thousand dollars}$$

It measures the extent to which the project net revenues are able to repay the investment, regardless of the sources or methods of financing. At this discount rate, the returns on investment are positive. When the discount rate is 16%, we have instead:

$$FNPV_I(16\%) = -30,000 + \frac{9600}{1.16} + \frac{13,000}{1.16^2} + \frac{13,500}{1.16^3} \approx -3414 \text{ thousand dollars}$$

At a 16% rate, the project is not attractive anymore.

The second part of Fig. 7.7 assesses whether the public and private investors will be willing to participate in the project. The focus is on row R6 where the private and public contributions, as well as the reimbursement of the loan, have been taken into account to compute the private net cash flow. The external contributions now appear with a negative sign as they are expenditures from the point of view of investors.

7.5 Real Versus Nominal Values

Inflation denotes an increase in the average price level, which thereby reduces the purchasing power of the money in question. When the price level rises, each unit of currency buys fewer goods and services. This effect is of high importance in project appraisal. Assume for instance that one earns \$1000 in 20 years. The inflation observed during this period is going to make this amount worth less. As such, the discount factors should be adjusted, to remove the effect of price changes. This is why, in practice, a nominal interest rate is used when cash flows are expressed in terms of their current value.

Inflation, discount and interest rates are closely related. Formally, the equation that links nominal and real interest rates is the following:

$$\left(1 + r^{\text{nominal}}\right) = \left(1 + r^{\text{real}}\right)(1 + \pi)$$

or equivalently:

$$r^{\text{nominal}} = r^{\text{real}} + \pi + \left(r^{\text{real}} \times \pi\right)$$

where r^{nominal} and r^{real} stand for the nominal and real interest rates, respectively. The inflation rate is denoted by π. It is the expected average annual rate of increase in the price of goods. Assume for example that the (constant) annual inflation rate is 5%. What is costing \$100 on average today will cost \$100(1+5%) =\$105 next year, and $\$100(1+5\%)^2 = \$110.25$ the year afterwards, etc.

If the nominal interest rate and inflation rate are sufficiently low, the previous formula can be approximated as follows:

$$r^{\text{nominal}} \approx r^{\text{real}} + \pi$$

Nominal rates thus comprise two components: a portion that represents expected inflation (known as the inflation premium) and a portion that represents the real rate of return. For example, with a nominal interest rate of 5% and an expected inflation rate of 2%, the real rate of interest is 3% approximately.

The existence of inflation raises the question of whether the analysis requires a nominal or a real discount rate. The rule is simple: if cash flows are measured in nominal (or current) terms, then they should be discounted with a nominal discount rate. If they are expressed in real (or constant) terms, they should be discounted with a real discount rate.

To obtain values expressed in real terms, one needs to adjust for changes in prices level, i.e. to "deflate" the cash flows. The consumer price index (CPI) is commonly used in this purpose. However, it may be relevant to use more specific indices (e.g., the medical care component of the CPI for a public health program), depending on the nature of the project. Generally speaking, the CPI measures changes in the price level of a representative basket of consumer goods and services purchased by households, multiplied by 100. It is defined by a base year for which the index is equal to 100 and different values for the following years. To convert cash flows in real terms for the base year, one simply divides each cash flow by the CPI for that year:

$$CF_t^{\text{real}} = \frac{CF_t^{\text{nominal}}}{CPI_t} \times 100$$

In this particular situation, the CPI is also termed a "deflator". Assuming that prices rise at the same rate π during inflationary periods, the previous expression can also be defined in terms of the base year only:

$$CF_t^{\text{real}} = \frac{CF_t^{\text{nominal}}}{(1+\pi)^t}$$

It is important to understand that we are not computing a present value. Instead, we are converting nominal values to real values by relating the future sum of money to the purchasing power observed at year 0. The approach is thus different from discounting.

Real interest rates better reflect the real return to a lender and the real cost to a borrower and are thus more relevant than nominal rates for economic decisions. However, with respect to net present value computations, both approaches will yield the same result. To illustrate, consider a future cash flow expressed in real terms CF_t^{real} at year t. The present value is computed as $CF_t^{\text{real}}/\left(1+r^{\text{real}}\right)^t$. Now, the inflated value of the cash flow is computed as $CF_t^{\text{nominal}} = CF_t^{\text{real}}(1+\pi)^t$. Using the nominal rate to compute the net present value of the nominal cash flow, we obtain:

$$\frac{CF_t^{\text{nominal}}}{\left(1+r^{\text{nominal}}\right)^t} = \frac{CF_t^{\text{real}}(1+\pi)^t}{\left(1+r^{\text{nominal}}\right)^t} = \frac{CF_t^{\text{real}}}{\left(1+r^{\text{real}}\right)^t}$$

The approaches are strictly equivalent. What should be retained here is that when the analysis is carried out at current prices (resp. constant), then the discount rate should be expressed in nominal terms (resp. real). Moreover, when the inflation rate is unstable through time, then the discount rate must be modified accordingly to compute the relevant discount factors.

Table 7.9 illustrates the approach through a simple example. The base year for the CPI is year 0. The inflation rate is assumed to be $\pi = 2\%$ over the whole period. This implies that the CPI for year 1 is 102, then 104 at year 2, and so on. Because of inflation, any revenue obtained in the future is worth less than if it were obtained in year 0. The third column describes the cash flows of a project under evaluation. Those cash flows are expressed in current prices. In the fourth column, they are converted in real terms. Each value is obtained by dividing the nominal values (multiplied by 100) by the CPI for the relevant period. Those values are lower because the effect of inflation has been accounted for.

Table 7.9 Nominal and real cash flow: example 6

Year	CPI	Nominal cash flow	Real Cash flow
0	100	–$100,000	–$100,000
1	102.0	$30,000	$29,412
2	104.0	$30,000	$28,835
3	106.1	$30,000	$28,270
4	108.2	$30,000	$27,715
5	110.4	$30,000	$27,172
6	112.6	$30,000	$26,639
7	114.9	$30,000	$26,117
8	117.2	$30,000	$25,605
9	119.5	$30,000	$25,103
10	121.9	$30,000	$24,610
	Discount rate=	**5.06%**	**3%**
	FNPV=	**$130,979**	**$130,979**

In Table 7.9, the nominal and real discount rates are set to 5.06% and 3%, respectively. Those values are in accordance with the inflation rate observed over the period. We have $5.06\% = 3\% + 2\% + (3\% \times 2\%)$. If the cash flows are expressed in nominal terms, then the nominal rate should be used accordingly:

$$FNPV(5.06\%) = -\$100,000 + \frac{\$30,000}{1.0506} + \ldots + \frac{\$30,000}{1.0506^{10}} = \$130,979$$

If, instead, one prefers to express the cash flow in real terms, we have:

$$FNPV(3\%) = -100,000 + \frac{29,412}{1.03} + \ldots + \frac{24,610}{1.03^{10}} = \$130,979$$

The results are equivalent.

Why the trouble of deflating cash flows if the approaches yield similar results? In the private sector, it is more common to work in values expressed in nominal terms. The use of current prices places the study in actual values, thereby making easier the planning of annual budgets. For the analysis of public policy projects, however, the use of real terms has the advantage to facilitate the comparison of cash flows, especially when the projects take place in different countries facing their own inflation rate. It is up to the evaluator to decide whether the benefits of using constant prices are worth the trouble.

7.6 Ranking Investment Strategies

Several financial techniques can be used to rationalize investment decisions. They include indicators such as the net present value, the accounting rate of return, the payback period, the discounted payback period, the break-even point or the break-even sales. They are essential to the evaluation of public works, especially when the authority in question does not have sufficient funds to undertake several projects at the same time. Table 7.10 provides a general definition of these items as well as the context in which they are used. Those methods are successively detailed below.

First, as already stated, discounted cash flow analysis allows competing strategies to be compared according to their financial profitability. The conclusions are summarized by indicators such as the *FNPV*, which evaluates the return of a project by comparing the inflows and outflows that result from its implementation. Consider for example Table 7.11 where four competing strategies are evaluated, each of the same time length. The indicator of profitability is provided at the bottom of the table. For a discount rate set to 4%, one would rank first strategy S_2, then strategy S_3, S_1 and S_4. As can be seen, strategy S_4 yields a negative net present

Table 7.10 Overview of performance indicators

Indicator	Acronym	Definition	Context
Financial net present value	*FNPV*	Total amount of gain or loss a project will produce in terms of present value	Profitability
Accounting rate of return	*ARR*	Average return on investment	Profitability
(Discounted) payback period	(*D*)*PB*	Number of periods for the cumulated net cash flows (expressed in their present value) to equal the initial investment	Risk in terms of opportunity cost
Break-even point	*BEP*	Output required to cover all fixed expenses	Risk in terms of sustainability
Break-even sales	*BES*	Amount of sales required to cover all fixed expenses	Risk in terms of sustainability

Table 7.11 Discounted cash flow analysis: example 7

Year	Cash flow statement			
	Strategy S_1	Strategy S_2	Strategy S_3	Strategy S_4
0	–$140,000	–$140,000	–$89,000	–$130,000
1	$17,500	$15,000	$11,400	$16,000
2	$17,500	$15,600	$11,400	$16,000
3	$17,500	$16,224	$11,400	$16,000
4	$17,500	$16,873	$11,400	$16,000
5	$17,500	$17,548	$11,400	$16,000
6	$17,500	$18,250	$11,400	$16,000
7	$17,500	$18,980	$11,400	$16,000
8	$17,500	$19,739	$11,400	$16,000
9	$17,500	$20,529	$11,400	$16,000
10	$17,500	$21,350	$11,400	$16,000
Discount rate=	**4%**	**4%**	**4%**	**4%**
FNPV=	**$1941**	**$4231**	**$3464**	**–$226**

value, which means that this strategy is not financially acceptable. Last, S_3 involves a lower investment outlay ($89,000) compared to strategy S_2 ($140,000). In that context, the evaluator must weight those pros and cons and take into account the budget constraint of the public authority.

The accounting rate of return (*ARR*) measures the extent to which the return on investment compensates the initial outlay. It expresses the average accounting profit as a percentage of the investment cost. The measure depends on the depreciation method, usually a straight-line method. In our setting, we have:

$$ARR = \frac{\text{Average accounting profit}}{\text{Initial investment cost}}$$

The average accounting profit is defined as:

$$\text{Average accounting profit} = \text{Average net cash flows} - \text{Depreciation per annum}$$

Consider again example 7 (Table 7.11). Strategy S_1 is characterized by an initial outlay of $CF_0 = -\$140{,}000$. As there is no residual value, if one uses a straight line depreciation method to write off this cost (see Sect. 2), then the depreciation per annum amounts to $CF_0/T = -\$14{,}000$. Moreover, strategy S_1 generates annual net cash flows of $\sum_{t=1}^{T} CF_t/T = \$17,500$, which finally yields an average accounting profit of $17,500–$14,000=$3500. The accounting ratio of return is thus computed as ($17,500–$14,000)/$140,000=2.50% (see Table 7.12). Using this criterion, strategy S_2 offers the highest return (2.86%). Strategy S_3 then comes second (2.81%). The profit potential of the different competing strategies is easily compared. The approach, however, does not account for the time value of money. Any return observed at the end of the project's life is worth as much as those observed earlier.

The payback method assesses the time it takes for a strategy to earn back the money initially invested. There are two cases depending upon whether the net cash flows are identical or not over the project's life. When they are equally distributed over the project's life ($CF_1 = \ldots = CF_T = CF$), the following formula is used:

$$PB = \frac{\text{Initial investment cost}}{\text{Annual net cash flow}} = \frac{-CF_0}{CF}$$

The payback period *PB* is obtained by dividing the initial investment outlay by the periodic cash inflow observed in the following periods. For projects with unequal cash flows, a similar indicator is computed:

$$PB = (\tau - 1) + \frac{-\sum_{t=0}^{\tau-1} CF_t}{CF_\tau}$$

Table 7.12 Accounting rate of return: example 7

	Strategy S_1		Strategy S_2		Strategy S_3		Strategy S_4	
Year	Depreciation	Profit	Depreciation	Profit	Depreciation	Profit	Depreciation	Profit
1	–$14,000	$3500	–$14,000	$1000	–$8900	$2500	–$13,000	$3000
2	–$14,000	$3500	–$14,000	$1600	–$8900	$2500	–$13,000	$3000
3	–$14,000	$3500	–$14,000	$2224	–$8900	$2500	–$13,000	$3000
4	–$14,000	$3500	–$14,000	$2873	–$8900	$2500	–$13,000	$3000
5	–$14,000	$3500	–$14,000	$3548	–$8900	$2500	–$13,000	$3000
6	–$14,000	$3500	–$14,000	$4250	–$8900	$2500	–$13,000	$3000
7	–$14,000	$3500	–$14,000	$4980	–$8900	$2500	–$13,000	$3000
8	–$14,000	$3500	–$14,000	$5739	–$8900	$2500	–$13,000	$3000
9	–$14,000	$3500	–$14,000	$6529	–$8900	$2500	–$13,000	$3000
10	–$14,000	$3500	–$14,000	$7350	–$8900	$2500	–$13,000	$3000
	Profit=	**$3500**	**Profit=**	**$4009**	**Profit=**	**$2500**	**Profit=**	**$3000**
	Investment=	**$140,000**	**Investment=**	**$140,000**	**Investment=**	**$89,000**	**Investment=**	**$130,000**
	ARR=	**2.50%**	**ARR=**	**2.86%**	**ARR=**	**2.81%**	**ARR=**	**2.31%**

where τ is the period in which the initial investment is fully recovered; $(\tau - 1)$ denotes the number of periods before full recovery; $-\sum_{t=0}^{\tau-1} CF_t$ represents the remaining uncovered cost in period τ; CF_τ is the net cash flow observed in period τ.

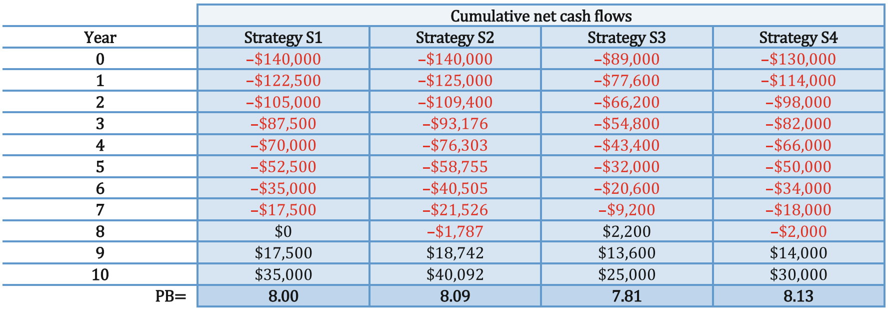
The payback period method provides a measure of a project's riskiness. It indicates the number of periods required for the proceeds of the project to recoup the original investment outlay. The approach is particularly useful under capital rationing, when the budget must be prioritized, or under the threat of expropriation. It also has the advantage to be easily understood, and can be used to show the time commitment of funds. The lower is PB, the sooner the initial cost can be recovered, and the sooner the inflows can be reinvested in new projects. This indicator of performance, however, does not measure profitability. For instance, any inflow received beyond the payback period is not considered. The method may thus cause rejection of a highly profitable source of earnings if those earnings are generated at the end of the project's life. Another issue is that this approach ignores the time value of money. The method can however be extended using cash flows expressed in their present values.

Consider for instance Fig. 7.8. Strategies S_1, S_3 and S_4 generate equally distributed cash flows over the projects' life. The payback periods can be computed as follows:

$$PB(S_1) = \frac{\$140,000}{\$17,500}; PB(S_3) = \frac{\$89,000}{\$11,400}; PB(S_4) = \frac{\$130,000}{\$16,000}$$

The values obtained (respectively, 8.00, 7.81 and 8.13) represent the number of years it takes to recover the initial investment.

The method is slightly more complex for strategy S_2 since the net cash flows are unequally distributed. One needs to determine the cumulative cash flows. The year of full recovery is $\tau = 9$ for strategy S_2. At year $\tau - 1 = 8$, the uncovered costs amounts to $-\sum_{t=0}^{8} CF_t = \$1787$. From Table 7.11, the net cash flow at year 9 amounts to $CF_9 = \$20,259$. Hence, we have:



	Cumulative net cash flows			
Year	Strategy S1	Strategy S2	Strategy S3	Strategy S4
0	-$140,000	-$140,000	-$89,000	-$130,000
1	-$122,500	-$125,000	-$77,600	-$114,000
2	-$105,000	-$109,400	-$66,200	-$98,000
3	-$87,500	-$93,176	-$54,800	-$82,000
4	-$70,000	-$76,303	-$43,400	-$66,000
5	-$52,500	-$58,755	-$32,000	-$50,000
6	-$35,000	-$40,505	-$20,600	-$34,000
7	-$17,500	-$21,526	-$9,200	-$18,000
8	$0	-$1,787	$2,200	-$2,000
9	$17,500	$18,742	$13,600	$14,000
10	$35,000	$40,092	$25,000	$30,000
PB=	8.00	8.09	7.81	8.13

Fig. 7.8 Payback period method: example 7

$$PB(S_2) = 8 + \frac{\$1787}{\$20,259} = 8.09$$

Overall, strategy S_3 is the strategy that yields the shortest payback period, then come S_1, S_2 and S_4.

Break-even analysis is another method for isolating investment risk. It is used to determine the volume of activity that is needed to cover the related costs. The approach is particularly relevant for evaluating self-financing projects, when sales revenue and user charges should ideally cover the project's costs. The method consists in dividing all costs into fixed and variable cost categories. By definition, a fixed cost is any expense that remains constant regardless of the level of output. Those costs occur on a periodic basis irrespective of whether the production takes place. Examples include rent, insurance premiums, and loan payments. Variable costs are expenses that fluctuate directly with changes in the level of output. They occur only when activity takes place. They are usually expressed as cost per unit. Examples include labor costs, material and packaging.

The break-even methodology is strongly related to the concept of sustainability. Formally, a break-even point is computed as:

$$BEP = \frac{\text{Fixed costs per year}}{\text{User charge} - \text{Variable cost per unit}}$$

This ratio represents the number of units that must be sold to make no profit or no loss. In this expression, the denominator is called the contribution. It is the amount each sold unit contributes towards profit. Alternatively, we can compute the break-even sales:

$$BES = BEP \times \text{User charge}$$

Assume for instance that the fixed costs amount to \$100,000 per year. If the variable cost per unit is \$2 while the user charge is \$10, then the break-even point is:

$$BEP = \frac{\$100,000}{\$10 - \$2} = 12,500 \text{ units}$$

It is the minimum number of units that must be sold each year on average. If the number of units sold is lower than 12,500, then the project will make a loss. This equivalently means that the sales should amount at least to $BES = 12{,}500 \times \$10 = \$125{,}000$ per year.

Let us now illustrate the approach for strategy S_2 (example 7). On average, let us assume a variable cost per unit of \$0.2 and fixed operating costs of \$1000 per year. The user charge is set to \$2. To compute the break-even point, loan payments must be included as a fixed cost. Imagine that the initial outlay (\$140,000) has been borrowed from the bank at an interest rate of 3%. The *PMT* formula from Excel can be used to compute the periodic payment (see Sect. 7.2). We have:

$$F = \$140,000 \times \frac{3\%}{1-(1+3\%)^{-10}} = \$16,412$$

This yields a break-even point equal to:

$$BEP(S_2) = \frac{\$1,000 + \$16,412}{\$2 - \$0.2} = 9673$$

The break-even sales amount to:

$$BES(S_2) = BEP(S_2) \times \$2 = \$19,347$$

A similar approach can be used for the other strategies.

7.7 Sensitivity Analysis

The values included in a cash flow statement are estimated based on the most probable forecasts (also known as the most-likely scenario). Their choice can be influenced by many factors and the assumptions made can be subject to error. Sensitivity analysis investigates those assumptions by assessing the effects of a variation in one or several variables on the project's key indicators. For instance, a sensitivity analysis can make use of different discount rates to check how the results are critical to the value of time. Each sensitive variable such as sales can also be changed to assess the investment's desirability.

Several types of sensitivity analysis exist. The easiest approach is to examine the most-likely scenario and consider the impact of changes in project variables individually. Ideally, the analysis examines all the independent variables of the project. Any aggregated variable such as total fixed assets or total operating expenditures are usually disregarded. In that context, the European Commission advocates the computation of what are termed switching values and elasticities. By definition a switching value is the value at which a project becomes acceptable. In other words, it is the value that a variable would have to take in order for the *FNPV* of the project to become zero. For simplicity of exposition, this value is expressed in terms of percentage change, based on the initial scenario. For instance, a 20% increase in the cost of raw materials may reduce the project's profitability to zero. If it takes a 60% increase in the cost of labor to achieve the same target, one would conclude that the cost of raw materials is a more sensitive variable.

The calculation of switching points can be done using trial and error. To illustrate, let us consider again the bridge project (example 4) and the profitability of the investment ($FNPV_I$ in Table 7.6). Assume that the discount rate is equal to 4%. Starting from the initial cash flow statement, one varies each variable independently from the others until one reaches the required threshold, i.e. $FNPV_I = 0$. To illustrate, the method is used to compute the switching value of raw materials (see row R7 of Fig. 7.9). The percentage increase required to zero out the profitability of a

		Year 0	Year 1	Year 2	Year 3
R1	Lands	–7000			
R2	Bridge infrastructure	–15000			
R3	Equipment	–4000			
R4	Start-up costs	–1500			
R5	Road network	–2500			
R6	**Total investment costs: R1+...+R5**	**–30000**	**0**	**0**	**0**

b

R7	Raw materials × (1+52.07%)		–3422	–3422	–3422
R8	Labor		–750	–750	–750
R9	Electric power		–300	–300	–300
R10	Maintenance		–450	–450	–450
R11	Administrative costs		–80	–80	–80
R12	Sales expenditures		–170	–170	–170
R13	**Total operating costs: R7+...+R12**	**0**	**–5172**	**–5172**	**–5172**
R14	Sales		13600	17000	17500
R15	**Total operating revenues: R14**	**0**	**13600**	**17000**	**17500**
R16	**Net operating revenues: R13+R15**	**0**	**8428**	**11828**	**12328**
R17	**Net cash flows (regardless financing & tax): R6+R16**	**–30000**	**8428**	**11828**	**12328**
	Discount rate=	4%			
	FNPV$_I$=	0			

Fig. 7.9 Switching value of raw materials: example 4. (**a**) Total Investments—thousands of dollars. (**b**) Operating revenues and costs—thousands of dollars

a

		Year 0	Year 1	Year 2	Year 3
R1	Lands	–7000			
R2	Bridge infrastructure	–15000			
R3	Equipment	–4000			
R4	Start-up costs	–1500			
R5	Road network	–2500			
R6	**Total investment costs: R1+...+R5**	**–30000**	**0**	**0**	**0**

b

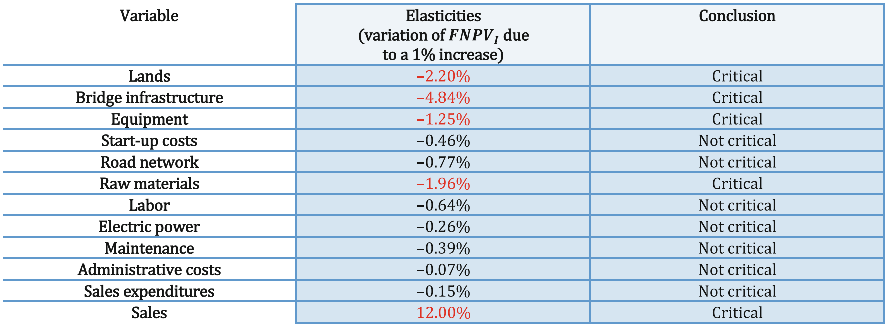
R7	Raw materials		–2250	–2250	–2250
R8	Labor		–750	–750	–750
R9	Electric power		–300	–300	–300
R10	Maintenance		–450	–450	–450
R11	Administrative costs		–80	–80	–80
R12	Sales expenditures		–170	–170	–170
R13	**Total operating costs: R7+...+R12**	**0**	**–4000**	**–4000**	**–4000**
R14	Sales × (1-7.33%)		12603	15754	16217
R15	**Total operating revenues: R14**	**0**	**12603**	**15754**	**16217**
R16	**Net operating revenues: R13+R15**	**0**	**8603**	**11754**	**12217**
R17	**Net cash flows (regardless financing & tax): R6+R16**	**–30000**	**8603**	**11754**	**12217**
	Discount rate=	4%			
	FNPV$_I$=	0			

Fig. 7.10 Switching value of sales: example 4. (**a**) Total Investments—thousands of dollars. (**b**) Operating revenues and costs—thousands of dollars

the investment is 52.07%. In Fig. 7.9, we have $3422 = 2250 \times (1 + 52.07\%)$ where 2250 represents the value initially described in the most-likely scenario (Table 7.6). Similarly, we can see in Fig. 7.10 that if the sales decrease by more than 7.33%, then the project falls below the minimum level of acceptability. Table 7.13 extends the approach to other variables. It can be seen that the most sensitive variables are the value of sales, the cost of infrastructure and raw materials. Administrative costs and sales expenditures appear as the less sensitive ones.

Table 7.13 Sensitivity analysis and switching values: example 4

Variable	Switching values
	Maximum increase before the *FNPV* equals 0
Lands	46.45%
Bridge infrastructure	21.68%
Equipment	81.29%
Start-up costs	216.75%
Road network	130.05%
Raw materials	52.07%
Labor	156.20%
Electric power	390.50%
Maintenance	260.40%
Administrative costs	1464.50%
Sales expenditures	689.30%
	Maximum decrease before the *FNPV* equals 0
Sales	−7.33%



Variable	Elasticities (variation of $FNPV_I$ due to a 1% increase)	Conclusion
Lands	−2.20%	Critical
Bridge infrastructure	−4.84%	Critical
Equipment	−1.25%	Critical
Start-up costs	−0.46%	Not critical
Road network	−0.77%	Not critical
Raw materials	−1.96%	Critical
Labor	−0.64%	Not critical
Electric power	−0.26%	Not critical
Maintenance	−0.39%	Not critical
Administrative costs	−0.07%	Not critical
Sales expenditures	−0.15%	Not critical
Sales	12.00%	Critical

Fig. 7.11 Sensitivity analysis and elasticities: Example 4

Once the most sensitive variables are identified, one can also calculate the effect of possible changes in these variables (usually from 1% to 15%) on the *FNPV*. As a matter of fact, the European Commission suggests calculating the effect of a 1% increase in each of the variables on the *FNPV*. The recommendation is to consider critical those variables for which such a variation gives rise to a variation of more than 1% in the value of the *FNPV*. Figure 7.11 presents the different elasticities computed for the bridge project. An "elasticity" is defined as the percentage change in the FNPV indicator for a 1% change in the variable. As previously, the sensitivity analysis shows that the amount of sales, the cost of lands, infrastructures, equipment and raw materials constitute critical variables.

Sensitivity tests can be misleading, especially when variables are correlated to each other. For instance, when preparing a cash flow statement, one may consider forecasts about sales revenues. Those sales may generate additional expenses, directly linked to the selling activity (sales expenditures), but also indirectly, for instance via maintenance actions. In that context, the basic techniques of varying one variable at a time, keeping the other variables constant, becomes unjustified. Instead, it is necessary to explore a change in a combination of variables. Several techniques have proven to be very helpful in this respect. Examples include scenario analysis and probabilistic analysis. A scenario analysis generally focuses on three alternatives: a most-likely scenario, a best-case scenario, and a worst-case scenario. The purpose is not to identify the exact conditions of each scenario but, instead, to draw attention to the main uncertainties involved in the project. A probabilistic analysis generalizes this approach to a continuous setting where each variable can take any value with some likelihood. The approach, also known as Monte Carlo analysis, examines those variations simultaneously and simulates thousands of scenarios, which results in a range of possible net present values with their probability of occurrence. Those approaches are presented in the next chapters.

Bibliographical Guideline

The concept of the time value of money is described in many textbooks. We can name in particular that of Van Horne and Wachowicz (2008). The book describes the basic principles of financial management and provides the reader with information about many topics in finance. Additionally, the reader can refer to Pirnot (2014), which introduces the main principles of consumer mathematics, including simple and compound interests, consumer loans, annuities and amortization.

Discounted cash flow analysis is fully described in Campbell and Brown (2003). The book closely integrates the theory and practice of cost benefit analysis using a spreadsheet framework. The reader can also refer to Boardman et al. (2010), which offers a practical introduction to cost benefit analysis. Furthermore, many textbooks provide a description of ranking methods such as the accounting rate of return, the payback period, or the break-even point. These references include for instance Alhabeeb (2014) and Ahuja et al. (2015).

Several institutional guides present the financial practices to be used in the context of public program evaluation. We may cite in particular the Guide to cost benefit analysis of investment projects of the European Commission. This guide provides a detailed presentation of the concept of sustainability, profitability, and sensitivity analysis. Additionally, the Project appraisal practitioners' guide by USAID, a US government agency, offers a review of the international best practices of project appraisal while approving a public sector project.

References

Ahuja, N. L., Dawar, V., & Arrawatia, R. (2015). *Corporate finance*. PHI Learning.

Alhabeeb, M. J. (2014). *Entrepreneurial finance: Fundamentals of financial planning and management for small business*. New York: Wiley.

Boardman, A., Greenberg, A., Vining, A., & Weimer, D. (2010). *Cost benefit analysis. The Pearson series in economics* (4th ed.). Old Tappan: Prentice Hall.

Campbell, H., & Brown, R. (2003). *Benefit-cost analysis: Financial and economic appraisal using spreadsheets*. Cambridge Books, Cambridge University Press: Cambridge.

European Commission. (2014). *Guide to cost-benefit analysis of investment projects. Economic appraisal tool for cohesion policy 2014-2020*.

Pirnot, T. (2014). *Mathematics all around*. Boston: Pearson.

USAID (2008). *The project appraisal practitioners' guide*. USAID/India Reform Project Compendium.

Van Horne, C., & Wachowicz, J. M. (2008). *Fundamentals of financial management*. Harlow: Prentice Hall.

Budget Impact Analysis

8

8.1 Introducing a New Intervention Amongst Existing Ones

Budget impact analysis examines the extent to which the introduction of a new strategy in an existing program affects an agency's budget. The focus is on the outcome achievement associated with the implementation of the program and the expected mid-run budget burden. Not only does the method provide information about the costs generated by a new intervention or treatment, but it also assesses how the new strategy will affect the overall supply of services and the amount of resources devoted to it. The approach may serve for instance to evaluate the impact of a new drug on the health care system, or be part of a budget planning process in order to analyze multiple scenarios. The adoption of the new strategy will affect positively or negatively the demand for other types of interventions and, thereby, will modify the costs associated with their supply. Overall, budget impact analysis provides a general tool for anticipating future changes in public expenditures associated with the launching of a new project.

The context of the analysis is a public or publicly monitored program providing services through existing supply to which the new strategy is going to be added. The initial supply frame consists of one or several mutually exclusive interventions that are already in effect. This initial framework is labeled the "current environment". The introduction of a new strategy will reshuffle demand and supply patterns, and will characterize the "new environment". The aim of budget impact analysis is to evaluate the budget and outcome changes initiated by the introduction of the additional strategy in the program. Figure 8.1 illustrates the approach. The adoption of a new strategy modifies the way the demand is addressed and may also divert the demand from other types of intervention, which finally alters the way supply is handled. The difference expected between the new environment and the current one represents the total impact of the new strategy on the total budget.

Let us take a first stylized example of a school district in a remote and poor area where primary education is delivered by two existing village schools. Due to travel costs and the lack of public transportation, those schools fail to reach children from

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_8

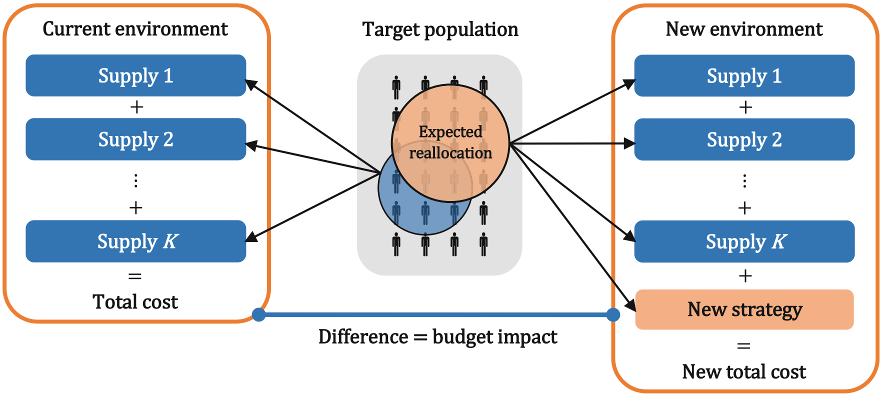


Fig. 8.1 Budget impact analysis

remote places. Public authorities intend to introduce a third one in this educational landscape, a district school with publicly organized transportation from home place to school. Will that arrangement attract yet out-of-school pupils? Will there be shift in demand from village schools to the district facility? What will the budget impact be for the public authority? The second example deals with the more widespread use of budget impact analysis in the field of public health. Admission to reimbursement of a new drug in addition to existing ones is likely to reshape the management of the disease and the ensuing budget burden. In the first example, the subjects of interest are school-age children in the district, the outcome targets are the children attracted to primary education. The suppliers of the service are the schools, from the two villages or from the district facility. In the second example, the population of interest consists of the people who have a disease likely to be treated by either the existing set of drugs or the new one. Suppliers are the drug manufacturers. The outcome target is the number of treated patients. A sound budget analysis should allow outcomes and the corresponding budget loads to be disaggregated by supplier.

Details of demand forecast and cost structure need to be provided. As a general principle, this information is gathered for a rather short time horizon, usually middle-run from 3 to 5 years. Cash flows are expressed in nominal value terms and, as a matter of fact, are not discounted. The project intends to add a new strategy to the set of strategies already in place. As a consequence, demand is rearranged. Variations may concern the flows of (1) incoming subjects, i.e. those attracted by the new strategy (for instance, children now carried by public transportation and sent to district school or patients eligible to the new drug) and (2) of out-going subjects because the new strategy may generate less dropouts among schoolchildren or less forced termination of treatment due to harmful side-effects. As was mentioned before, budget impact analysis restricts its span to the usual mid-run budget planning horizon. That fits the budget holder's constraints and objective but is not meant necessarily to reflect the situation of the users. For instance, in the management of chronic diseases, the patient's horizon is their life expectancy. Conversely, an educational program may extend over a shorter horizon, in which case the first generations of users would have left it before the budget impact horizon is reached.

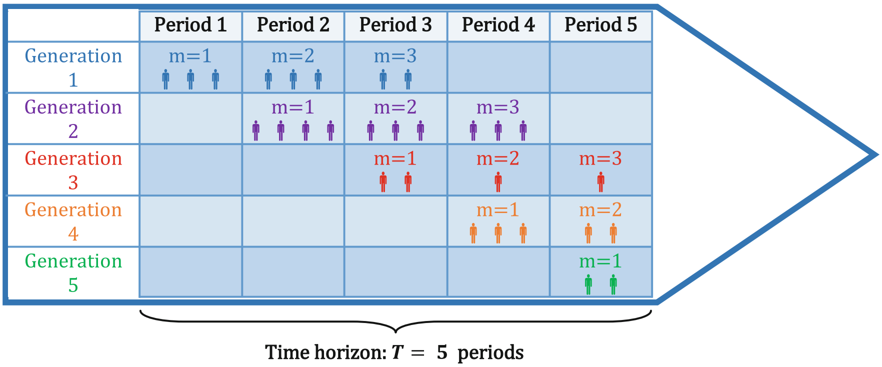
Finally, let us point out that budget impact analysis does not provide a decision rule: financial sustainability and the extent to which outcome targets are reached are to be appraised by the decision-maker according to their objectives and constraints. The method is thus descriptive rather than prescriptive.

The present chapter offers a simple analytical framework with the aim of presenting the general scheme of the method. Despite its simplicity, the method must be implemented with caution. Many difficulties may arise, especially when estimating demand flows. Demand projections are context-dependent and may require significant technical expertise (in education, medicine, epidemiology, sociology, demography, etc.). To illustrate, let us take the case of a health program. The first step consists in defining the "total population", i.e. those who live in the jurisdiction where the program is implemented. The "affected population" comprehends both the prevalent subjects from the previous periods that remain with the disease, and the incident subjects. Prevalence is the proportion of the total population that is affected with the disease in question at a specific time. Incidence is the rate of occurrence of new disease cases. Inclusion criteria applied to the affected population define the "eligible population", namely the population who is susceptible to treatment. Many factors also define exit rates of the eligible population from the interventions of the program (e.g., non-compliance, supply shortages, adverse events requesting treatment termination, general or disease-related mortality rate, migration to another jurisdiction). Building demand forecast is thus a step that should not be taken lightly, especially because it serves as the foundation for the rest of the analysis.

The remainder of the chapter is as follows. Sections 8.2 and 8.3 introduce the method. The analytical framework is first presented in the case of a single supply, e.g., one school, one drug or more generally one service or strategy (Sect. 8.2). The framework is then extended to several supplies and compares the current environment with a new one in which an additional strategy is added (Sect. 8.3). Section 8.4 provides a numerical example. Section 8.5 is dedicated to deterministic sensitivity analysis through the exploration of alternative scenarios of parameters values.

8.2 Analytical Framework

The purpose of budget impact analysis is to provide a simulation model that shows how the adoption of a new strategy will affect a budget, given that a specific range of services are already being supplied. The analytical framework is that of an overlapping-generations model. Figure 8.2 illustrates the mathematics behind this framework. The analysis is concerned with T generations of subjects, where T represents the time horizon of the budget impact analysis. Each generation is denoted $t = 1 \ldots T$. Those time periods are usually but not necessarily years. A



	Period 1	Period 2	Period 3	Period 4	Period 5
Generation 1	m=1 🧍🧍🧍	m=2 🧍🧍🧍	m=3 🧍🧍		
Generation 2		m=1 🧍🧍🧍🧍	m=2 🧍🧍🧍	m=3 🧍🧍🧍	
Generation 3			m=1 🧍🧍	m=2 🧍	m=3 🧍
Generation 4				m=1 🧍🧍🧍	m=2 🧍🧍
Generation 5					m=1 🧍🧍

Fig. 8.2 Cohorts of users over the budget impact horizon

strategy S_k is characterized by a set of intervention periods $m = 1 \cdots M(S_k)$ where M is the total number of intervention periods. We do not have necessarily $M < T$, or $M > T$, because the time horizon of the budget impact analysis, usually three to five years, is independent of the periods of intervention. In Fig. 8.2, for instance, the analysis is interested in $T = 5$ years and focuses on one single strategy. Each of the five generations benefit from $M = 3$ periods of intervention, denoted $m = 1$, 2 and 3 respectively.

As illustrated in Fig. 8.2, what matters is the flow of subjects who benefit from the interventions. Generations differ in their number and characteristics, which modifies the demand for the good or service in question. For instance, they are the children aged 6 enrolled for a 3-year educational program. The second year of implementation sees the coexistence of two generations of schoolchildren. The new one, aged 6, attends the school for their first-year program, along with their elders, now aged 7, who join the second-year program. Those generations of children aggregate themselves over the time horizon of the budget impact analysis. If that horizon is for instance five calendar years, the first generation will have left at the end of the third year, while at the end of the fifth year, the fifth generation will begin their first-year program, in parallel with the third generation (completing the program) and the fourth (in their second year). Note that the number of subjects in a given generation is not necessarily stable through time. For instance, a few pupils may have to repeat one school year, in which case they are treated as if they were belonging to the next generation. Some others may move with their parents to an area that is outside the scope of the considered program.

Formally, for a given strategy S_k, the quantity of good or service supplied to generation t is denoted by $x_t^m(S_k)$, where m denotes each intervention period. The x's can also refer directly to the number of subjects, in which case each subject is assumed to consume one single unit of the good. The supply frame is defined by a table as that of Fig. 8.3 made of T rows standing for the generations and T columns characterizing the time periods. We have $x_t^m(S_k) = 0$ either when $m > M(S_k)$ (the strategy has already been supplied) or when $m < 1$ (the strategy is not yet in effect).

	Time periods ($t = 1 \ldots T$)				
	Period 1	Period 2	Period 3	Period 4	Period 5
Generation 1	$x_1^1(S_k)$	$x_1^2(S_k)$	$x_1^3(S_k)$	0	0
Generation 2	0	$x_2^1(S_k)$	$x_2^2(S_k)$	$x_2^3(S_k)$	0
Generation 3	0	0	$x_3^1(S_k)$	$x_3^2(S_k)$	$x_3^3(S_k)$
Generation 4	0	0	0	$x_4^1(S_k)$	$x_4^2(S_k)$
Generation 5	0	0	0	0	$x_5^1(S_k)$
Total strategy S_k	$\boldsymbol{x_1(S_k)}$	$\boldsymbol{x_2(S_k)}$	$\boldsymbol{x_3(S_k)}$	$\boldsymbol{x_4(S_k)}$	$\boldsymbol{x_5(S_k)}$

Fig. 8.3 Supply from strategy S_k to overlapping generations

Figure 8.3 offers the example of a 5-year time horizon ($T = 5$). Strategy S_k is characterized by three intervention periods ($M = 3$). In period 1, the program involves an initial cohort or generation of users, which results in a total quantity supplied equal to $x_1^1(S_k)$. In period 2, the quantity supplied for that generation is $x_1^2(S_k)$. In the meantime, the second generation of users has arrived, generating an extra supply equal to $x_2^1(S_k)$. At time period 3, generation 1 completes the program, generation 2 enters the second year of intervention, generation 3 begins the program, and so on.

By using an intermediate time index i, we can express the annual supply at year t as:

$$x_t(S_k) = \sum_{i=1}^{t} x_i^{t+1-i}(S_k)$$

with $x_i^{t+1-i} = 0$ for $t + i - 1 > m$. Total supply over the whole horizon T is then defined as:

$$x(S_k) = \sum_{t=1}^{T} x_t(S_k) = \sum_{t=1}^{T} \sum_{i=1}^{t} x_i^{t+1-i}(S_k)$$

For instance, Fig. 8.3 is read column by column:

In period 1: $x_1(S_k) = \sum_{i=1}^{1} x_i^{2-i} = x_1^1$.

In period 2: $x_2(S_k) = \sum_{i=1}^{2} x_i^{3-i} = x_1^2 + x_2^1$.

In period 3: $x_3(S_k) = \sum_{i=1}^{3} x_i^{4-i} = x_1^3 + x_2^2 + x_3^1$.

In period 4: $x_4(S_k) = \sum_{i=1}^{4} x_i^{5-i} = 0 + x_2^3 + x_3^2 + x_4^1$.

In period 5: $x_5(S_k) = \sum_{i=1}^{5} x_i^{6-i} = 0 + 0 + x_3^3 + x_4^2 + x_5^1$.

The total supply is thus computed as follows:

	Time periods $t = 1 \ldots T$				
	Period 1	Period 2	Period 3	Period 4	Period 5
Generation 1	$uc^1(S_k)$	$uc^2(S_k)$	$uc^3(S_k)$	0	0
Generation 2	0	$uc^1(S_k)$	$uc^2(S_k)$	$uc^3(S_k)$	0
Generation 3	0	0	$uc^1(S_k)$	$uc^2(S_k)$	$uc^3(S_k)$
Generation 4	0	0	0	$uc^1(S_k)$	$uc^2(S_k)$
Generation 5	0	0	0	0	$uc^1(S_k)$

Fig. 8.4 Annual costs in supplying strategy S_k

$$x(S_k) = x_1(S_k) + x_2(S_k) + x_3(S_k) + x_4(S_k) + x_5(S_k)$$

Supply is thus conceived in an overlapping generation setting. Factors determining successive eligible demands for a given generation (inclusion criteria, compliance rate, discontinuation, etc.) are considered as exogenous since they are context-dependent and determined upstream of the supply and cost analysis.

We now move on to cost specification. We make the simplifying assumption that unit costs associated with a unit of supply are not t-dependent, i.e. do not vary from one generation to another. However, they are m-dependent, i.e. dependent on the period of intervention. For instance, there is an initial intervention cost in $m = 1$ then follow-up costs for $m \geq 2$. That allows various cost profiles. Furthermore, the costs are strategy-dependent, i.e. do vary from one strategy to another. Formally, unit cost is denoted uc^m. To illustrate, Fig. 8.4 provides cost details for strategy S_k and relates to the supply frame of Fig. 8.3.

For each period t, one must compute the total cost defined by the sum of products of quantities and unit costs. We have:

$$C_t(S_k) = \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k)$$

Over the whole time horizon, the cost of strategy S_k is:

$$C(S_k) = \sum_{t=1}^{T} C_t(S_k) = \sum_{t=1}^{T} \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k)$$

For instance, using information from both Figs. 8.3 and 8.4, we have:

In period 1: $C_1(S_k) = uc^1 \times x_1^1$.

In period 2: $C_2(S_k) = uc^2 \times x_1^2 + uc^1 \times x_2^1$.

In period 3: $C_3(S_k) = uc^3 \times x_1^3 + uc^2 \times x_2^2 + uc^1 \times x_3^1$.

In period 4: $C_4(S_k) = 0 + uc^3 \times x_2^3 + uc^2 \times x_3^2 + uc^1 \times x_4^1$.

In period 5: $C_5(S_k) = 0 + 0 + uc^3 \times x_3^3 + uc^2 \times x_4^2 + uc^1 \times x_5^1$.

The total cost of strategy S_k is $C(S_k) = C_1(S_k) + C_2(S_k) + C_3(S_k) + C_4(S_k) + C_5(S_k)$.

It is important to note that costs do not reflect the producer's cost function but rather the expense borne by the budget holder in charge of financing the project. For instance, in the case of public health, the cost does not stem from the analytical accounts of the manufacturer, in this case a pharmaceutical firm providing a new drug to be added to the existing treatment options. Rather, the cost taken into account is the one borne by the third-party payer(s), e.g., social security or mutual funds who bear reimbursement. The cost boundary is to be precisely defined and justified. For instance, does it include out-of-pocket costs for patients? Do parents have to pay a fee for sending their children in a particular school? The budget holder may or may not take those into account depending on the chosen cost perimeter.

8.3 Budget Impact in a Multiple-Supply Setting

When moving from a single supply to a multiple supply framework, one must make sure that the cost perimeter remains stable so as to allow comparisons between the current and the new environment. Simply put, we assume that the adoption of a new strategy does not affect the supply costs of the other strategies.

Once the demand has been estimated and the supply frame has been built, one can proceed to the budget impact analysis per se. The first step is to define the current environment, denoted e_0 hereafter. It involves a perimeter that encompasses the strategies $S_1, \ldots, S_K$ that are currently in effect: $e_0 = \{S_1, \ldots, S_K\}$. By construction, it does not include the new strategy. The introduction of a new strategy S_{K+1} brings in the new environment: $e_1 = \{S_1, \ldots, S_K, S_{K+1}\}$. The adoption of strategy S_{K+1} may modify the demand forecast, the supply frame and consequently the budget perimeter. Strategy S_{K+1} has unit costs $uc^m(S_{K+1})$. Unit costs are assumed to be unchanged for the existing strategies $S_1, \ldots, S_K$. However, demand patterns may change since the introduction of the new strategy is likely to reshuffle subjects amongst strategies, attract new ones, and prevent more (or less) from leaving the program.

Let $x_t^m(S_k|e_0)$, $k = 1 \ldots K$, and $x_t^m(S_k|e_1)$, $k = 1 \ldots K+1$, denote the quantities supplied for each strategy S_k under the current environment and the new environment, respectively. Total budget for the current environment is computed as:

$$C(e_0) = \sum_{k=1}^{K} C(S_k|e_0)$$

where

$$C(S_k|e_0) = \sum_{t=1}^{T} \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k|e_0)$$

The term $C(S_k| e_0)$ represents the total cost of strategy S_k under the current environment ($k = 1 \ldots K$). Similarly, total budget for the new environment is:

$$C(e_1) = \sum_{k=1}^{K+1} C(S_k|e_1)$$

with

$$\mathrm{C}(S_k|e_1) = \sum_{t=1}^{T} \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k|e_1)$$

Here, the term $C(S_k| e_1)$ represents the total cost of strategy S_k under the new environment ($k = 1 \ldots K + 1$). Using these equations, the budget impact over the whole time horizon is expressed as a simple difference:

$$\text{Budget impact} = C(e_1) - C(e_0)$$

If the impact is found to be positive, it means that the new strategy generate extra-cost. If the impact is found to be negative, then the new environment is cost-saving. As already stated in Sect. 8.1, budget impact analysis does not provide a decision rule. Extra-cost can be synonymous with higher quality of supply while extra-saving may imply a reduction in that quality. It is up to the decision-maker to decide whether the change in cost is acceptable.

In practice, total budgets can also be decomposed on an annual basis so as to assess the budget load year after year, which is more in line with the budget impact standpoint. The annual budget burden for the current environment is defined as:

$$C_t(e_0) = \sum_{k=1}^{K} \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k|e_0)$$

while for the new environment, it is

$$C_t(e_1) = \sum_{k=1}^{K+1} \sum_{i=1}^{t} uc^{t+1-i}(S_k) \times x_i^{t+1-i}(S_k|e_1)$$

The annual basis naturally provides a more detailed analysis:

$$\text{Budget impact (at } t) = C_t(e_1) - C_t(e_0), \quad t = 1 \cdots T$$

Admittedly, the chosen time horizon is to some extent arbitrary since the lifespan of the assessed facilities or drugs is likely to go much beyond the chosen time horizon. However, budget impact analysis is meant to assess middle run feasibility in terms of sustainable budget and ability to reach target outcomes. We now move on to a numerical example that will serve as the base case analysis.

8.4 Example

Budget impact analysis is highly context dependent so that building an all-use template would be a wager. For instance, the subjects of interest within a given general population can be school-age children eligible to attend an education institute or patients suffering from a disease that requires treatment. This explains why we present an example of analysis that cannot claim generality.

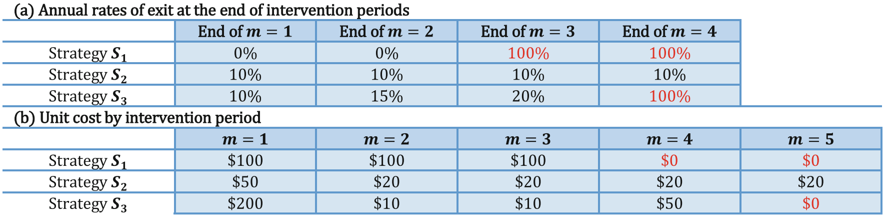
The example considers a time horizon that ranges from 2020 to 2024 ($T = 5$) and, as such, involves five generations of subjects. Table 8.1a displays forecasts about the beneficiary population for the five generations. For year 2020, 5000 subjects enter the model. This initial cohort comprehends subjects already treated (the prevalent population) and the newcomers to the treatment (the incident population). In 2021, a new generation enters the model with 1100 incident subjects, and so on.

The budget impact analysis is interested in a set of three strategies, namely S_1, S_2 and S_3. The current environment e_0 includes two strategies S_1 and S_2 while the new environment e_1 is characterized by strategies S_1, S_2 and S_3. Those strategies are assumed to be mutually exclusive interventions: a subject cannot benefit from more than one strategy at the same time. This does not mean that the strategies are jointly exhaustive, as one individual that has entered the program may also leave it. In other words, strategies S_1, S_2 and S_3 do not together exhaust all the possibilities faced by the subjects. Yet, only the strategies that directly influence the budget are analyzed.

Table 8.1b shows how the incident population is distributed among the strategies of the current environment. In period 1, the current environment allocates 40% of the incident population to strategy S_1 and 60% to strategy S_2. As can be seen, those shares evolve over time. In period 2 for instance, 30% of the incident population benefit from strategy S_1 while 70% benefit from strategy S_2. It can be seen in Table 8.1c that the new environment reallocates those demand shares. For instance, in 2022, the incident population counts 1200 subjects. The current environment shares them between strategy S_1 (which takes care of 360 subjects) and strategy S_2 (which treats 840 subjects). Should strategy S_3 be introduced through the new environment, that same cohort would then be allocated differently to strategy S_1

Table 8.1 Allocation of incident population among strategies

(**a**) Incident population					
Year	2020	2021	2022	2023	2024
Counts	5000	1100	1200	1300	1400
(**b**) Current environment ($\boldsymbol{e}_0$)					
Strategy $\boldsymbol{S}_1$	2000 (40%)	330 (30%)	360 (30%)	520 (40%)	420 (30%)
Strategy $\boldsymbol{S}_2$	3000 (60%)	770 (70%)	840 (70%)	780 (60%)	980 (70%)
(**c**) New environment ($\boldsymbol{e}_1$)					
Strategy $\boldsymbol{S}_1$	1500 (30%)	220 (20%)	240 (20%)	390 (30%)	280 (20%)
Strategy $\boldsymbol{S}_2$	3000 (60%)	550 (50%)	480 (40%)	260 (20%)	560 (40%)
Strategy $\boldsymbol{S}_3$	500 (10%)	330 (30%)	480 (40%)	650 (50%)	560 (40%)



(a) Annual rates of exit at the end of intervention periods

	End of $m = 1$	End of $m = 2$	End of $m = 3$	End of $m = 4$
Strategy S_1	0%	0%	100%	100%
Strategy S_2	10%	10%	10%	10%
Strategy S_3	10%	15%	20%	100%

(b) Unit cost by intervention period

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
Strategy S_1	\$100	\$100	\$100	\$0	\$0
Strategy S_2	\$50	\$20	\$20	\$20	\$20
Strategy S_3	\$200	\$10	\$10	\$50	\$0

Fig. 8.5 Exit and cost patterns

(240 subjects), strategy S_2 (480 subjects), and strategy S_3 (480 subjects). Note that, for simplicity of exposition, the incident population is assumed to be the same in the current and new environments (Table 8.1a). This assumption should be relaxed if the new strategy S_3 could attract new subjects. For instance, strategies S_1 and S_2 may be treatments that are contra-indicated for patients who could be now treated with strategy S_3 (e.g., a drug that they would tolerate).

Exit criteria for lost or censored subjects may differ from one strategy to another. For instance, termination may depend on the duration of intervention and on individual characteristics or circumstances (non-compliance, dropout, side effects, relocation, etc.). Exit rates are not affected by the introduction of the new strategy. Figure 8.5a provides information about the annual exit rates at the end of intervention periods for each strategy. Strategy S_1 is characterized by $M(S_1) = 3$ years of intervention, then those interventions fully stop (hence the 100% exit rate at the end of period $m = 3$). The assumption is that no subjects are lost or censored: the exit rate is set to 0%. Strategy S_2 involves at least 5 years of interventions with a 10% dropout each year. Strategy S_3 is characterized by $M(S_3) = 4$ years of intervention (100% of reached subjects leave the model after their 4-year treatment) and evidences a variable dropout during the intervention years.

In Fig. 8.5b, cost profiles depend on the strategies and their treatment length. Like exit rates, they are not affected by the introduction of the new strategy. Strategy S_1 shows a constant and relatively high unit cost over its duration. Strategy S_2 implies a rather smaller overall budget for a given subject, with a kind of initial investment and subsequent follow-up costs. Strategy S_3 has a quite high initial cost with decreasing follow-up costs, but its overall individual budget burden is still lower than that of strategy S_1.

Figure 8.6 describes the follow-up of subjects under the two scenarios. At the first year of analysis (2020) the program reaches 5000 subjects (patients with a given disease, children of school age in the targeted district). Part of them could be prevalent (they have already been treated during the previous year at least) and the rest is incident (they become eligible to treatment in the case of a disease or reach the age of 6 in the school example). During the second year (2021), the incident generation 2 with its 1200 subjects adds up to the prevalent generation 1 (apart from those who may have left the program, which does not happen for strategy S_1 in our example). During the third year (2022), generations 1 and 2 are now prevalent and are joined by the incident generation 3 (1200 subjects), etc.

(a) Current environment e_0

		2020	2021	2022	2023	2024
Strategy S_1	Generation 1	2000	2000	2000	0	0
	Generation 2	0	330	330	330	0
	Generation 3	0	0	360	360	360
	Generation 4	0	0	0	520	520
	Generation 5	0	0	0	0	420
	Total strategy S_1	**2000**	**2330**	**2690**	**1210**	**1300**
Strategy S_2	Generation 1	3000	2700	2430	2187	1968
	Generation 2	0	770	693	624	561
	Generation 3	0	0	840	756	680
	Generation 4	0	0	0	780	702
	Generation 5	0	0	0	0	980
	Total strategy S_2	**3000**	**3470**	**3963**	**4347**	**4892**
	Total $S_1 + S_2$	**5000**	**5800**	**6653**	**5557**	**6192**
(b) New environment e_1						
Strategy S_1	Generation 1	1500	1500	1500	0	0
	Generation 2	0	220	220	220	0
	Generation 3	0	0	240	240	240
	Generation 4	0	0	0	390	390
	Generation 5	0	0	0	0	280
	Total strategy S_1	**1500**	**1720**	**1960**	**850**	**910**
Strategy S_2	Generation 1	3000	2700	2430	2187	1968
	Generation 2	0	550	495	446	401
	Generation 3	0	0	480	432	389
	Generation 4	0	0	0	260	234
	Generation 5	0	0	0	0	560
	Total strategy S_2	**3000**	**3250**	**3405**	**3325**	**3552**
Strategy S_3	Generation 1	500	450	383	306	0
	Generation 2	0	330	297	252	202
	Generation 3	0	0	480	432	367
	Generation 4	0	0	0	650	585
	Generation 5	0	0	0	0	560
	Total strategy S_3	**500**	**780**	**1160**	**1640**	**1714**
	Total $S_1 + S_2 + S_3$	**5000**	**5750**	**6525**	**5815**	**6176**
	Difference e_1- e_0	**0**	**–50**	**–129**	**+258**	**–16**

Fig. 8.6 Follow-up of generations by scenario and by strategy

Let us consider for instance strategy S_2 under the current environment (Fig. 8.6a). The values are computed as follows. First, at year 2020, 60% of the incoming population (5000 from Fig. 8.6b) are treated with strategy S_2, namely 3000 subjects. At the end of year 2020, generation 1 faces an exit rate of 10% (Fig. 8.5a) which implies that 300 subjects from this generation will leave the program. As such, in 2021, the number of subjects of generation 1 benefiting from S_2 is 2700. In the meantime, the incident population (generation 2) amounts to 770 (60% of the 1100 incoming subjects). The total number of subjects treated with S_2 in 2021 is thus 3470. At the end of year 2021, both generations face a 10% exit rate, which means that generations 1 and 2 are reduced in 2022–2430 and 693 subjects, respectively. They are joined by 70% of generation 3 (840 subjects) so that 3963 subjects are now part of the program under the S_2 strategy. The procedure continues as such until the time horizon is reached. A similar method is used for the new environment.

The comparison of Fig. 8.6a with Fig. 8.6b highlights differences in the way attrition affects the cohorts of users. In the current environment, once subjects from a given generation (say generation 3 entering the program in 2022) have been

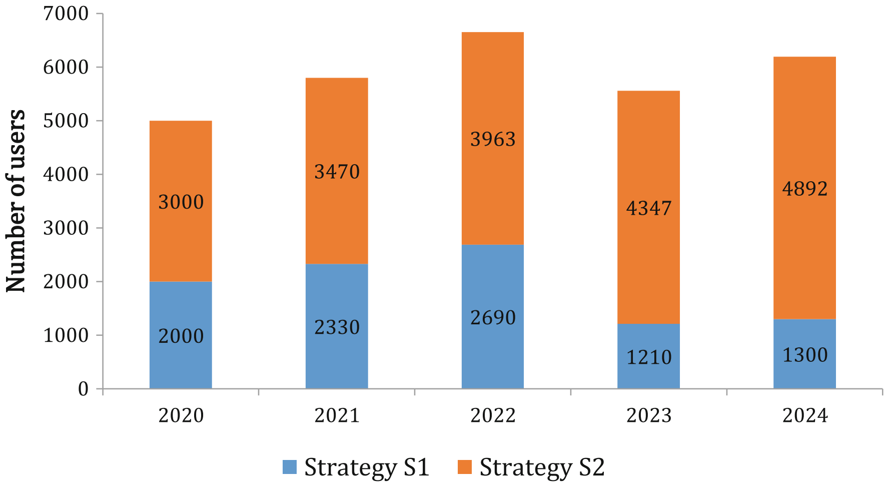


Fig. 8.7 Cohorts of users: current environment

allocated to a strategy (say strategy S_2, and namely 840 subjects), they get treated unless they exit at the previously defined termination rates. For an annual 10% rate, the number of subjects treated with strategy S_2 decreases to 756 in 2023 and to 680 in 2024. In the new environment, strategy S_2 follows a similar pattern but the initial number of subjects differs because the demand share of S_2 has decreased with the arrival of strategy S_3. The new environment would only allocate 480 subjects to strategy S_2 and attrition would reduce the treated subjects among them to 432 in 2023 and 389 in 2024.

Total rows of Figs. 8.6a and 8.6b indicate the population reached under a given environment. It is an indicator of the annual global outcome associated with that environment. Last row of Fig. 8.6 computes the difference between those totals. In 2020, this difference amounts to zero as the population in both environments is assumed to be the same. In 2021, the new environment has 50 less subjects than the current environment. This result is due to the reallocation of subjects among strategies and the fact that strategies differ in their termination rates. As can be seen, the differences in annual global outcome are very thin in this example. At year 2022, it amounts to −129, then +258 in 2023, and −16 in 2024. Yet, what matters is more the way the subjects are distributed among the strategies. Differences in allocation of subjects may generate high differences in terms of costs. For instance, the population reached by the strategies under the current and new environments is displayed in Figs. 8.7 and 8.8, respectively. It can be seen that an important share of those benefiting from strategy S_2 in the current environment are allocated to strategy S_3 in the new environment. This may strongly affect the total cost of the program.

Costs patterns are displayed in Fig. 8.9. They are decomposed by scenario, strategy and year. The values have been generated using information from

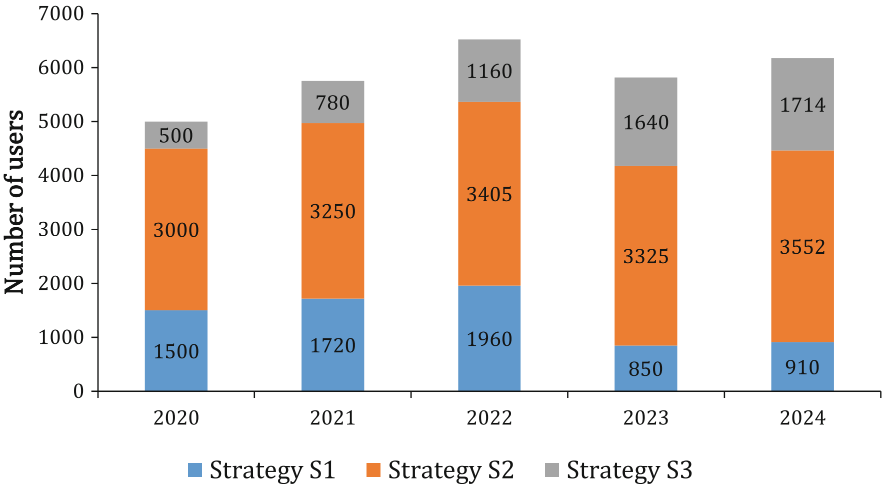


Fig. 8.8 Cohorts of users: new environment

Figs. 8.6 and 8.5b. Consider for instance strategy S_2 in the current environment. The total row of Fig. 8.9a has been computed as follows:

At year 2020: $C_1(S_2) = \$50 \times 3000 = \$150{,}000$;
At year 2021: $C_2(S_2) = \$20 \times 2700 + \$50 \times 770 = \$92{,}500$;
At year 2022: $C_3(S_2) = \$20 \times 2430 + \$20 \times 693 + \$50 \times 840 = \$104{,}460$;
and so on.

The aim is to generate annual costs by environment and to compare them in order to calculate the annual budget impact of strategy S_3. This is done in Fig. 8.9 where the last row shows the difference in costs between the current environment and the new environment. As can be observed, the annual budget impact is positive for all years except in 2021 and amounts to $50,000 in 2020, −$1500 in 2021, $7835 in 2022, $80,100 in 2023 and $53,220.4 in 2024. Summing those figures yields the total budget impact of introducing strategy S_3 as a new intervention in the program, namely $189,666. Figure 8.10 displays the annual impacts over the time horizon of the analysis. With this particular example, the impact is mainly positive which means that the budget holder will have to commit additional financial resources if the new strategy is to be introduced in the supply frame.

8.5 Sensitivity Analysis with Visual Basic

Budget impact analysis provides a quantitative assessment of the annual allocation of demand to the strategies and of the cost burden associated with it. Those results are quite dependent on the assumptions on demand flows and shares amongst strategies (as in Table 8.1) as well as on the exit rates and cost structure (as in

(a) Current environment (e_0)		2020	2021	2022	2023	2024
Strategy S_1	Generation 1	$200,000	$200,000	$200,000	$0	$0
	Generation 2	$0	$33,000	$33,000	$33,000	$0
	Generation 3	$0	$0	$36,000	$36,000	$36,000
	Generation 4	$0	$0	$0	$52,000	$52,000
	Generation 5	$0	$0	$0	$0	$42,000
	Total strategy S_1	**$200,000**	**$233,000**	**$269,000**	**$121,000**	**$130,000**
Strategy S_2	Generation 1	$150,000	$54,000	$48,600	$43,740	$39,366
	Generation 2	$0	$38,500	$13,860	$12,474	$11,227
	Generation 3	$0	$0	$42,000	$15,120	$13,608
	Generation 4	$0	$0	$0	$39,000	$14,040
	Generation 5	$0	$0	$0	$0	$49,000
	Total strategy S_2	**$150,000**	**$92,500**	**$104,460**	**$110,334**	**$127,241**
	Total $S_1 + S_2$	**$350,000**	**$325,500**	**$373,460**	**$231,334**	**$257,241**
(b) New environment (e_1)						
Strategy S_1	Generation 1	$150,000	$150,000	$150,000	$0	$0
	Generation 2	$0	$22,000	$22,000	$22,000	$0
	Generation 3	$0	$0	$24,000	$24,000	$24,000
	Generation 4	$0	$0	$0	$39,000	$39,000
	Generation 5	$0	$0	$0	$0	$28,000
	Total strategy S_1	**$150,000**	**$172,000**	**$196,000**	**$85,000**	**$91,000**
Strategy S_2	Generation 1	$150,000	$54,000	$48,600	$43,740	$39,366
	Generation 2	$0	$27,500	$9,900	$8,910	$8,019
	Generation 3	$0	$0	$24,000	$8,640	$7,776
	Generation 4	$0	$0	$0	$13,000	$4,680
	Generation 5	$0	$0	$0	$0	$28,000
	Total strategy S_2	**$150,000**	**$81,500**	**$82,500**	**$74,290**	**$87,841**
Strategy S_3	Generation 1	$100,000	$4,500	$3,825	$15,300	$0
	Generation 2	$0	$66,000	$2,970	$2,524.5	$10,098
	Generation 3	$0	$0	$96,000	$4,320	$3,672
	Generation 4	$0	$0	$0	$130,000	$5,850
	Generation 5	$0	$0	$0	$0	$112,000
	Total strategy S_2	**$100,000**	**$70,500**	**$102,795**	**$152,144.5**	**$131,620**
	Total $S_1 + S_2 + S_3$	**$400,000**	**$324,000**	**$381,295**	**$311,434.5**	**$310,461**
	Budget impact	$50,000	–$1,500	$7,835	$80,100.5	$53,220.4

Fig. 8.9 Costs by scenario and by strategy

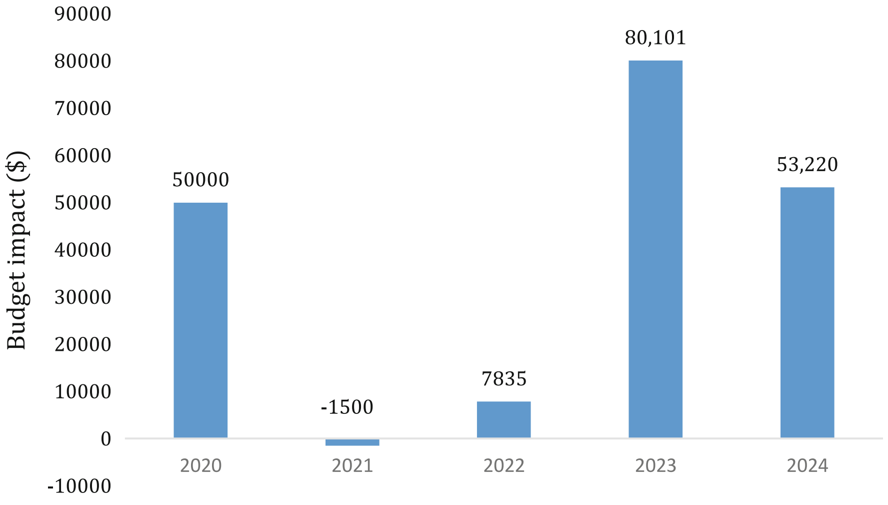


Fig. 8.10 Budget impact of introducing strategy S_3

Fig. 8.5). This is why exploring uncertainty through sensitivity analysis, far from being an appendix to the assessment, is a constitutive part of budget impact analysis.

The first type of uncertainty relevant to that analysis is parameter uncertainty surrounding the values of input. Parameter uncertainty relates to the estimation of generational demands, demand shares amongst strategies and cost parameters. This form of uncertainty is associated with factors such as the existence of several but conflicting studies or expert opinions documenting data or, conversely, is related to a lack of data for parameters of interest. Because of limited or poor information, much of parameter uncertainty cannot be quantified with relevance.

The second type of uncertainty is structural and is conveyed by the assumptions that prevailed in framing the budget impact analysis: they relate to inflexions in the current environment patterns induced by the introduction of the new strategy, which are not easy to anticipate. Specifically, the allocation of the target population among strategies as well as their exit rates are likely to be structurally distorted by the inclusion of the new strategy. With structural uncertainty, the interrogation is not about doubts on the value of a given specific parameter (for instance the demand share of strategy S_1 at year $t = 2023$ in the new environment may not be 30% with certainty, thus there is parameter uncertainty around that value). What is questioned with structural uncertainty is the parameter pattern for demand shares and exit rates and cost structure as a whole. By essence, structural uncertainty in budget impact analysis strongly resists quantification.

In this context, sensitivity exploration in budget impact analysis is best carried out through scenario analyses. The usual approach consists in using a most-likely scenario, also known as base case, as a benchmark. The sensitivity analysis then can change selected input parameter values and structural assumptions so as to generate plausible outcomes and budget burdens alternative to the base-case.

Let us consider again the example of Sect. 8.4. There are many ways of implementing sensitivity simulations using templates. Excel can for instance be used in this purpose. The analysis can be separated in three worksheets as described in Table 8.2. The first worksheet is labeled "Parameters" and describes input parameters (incident population, exit rates, unit costs), i.e. contains information similar to that presented in Table 8.1 and Fig. 8.5. The second worksheet is named "Base case" and displays all the results of the budget impact analysis as in Figs. 8.6 and 8.9 as well as Figs. 8.1 and 8.2. The third worksheet is labeled "Sensitivity analysis" and simulates how parameters and structural uncertainty influence the impact on costs and outcomes of the introduction of the new strategy. We suggest here a very simple macro command under Excel that allows modifying any subset of parameters and then restores the initial values.

Figure 8.11 provides the corresponding code. Macros are written in the Visual Basic programming language and are stored in a separate Visual Basic editor, nevertheless linked to the main workbook. By definition, a macro is an action or a set of actions that is used to automate tasks. In Excel, a few settings may have to be changed before running the macro. First, the Developer tab is not displayed by default. It can be added to the ribbon via the following steps: (1) click the Microsoft

Table 8.2 Organization of budget impact analysis in Excel

Worksheet 1 “Parameters”	Worksheet 2 “Base case”	Worksheet 3 “Sensitivity analysis”
For both the current and new environments, information about the incident populations and their allocation among strategies.	Follow-up of generations by scenario and by strategy.	Simulates how parameters and structural uncertainty influence the supply.
For all strategies, information about annual rates of exit at the end of intervention periods, and unit costs by intervention period.	Costs by scenario and by strategy.	Simulates how parameters and structural uncertainty influence the cost of strategies.

```
Sub Macro1()

'Macro1: restore initial values of parameters

    'Selection of the worksheet displaying initial parameters
    Sheets("Parameters").Select
    'Selection of the range of cells displaying initial parameters
    Range("B3:G27").Select
    'Copy of initial values
    Selection.Copy
    'Selection of the worksheet for sensitivity analysis
    Sheets("Sensitivity Analysis").Select
    'Selection of cell range for new values
    Range("B4").Select
    'Copy of values of initial parameters
    ActiveSheet.Paste

End Sub
```

Fig. 8.11 Visual basic macro for restoring initial values

Office Button, (2) click Excel Options and (3) in the Popular category, under Top options for working with Excel, select the “Show Developer tab in the Ribbon” check box, and click OK.

Once the Developer tab is displayed, click on Visual Basic in the Developer tab, then click on Insert and Module. The next step is to copy and past the code of Fig. 8.11. Once saved, the macro can be accessed from the Developer tab, by clicking the Macros command on the ribbon (Developer tab, Code group). It is also possible to add a button and to assign a macro to it in a worksheet. On the Developer tab, in the Controls group, one needs to click Insert, and then under Form Controls, one must click Button. The final step is to choose the location of the button on the worksheet, and to assign a macro to it.

In Fig. 8.11, the macro makes use of two worksheets only, the one that contains the parameters (worksheet 1 labeled “Parameters”) and the one that performs the sensitivity analysis (worksheet 3 labeled “Sensitivity analysis”). Note that worksheet 3 is an exact copy of worksheet 2 (same calculations) and will differ only with respect to the initial parameters. In Fig. 8.11, command lines begin with the ′ symbol. The first step is to select the initial set of parameters by first stating the

Table 8.3 Parameter sensitivity analysis: example of a cost decrease

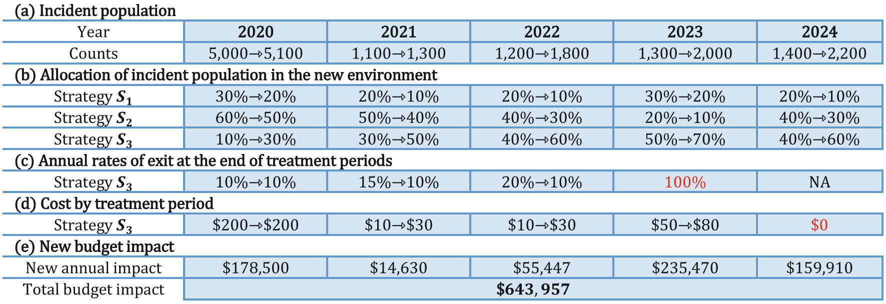
(**a**) Base case					
Year	2020	2021	2022	2023	2024
Budget impact	\$50,000	−\$1500	\$7835	\$80,100.5	\$53,220.4
Total budget impact	\$189,666				
(**b**) Impact of a 25% decrease in the initial annual cost of strategy S_3					
New annual impact	\$25,000	−\$18,000	−\$16,165	\$47,600.5	\$25,220.4
Total budget impact	\$63,655				

corresponding sheet (worksheet 1 labeled "Parameters") then identify in that sheet the relevant cells (in the example, these are the cells ranging from B3 to G27). Initial values are then saved through the *Selection*.*Copy* code. Those values are then displayed in the Sensitivity analysis worksheet from cell B4 in the example. In other words, the macro is used to copy the parameters of the first worksheet (cells B3 to G27) in the third worksheet (cells B4 to G28). We can then proceed to the budget impact analysis in the third worksheet by modifying those values. The advantage of the macro is that we can examine multiple scenarios, and then restore the parameters very easily.

Even in a simple example such as the one presented here, the number of parameter combinations is such that they cannot be explored systematically, all the more so that we do not have information on the likely variation range of parameters. One should then select a number of plausible scenarios, relevant to the concerns of the policy-maker, and examine their outcome and cost consequences. The following examples are thus purely illustrative and would require an explicit contextualization in order to receive a fully pertinent interpretation.

We first investigate variations in a single parameter (see Table 8.3) using the analysis of Sect. 8.5 as the base case. For instance, the consequences of a 25% decrease in the initial annual cost of strategy S_3 ($uc^1(S_3) = 150$ instead of 200) are such that the budget impact of introducing the new strategy is approximately three times less, without any repercussion on demand allocation (Table 8.3b).

Structural sensitivity analysis can also be explored through various scenarios involving several simultaneous changes in parameters that express for instance a best case scenario or conversely a worst case scenario. Defining what is worst or best is of course dependent upon the context and the hierarchy between consequences. For instance, the public decision-maker may give priority to cost (that should be contained) or to demand (that should be directed to the new strategy). Figure 8.12 illustrates the methodology using the framework of Sect. 8.4 as the base case. We examine the case of a chronic disease initially treated by either strategy S_1 or strategy S_2. The new and innovative treatment is strategy S_3. A "stress-test" sensitivity scenario would consider the following deviations in the model structure. First, public health authorities may anticipate a steady increase in the prevalence of the disease due for instance to ingrained harmful diet and lifestyle



(a) Incident population					
Year	**2020**	**2021**	**2022**	**2023**	**2024**
Counts	5,000→5,100	1,100→1,300	1,200→1,800	1,300→2,000	1,400→2,200
(b) Allocation of incident population in the new environment					
Strategy S_1	30%→20%	20%→10%	20%→10%	30%→20%	20%→10%
Strategy S_2	60%→50%	50%→40%	40%→30%	20%→10%	40%→30%
Strategy S_3	10%→30%	30%→50%	40%→60%	50%→70%	40%→60%
(c) Annual rates of exit at the end of treatment periods					
Strategy S_3	10%→10%	15%→10%	20%→10%	100%	NA
(d) Cost by treatment period					
Strategy S_3	$200→$200	$10→$30	$10→$30	$50→$80	$0
(e) New budget impact					
New annual impact	$178,500	$14,630	$55,447	$235,470	$159,910
Total budget impact	**$643, 957**				

Fig. 8.12 Structural sensitivity analysis: a stress-test scenario

habits (Fig. 8.12a). Second, the innovative strategy S_3 would attract every year additional subjects from each strategy (compared with the base case), thus generating a reduction of 10 percentage points in the demand for S_1 and S_2 (Fig. 8.12b). Third, treatment compliance may be better than expected so that exit rates would go down to 10% each year (Fig. 8.12c). Finally, follow-up costs may be greater than initially expected (Fig. 8.12d). As can be seen from Fig. 8.12e, compared to the base case, this scenario induces an important increase in the budget burden.

Figures 8.13, 8.14 and 8.15 show the distribution of demands and budget impacts of the "stress-test" scenario. The new budget impact evidences a strong effect on the cost burden. Whether that burden is bearable and relevant is a question that cannot be answered by the budget impact analysis. The latter nevertheless illuminates the magnitude of the effects. In this example, the public decision-maker faces a real public health concern with the constant rise of a chronic disease, due to lifestyle habits, environmental factors, etc. If the policy answer is to introduce a new and expensive drug, then it will have a considerable budget impact and opportunity cost. The sensitivity analysis may prompt reflections about alternative ways of dealing with the problem.

Bibliographical Guideline

Budget impact analysis is to our knowledge mostly if not solely used in health economics, although its design should make it a rather general tool for assessing the consequences of introducing a new strategy amongst existing ones in the implementation of a public project. The main proponent of that method has been Mauskopf (1998) with a seminal analysis that has generated a lot of interest both among scholars and practitioners of public health decision making (Trueman et al. 2001; Garattini and van de Vooren 2011). Budget impact analysis is now recommended by several national health technology assessment agencies when a new innovative treatment claims reimbursement from social security or mutual funds. It is also recognized at the international level by scientific authorities in the field (ISPOR 2014).

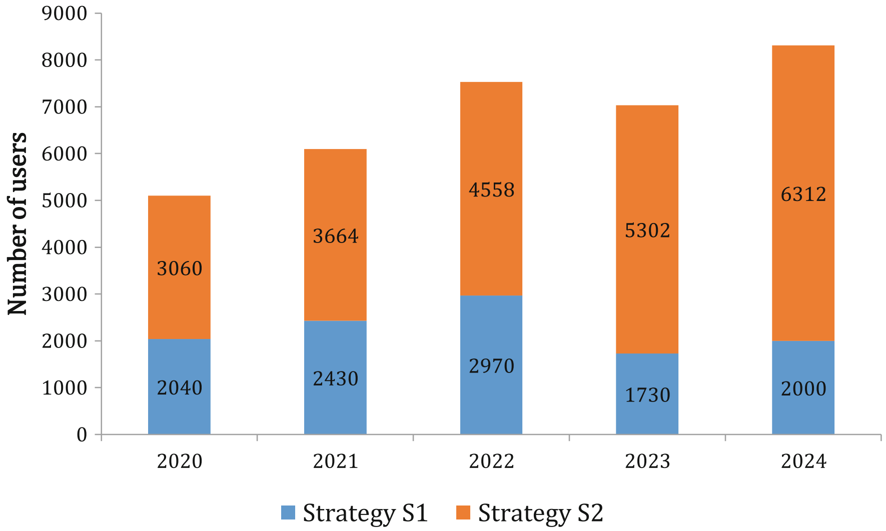


Fig. 8.13 Cohorts of users under the stress-test scenario: current environment

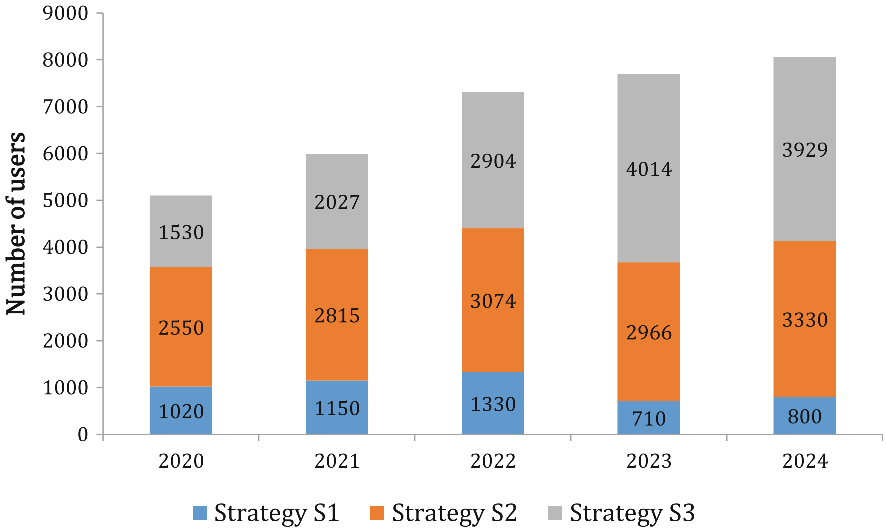


Fig. 8.14 Cohorts of users under the stress-test scenario: new environment

Uncertainty related to the choice of parameter pattern and values cannot be dealt with by using partial deterministic analysis and probabilistic sensitivity analysis. Budget impact analysis rather uses scenarios involving either variations of a small set of parameters or larger scale deviations from the base case in order for instance

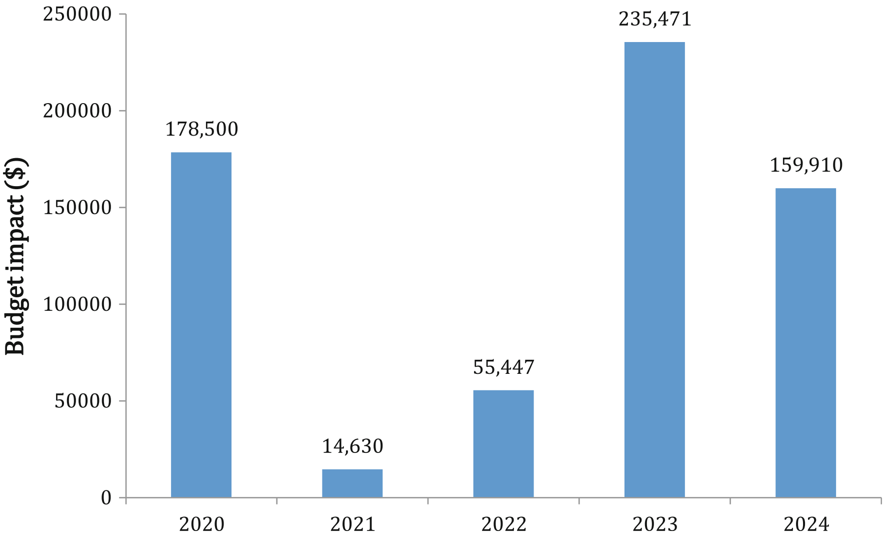


Fig. 8.15 Budget impact of the stress-test scenario

to account for optimistic or pessimistic stances (Nuijten et al. 2011; Mauskopf 2014).

Applications to fields of public action other than pharmacoeconomics and health economics could be developed by complementing the budget impact model with a detailed argumentation of what the target population consists of, how it evolves according to relevant parameters such as socio-economic, epidemiologic, demographic parameters. Our presentation has eluded that dimension of budget impact analysis in order to remain as all-purpose as possible. Demand analysis necessitates specific knowledge about the assessed field of public action and, in practical terms, requires a preliminary worksheet generating demand shares and their evolution in time). In the case of health policies, national health technology assessment agencies often provide templates helping with such calculations (e.g., in France: Ghabri et al. 2017).

Finally, our model of budget impact analysis pays tribute to the overlapping generations setting that was initially conceived by Samuelson (1958) and further formalized by Gale (1973). That allows a systematic presentation of the cost and outcome consequences of the introduction of a new strategy in an existing field of action, both over the lifecycle of a generation of users and across generations.

References

Gale, D. (1973). Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory*, *6*, 12–36.

Garattini, L., & van de Vooren, K. (2011). Budget impact analysis in economic evaluation: A proposal for a clearer definition. *European Journal of Health Economics, 12*, 449–502.

Ghabri, S., Poullié, A.-I., Autin, E., & Josselin, J.-M. (2017). *Guidelines for budget impact analysis in economic evaluations of drugs and medical devices*. French national authority for health [HAS].

ISPOR (International Society for Pharmacoeconomics and Outcomes Research), Sullivan, S. D., Mauskopf, J., Augustovski, F., Jaime Caro, J., Lee, K. M., et al. (2014). Budget impact analysis-principles of good practice: Report of the ISPOR 2012 budget impact analysis good practice II task force. *Value in Health, 17*, 5–14.

Mauskopf, J. (1998). Prevalence-based economic evaluation. *Value in Health, 1*, 251–259.

Mauskopf, J. (2014). Budget-impact analysis. In A. Culyer (Ed.), *Encyclopedia of health economics* (Vol. 1, pp. 98–107). San Diego: Elsevier.

Nuijten, M. J., Mittendorf, T., & Persson, U. (2011). Practical issues in handling data input and uncertainty in a budget impact analysis. *European Journal of Health Economics, 12*, 231–241.

Samuelson, P. A. (1958). An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy, 66*, 467–482.

Trueman, P., Drummond, M., & Hutton, J. (2001). Developing guidance for budget impact analysis. *Pharmacoeconomics, 19*, 609–621.

9 Cost Benefit Analysis

9.1 Rationale for Cost Benefit Analysis

The aim of cost benefit analysis is to determine whether a public policy choice is globally beneficial or detrimental to society's welfare. The method allows several projects to be compared not only on the basis of their financial flows (investments and operating costs) but also on the basis of their economic effects (variation in consumption of public services, change in the satisfaction derived from them). For instance, the construction of a highway not only modifies the financial flows of the authority at stake, but affects traffic congestion, reduces travel time, improves safety, influences positively or negatively real estate in the areas served by the project. All these consequences must be properly accounted for. To do so, cost benefit analysis expresses these impacts in terms of a common metric, their equivalent money value. Then the method confronts the financial flows with the economic impacts to measure the total contribution of the project to the well-being of the society.

Cost benefit analysis is primarily used for assessing the value of large capital investment projects (transportation infrastructures, public facilities, recreational sites) but can be employed for appraising smaller projects, public regulations or a change in operating expenditures, especially where other methods fail to account for welfare improvements. It became popular in the 1960s with the growing concern over the quality of the environment, not only in the USA but also in many other countries. Nowadays, institutions such as the European Commission, the European Investment Bank or the World Bank require cost benefit analysis studies for extra funding and loan requests. In the USA, the approach is frequently used when a public policy choice imposes significant costs and economic impacts. In Europe, more and more governments use the approach to justify a particular policy to taxpayers. This was for instance the case with HS2, a controversial high-speed rail link between London, Birmingham and Manchester.

Cost benefit analysis is concerned with two types of economic consequences. First, a project produces "direct effects" affecting the welfare of those who

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_9

primarily benefit from the public service. They are related to the main objectives of the program. Table 9.1 provides several examples. They include for instance the satisfaction from using recreational parks or the time saved by projects in the transport sector. Assessing the direct contribution of a public policy to society's welfare is at the heart of cost benefit analysis. Second, a project may generate "negative and positive externalities", so-called "external effects". Those are costs and benefits that manifest themselves beyond the primary objectives of the program. They appear when a policy affects the consumption and production opportunities of third parties. Typical examples are deterioration of landscape (negative externality) and economic development (positive externality), as described in Table 9.1. Given their potential impact on society's welfare, externalities are also monetized, including social and environmental effects.

The cost benefit analysis methodology also ensures that the price of inputs and outputs used in the analysis reflect their true economic values, and not only the values observed in existing markets. Government intervention may divert the factors of production (land, labor, capital, entrepreneurial skill) from other productive use. This is particularly true where markets are distorted by regulated prices or where taxes or subsidies are imposed on imports or exports (see Table 9.1). For instance, a land made available for free by a public body generates an implicit cost to the taxpayers: the opportunity cost of not renting the land to another entity. A tax on imports which affects the price of inputs also has to be deduced to better reflect the true costs to taxpayers. To account for these distortions, cost benefit analysis makes use of what is called conversion factors. They represent the weights by which market prices have to be multiplied to obtain cash flows valued at their true price from the society's point of view.

Public policies involve many objectives, concern different beneficiary groups and differ with respect to their time horizon. The cost benefit analysis methodology has the advantage to simplify the multidimensionality of the problem by calculating a monetary value for every main benefit and cost. To make those items fully comparable, the approach converts all the economic and financial flows observed at different times to present-day values. This approach, known as discounting, is essential to cost benefit analysis. It enables the projects to be evaluated based on how the society values the well-being of future generations. The idea is that a benefit, or a cost, is worth more when it has an immediate impact. As a consequence, activities imposing large costs or benefits on future generations may appear of less importance. The use of an appropriate discount factor, which weights the time periods, is here decisive.

Despite its popularity among many government agencies and the fact that cost benefit analysis provides an all-encompassing tool for evaluating public policies, the approach has been intensively decried. One of the major drawbacks is the fact that it often struggles to put monetary values on items such as aesthetic landscapes, human health, economic growth, environmental quality, time or even life. The core of the problem is that public projects mostly affect the value of goods that are not bought and sold on regular markets (non-market goods). For this reason, variations in society's welfare cannot be valued based on the usual observation of market

Table 9.1 Examples of economic impacts

Type of project	Direct benefits	Positive externalities	Negative externalities	Price distortions
Rail investment	Time savings, additional capacity, increased reliability	Economic development, reduced negative externalities from other transport modes	Aesthetic and landscape impacts, impacts on human health	Opportunity costs of raw materials due the diversion of them from the best alternative use, land made available free of charge by a public body while it may earn a rent, tariff subsidized by the public sector, labor market distortions due to minimum wages or unemployment benefits
Waste treatment	Treatment of waste which minimizes impacts on human health	Environmental benefits	Aesthetic and landscape impacts, other impacts on human health, increase in local traffic for the transport of waste	
Production of electricity from renewable energy sources	Reduction in greenhouse gases	Amount of fossil fuels or of other non-renewable energy sources saved	Aesthetic and landscape impacts, negative effects on air, water and land	
Telecommunication infrastructures	Time saved for each communication, new additional services	Economic development induced by the project	Aesthetic and landscape impacts, Impacts on human health	
Parks and forests	Recreational benefits, utilization and transformation of wood	Improvement of the countryside, environmental protection, increased income for the tourist sector	Increased traffic	

prices. One has to rely instead on non-market valuation methods, such as revealed preference methods or stated preference techniques. Obtaining accurate estimates of costs and benefits is all the more challenging that many factors may affect the conclusions of the study, such as the time horizon, whether people have prior experiences of using the service or whether they have perfect information about the consequences of the project, etc. To overcome these issues, a sound cost benefit analysis generally ends with a sensitivity analysis which examines how the conclusions of the study change with variations in cash flows, assumptions, or the manner in which the evaluation is set up. This sensitivity analysis can be implemented in a deterministic way. In this case, the uncertainty of the project is described through simple assumptions, for instance by assuming lower and upper bounds in economic flows. The sensitivity analysis can also be probabilistic. Each economic flow is assigned a random generator based on a well-defined probability distribution. A mean-variance analysis then helps the decision-maker to fully compare the risk and performance of each competing strategy.

The remainder of the chapter is organized as follows. Section 9.2 details the theoretical foundations of the method. Section 9.3 provides a tutorial describing the discounting approach. Section 9.4 explains how to convert market prices to economic prices. Sections 9.5 and 9.6 show how to implement a deterministic and probabilistic sensitivity analysis. Section 9.7 finally explains the methodology of mean-variance analysis.

9.2 Conceptual Foundations

Government intervention affects the satisfaction of the agents composing society in many different ways. While some agents will benefit from a change in the public good-tax structure, other agents can be worse off. An investment decision may also generate positive and negative externalities that affect the well-being of particular agents in specific areas. Due to the presence of government-imposed regulations, project cash flows may be distorted and should be valued at their opportunity costs. Cost benefit analysis captures all these economic consequences by expressing them in terms of a common currency, money. How is that possible? How can a change in the society’s welfare be related to dollar value? The approach is actually based on the observation that individuals are often willing to pay more for a good than the price they are charged for it. For instance, if one is willing to spend $10 at most for a service, but pays only $2, one achieves some surplus of $8. This is what cost benefit analysis aims at measuring. If there is a public policy choice for which the net benefits are greater than those of the competing strategies, then society should go ahead with it.

More specifically, cost benefit analysis relies on estimating what is called the economic surplus, a measure of welfare that we find in microeconomics, a branch of economics that studies the behavior of agents at an individual level. Under this framework, the surplus is computed as the sum of two elements: (1) the consumer surplus, measured as the monetary gain obtained by agents being able to purchase a good for a price that is less than the highest price they are willing to pay, and (2) the producer surplus, defined as the difference between the price producers would be willing to supply a good for and the price actually received. While a reasonable measure of benefit to a producer is the net profit, the task is much more difficult with respect to the consumer surplus. One has to identify here the demand for the good. All the difficulty and ingenuity of cost benefit analysis lies there. This section provides a simple microeconomic framework serving to better describe the underlying assumptions.

Formally, let $u(x, z)$ denote the utility (or satisfaction) an agent derives from consuming a private good in quantity x and a public good in quantity z. The public good is provided by the government and financed through taxes. The budget constraint of the agent is defined as $p_x x + br = y$ where p_x denotes the price of the private good, b is the tax base of the agent, r stands for the tax rate chosen by the government and y represents the agent's income. Although we could relax this assumption, the public budget is assumed to be balanced. We have $rB = cz$, where B is the total tax base upon which the tax rate is applied, and c denotes the marginal cost of production. Using $r = cz/B$ in the budget constraint of the agent, we obtain $p_x x + p_z z = y$ where $p_z = cb/B$ represents the (tax) price of the public good. The demand of the agent for the public good is then determined by the maximization of u given this budget constraint:

$$\max_{\{\mathrm{x,z}\}} u(x, z) \text{ subject to } p_x x + p_z z = y$$

To simplify the exposition, we set $p_x = 1$ and assume that the agent has quasi-linear preferences: $u(x, z) = x + v(z)$. Function v is an increasing and concave function of z while x enters the utility function as a simple additive term. Solving the optimization problem by substitution yields:

$$\max_{\{\mathrm{z}\}} u(y - p_z z, z) \Leftrightarrow v'(z) = p_z$$

The solution is obtained by taking the derivative of $y - p_z z + v(z)$ with respect to z. The derivative of v represents the inverse demand function. The lower the price (for instance due to a decrease in b or c), the higher the demand for the good, as illustrated in Fig. 9.1. This generalizes the usual "law of demand" to the case of a publicly provided good.

The welfare the agent derives from z is directly linked to the shape of the inverse demand curve. If the public good is not produced at all, the level of satisfaction obtained by the agent is $u(x, 0) = y$. If z units are produced, the maximum amount A the agent would be willing to pay is determined by the condition $u(y - A, z) > u(x, 0)$. Equivalently, we have $A < v(z)$. In other words, $v(z)$ represents the willingness to pay (WTP) of the agent. Similarly, $v'(z)$ is what is termed the marginal WTP. The consumer surplus is thus defined as:

$$s(z) = v(z) - p_z z$$

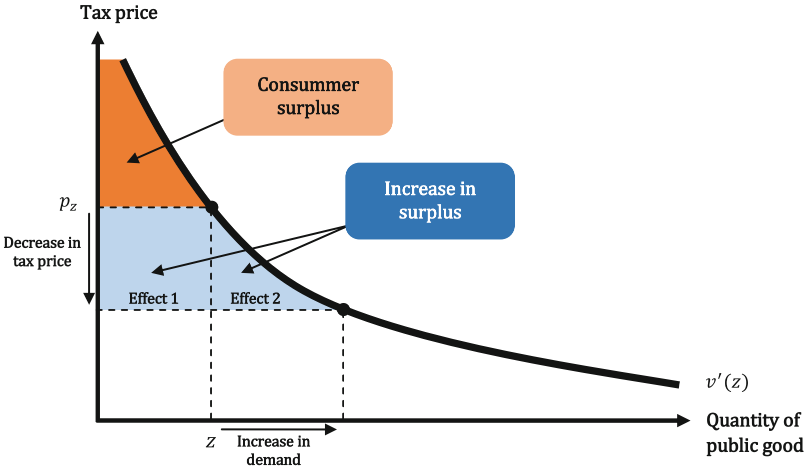


Fig. 9.1 Demand for public good and consumer surplus

where $p_z z$ represents the cost of the policy to the agent. This expression can also be written as:

$$s(z) = \int_0^z v'(z)dz - p_z z$$

Graphically, this means that the consumer surplus is the area comprised between the inverse demand curve $v'(z)$ and the horizontal line at the price p_z. As shown in Fig. 9.1, a decrease in the price generates a direct effect on surplus (effect 1), but also a change in the demand and, if this demand is satisfied, an additional increase in surplus (effect 2).

Consider for instance an agent who enjoys visiting a recreational park. Assume that one has information about his/her willingness to pay for visiting the site. As shown in Fig. 9.2, the marginal willingness to pay is decreasing. For a first visit, the agent is willing to pay \$9 at most. For a second visit, the agent is willing to pay \$4, and so on. This illustrates the "law of diminishing marginal utility". There is a decline in the marginal welfare the agent derives from consuming each additional unit. At some point, and for a given price, the agent will stop visiting the site which will determine his/her final demand. To illustrate, imagine that the price is \$2. The agent will visit the site three times and obtain a surplus of \$9. This information is essential to cost benefit analysis. It allows the net welfare of the agent to be quantified in monetary terms. For instance, if a public body aims at providing a better access to the recreational park, thereby reducing the travel expenses of the agent by \$1, then the surplus will increase by \$3. This means that the agent does not

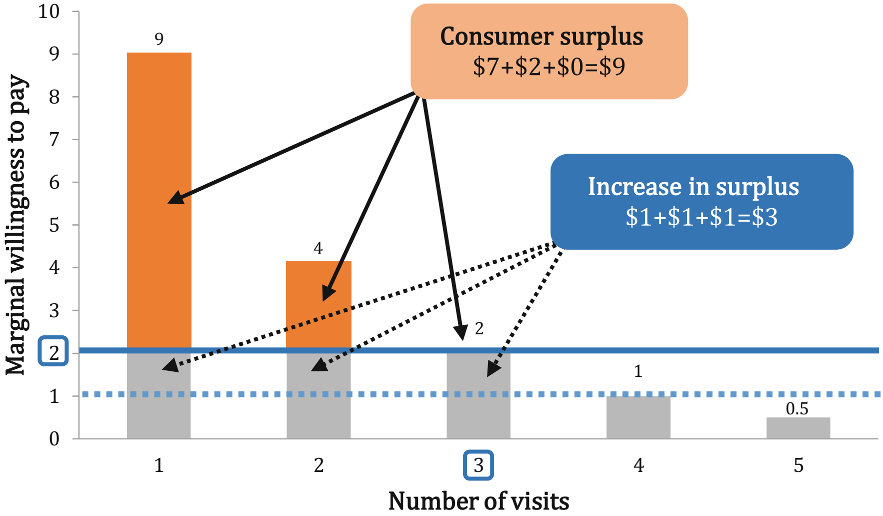


Fig. 9.2 Consumer surplus in a discrete choice setting: example 1

want to pay more than \$3 for the project. This amount represents the willingness to pay of the agent for a better access to the park.

Let us now consider a framework with two agents, indexed 1 and 2, respectively. They differ in their tax base (b_1 and b_2) and consequently in their tax price, $p_1 = cb_1/(b_1 + b_2)$ and $p_2 = cb_2/(b_1 + b_2)$, with $p_1 + p_2 = c$. They have different preferences about the public good: $u_1 = x + v_1(z)$ and $u_2 = x + v_2(z)$. Cost benefit analysis aims at selecting one single level of public good by maximizing the total surplus $s_1(z) + s_2(z)$. This amounts to differentiate $v_1(z) + v_2(z) - cz$ with respect to z. We obtain:

$$v_1'(z) + v_2'(z) = c$$

In equilibrium, the sum of the marginal WTP is equal to the marginal cost of production. In economics, this is known as the Pareto-optimality condition for the supply of a public good. For instance, in Fig. 9.3, agent 1 desires a lower quantity of public good than agent 2. To maximize the surplus, cost benefit analysis considers the demand of the whole society (by summing the two demand curves) and chooses a solution that lies between the ideal points of the agents (such that the budget constraint $p_1 + p_2 = c$ is fulfilled). By doing so, we are guaranteed to reach a situation of Pareto-optimality which, by definition, is a state of allocation of resources in which it is impossible to make any one individual better off without making at least one individual worse off.

The concept of Pareto-optimality should not be mistaken with that of Pareto-improvement, which denotes a "move" that benefits an individual or more without

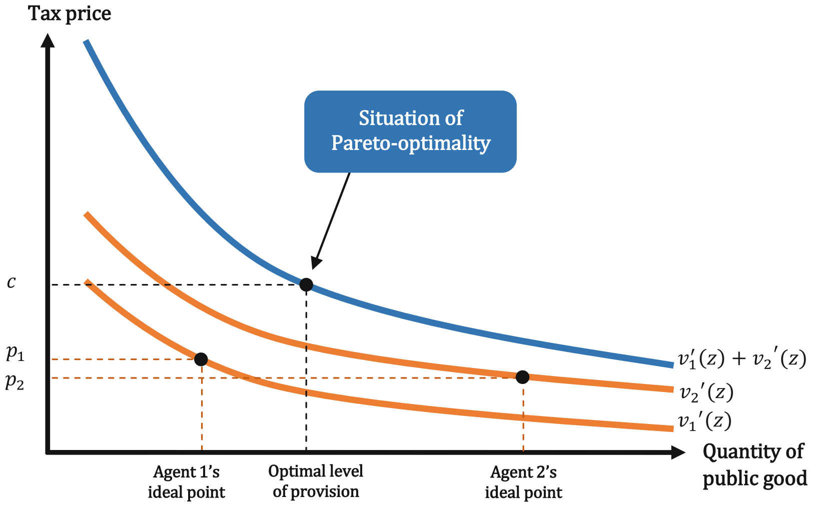


Fig. 9.3 The optimal provision of public goods

Table 9.2 The concept of Pareto-optimality: example 2

Project	Agent 1's surplus	Agent 2's surplus	Agent 3's surplus	Total surplus
Strategy S_1	$5	$25	$75	$105
Strategy S_2	$0	$24	$84	$108

reducing any other individual's well-being. To illustrate, consider three agents who have preferences about two competing strategies. In Table 9.2, those preferences are expressed in monetary terms. We can see that agents 1 and 2 prefer strategy S_1, while agent 3 prefers strategy S_2. In this example, the concept of Pareto-improvement is not useful as it is impossible to implement a policy change without making at least one individual worse off. A change from S_1 to S_2 is not Pareto-improving but neither is a move from S_2 to S_1. Then, how can we reach a situation of Pareto-optimality? The answer is through reallocation. One has to rely on what is termed a Kaldor-Hicks improvement. A move is more efficient as long as everyone can be compensated to offset any potential loss. Using this criterion, one would typically select strategy S_2. To clarify the whys and wherefores, one needs at this stage to understand that what matters in cost benefit analysis is welfare. In Table 9.2, agent 3 derives a high level of satisfaction from S_2 and is willing to pay a lot for it, even if this means to compensate the welfare loss of the other agents. For instance, if agent 3 gives $5 to agent 1 and $1 to agent 2, a move from S_1 to S_2 would benefit the whole society.

Cost benefit analysis provides public managers with a decision criterion based on the Kaldor-Hicks criterion. A project is an improvement over the status quo if the sum of welfare variations is positive. If the gains exceed the losses, then the winners could in theory compensate the losers so that policy changes are Pareto-improving. Yet, the compensation scheme must be chosen carefully. Cost benefit analysis is in this respect not equipped to conceive and implement redistributive schemes. Instead, the approach aims at selecting a particular policy among competing strategies. It nevertheless remains that redistribution, if carried out, can strongly affect work incentives, induce mobility, generate tax evasion or encourage tax fraud.

9.3 Discount of Benefits and Costs

Project selection starts with an option analysis discussed at the level of a planning document such as a master plan. The set of strategies is generally reduced so that at least three alternatives are examined: (1) a baseline strategy (or status quo) which is a forecast of the future without investment; (2) a minimum investment strategy and (3) a maximum investment strategy. Once the strategies are identified, a financial analysis is implemented to determine whether they are sustainable and profitable (a chapter has been dedicated to this step). If the financial analysis is not conclusive, then cost benefit analysis must outline some rationale for public support by demonstrating that the policy generates sufficient economic benefits. This step, termed economic appraisal, aims to assess the viability of a project from the society perspective. To do so, all economic impacts are expressed in terms of equivalent money value. Discounting then renders these items fully comparable by multiplying all future cash flows by a discount factor.

Formally, let $NB_t{=}B_t - C_t$ denote for each year t the net economic benefit, defined as the difference between total benefits B_t and total costs C_t. Discounting is accomplished by computing the economic net present value (*ENPV* hereafter):

$$ENPV = NB_0 + \delta_1 NB_1 + \cdots + \delta_T NB_T$$

where T represents the time horizon of the project and δ_t $(t = 1 \ldots T)$ are the discount factors by which the net benefits at year t are multiplied in order to obtain the present value. The discount factors are lower than one and decreasing with the time period. They are defined as:

$$\delta_t = \frac{1}{(1+r)^t}, \quad \text{for } t = 1 \ldots T$$

where r denotes the economic discount rate. This rate is different from the discount rate used in the financial appraisal. It does not represent some opportunity cost of capital (the return obtained from a best alternative strategy). It reflects instead the society’s view on how future benefits and costs should be valued against present ones. This rate is generally computed and recommended by government agencies such as the Treasury, or upper authorities such as the European Union. Their values may differ from 2 to 15% from one country to another. They are usually expressed in annual terms. When the analysis is carried out at current prices (resp. constant), the discount rate is expressed in nominal terms (resp. real) accordingly.

A positive economic net present value provides evidence that the project is desirable from a socio-economic point of view. In Excel, the computation can be accomplished with the *NPV* function. This command calculates the net present value via two entries: (1) a discount rate and (2) a series of future payments. One needs to enter the following formula in a cell:

$$= value0 + NPV(rate, value1, value2, \ldots)$$

where *rate* is the discount rate, *value*0 represents the first cash flow. This cash flow is excluded from the *NPV* formula because it occurs in period 0 and should not be discounted. Last, "value1, value2, … " is the range of cells containing the subsequent cash flows.

The following statement holds true in many occasions: the higher is the discount rate, the less likely the net present value is to reach positive values. The reason behind this is that most policy decisions involve large immediate outlays for building the infrastructure. Benefits on the other hand are observed in the future all along the project's life. When the discount rate increases, the value of future inflows decrease, which thereby reduces the *ENPV*. At some point, known as the internal rate of return (*EIRR*), the net present value reaches negative values. We have:

$$EIRR = r \text{ such that } ENPV(r) = 0.$$

The internal rate of return is defined as the discount rate that zeroes out the economic net present value of an investment. It cannot be determined by an algebraic formula but can be approximated in Excel using the *IRR* function:

$$= IRR(value0, value1, value2, \ldots)$$

The formula yields the internal rate of return for a series of cash flows (here *value*0 , *value*1 , *value*2), starting from the initial period. Values must contain at least one positive value and one negative value.

The *EIRR* is an indicator of the relative efficiency of an investment. It should however be used with caution as multiple solutions may be found, especially when large cash outflows appear during or at the end of the project. An example is provided in Fig. 9.4. While strategy S_1 is characterized by a negative net benefit observed only in period 1, strategy S_2 induces negative values both at the beginning and at the end of the project. As a consequence, strategy S_2 yields two internal rates of return, around 1% and 7% respectively, while strategy S_1 generates only one *EIRR*, around 4%. Given these difficulties, the net present value is often considered as a more suitable criterion for comparing alternative strategies. The strategy with the highest *ENPV* is designated as the most attractive and chosen first. For instance,

t	Net benefits (strategy S_1)	Net benefits (strategy S_2)
1	−100000	−70000
2	7000	10000
3	7500	10500
4	8000	10500
5	8000	10500
6	8000	10500
7	8000	10500
8	8000	10500
9	8000	10500
10	8000	10500
11	8000	10500
12	8000	10000
13	8000	10000
14	8000	10000
15	8000	10000
16	8000	10000
17	8000	10000
18	8000	10000
19	8000	10000
20	8000	−120000

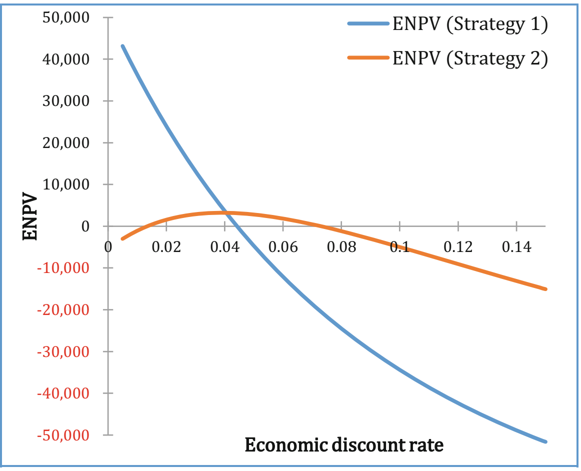


Fig. 9.4 Multiple internal rate of returns: example 3

we can see from Fig. 9.4 that strategy S_1 prevails over strategy S_2 only for small discount rates, lower than 4% approximately.

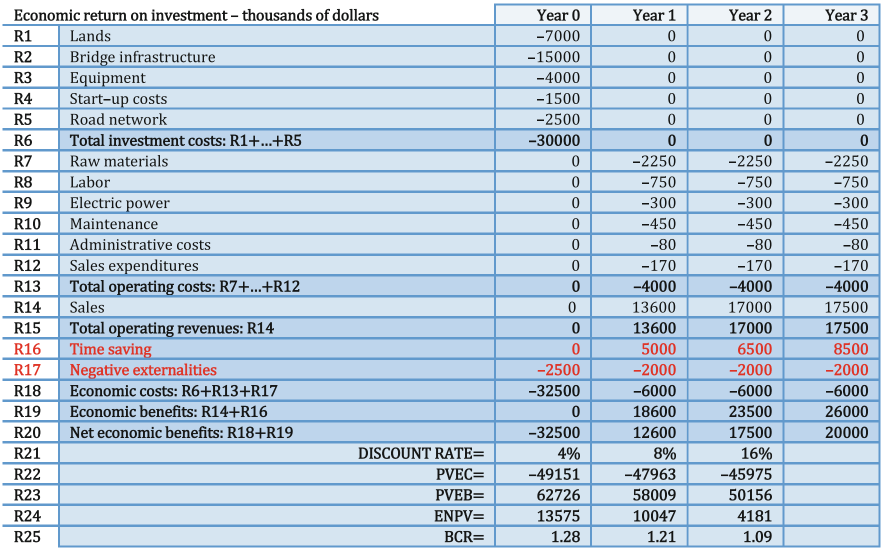
An alternative approach to assess the relative efficiency of an investment is the benefit-cost ratio. It is defined as:

$$BCR = \frac{PVEB}{PVEC} \text{ with } PVEB = \sum_{t=0}^{T} \frac{B_t}{(1+r)^t} \text{ and } PVEC = \sum_{t=0}^{T} \frac{C_t}{(1+r)^t}$$

The *BCR* is computed from the present value of the benefits (*PVEB*) and the present value of the costs (*PVEC*). Ideally, it should be higher than 1 and maximized. Unlike the *EIRR*, the *BCR* has the advantage of being always computable.

When comparing two alternative investments of different size, the *BCR* and the *ENPV* may reach different conclusions. The reason is that the *BCR* is a ratio and, therefore, like the *EIRR*, it is independent of the amount of the investment. Consider for instance two alternative strategies: a small investment project versus a large one. The smaller project induces net benefits and costs which amount to $PVEB = 10$ and $PVEC = 5$ million dollars respectively. This yields a *BCR* of 2 and an *ENPV* of 5 million dollars. The larger project generates benefits and costs which amount to $PVEB = 50$ and $PVEC = 40$ million dollars, respectively. This generates a *BCR* of 1.25 and an *ENPV* of 10 million dollars. As can be seen, while the *BCR* is higher for the smaller project, the *ENPV* is higher for the larger project.

Direct effects and externalities, when they are quantified, are directly included as new items in the financial analysis, thus filling the cash flow statement with additional rows. To illustrate, Fig. 9.5 provides a detailed presentation of costs



Economic return on investment – thousands of dollars		Year 0	Year 1	Year 2	Year 3
R1	Lands	–7000	0	0	0
R2	Bridge infrastructure	–15000	0	0	0
R3	Equipment	–4000	0	0	0
R4	Start–up costs	–1500	0	0	0
R5	Road network	–2500	0	0	0
R6	**Total investment costs: R1+...+R5**	**–30000**	**0**	**0**	**0**
R7	Raw materials	0	–2250	–2250	–2250
R8	Labor	0	–750	–750	–750
R9	Electric power	0	–300	–300	–300
R10	Maintenance	0	–450	–450	–450
R11	Administrative costs	0	–80	–80	–80
R12	Sales expenditures	0	–170	–170	–170
R13	**Total operating costs: R7+...+R12**	**0**	**–4000**	**–4000**	**–4000**
R14	Sales	0	13600	17000	17500
R15	**Total operating revenues: R14**	**0**	**13600**	**17000**	**17500**
R16	Time saving	0	5000	6500	8500
R17	Negative externalities	–2500	–2000	–2000	–2000
R18	Economic costs: R6+R13+R17	–32500	–6000	–6000	–6000
R19	Economic benefits: R14+R16	0	18600	23500	26000
R20	Net economic benefits: R18+R19	–32500	12600	17500	20000
R21	DISCOUNT RATE=	4%	8%	16%	
R22	PVEC=	–49151	–47963	–45975	
R23	PVEB=	62726	58009	50156	
R24	ENPV=	13575	10047	4181	
R25	BCR=	1.28	1.21	1.09	

Fig. 9.5 Economic return on investment: example 4

and benefits for a bridge project. For simplicity of exposition, the time horizon is set to 3 years. The project involves an immediate outlay of $30 million (for the lands, bridge infrastructure, equipment, etc.) and is followed by annual operating expenditures of $4 million (raw materials, labor, etc.). It generates annual revenues via a toll, which are expected to amount to $13.6 million for the first year, $17 million for the second year, and $17.5 million for the third year. The main economic benefit of the bridge project is the time saved by users (row R16). Those benefits may have been valued for instance at the opportunity cost of time, which is the fraction of salary the users could earn with the time saved, net of toll fees. Furthermore, the bridge generates negative externalities due to noise and aesthetic impacts, both at the time of construction and during the following years (row R17). Those external effects may have been estimated using revealed preferences techniques. To avoid double-counting, the economic appraisal does not account for external sources of financing (interest and principal repayment, private equity) as they are supposed to cover the initial investment costs.

In practice, a road connection induces several economic effects. It generates savings in travel time for the users. In the meantime, if the infrastructure is equipped with a toll gate, it provides a revenue stream for the operator. In that context, the toll is a cost from the point of view of the users and a revenue from the point of view of the operator. The effects cancel each other out. To better assess the impact on welfare for the whole society (users and operator), the users' benefits must be expressed net of fees. In other words, the consumer surplus (e.g., time saving in Fig. 9.5) is computed as the difference between the gains in terms of time saved and the toll fees. By doing so, toll revenues/fees are finally excluded from the economic appraisal:

$$\underbrace{\begin{pmatrix}\text{Gains in terms}\\ \text{of time saved}\end{pmatrix} - \begin{pmatrix}\text{Toll}\\ \text{fees}\end{pmatrix}}_{\text{Time saving (R16)}} + \underbrace{\begin{pmatrix}\text{Toll}\\ \text{fees}\end{pmatrix}}_{\text{Sales (R14)}} = \begin{pmatrix}\text{Gains in terms}\\ \text{of time saved}\end{pmatrix}$$

The approach is thus different from that of a financial appraisal, where toll revenues are included as a positive cash flow, to assess the financial sustainability and profitability of the investment strategy.

In Fig. 9.5, the most important row is R20, which represents the net benefits of the project for each time period. Summing all the flows (−32,500 + 12,600 + 17,500 + 20,000) to evaluate the suitability of the project would make us assume that future flows matter as much as present flows. In practice, however, one usually prefers to give lower importance to future cash flows. Discounting is accomplished by applying each year a discount factor that reflects the value of future cash flows to society. Consider for instance Fig. 9.5 where several discount rates (row R21) have been considered to assess the desirability of the project. With a discount rate equal to 4%, the economic net present value is computed as:

$$ENPV(4\%) = -32,500 + \frac{12,600}{1.04} + \frac{17,500}{1.04^2} + \frac{20,000}{1.04^3} \approx 13,575$$

Similarly, using information from total costs (row R18) and total benefits (R19), we have:

$$PVEC(4\%) = (-)32,500 + \frac{(-)6000}{1.04} + \frac{(-)6000}{1.04^2} + \frac{(-)6000}{1.04^3} \approx (-)49,151$$

$$PVEB(4\%) = 0 + \frac{18,600}{1.04} + \frac{23,500}{1.04^2} + \frac{26,000}{1.04^3} \approx 62,726$$

Equivalently, the difference between these two expressions yields the net present value:

$$ENPV(4\%) = PVEB(4\%) - PVEC(4\%) \approx 13,575$$

The *ENPV* decreases to 10,047 when the discount rate increases to 8%, and to 4181 when it equals 16% (row R24). Those values are positive and provide evidence that the project is desirable.

The benefit-cost ratios are provided in row R25 of Fig. 9.5. They are higher than one, which points out again the attractiveness of the project, no matter what the value of the discount rate is. They are computed from rows R22 and R23. For instance, for a 4% discount rate, we have:

$$BCR(4\%) = \frac{PVEB(4\%)}{PVEC(4\%)} = \frac{62,726}{49,151} \approx 1.28$$

Statistical programming languages such as R-CRAN have proven to be very handy when it comes to assigning probability distributions to particular events. In the bridge example, cash flows may for instance be characterized in a less deterministic manner to better assess the uncertainty of the project. This analytical technique, known as probabilistic sensitivity analysis, is further detailed in Sect. 9.6. As a preliminary step, Fig. 9.6 provides the codes to be used if one wants to calculate the *ENPV* and other performance indicators in R-CRAN.

Figure 9.6 reproduces the results obtained in Fig. 9.5. The first step consists in downloading the Excel worksheet in R-CRAN. For this purpose, the table has been modified and consists only of raw data, cleaned of totals, and rearranged so that the costs and benefits are presented successively. The command *read*.*table* reads the file (saved as a *.csv* file on disc *C*:) and creates a data frame from it, renamed *D*. This yields a table equivalent to Fig. 9.5. The object *D* has the properties of a matrix. An element observed in row *i* and column *j* is referred to as $D[i,j]$.

Elements of *D* are summable. The cost vector *C* is for instance obtained by summing rows 2 to 13, while the benefit vector *B* is the result of the sum of rows 14 and 15. The command *colSums* is used to ensure that only the rows are summed over columns 2 to 5. Last, package *Fincal* and its function *npv* are used to compute the different performance indicators. The entry *Disc*.*Rate* specifies the vector of discount rates to be applied in the *npv* function. As the costs are already expressed in negative values, the analysis does not need to subtract them. For the *BCR*, one needs to use the function *abs*() to express the costs in absolute value.

Under special circumstances, an indicator known as the net benefit investment ratio (*NBIR*) can also be examined. It is defined as the ratio of the present value of the benefits (*PVEB*), net of operating costs ($PVEC - PVK$), to discounted investment costs (*PVK*):

$$NBIR = \frac{PVEB - (PVEC - PVK)}{PVK}$$

For instance, in Fig. 9.5, the investment outlay is $PVK = 30{,}000$. For a discount rate equal to 4%, the present value of operating costs amounts to $PVEC - PVK = 49{,}151 - 30{,}000 = 19{,}151$. We also have $PVEB = 62{,}726$. This yields a NBIR equal to $(62{,}726 - 19{,}151)/30{,}000 = 1.45$. This ratio assesses the economic profitability of a project per dollar of investment. It is very useful when an authority is willing to finance several projects but has to face a budget constraint. The method allows the best combination of projects to be selected.

```
> D=read.table("C://mydataCBA1.csv",head=FALSE,sep=";")
> D
                        V1     V2    V3    V4    V5
1                     Year      0     1     2     3
2                    Lands  -7000     0     0     0
3    Bridge infrastructure -15000     0     0     0
4                Equipment  -4000     0     0     0
5           Start-up costs  -1500     0     0     0
6             Road network  -2500     0     0     0
7            Raw materials      0 -2250 -2250 -2250
8                    Labor      0  -750  -750  -750
9           Electric power      0  -300  -300  -300
10             Maintenance      0  -450  -450  -450
11    Administrative costs      0   -80   -80   -80
12      Sales expenditures      0  -170  -170  -170
13 Negative externalities  -2500 -2000 -2000 -2000
14                   Sales      0 13600 17000 17500
15             Time saving      0  5000  6500  8500

> # Total costs and benefits
> C=colSums(D[2:13,2:5])
> C
    V2     V3     V4     V5
-32500  -6000  -6000  -6000
> B=colSums(D[14:15,2:5])
> B
   V2    V3    V4    V5
    0 18600 23500 26000
> # Discounted values
> library(FinCal)
> Disc.RATE=c(0.04,0.08,0.16)
> PVEC=npv(Disc.RATE,C)
> PVEC
[1] -49150.55 -47962.58 -45975.34
> PVEB=npv(Disc.RATE,B)
> PVEB
[1] 62725.59 58009.32 50155.91
> ENPV=npv(Disc.RATE,B+C)
> ENPV
[1] 13575.046 10046.741  4180.573
> BCR=PVEB/abs(PVEC)
> BCR
[1] 1.276193 1.209470 1.090931
> irr(B+C)
[1] 0.231104
```

Fig. 9.6 Discounting cash flows with R CRAN: example 4

To illustrate the advantages of the *NBIR*, let us consider an authority that has a budget of $1,000,000. Information about the competing strategies is provided in Table 9.3 (amounts in thousands of dollars). If one were to compare the alternatives using net present values, strategies S_1 and S_2 would appear as the best alternatives. The budget constraint would be fulfilled and, overall, one would obtain a total

Table 9.3 Ranking of project under capital rationing: example 5

Strategy	Investment costs *PVK*	Present value of benefits net of operating costs $PVEB - (PVEC - PVK)$	Economic net present value *ENPV*	Profitability ratio *NBIR*
S_1	400	500	100	1.30
S_2	600	680	80	1.10
S_3	300	370	70	1.25
S_4	200	260	60	1.23
S_5	500	530	30	1.06
S_6	200	220	20	1.13

ENPV equal to 100+80=180 thousand dollars. The *ENPV*, however, does not assess accurately the return on investment. With this criterion, larger projects are more likely to be selected. In contrast, the net benefit investment ratio assesses the profitability of the investment independently of its size (like the *BCR*). Using this criterion, we can see from the last column of Table 9.3 that the approach would rank strategy S_1 first, then S_3, S_4 and S_6. Assuming that only part of S_6 is financed (50% of it), this combination of strategies would yield a *ENPV* equal to 100+70+60+20/2=240 thousands of dollars. Thus, under capital rationing, the *NBIR* appears as a very performing selection criterion.

9.4 Accounting for Market Distortions

Conversion factors are related to the concept of shadow prices (also termed accounting or economic prices). They reflect the cost of an activity when prices are unobservable or when they do not truly reflect the real cost to society. Shadow prices do not relate to real life-situations. They correspond instead to the prices that would prevail if the market was perfectly competitive. For instance, the prices used in the financial appraisal (which are usually referred to as market prices) are likely to include taxes or government subsidies. Prices have to be adjusted in consequence, to better reflect trading values on a hypothetical free market. Yet, the term market prices can be misleading. In cost benefit analysis, it stands for the actual price of transaction subject to market distortions, while in common language a "market economy" denotes an economy that is little planned or controlled by government. To avoid any confusion in the remaining of the chapter, we will use the term "financial prices" as a synonym for "market prices".

What is the true economic cost of inputs and outputs for cost benefit analysis use? To answer this question, it is convenient to appeal to the concept of opportunity cost. We should ask ourselves "what would be the value of inputs and outputs if they were employed somewhere else?" For instance, the Little-Mirrlees-Squire-van der Tak method advocates the use of international prices (converted to domestic currencies), for evaluating and comparing public projects in different countries. The rationale for the method is that world prices more accurately reflect the opportunities that are available to the countries. The method has for instance been frequently used for calculating shadow prices for aid projects in developing countries, where markets are often considered more distorted.

In some cases, it is easy to correct for price distortions *ex ante*. For instance, the cash flows should be net of VAT (value added tax). The value of land and buildings provided by other public bodies can be included directly at their true costs. In some other cases, however, the use of a conversion factor may be necessary. Formally, a shadow price is defined as:

$$\text{Shadow price} = \text{Financial price} \times CF$$

The conversion factor CF approximates the degree of perfection of the market. In an undistorted economy, the conversion factor is equal to one and shadow prices are identical to financial prices. Should CF be higher than one, then the financial prices would yield an underestimation of the true value of inputs and outputs. If lower than one, they yield instead an overestimation. Several examples are provided below.

Regulated Price This situation occurs when an input is made available at a lower price by the public sector, hence yielding an underestimation of the true costs to taxpayers. Common examples are a land proposed at a reduced price by a public authority while it may earn a higher rent or price otherwise; or an energy sold at a regulated tariff. The conversion factor should reflect these opportunity costs:

$$CF = \frac{\text{Shadow price}}{\text{Financial price}}$$

Consider for instance row R1 of Fig. 9.5. Assume that the land has been sold at 80% of the usual price. The conversion factor is computed as $CF = 1/0.80 = 1.25$. In this case, the shadow value of lands amounts to $7000 \times 1.25 = 8750$ thousand dollars. Similarly, assume that electricity (row R9 of Fig. 9.5) is produced at a tariff that covers only 60% of marginal cost. The true cost to society is defined as $300 \times (1/0.60) = 500$ thousand dollars.

Undistorted Labor Market When the labor market is perfectly competitive, the economy reaches an equilibrium where only "voluntary unemployment" prevails. People have chosen not to work solely because they do not consider the equilibrium wages as sufficiently high. If this is to be the case, the project would only divert the labor force from their current use, at its market value (assuming that the project is not large enough to influence wages). The conversion factor is defined as $CF = 1$.

Distorted Labor Market When minimum wages are adopted in a given labor market, the quantity of labor supplied increases (the number of workers who wish to work) while the demand for labor decreases (the number of positions offered by employers). This creates a situation of involuntary unemployment, where labor supply exceeds demand. Some individuals, in particular unskilled workers, are willing to work at the prevailing wage but are unable to find employment. In such markets, the opportunity costs of hiring these individuals is lower because the project would hire agents who would have been unemployed otherwise. In that context, the conversion factor is lower than one (but not necessarily zero as the new workers could also work in the informal economy or just enjoy leisure). For instance, in its "Guide to Cost-Benefit Analysis of Investment", the European Commission advocates the use of the regional unemployment rate (u) as a basis for the determination of the shadow wage:

$$CF = (1 - u)(1 - s)$$

where s is the rate of social security payments and relevant taxes that should be excluded from the financial prices as they also represent a revenue for the public sector. Assume for instance that the unemployment rate is $u = 9\%$ and $s = 15\%$. Row R8 of Fig. 9.5 would be replaced with $750 \times (1 - 9\%)(1 - 15\%) \approx 522$ thousand dollars.

Import Tariffs A tariff on imports increases the costs of inputs used in the project and, at the same time, induces additional revenue for the central government which can be used for other purposes. The true cost of inputs would be overestimated if no adjustment of cash flows were to be made. To better adjudge the true cost to society, any tariff should be excluded from the financial statement. This is equivalent to say that only the CIF price (cost plus insurance and freight) should be used for valuing the imported inputs. Let t_m denotes the proportional tax rate on imports. The conversion factor is defined as:

$$CF = \frac{\text{Shadow (world) price}}{\text{Financial price}} = \frac{\text{CIF price}}{\text{CIF price} \times (1 + t_m)} = \frac{1}{(1 + t_m)}$$

Assume in row R7 of Fig. 9.5 that the raw materials have been imported. If the tax rate is equal to $t_m = 20\%$, then the value to be used in the economic appraisal is $2250 \times (1/1.2) = 1875$ thousand dollars. The tariff has been removed from the price.

Export Subsidies If the project receives an additional payment for exporting, it would be at the expense of the domestic taxpayers. The reasoning is thus similar to that previously made regarding an import tariff. Any subsidy should be excluded from the financial flows. This amounts to consider the "free on board" FOB price only (before insurance and freight charges):

$$CF = \frac{\text{Shadow (world) price}}{\text{Financial price}} = \frac{\text{FOB price}}{\text{FOB price} \times (1 + s_x)} = \frac{1}{(1 + s_x)}$$

where s_x denote the rate of subsidy.

Major Non-traded Goods The fact that an input is not imported or an output not exported does not necessarily mean that they are not subject to trade distortions. The existence of a tariff that aims to protect the domestic market may explain for instance the current use of local inputs. In such situations, the tax on imports is also reflected in the domestic prices. Similarly, a subsidy that encourages the producers to export may generate an increase in the price level. When data are available, these distortions can be valued using the same approach as previously. Assume for instance that the government has imposed an import tax of 25% on equipment and infrastructure (rows R2 and R3 of Fig. 9.5). The conversion factor to be used is $1/(1+25\%)=0.8$, even if those goods are bought in the domestic market.

Minor Non-traded Goods For minor items, or when data are not easily available, the European Commission advocates the use of a "standard conversion factor". The latter is specified as:

$$SCF = \frac{M + X}{M + T_m - S_m + X - T_x + S_x}$$

where M denotes the total imports valued at the CIF price, X the total exports valued at the FOB price, T_m and S_m the total import taxes and subsidies, and T_x and S_x the total export tax and subsidies. In simple words, SCF is the ratio of the value at world prices of all imports and exports to their value at domestic prices. It generalizes the previous formulas and provides a general proxy of how international trade is distorted due to trade barriers. It assesses how the prices would change on average if such barriers were removed. For instance, should $T_m + S_x$ be larger than $S_m + T_x$, then, on average, the country would support the domestic producers. In that context, the standard conversion factor would be lower than one, meaning equivalently that the trade balance is artificially increased, or that the domestic prices are overvalued. The inverse of the standard conversion factor ($1/SCF$) is also termed "shadow exchange rate factor". In practice, the approach is used when one wants to compare the economic performance of competing projects in different developing countries.

The conversion factors previously determined are applied to the cash flows of the bridge project (Fig. 9.6). They are displayed in the last column of Fig. 9.7. For all minor traded items (for which data on trade distortions were not accurately available) a standard conversion factor equal to 0.9 has been considered (emphasized in red). As can be seen from the *ENPV* and *BCR* criteria, the project remains economically viable, even at a discount rate of 16%. For this rate, the performance indicators have been computed has follows, and rounded for presentation purposes:

Efficiency cash flows – thousands of dollars		Year 0	Year 1	Year 2	Year 3	CF
R1	Lands	−7875				(1/0.80)×0.9
R2	Bridge infrastructure	−12000				0.8
R3	Equipment	−3200				0.8
R4	Start-up costs	−1350				0.9
R5	Road network	−2250				0.9
R6	**Total investment costs: R1+R5**	**-26675**	**0**	**0**	**0**	
R7	Raw materials		−1875	−1875	−1875	(1/1.20)
R8	Labor		−522	−522	−522	0.91×0.85×0.9
R9	Electric power		−450	−450	−450	(1/0.60)×0.9
R10	Maintenance		−405	−405	−405	0.9
R11	Administrative costs		−72	−72	−72	0.9
R12	Sales expenditures		−153	−153	−153	0.9
R13	**Total operating costs: R7+...+R12**	**0**	**-3477**	**-3477**	**-3477**	
R14	Sales		12240	15300	15750	0.9
R15	**Total operating revenues: R14**	**0**	**12240**	**15300**	**15750**	
R16	**Time saving**	**0**	**5000**	**6500**	**8500**	
R17	**Negative externalities**	**-2500**	**-2000**	**-2000**	**-2000**	
R18	**Economic costs: R6+R13+R17**	**-29175**	**-5477**	**-5477**	**-5477**	
R19	**Economic benefits: R15+R16**	**0**	**17240**	**21800**	**24250**	
R20	**Net economic benefits: R19+R18**	**-29175**	**11763**	**16323**	**18773**	
R21	**DISCOUNT RATE=**	**4%**	**8%**	**16%**		
R22	**PVEC=**	**-44374**	**-43290**	**-41476**		
R23	**PVEB=**	**58290**	**53903**	**46599**		
R24	**ENPV=**	**13916**	**10614**	**5123**		
R25	**BCR=**	**1.31**	**1.24**	**1.12**		

Fig. 9.7 Corrections for market distortions: example 4

$$PVEC(16\%) = (-)29,175 + \frac{(-)5477}{1.16} + \frac{(-)5477}{1.16^2} + \frac{(-)5477}{1.16^3} \approx (-)41,476$$

$$PVEB(16\%) = 0 + \frac{17,240}{1.16} + \frac{21,800}{1.16^2} + \frac{24,250}{1.16^3} \approx 46,599$$

$$ENPV(16\%) = PVEB(16\%) - PVEC(16\%) \approx 5123$$

$$BCR(16\%) = PVEB(16\%)/PVEC(16\%) \approx 1.12$$

In theory, conversion factors should provide the evaluator with a better decision tool. However, in practice, they are unique to the context and the methods used to approximate those weights are often based on rough calculations. While time-consuming, those adjustments can also be of minor importance for the investment decision. Therefore, conversion factors should be used with caution or in exceptional cases only. It is also possible to provide the results of the analysis both with and without shadow prices.

At this stage of the analysis, the net benefits of the bridge project should be analyzed against those of larger and smaller investments. Moreover, it may be useful to check whether some excluded items are likely to compromise or reinforce the decision made, especially if the *ENPV* reaches surprisingly high values. For instance a *BCR* approaching 2 would mean that the economic benefits are twice as high as the economic costs. Then why was the project not implemented before? If data is available, it can also be useful to compare the economic return of the investment with that of an already existing project. Last, if some important costs and benefits have been excluded from the analysis because it was not possible to monetize them, a description of these items should at least be provided in physical terms. In this respect, a multi-criteria analysis can be used to combine both physical terms and monetary terms (e.g., the *ENPV*) into a single indicator.

9.5 Deterministic Sensitivity Analysis

Cost benefit analysis generally goes further by questioning the valuation of the costs and benefits themselves. When some important items are difficult to estimate but yet quantified, or when some degree of uncertainty is inherent to the study, a sensitivity analysis can be used to examine the degree of risk in the project. This can take the form of a partial sensitivity analysis, a scenario analysis, a Monte Carlo analysis or a mean-variance analysis.

Uncertainty not only refers to variations in the economic environment (uncertain economic growth, natural hazards, modification of relative prices), but also to sampling errors resulting from data collection. Consider for instance the bridge example (Fig. 9.7). The study makes assumptions about the cost of inputs, about the amount of sales, and uses estimates to calculate the economic effects (time saving, externalities). Those cash flows can only be assessed or forecasted imprecisely. This affects in return the precision of the *ENPV*. The purpose of a sensitivity analysis is to identify these sources of variability and assess how sensitive the conclusions are to changes in the variables in question.

In a partial sensitivity analysis, only one single variable is modified while holding the other variables constant. Figure 9.8 illustrates the approach by assessing first the effect of a change in energy costs (row R9 of Fig. 9.7) on the *ENPV* of the bridge project. A 15 percent increase in costs yields for instance a net benefit of $13,699 for a 4% discount rate, $10,412 for a 8% discount rate and $4947 for a 16% discount rate. The results and conclusions (suitability of the project) are not really sensitive to those variations. Similar results are obtained when the costs of raw materials vary from −15% to +15% (see Fig. 9.8).

A partial sensitivity analysis has its limits as the approach does not consider variations in more than one variable. To solve this issue, an alternative approach known as scenario analysis is commonly used. It combines the results in a three-

	Change in energy costs						
	−15%	−10%	−5%	0%	5%	10%	15%
ENPV (4%)	14134	14061	13989	13916	13844	13771	13699
ENPV (8%)	10815	10748	10681	10614	10546	10479	10412
ENPV (16%)	5299	5240	5182	5123	5065	5006	4947
	Change in raw materials						
	−15%	−10%	−5%	0%	5%	10%	15%
ENPV(4%)	14697	14437	14176	13916	13656	13396	13136
ENPV(8%)	11338	11097	10855	10614	10372	10130	9889
ENPV(16%)	5755	5544	5334	5123	4913	4702	4492

Fig. 9.8 Partial sensitivity analysis: example 4

Table 9.4 Scenario analysis: example 4

	Best-case Energy costs: −15% Raw materials: −15%	Most-likely Energy costs: 0% Raw materials: 0%	Worst-case Energy costs: +15% Raw materials: +15%
ENPV(4%)	14,914	13,916	12,918
ENPV(8%)	11,540	10,614	9687
ENPV(16%)	5931	5123	4316

Table 9.5 Scenario analysis and project selection: example 6

ENPV	Worst-case	Most-likely	Best-case
Strategy S_1	1000	1500	2000
Strategy S_2	−500	1000	3000
Strategy S_3	−500	1000	2000

scenario framework with two extreme scenarios and one most-likely scenario. These scenarios draw attention to the main uncertainties involved in the project. The idea is to focus on the upper boundaries (best-case scenario) and lower boundaries (worst-case scenario) of the study's results. For instance, considering variations that are similar to those of Fig. 9.8, the approach would combine the assumed lower and upper values (–15% and +15%) of both variables simultaneously to characterize the two extreme scenarios. By doing so, we obtain Table 9.4. We can see that the *ENPV* never reaches negative values, which gives support to the strategy at stake. The results can then be compared to those obtained for competing strategies.

Consider now Table 9.5 where the results of a scenario analysis are displayed for three competing strategies, in thousands of dollars. It can be seen that strategy S_3 is dominated by the other strategies as its net present value is always equal to or lower than the net present values of the other projects. Strategy S_3 is thus eliminated. If no other dominant strategy is apparent, different decision-rules can be employed:

1. The maximin rule consists in selecting the alternative that yields the largest minimum payoff. This rule would be typically used by a risk-averse decision-maker. In Table 9.5 for instance, the minimum payoff is 1000 for strategy S_1, while it is –500 for S_2. Strategy S_1 is thus selected. This ensures that the payoff will be of at least 1000 whatever happens.
2. The minimax regret rule consists in minimizing the maximum regret. Regret is defined as the opportunity cost incurred from having made the wrong decision. In Table 9.5, if one chooses strategy S_1 we have no regret when the worst-case or most-likely scenarios occur. On the other hand, if the best-case scenario occurs, the regret is $2000 - 3000 = -1000$. If one chooses strategy S_2, the regret amounts to $-500 - 1000 = -1500$ for the worst-case scenario, $1000 - 1500 = -500$ for the most-likely scenario, and zero for the best-case scenario. Overall, the maximum regret is –1000 for strategy S_1 and –1500 for

strategy S_2. Strategy S_2 is thus eliminated. The approach is relatively similar to the maximin approach as it favors less risky alternatives. The approach however accounts for the opportunity cost of not choosing the other alternatives. It is used when one wishes not to regret the decision afterwards.

3. The maximax rule involves selecting the project that yields the largest maximum payoff. For project S_1, the maximum payoff is 2000, while for S_2, it is 3000. A risk-lover decision-maker would favor project S_2.
4. The Laplace rule implies maximizing the expected payoffs assuming equiprobable scenarios. In our example, we would assign a probability of 1/3 to each scenario. The expected payoff for project S_1 is defined as: $(1/3) \times 1000 + (1/3) \times 1500 + (1/3) \times 2000 = 1500$. For project S_2, we have: $(1/3) \times (-500) + (1/3) \times 1000 + (1/3) \times 3000 = 1166$. Based on this decision rule, strategy S_1 is selected.

The two methods of sensitivity analysis described above are often considered as deterministic as they assess how costs are sensitive to pre-determined change in parameter values through upper and lower bounds. They may be sufficient to assess roughly the risk associated with a project, but they do not account for all the uncertainty involved. In particular, they do not evaluate precisely the probability of occurrence of each possible outcome. A probabilistic sensitivity analysis is a useful alternative in this respect. Instead of focusing only on extreme bounds, the approach examines a simulation model that replicates the complete random behavior of the sensitive variables. It allows the distribution of the *ENPV* to be fully examined.

9.6 Probabilistic Sensitivity Analysis

A probabilistic sensitivity analysis assigns a probability distribution to all sensitive variables. These variables thereby become random in the sense that their value is subject to variations due to chance. The approach, also known as Monte Carlo analysis, examines those variations simultaneously and simulates thousands of scenarios, which results in a range of possible *ENPV* with their probabilities of occurrence. The method requires detailed data about the statistical distribution of the variables in question. One can for instance use information about similar existing projects, observed variations in input prices or, if the direct benefits and externalities have been estimated in the context of the project, the estimated standard deviations.

Among the most common distribution patterns that are used in probabilistic sensitivity analysis, we may name the uniform distribution, the triangular distribution, the normal distribution, and the beta distribution. The triangular and beta distributions are frequently used because both can be characterized by the worst-case, most likely, and best-case parameters. The normal distribution on the other hand has proven to be very handy provided that information on standard deviations is available. In what follows, R-CRAN is used to illustrate the differences among

```
> # Uniform distribution
> energy=runif(1000000,-500,-400)
> plot(density(energy),main="5.1 Uniform distribution")

> # Triangular distribution
> library(triangle)
> energy=rtriangle(1000000,-500,-400,-450)
> plot(density(energy),main="5.2 Triangle distribution")

> # Normal distribution
> energy=rnorm(1000000,-450,100)
> plot(density(energy),main="5.3 Normal distribution")

> # (Generalised) Beta distribution
> library(mc2d)
> energy=rbetagen(1000000,1.5,1.5,-500,-400)
> plot(density(energy),main="5.4 Beta distribution (1.5,1.5)")

> energy=rbetagen(1000000,5,5,-500,-400)
> plot(density(energy),main="5.5 Beta distribution (5,5)")

> energy=rbetagen(1000000,2,5,-500,-400)
> plot(density(energy),main="5.6 Beta distribution (2,5)")
```

Fig. 9.9 Probability distributions in R-CRAN

them (see Figs. 9.9 and 9.10). Many other probability distributions exist. For simplicity of exposition, they are not presented here.

The uniform distribution assigns an equal probability of occurrence to each possible value of the random variable. It is used when little information is available about the distribution in question. Consider for instance the energy costs (row R9) of Fig. 9.7. They are initially set to –450 thousand dollars. We could assign instead a range of values between –500 and –400 with an equal probability of occurrence. With R-CRAN (Fig. 9.9), this amounts to use the *runif* command. The first entry denotes the number of randomly generated observations (here 1,000,000), while the second and third entries stand for the lower and upper limits of the distribution. Figure 9.10a provides the probability density function estimated with *plot*(*density*()). As can be seen, the uniform distribution yields a rectangular density curve (which would be perfectly rectangular with an infinite number of observations). Each value has an equal probability of occurring inside the range [−500, −400]. In Fig. 9.10, the bandwidth relates to the precision of the local estimations used to approximate the shape of the density curve and is of no interest for the present purpose.

The triangular distribution is a continuous probability distribution with a minimum value, a mode (most-likely value), and a maximum value. In contrast with the uniform distribution, the triangular distribution does not assign the same probability of occurrence to each outcome. The probability of occurrence of the lower and upper bounds is zero, while the maximum of the probability density function is obtained at the mode. This distribution is used when no sufficient or reliable information are available to identify the probability distribution more rigorously. In R-CRAN (Fig. 9.9), we can use the *rtriangle* function from the *triangle* package.

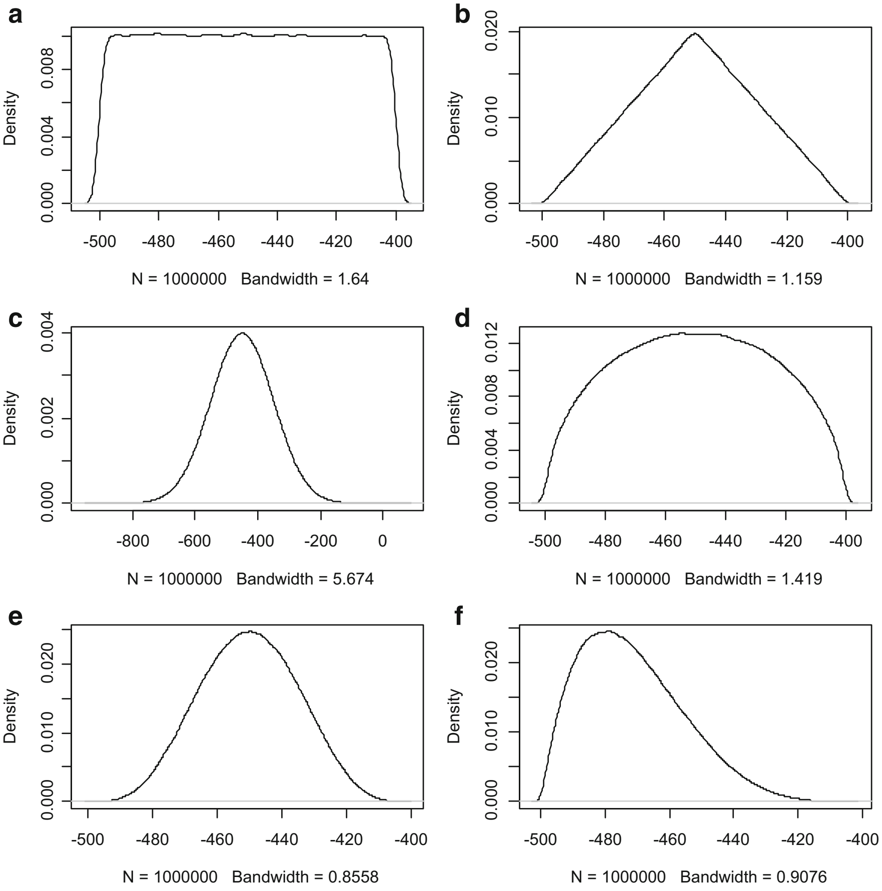


Fig. 9.10 Examples of probability distributions. (**a**) Uniform distribution. (**b**) Triangle distribution. (**c**) Normal distribution. (**d**) Beta distribution (1.5,1.5). (**e**) Beta distribution (5,5). (**f**) Beta distribution (2,5)

The first entry specifies the number of observations; the second and third entries are the lower and upper limit of the distribution; last entry stands for the mode of the distribution. As can be seen from Fig. 9.10b, the program yields a triangular probability density function with minimum and maximum values obtained at –500 and –400, respectively.

With the normal distribution, the density curve is symmetrical around the mean and has a bell-shaped curve (see Fig. 9.10c). The random variable can take any value from $-\infty$ to $+\infty$. Train punctuality is an example of a normal probability density function. A train arrives frequently just in time, less frequently 1 min earlier or late and very rarely 20 min earlier or late. In R-CRAN (Fig. 9.9), the command in use is *rnorm*. The first entry specifies the number of observations; the second is the mean; the third is the standard deviation. The simplest form of this distribution is known as the standard normal distribution. It has a mean of 0 and a variance of 1. In Fig. 9.9, the standard deviation is set arbitrarily to 100. In Fig. 9.10c, the values tend to distribute in a symmetrical hill shape, with most of the observations near the specified mean. Provided that information about the standard deviation is available, the normal distribution may provide a reasonable approximation to the distribution of many events. Meanwhile, a lot of variables are likely to be not normally distributed. They may exhibit skweness (asymmetry), or have a different kurtosis (differently curved). A way to capture these differences is to rely instead on the beta distribution.

The beta distribution is determined by four parameters: a minimum value, a maximum value and two positive shape parameters, denoted α and β. Depending on those parameters, the beta distribution takes different shapes. Like the uniform and triangular distributions, it models outcomes that have a limited range. In Fig. 9.9, energy costs are randomized using the *rbetagen* command available with the package *mc*2*d*. The first entry is the number of randomly generated observations; the second and third entries stand for parameters α and β; the last two entries represent the lower and upper bounds, respectively. When α and β are the same, the distribution is symmetric (Figs. 9.10d, e). As their values increase, the distribution becomes more peaked (Fig. 9.10e). When α and β are different, the distribution is asymmetric. For instance, in Fig. 9.10f, the curve is skewed to the right because α is set to be lower than β.

So far, we have assumed that there was uncertainty about the project's variables independently of each other. Those variables were considered uncorrelated. When this is the case, one can easily assign a random generator to each variable without worries. In some other cases however, the use of joint probability distributions is required. Consider for instance time saving (row R16 of Fig. 9.7) and sales (row R14). To some extent, those variables are likely to be correlated. The higher is the traffic, the higher are the sales revenues and the time saved by users (if there is induced traffic congestion, time saved may decrease and correlation would be negative). Traffic may also significantly affect maintenance costs (raw materials, labor, electric power, etc.). If one were to assign separately a random generator to these variables they would be varying independently from each other. We could for instance observe a decrease in sales and, in the meanwhile, a significant increase in time saving. To avoid those situations, it is generally advised to use a multivariate probability distribution. This can be done for instance with a multivariate normal distribution, provided that information about the covariance matrix is available.

A covariance matrix contains information about the variance and covariance for several random variables:

$$V = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \ldots & \mathrm{Cov}(x_1, x_K) \\ \mathrm{Cov}(x_1, x_2) & \mathrm{Var}(x_2) & \ldots & \mathrm{Cov}(x_2, x_K) \\ \vdots & \vdots & \ldots & \vdots \\ \mathrm{Cov}(x_1, x_K) & \mathrm{Cov}(x_2, x_K) & \ldots & \mathrm{Var}(x_K) \end{bmatrix}$$

Table 9.6 Dataset for comparison: example 4

Sales	Time
12,652	2310
11,688	2441
13,044	2408
12,343	2466
11,753	2092
14,292	2595
13,802	2810
12,249	2659
11,598	2485
12,239	2147

The variances appear along the diagonal and covariances appear in the off-diagonal elements. This matrix is symmetric (i.e. entries are symmetric with respect to the main diagonal). If a number outside the diagonal is large, then the variable that corresponds to that row and the variable that corresponds to that column change with one another.

The covariance matrix can be used to generate random observations that are dependent from each other. To illustrate, let us examine again the bridge project (example 4). Assume that we have obtained data about a bridge of equivalent size. We can use information about the existing infrastructure to build a random generator for both sales (variable *Sales*) and time saving (variable *Time*). Table 9.6 provides the dataset (10 years) and Fig. 9.11 illustrates the simulation method. The mean values are computed using the *mean* command, while the covariance matrix is directly obtained using function *cov*. Those values are used afterwards in the *rmvnorm* function of the package *mvtnorm*. Basically speaking, the *rmvnorm* function randomly generates numbers using the multivariate normal distribution. The first entry is the number of observations; the second entry is the vector of means; last entry specifies the covariance matrix. The *rmvnorm* command generates a matrix made of (1) as many columns as there are variables and (2) as many rows as they are observations. For illustrative purposes, the number of randomly generated observations is set to 10,000. We thus have two columns (one for *Sales* and one for *Time*) and ten thousands rows. If one were to use the resulting random generator in a Monte Carlo framework, we would generate instead a matrix made of as many columns as there are variables and as many rows as they are time periods.

Figure 9.11 ends with assessing the quality of the model by comparing the real distributions against the simulated ones. The first *plot* command provides a scatter plot of the relationship between sales and time saving using the existing dataset. As expected, some correlation between the variables is highlighted. As shown through a simple linear regression (*abline* command), this correlation is accurately taken into account by the simulated data (displayed in red in Fig. 9.12). To draw this regression line, we have made use of the ten thousand randomly generated observations. For simplicity of exposition, those observations are not displayed

```
> E=read.table("C://mydataCBA2.csv",head=TRUE,sep=";")

> mean(E$Sales)
[1] 12611
> mean(E$Time)
[1] 4929.833
> cov(E)
          Sales     Time
Sales 1086229.2 348657.1
Time   348657.1 323303.3

> library(mvtnorm)
> simu=rmvnorm(n=10000, mean=c(mean(E$Sales),mean(E$Time)),
+ sigma=cov(E))
> Simu.Sales=simu[,1]
> Simu.Time=simu[,2]

> plot(E$Sales~E$Time,main="6.1 Scatter plot")
> abline(lm(Simu.Sales~Simu.Time),col="red")
> plot(density(E$Sales),main="6.2 Probability density of sales")
> points(density(Simu.Sales),col="red",type="l")
> plot(density(E$Time),main="6.3 Probability density of time saving")
> points(density(Simu.Time),col="red",type="l")
```

Fig. 9.11 Simulating joint distributions with R-CRAN: example 4

on the graph. Last, Fig. 9.12b, c compare the probability density of the real observations (in black) with that of the simulated ones (in red). As can be seen, the model provides an accurate estimation of the phenomena in question.

Note that the simulations presented in Fig. 9.11 can be easily extended to more than two variables. The resulting random generator can also be used for simulating cash flows in a Monte Carlo framework. In this purpose, the randomly generated data can be multiplied by a trend variable, e.g., $1\,,(1+x)\,,(1+x)^2\,,\ldots$, to account for a possible automatic increase in sales in the first years of the project (for instance by x% each year).

A Monte Carlo simulation consists in assigning a probability distribution to each of the sensitive cash flows and running the model a high number of times to generate the probability distribution of the *ENPV*. A loop is created in which (1) sensitive cash flows are assigned a random number generator, (2) the *ENPV* is computed using the randomly generated cash flows and (3) the *ENPV* is saved for subsequent analysis. The simulation is repeated a large number of times (10,000 times or more) to provide a range of possible *ENPV*. This range is then used to estimate a probability distribution. One can then decide whether the probability that the *ENPV* is negative is an acceptable risk or not.

Figure 9.13 applies a Monte Carlo simulation in order to estimate the risk of the bridge project (example 4). For sake of simplicity, the dataset still corresponds to that of Fig. 9.5. First, the program defines a loop that goes from $i=1$ to 10,000, this last number representing the number of times the *ENPV* will be randomly generated. Although we could extend the simulations to a larger number of variables, only “electric power”, “raw materials”, “Sales” and “time saving” are assigned a random number. The uniform and triangular distributions (*runif*,

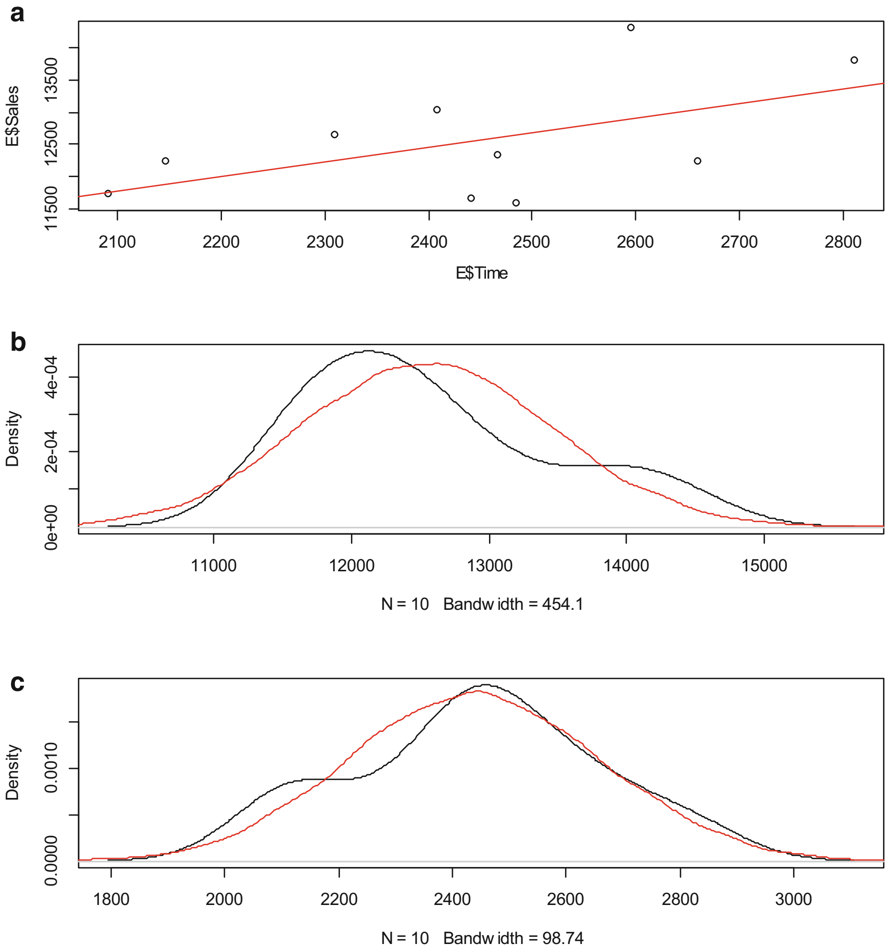


Fig. 9.12 Bivariate normal distribution: example 4. (**a**) Scatter plot. (**b**) Probability density of sales. (**c**) Probability density of time saving

rtriangl e) as well as the multivariate normal distribution (*mvtnorm*) are used in this purpose.

The cash flows for electric power, raw materials, sales and time saving are observed at year 1, 2 and 3. As such, three observations per variable must be generated. A joint distribution similar to that of Fig. 9.11 (database *E*) is used to simulate sales and time saving. To ensure that the difference in means between Table 9.6 and Fig. 9.5 are fully assessed, two additional variables are created: *Weights.Sales* and *Weights.Time*. They are used to weight the random generator (*simu*) so that the average values of sales and time saving correspond to those of Fig. 9.5. Mathematically, we rely on the fact that the standard deviation of the

```
> D=read.table("C://mydataCBA1.csv",head=FALSE,sep=";")
> D
                        V1     V2    V3    V4    V5
1                     Year      0     1     2     3
2                    Lands  -7000     0     0     0
3    Bridge infrastructure -15000     0     0     0
4                Equipment  -4000     0     0     0
5           Start-up costs  -1500     0     0     0
6             Road network  -2500     0     0     0
7            Raw materials      0 -2250 -2250 -2250
8                    Labor      0  -750  -750  -750
9           Electric power      0  -300  -300  -300
10             Maintenance      0  -450  -450  -450
11    Administrative costs      0   -80   -80   -80
12      Sales expenditures      0  -170  -170  -170
13 Negative externalities  -2500 -2000 -2000 -2000
14                   Sales      0 13600 17000 17500
15             Time saving      0  5000  6500  8500

> # Correspondence between Tables 11 and 3
> Weights.Sales=D[14,3:5]/mean(E$Sales)
> Weights.Time=D[15,3:5]/mean(E$Time)

> # Monte Carlo simulations
> library(triangle)
> library(FinCal)
> ENPV=NA
> for(i in 1:10000){
+ D[7,3:5]=runif(3, min=-2350, max=-2150)
+ D[9,3:5]=rtriangle(3,-350,-250,-300)
+ simu=rmvnorm(n=3, mean=c(mean(E$Sales),mean(E$Time)),sigma=cov(E))
+ D[14,3:5]= simu[,1]*Weights.Sales
+ D[15,3:5]= simu[,2]*Weights.Time
+ C=colSums(D[2:13,2:5])
+ B=colSums(D[14:15,2:5])
+ ENPV[i]=npv(0.04,B+C)
+ }
> plot(density(ENPV))
> quantile(ENPV,c(.025, .975))
     2.5%     97.5%
 8624.446 18455.827
> mean(ENPV)
[1] 13573.46
> var(ENPV)
[1] 6350261
```

Fig. 9.13 Monte Carlo simulation with R-CRAN: example 4

product of a constant a with a random variable X is equal to the product of the standard deviation of X with the constant: $\sqrt{\mathrm{Var}(aX)} = a\sqrt{\mathrm{Var}(X)}$. In other words, we consider the possibility that the new bridge (aX, i.e. database D) is not of the same size as the existing bridge (X, i.e. database E) and, consequently, that the standard errors are proportionally different. From the previous mathematical formula, we can see that we can indifferently include the weights (*Weights.Sales* and *Weights.Time*) in the random generator (in which case each generated value accounts for that modification) or after as it is done in Fig. 9.13.

Then the discounted cash flows are computed in a manner similar to that presented in Fig. 9.6. The analysis ends with drawing the probability density

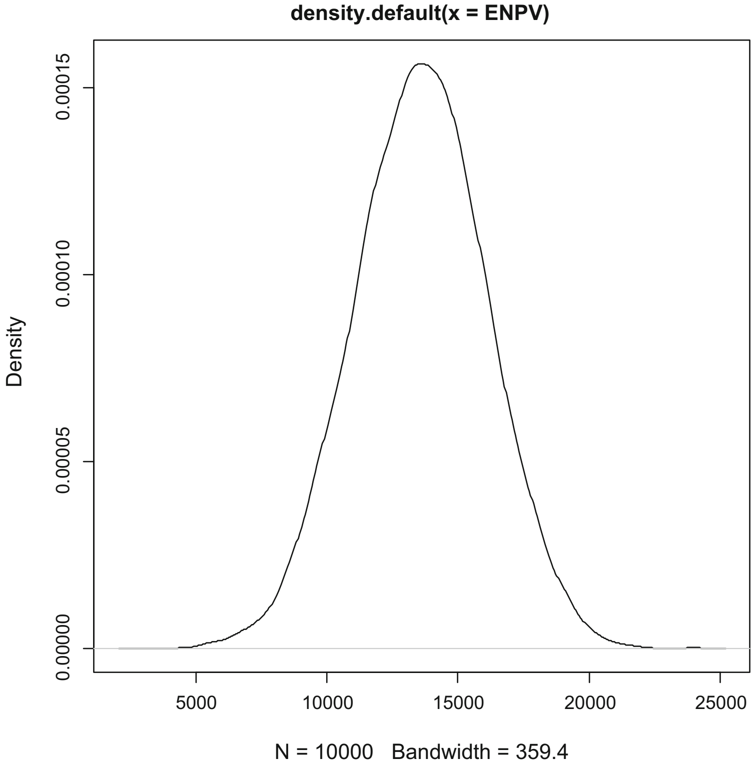


Fig. 9.14 Estimated distribution of the ENPV: example 4

function of the randomly generated *ENPV* (see Fig. 9.14). What matters here is the 95%-confidence interval obtained using the function *quantile*. The latter yields positive values, ranging from 8643 to 18,561 thousand dollars, which gives support to the bridge project. Roughly speaking, the probability that the *ENPV* falls outside this interval is lower than 5%. Notice that the mean is approximately 13,568 thousand dollars, i.e. quite similar to the *ENPV* obtained in Fig. 9.5. This result is not surprising as the mean of the distributions used for the simulations are set to the same values as those of the raw dataset (for instance –300 for electricity and –2250 for raw materials). More interesting is the variance, which indicates the risk of the project.

9.7 Mean-Variance Analysis

Once the probability distributions have been calculated for several strategies, the mean and variance of each *ENPV* distribution can be compared. The approach, known as mean-variance analysis plots the different strategies in the mean-variance plane and selects them based on their position in this plane. Under this framework, the mean of the *ENPV* represents the expected return of the project from the society point of view. The variance, on the other hand, represents how spread out the economic performance is. It measures the variability or volatility from the mean and, as such, it helps the decision-maker to evaluate the risk behind each strategy. A variance value of zero means for instance that the chances of achieving the most-likely scenario are 100%. On the contrary, a large variance implies uncertainty about the final outcome. If two strategies have the same expected *ENPV*, but one has a lower variance, the one with the lower variance is generally preferred.

Imagine that a decision-maker must choose one alternative out of four possible strategies. The plane in question is presented in Fig. 9.15. Strategy S_1 and strategy S_2 have similar variance. In other words, they have the same risk. However, the distribution of S_2 yields a higher *ENPV* on average, and is thus preferable. Consider now S_2 versus S_3. They have the same mean, but the *ENPV* of S_3 is more dispersed. Strategy S_3 is what is called a "mean preserving spread" of strategy S_2 and, as such, is riskier. A risk-averse decision-maker would typically prefer strategy S_2 since the likelihood that the *ENPV* falls below zero is lower. Comparing strategy S_2 with

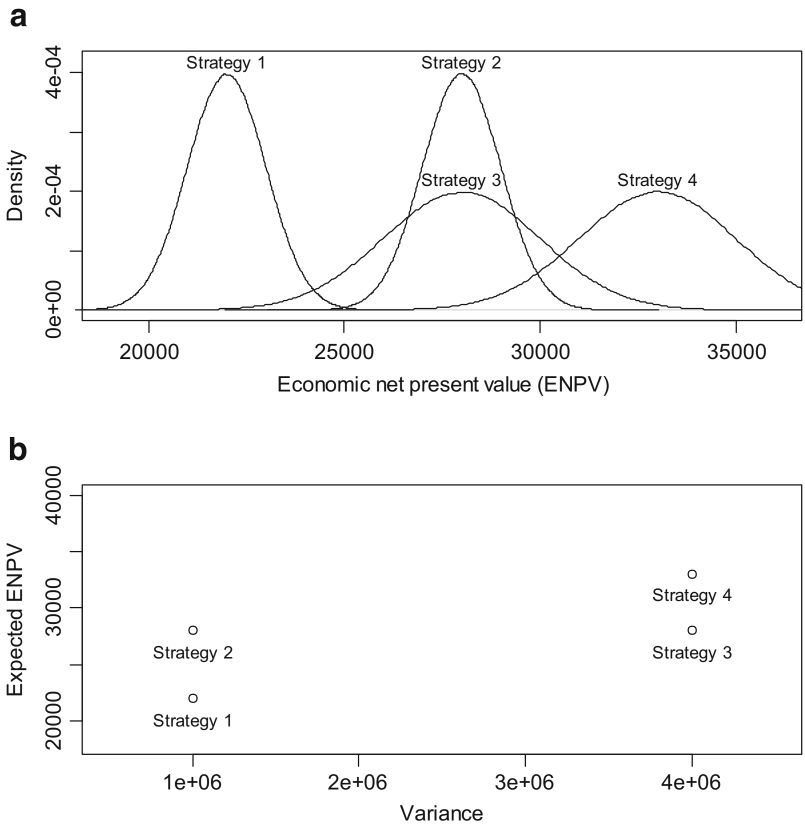


Fig. 9.15 Mean-variance analysis: example 4. (**a**) Probability distributions. (**b**) Mean-variance plane

strategy S_4 is less clear-cut. While S_4 has a much higher mean, its variance is also greater. In this situation, it is up to the decision-maker to decide what matters most, whether it is an increase in the expected *ENPV* or a decrease in the variance.

Bibliographical Guideline

The concept of consumer surplus is attributed to Dupuit (1844, 1849), an Italian-born French civil engineer who was working for the French State. His articles "On the measurement of the utility of public works" and "on tolls and transport charges" provide a discussion on the concept of marginal utility. They point out that the market price paid for consuming a good does not provide an accurate measure of the utility derived from its consumption. If one wants to construct a public infrastructure, it is instead the monetary value of the absolute utility, i.e. the willingness to pay, that matters.

The theory of externalities was initially developed by Pigou (1932) who demonstrated that, under some circumstances, the government could levy taxes on companies that pollute the environment or create economic costs.

Shadow prices have been intensively used after the proposal of Little and Mirrlees (1968, 1974) to use world market prices (and standard conversion factors) in project evaluation. Their approach was subsequently promoted by Squire and van der Tak (1975) in a book commissioned by the World Bank.

The modern form of Monte Carlo simulation was developed by von Neumann and Ulam for the Manhattan Project, in order to develop nuclear weapons. The method was named after the Monte Carlo Casino in Monaco.

The mean-variance portfolio theory is attributed to Markowitz (1952, 1959), which assumes that investors are rational individuals and that for an increased risk they expect a higher return.

Many guides provided by government agencies are available online. We may cite in particular the "Guide to Cost-Benefit Analysis of Investment Projects" of the European Commission which describes project appraisal in the framework of the 2014–2020 EU funds, as well as the agenda, the methods and several case studies. The European Investment Bank (EIB) proposes as a complement a document that presents the economic appraisal methods that the EIB advocates to assess the economic viability of projects. Additional guides are available, such as the "Canadian cost benefit analysis guide" provided by the Treasury Board of Canada Secretariat, the "Cost benefit analysis guide" prepared by the US Army, the guide to "Cost-benefit analysis for development" proposed by the Asian Development Bank, the "Cost benefit analysis methodology procedures manual" by the Australian Office of Airspace Regulation. Those documents review recent developments in the field and provide several examples of application with the purpose to make CBA as clear and as user-friendly as possible.

Several textbooks can also be of interest for the reader. We may in particular cite Campbell and Brown (2003). Their book illustrates the practice of cost benefit analysis using a spreadsheet framework, including case studies, risk and alternative scenario assessment. We may also cite Garrod and Willis (1999) and Bateman et al.

(2002) who provide an overview of the theory as well as the different methods to estimate welfare changes.

Bibliography

Australian Office of Airspace Regulation. (2007). *Cost benefit analysis methodology procedures manual*. Civil Aviation Safety Authority.

Asian Development Bank. (2013). *Cost-benefit analysis for development: A practical guide*. Mandaluyong City, Philippines: Asian Development Bank.

Bateman, I., Carson, R., Day, B., Haneman, M., Hanley, N., Hett, T., et al. (2002). *Economic valuation with stated preference techniques: A manual*. Cheltenham: Edward Elgar.

Campbell, H., & Brown, R. (2003). *Benefit-cost analysis: Financial and economic appraisal using spreadsheets*. Cambridge University Press: Cambridge Books.

Dupuit, J. (1844). De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées, 2*, 332–375 [English translation: Barback, R. H. (1952). On the measurement of the utility of public works. International Economic Papers, 2, 83–110].

Dupuit, J. (1849). De l'influence des péages sur l'utilité des voies de communication. *Annales des Ponts et Chaussées, 2*, 170–248 [English translation of the last section: Henderson, E. (1962). On tolls and transport charges. International Economic Papers 11: 7–31].

European Commission. (2014). *Guide to cost-benefit analysis of investment projects. Economic appraisal tool for cohesion policy 2014-2020*.

European Investment Bank. (2013). *The economic appraisal of investment projects at the EIB*.

Garrod, G., & Willis, K. G. (1999). *Economic valuation of the environment: Methods and case studies*. Cheltenham: Edward Elgar.

Little, I. M. D., & Mirrlees, J. A. (1968). *Manual of industrial project analysis in developing countries, 2: Social cost–benefit analysis*. Paris: OECD.

Little, I. M. D., & Mirrlees, J. A. (1974). *Project appraisal and planning for developing countries*. New York: Basic Books.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance, 7*, 77–91.

Markowitz, H.M. (1959). Portfolio selection: Efficient diversification of investments. New York: Wiley (Reprinted by Yale University Press, 2nd ed. Basil Blackwell, 1991).

Pigou, A. C. (1932). *The economics of welfare* (4th ed.). London: Mac Millan.

Squire, L., & van der Tak, H. G. (1975). *Economic analysis of projects*. Baltimore: Johns Hopkins, University Press for the World Bank.

Treasury Board of Canada Secretariat. (2007). *Canadian cost-benefit analysis guide: Regulatory proposals*.

US Army. (2013). *Cost benefit analysis guide*. Prepared by the Office of the Deputy Assistant Secretary of the Army.

Cost Effectiveness Analysis 10

10.1 Appraisal of Projects with Non-monetary Outcomes

Cost effectiveness analysis is a decision-making tool that compares the relative costs and outcomes of two or more strategies competing for the implementation of a public program. The method has initially been developed in the field of public health and, since then, is more and more used in other fields of public action. The keystone of the approach is that it does not require estimating the equivalent money value of the outcomes. It rather uses physical or arithmetical units for their measurement. Examples of applications are education, public health programs or policies for a safer environment, wherever the decision-maker would feel it inappropriate or unnecessary to monetize outcomes. Measures of effectiveness are for instance the rate of school completion, the number of premature births averted, composite indicators of safe environment, etc.

In its simpler form, cost effectiveness analysis consists in comparing incremental cost effectiveness ratios (*ICER*) where the numerator is the cost difference when shifting from one strategy to another, and the denominator is the effectiveness difference. The principle of confronting competing and mutually exclusive strategies through differential costs and differential effectiveness builds on the concept of opportunity cost. By selecting a particular strategy, you give up the net advantages of the waived strategy. The perspective is thus comparative (the goal is to find out the differences among strategies) and consequentialist (the maximization of effectiveness from available resources is the only relevant consideration). We have:

$$ICER = \frac{\text{Difference in cost}}{\text{Difference in effectiveness}}$$

The ratio represents the incremental cost associated with one additional unit of effectiveness. For instance, if a new strategy costs \$20 extra dollars and brings in 10 units of effectiveness, then the *ICER* amounts to \$20/10=\$2. Each additional

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_10

unit of effectiveness thus costs \$2. It is up to the decision-maker to decide whether this extra cost is beneficial to society or not.

Once data on cost and effectiveness have been gathered, the *ICER* is compared to the marginal willingness to pay of the decision-maker for an additional unit of effectiveness. The decision-maker represents the community for which the project will be implemented and marginal willingness to pay is thus collective. The choice rule is that a competing strategy is preferred over a reference strategy if its incremental cost effectiveness ratio is lower than a given level of collective willingness to pay:

$$\text{Competing strategy} \succ \text{Reference strategy} \Leftrightarrow ICER < WTP$$

Coming back to our example, if the *WTP* is \$3, then the new strategy is considered as more cost effective and thereby selected. On the other hand, if the willingness to pay is lower than \$2, then implementing the new strategy is considered detrimental to society.

The previous choice rule can be generalized to any level of willingness to pay when one moves to the incremental net benefit (*INB*) approach, which is a linear rearrangement of the incremental cost effectiveness ratio:

$$INB = WTP \times (\text{Difference in effectiveness}) - (\text{Difference in cost})$$

The decision rule is such that the switch to a new strategy is accepted if $INB > 0$. In our previous example, we have $INB = 10 \times WTP - \$20$. The *INB* thus amounts to –\$10 if the willingness to pay is \$1, \$0 if the willingness to pay is \$2 and \$10 if the willingness to pay is 3\$. In other words, the approach offers a way to measure welfare changes expressed in dollar values without presupposing any measure of willingness to pay. If more than two strategies are in competition, a simple comparison of the *INB* yields the most efficient policies.

The *ICER* and *INB* indicators are based on pairwise comparisons which can be a shortcoming when the evaluator is faced with more than two strategies. The construction of an efficiency frontier overcomes this problem. It is a graphical tool that divides strategies into two categories: the first comprehends the most efficient strategies (the frontier) and the second displays strategies (out of the frontier) that are not cost-effective relative to the strategies of the first category. The method relies on two concept of dominance, simple dominance and extended dominance. A strategy is subject to simple dominance if it yields higher cost and lower effectiveness than another strategy. A strategy is subject to extended dominance if its *ICER*, i.e. its incremental cost, is higher than that of the next more effective strategy. All the strategies that are subject to dominance, be it simple or extended, are excluded from the frontier. The approach thus provides a sorting of strategies, for any value of collective marginal willingness to pay.

A strategy may not only affect outcomes in the next period, but may also induce multiple consequences in the subsequent periods. Therefore, when selecting among strategies, it is necessary to consider a multiple-state setting (situations that individuals can be in), transitions between states (when individuals move from one state to another), transition probabilities (how likely such moves are), and a relevant time horizon. The approach, also known as decision analytic modeling, generally relies on Markov chains to define this setting. A Markov model considers a finite number of states representing the different situations induced by the strategies (e.g., different stages of a disease, death or recovery). They are assumed to be mutually exclusive since a subject cannot be in more than one state at the same time. Each time period or Markov cycle is associated with cost and effectiveness measures that depend on the allocation of subjects among the Markov states. Decision analytical modeling then computes the total cost and total effectiveness of the competing options with the aim to provide decision-makers with relevant information about the resources required to operate the strategies as well as their consequences.

Cost effectiveness analysis is based on a set of parameters that can strongly affect the accuracy of the conclusions if their value is not carefully established. It is the task of the evaluator to demonstrate to what extent those conclusions are robust to the assumptions made. Several methods exist in this respect. Among the most sophisticated ones, one may name Monte Carlo simulations (aka probabilistic sensitivity analysis) which offer a way to measure and evaluate the uncertainty inherent to cost and effectiveness measurements. Each parameter of the model (for example, the probability of an intervention being successful or the cost associated with that intervention) is assigned a random generator that is based on a pre-defined probability distribution. The way data has been collected (sample data or non-sampled secondary data source) generally determines the shape of the random generators. The model is then iterated a high number of times until a sufficient number of simulations is obtained. Those simulated data are finally examined in order to compute a 95% confidence interval on the cost and effectiveness measurements. Furthermore, each iteration can be used to compute cost effectiveness indicators such as the *ICER* or the *INB*. It is also possible to plot the simulated data in the differential cost and effectiveness plane to assess, for instance, whether a strategy yields on average a decrease in cost and/or an increase in effectiveness. Last, one may rely on cost-effectiveness acceptability curves which indicate the number of times each strategy is optimal in those simulations.

To sum up, cost effectiveness analysis data are obtained in three related sequences. The first one involves primary or initial data on costs and effectiveness (for instance the direct cost of cleansing the soil from a pollutant and the prevalence of infectious diseases related to that pollutant), usually gathered from experiments or previous case studies. The second step consists in simulations, based on those primary data, using decision-analytic modeling (Markov chains). The latter calculates period after period vectors of projected cost and effectiveness and cumulates them over a chosen time horizon. Those vectors, once aggregated, form the cost and effectiveness measurements. In the above example, the effectiveness measure can be the decrease in the number of individuals affected by diseases or in the number of deaths due to pollution; costs can be cleansing expenses as well as expenditures associated with taking care of diseased people. Third, Monte Carlo simulations can be implemented to assess the robustness of the results. This step is essential as the uncertainty surrounding decision analytic models makes deterministic analyses rather fragile tools when they are not accompanied by a full exploration of methodological, parameter and structural uncertainty. If for instance the reference strategy is "doing nothing", one can compare its net merit with those of competing strategies, say, "proceed to an immediate and massive cleansing" or "proceed to sequential and progressive cleansing", over a large set of simulated data.

The chapter is organized as follows. Section 10.2 describes the usual cost effectiveness indicators, namely the incremental cost effectiveness ratio and the incremental net benefit, which can be used for pairwise comparisons of mutually exclusive strategies. Section 10.3 moves on to defining cost-effective strategies through the construction of the efficiency frontier in presence of several strategies. Section 10.4 introduces decision analytic modeling and shows how to obtain cost and effectiveness data from a Markov model. Section 10.5 implements those calculations in R-CRAN and Sect. 10.6 extends the approach to QALYs (Quality Adjusted Life Years). Section 10.7 discusses the various forms of uncertainty in decision analytic models and investigates parameter uncertainty through Monte Carlo simulations. Last, Sect. 10.8 explains how to analyze and interpret those simulation outputs.

10.2 Cost Effectiveness Indicators

The aim of this section is to assess a public project through pairwise comparisons of alternative strategies of implementation. Let us consider for instance dropping out of school as a policy problem. Alternative educational strategies intended to reduce dropouts address the problem of early school leaving in different ways, generating different costs and effectiveness levels. It is thus necessary to reason in terms of cost effectiveness differentials. In what follows, let us assume that the evaluator faces a set of strategies that provide support to schoolchildren at risk of dropping out, e.g., mandatory attendance, compulsory add-on programs that complement existing schooling, limited-period training programs, social assistance, relocation to another school, etc. The situation of reference is the currently implemented strategy S_0, or status quo. Alternatives to S_0 are the competing strategies S_1, $S_2, \cdots, S_K$, which are assumed to be mutually exclusive.

Strategies are associated with production functions involving capital inputs (technical, physical, and human), labor (teachers, assistants, social workers), and space. Costs may include those for the public sector or the agency providing the program, for the individuals to whom the program is dedicated, for the other parties involved (mutual funds, donors, other agencies). The cost perimeter is defined by the decision-maker. The unit of measurement for costs is monetary, with total cost for each strategy denoted $C_0, C_1, C_2, \ldots, C_K$. Moving from the reference strategy to a competing strategy k generates differential cost $\Delta C_{0 \rightarrow k} = C_k - C_0$.

What are the outputs of those production functions? What would their unit of account be? Cost effectiveness analysis takes the stance that the arithmetic or physical consequences of the strategy account for its outcomes. In our example, the number of children completing high school would be a measure of the effectiveness of the implemented program, but one could argue that the number of children completing high school with marks above a specified level would be a better measure. Again, the task of defining those measures goes to the decision-maker. The key point here is that the unit of measurement of effectiveness is non-monetary. If we denote $E_0, E_1, E_2, \ldots, E_K$ the respective effectiveness of the competing strategies, then differential effectiveness is given by $\Delta E_{0\rightarrow k} = E_k - E_0$.

Consider the various possible locations of strategies on the $[\Delta E, \Delta C]$ mapping with reference strategy S_0 as the origin. Figure 10.1 displays the four cases that may arise. Since we have just plotted the differential effectiveness and cost coordinates, the cost effectiveness mapping is at this stage simply descriptive. The situations in the South East and North West quadrants are obvious. North West strategies S_{NW} are subject to simple dominance with respect to S_0. South East strategies S_{SE} are simply dominant. In contrast, for the North East and South West quadrants, cost effectiveness indicators as well as information on collective marginal willingness to pay (*WTP*) are required in order to make a decision to switch or not from the reference to one of its alternatives.

To illustrate the interest of mapping strategies on the $[\Delta E, \Delta C]$ plan, consider Table 10.1 where cost and effectiveness values are provided for six competing strategies. We distinguish the denomination of a strategy $(a, b, c, \ldots)$ from its notation $(S_1, S_2, S_3, \ldots)$ since while the former always remains (e.g., *a* corresponds to "mandatory attendance"), its notation is likely to vary during the implementation of the efficiency analysis, as we shall see later. The plotting of cost and effectiveness data on the $[E, C]$ mapping is rather uninformative, as depicted in Fig. 10.2. However, should for instance strategy *c* be the status quo (S_0), then a

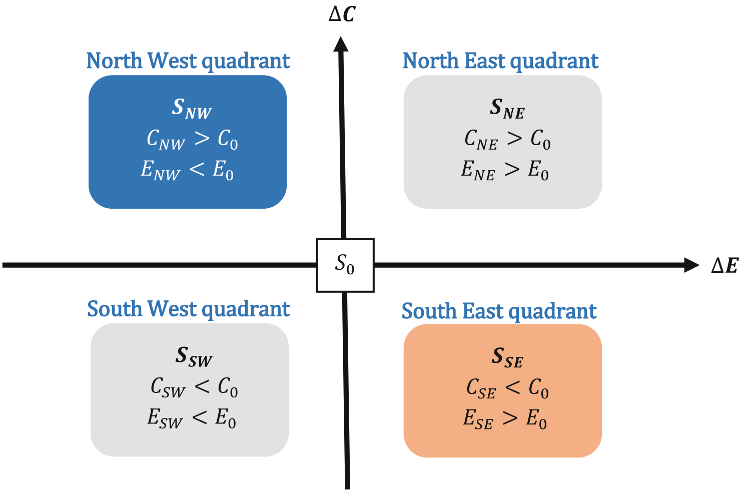


Fig. 10.1 Differential cost effectiveness mapping

Table 10.1 Cost and effectiveness of competing strategies: example 1

Denomination of strategy	Notation of strategy	Effectiveness E_k	Cost ($) C_k	$ICER\ (S_3 \rightarrow S_k)$
a	S_1	37	1500	20
b	S_2	39	1650	−16.7
c	S_3	42	1600	NA
d	S_4	46	1700	25
e	S_5	48	1900	50
f	S_6	50	2000	50

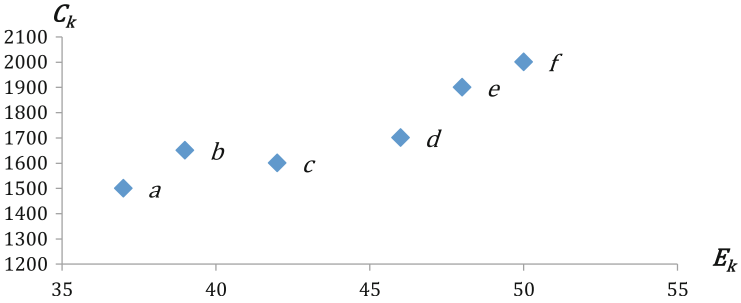


Fig. 10.2 Cost effectiveness mapping: example 1

pairwise comparison would actually lead to a scatter plot with strategy *c* as the origin, as shown in Fig. 10.3. The differential cost-effectiveness mapping thus illuminates the analysis. Indeed, from Fig. 10.3, it is easy to see that *b* is subject to simple dominance with respect to *c*, i.e. strategy *b* is more costly and less effective. However, from Fig. 10.3, it can be seen that the concept of simple dominance is not sufficient to select one single strategy. Strategies *a*, *d*, *e* and *f* are for instance not dominated by *c* and do not dominate *c* either. One has then to define indicators that sort out the relative cost effectiveness of the competing strategies.

The first of cost effectiveness indicators is the incremental cost effectiveness ratio (*ICER*). It describes the additional investment of resources required for each additional unit of outcome improvement expected to result from investing in S_k rather than S_0. The corresponding formula is:

$$ICER(S_0 \rightarrow S_k) = \frac{\Delta C_{0 \rightarrow k}}{\Delta E_{0 \rightarrow k}} = \frac{C_k - C_0}{E_k - E_0}$$

The *ICER* is a slope as depicted in Figs. 10.4 and 10.5, respectively in the North East (strategy S_{NE}) and South West (strategy S_{SW}) quadrants. The greater is the

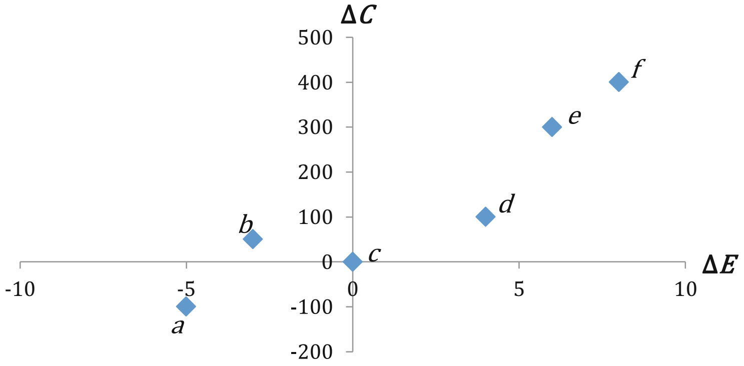


Fig. 10.3 Differential cost and effectiveness: example 1

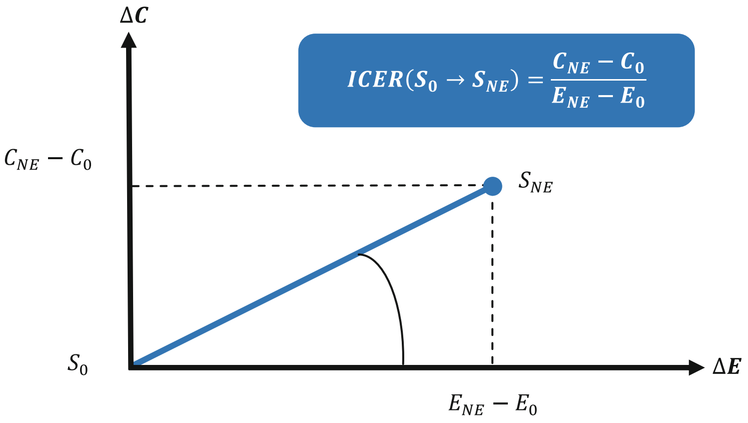

ΔC
$ICER(S_0 \rightarrow S_{NE}) = \frac{C_{NE} - C_0}{E_{NE} - E_0}$
$C_{NE} - C_0$
S_{NE}
S_0
ΔE
$E_{NE} - E_0$

Fig. 10.4 *ICER* in the North East quadrant

slope, the higher is the incremental cost effectiveness ratio compared to the status quo.

In Table 10.1, when the reference strategy is *c*, the *ICER* ranges from −16.7 to 50 (last column of Table 10.1). The negative sign on −16.7 is related to the concept of simple dominance (strategy *b* is simply dominated). If the focus were merely on cost, strategy *a* would then appear as the most efficient strategy as it minimizes the *ICER*. However, strategy *a* is also the strategy that reaches the lowest level of effectiveness (only 1500 while *e* and *f* reach 1900 and 2000 respectively) and deteriorates the situation with regard to the existing status quo. As for the value of 25 (corresponding to the move from strategy *c* to strategy *d*), it means that starting from the status quo, an increase of one unit in the effectiveness measure will cost $25 if that effectiveness improvement is implemented through strategy *d*. Strategies *e* and *f* generate the same *ICER* of 50 so that this cost effectiveness

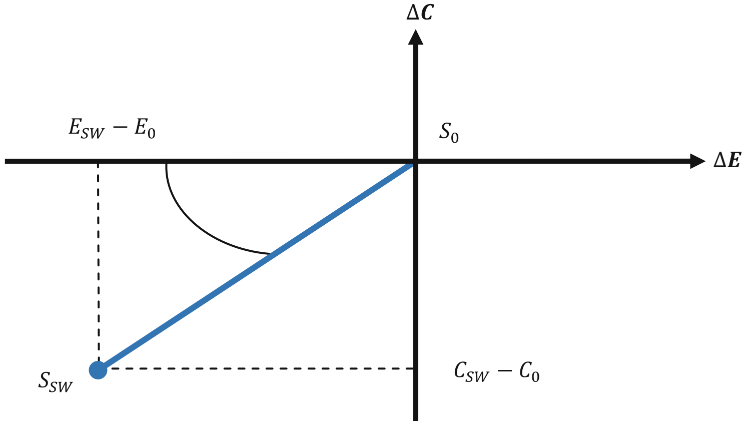


Fig. 10.5 *ICER* in the South West quadrant

indicator cannot sort them out. We will come back to that point later but one can still observe that the monetary effort asked from the decision-maker is twice as much if strategy *e* or *f* are chosen than if strategy *d* is selected. If solving the dropout problem is a priority in education reform, then the decision-maker may decide to dedicate extra money to reach a higher level of effectiveness than what is currently obtained through the implementation of strategy *c*. A decision based on (positive) *ICER* only is therefore not sufficient to assess the cost effectiveness of the strategies. Positive *ICER*s have to be assessed against the willingness to allocate extra resources in order to move from status quo to the chosen strategy.

Cost effectiveness analysis thus requires information on the collective marginal willingness to pay (*WTP*) for extra outcome or willingness to accept (*WTA*) outcome loss. We assume afterwards that the *WTP*/*WTA* line has no kink at the origin, meaning that the cost-saving required to accept the loss of one unit of effectiveness is no greater than the extra cost consented to increase effectiveness by one unit. The decision rule to accept or reject the switch from S_0 to S_{NE} or S_{SW} is then very simply described in Figs. 10.6 (North East quadrant) and 10.7 (South West quadrant). To sum up, in the North-East quadrant:

$$ICER(S_0 \to S_{NE}) > 0 : C_{NE} > C_0, E_{NE} > E_0,$$
$$\text{switch accepted if } ICER < WTP.$$

In the North-west quadrant:

$$ICER(S_0 \to S_{NW}) < 0 : C_{NW} > C_0, E_{NW} < E_0,$$
$$\text{switch rejected (simply dominated).}$$

In the South-East quadrant:

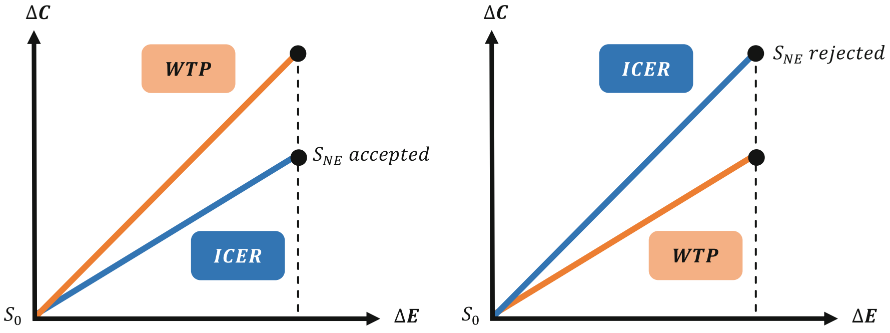


Fig. 10.6 Decision rule with *ICER* in the North East quadrant

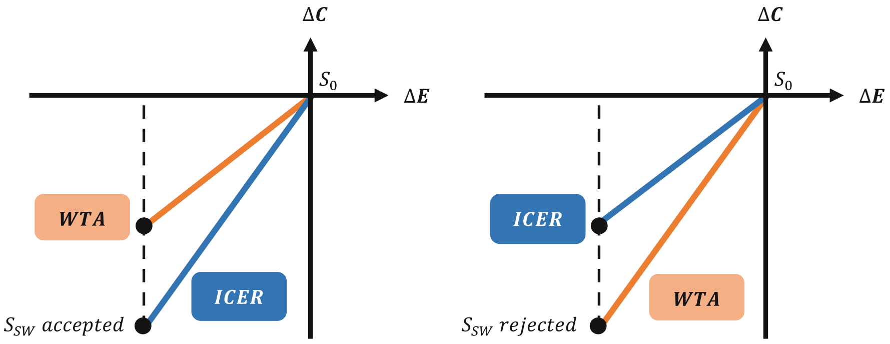


Fig. 10.7 Decision rule with *ICER* in the South West quadrant

$$ICER(S_0 \rightarrow S_{SE}) < 0 : C_{SE} < C_0, E_{SE} > E_0,$$
$$\text{switch accepted (simply dominant)}.$$

In the South-West quadrant:

$$ICER(S_0 \rightarrow S_{SW}) > 0 : C_{SW} < C_0, E_{SW} < E_0,$$
$$\text{switch accepted if } ICER < WTA.$$

Figure 10.8 synthesizes the various situations.

Several questions have been raised about the use of *ICER* for decision-making. First, being a slope brings about problems such as the one described in Table 10.2. *ICER*s can be similar (Case 1 considers two North East strategies for which the *ICER* is 4; case 2 has one North East strategy and the other is South West for the same *ICER* of 5). From Table 10.1, we can see that example 1 also experienced that with strategies *e* and *f*. The *ICER* may thus hide differences in strategies though admittedly a closer look at data should overcome that problem. The ratio nature of the *ICER* is also of concern when differences in effectiveness are small enough to generate extremely high values that are awkward to interpret. Finally, the *ICER* is

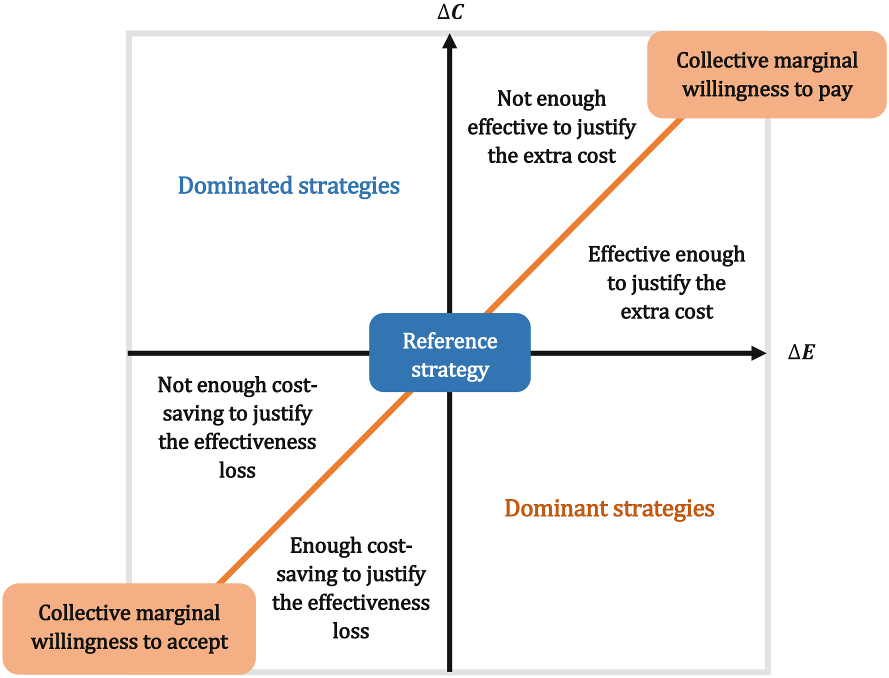

ΔC
ΔE
Collective marginal willingness to pay
Not enough effective to justify the extra cost
Dominated strategies
Effective enough to justify the extra cost
Reference strategy
Not enough cost-saving to justify the effectiveness loss
Dominant strategies
Enough cost-saving to justify the effectiveness loss
Collective marginal willingness to accept

Fig. 10.8 Decision in the differential cost effectiveness mapping

Table 10.2 Cases of indetermination with *ICER*s: example 2

Case 1	S_0	S_1	S_2
Cost	$C_0 = 100$	$C_1 = 120$	$C_2 = 200$
Effectiveness	$E_0 = 0$	$E_1 = 5$	$E_2 = 25$
Case 2	S_0	S_3	S_4
Cost	$C_0 = 100$	$C_3 = 125$	$C_4 = 90$
Effectiveness	$E_0 = 0$	$E_3 = 5$	$E_4 = -2$

by construction a parameter usually obtained from running a decision analytic model and this parameter is to be confronted with a single value of *WTP*. That value may not be readily available or may not be defined at all. Fortunately, the second cost effectiveness indicator, the incremental net benefit (*INB*), is able to overcome those problems.

The *INB* is defined as the difference between the variation in effectiveness associated with the shift from strategy S_0 to strategy S_k, valorized by what the community is ready to pay (allocate) for each extra unit of effectiveness, i.e. the collective marginal willingness to pay, and the variation in cost induced by the change of strategy:

$$INB_{0\to k}(WTP) = WTP \times \Delta E_{0\to k} - \Delta C_{0\to k}$$

The decision rule is such that the switch is accepted if $INB_{0 \to k}(WTP) > 0$, which is equivalent to $WTP > ICER(S_0 \to S_k)$. Consider for instance strategy S_2 from Table 10.2. The unit cost of an increase in effectiveness is:

$$ICER(S_0 \to S_2) = \frac{\$200 - \$100}{25 - 0} = \$4$$

Should marginal willingness to pay be lower than \$4, the decision-maker would actually reject strategy S_2. Using the *INB* criterion yields more general results. For *WTP* equal to 1, we obtain $INB_{0 \to 2}(1) = 1 \times (25 - 0) - (200 - 100) = -\$75$, i.e. a move from strategy S_0 to strategy S_2 would yield an incremental net loss of \$75. For *WTP* equal to \$5, moving from S_0 to strategy S_2 would yield instead an incremental net benefit of $INB_{0 \to 2}(5) = \$25$. By construction, the *INB* equals 0 when $WTP = \$4 = ICER$. The final decision thus depends on the marginal willingness to pay of the decision-maker. The higher the willingness to pay, the more likely a strategy in the North East quadrant is to be selected.

Figure 10.9 shows how to represent the *INB* of an accepted strategy in the North East quadrant for a given value of *WTP*. Since the *INB* does not require specifying a value of *WTP*, it is also possible to represent it as a function in the mapping [*WTP*, *INB*]. For instance, Fig. 10.10 assumes that alternative strategy S_k belongs to the North-East quadrant. For values of *WTP* greater than the *ICER*, the strategy is preferred to the status quo. Thus, with the *INB*, the decision-maker is provided with a tool that allows discussion and debate about the extent to which the community can or wishes to allocate resources to a more costly yet more effective strategy.

Cost effectiveness indicators provide pairwise comparisons of the strategies competing for the implementation of a public project. The *ICER* is built as a ratio of differential costs and outcomes from a reference strategy to a comparator. It

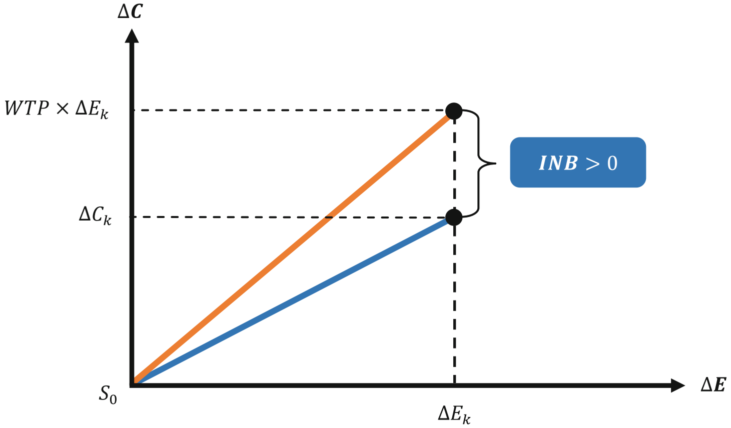


Fig. 10.9 *INB* of an accepted strategy in the North East quadrant

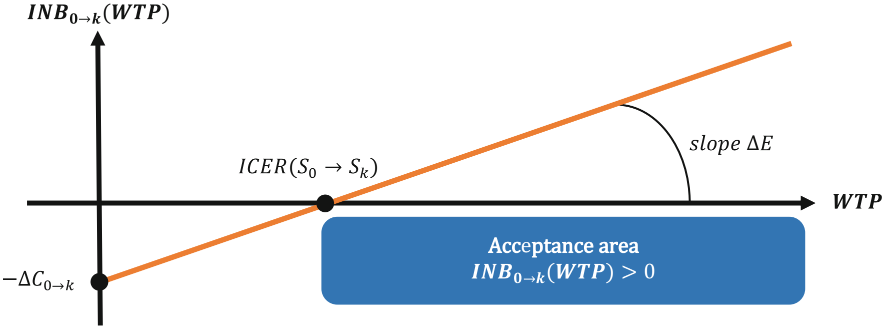


Fig. 10.10 Decision rule with the *INB* as a function of *WTP*

provides a single value of extra cost per extra unit of effectiveness gained when moving away from the reference. This single value is to be compared to the marginal willingness to pay of the community for that extra unit of effectiveness. *WTP* then serves as a threshold below which strategies are acceptable. If the community cannot or does not wish to provide that threshold value, then the *INB* allows to consider the whole range of potential values of *WTP*. The *ICER* is a choice parameter to be confronted with *WTP*, the *INB* is a function of *WTP*. However, comparing strategies in pairs does not provide a full picture of the decision problem, particularly when several strategies are in competition. This is why the next step of the analysis moves on to the construction of the efficiency frontier.

10.3 The Efficiency Frontier Approach

When several mutually exclusive strategies compete for the implementation of a public project, building an efficiency frontier allows non dominated and collectively acceptable strategies to be identified. The approach distinguishes two concept of dominance: simple versus extended dominance. In the case of simple dominance, a strategy is considered inefficient if there is at least one strategy that has higher effectiveness and lower cost. Extended dominance indicates that there exists a combination of strategies that yields higher effectiveness and lower cost. The efficiency frontier links the strategies that are not subject to simple or extended dominance.

Formally, when K strategies are in competition, the algorithm for finding the frontier is as follows. The first step is the ranking and indexing of strategies by increasing effectiveness:

$$\text{For all } k \text{ in } \{2, \ldots, K-1\} \text{ we must have } E_{k-1} < E_k < E_{k+1}$$

The second step excludes strategies subject to simple dominance (SSD):

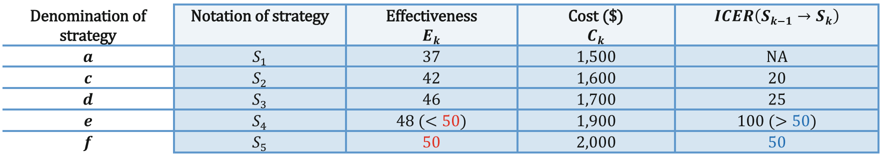
$$S_k \text{ is SSD if } C_k > C_{k+1}$$

One must replicate the exclusion criterion until no more strategies are SSD. The third step rests on the calculation of $ICER(S_{k-1} \rightarrow S_k)$ which allows the exclusion of strategies subject to extended dominance (SED):

$$S_k \text{ is SED if } ICER(S_{k-1} \rightarrow S_k) > ICER(S_k \rightarrow S_{k+1})$$

Again, calculations should be replicated until all SED strategies have been ruled out. The fourth step consists in (1) drawing the (convex) cost-effectiveness frontier on the $[\Delta E, \Delta C]$ mapping, by linking non-dominated strategies and (2) selecting interventions meeting *WTP* requirements. That can be done by calculating the *ICER* from strategy 0 as the reference to the efficient alternative strategies, and by checking it graphically.

Now, let us reconsider the example of school dropout (Example 1, Table 10.1) where six strategies are in competition. Strategy *a* yields the lowest effectiveness and the next strategies are already ordered by increasing effectiveness. The second step excludes SSD strategies: intervention *k* is simply dominated by intervention $k + 1$ if intervention $k + 1$ is more effective and less expensive. As already shown, strategy *b* is SSD and thus eliminated. One checks that no other strategy is now SSD, which is verified from Table 10.1. The third step calculates *ICER*s of non-excluded strategies, relative to the preceding intervention, in order to identify SED strategies (if any). Figure 10.11 shows the result and allows detecting that strategy *e* induces a greater cost per extra unit of effectiveness: its *ICER* is greater than the *ICER* of the next effective strategy *f*, while it is less effective. Here, strategy *e* is SED and thus eliminated. Then, replication of the third step does not identify anymore SED strategies as shown in Table 10.3 which delineates strategies belonging to the efficiency frontier (Fig. 10.12). The graph confirms that strategy



Denomination of strategy	Notation of strategy	Effectiveness E_k	Cost ($) C_k	$ICER(S_{k-1} \rightarrow S_k)$
a	S_1	37	1,500	NA
c	S_2	42	1,600	20
d	S_3	46	1,700	25
e	S_4	48 (< 50)	1,900	100 (> 50)
f	S_5	50	2,000	50

Fig. 10.11 Elimination of SED strategies: example 1

Table 10.3 Strategies on the efficiency frontier: example 1

Denomination of strategy	Notation of strategy	Effectiveness E_k	Cost ($) C_k	*ICER* $(S_{k-1} \rightarrow S_k)$
a	S_1	37	1500	NA
c	S_2	42	1600	20
d	S_3	46	1700	25
f	S_4	50	2000	75

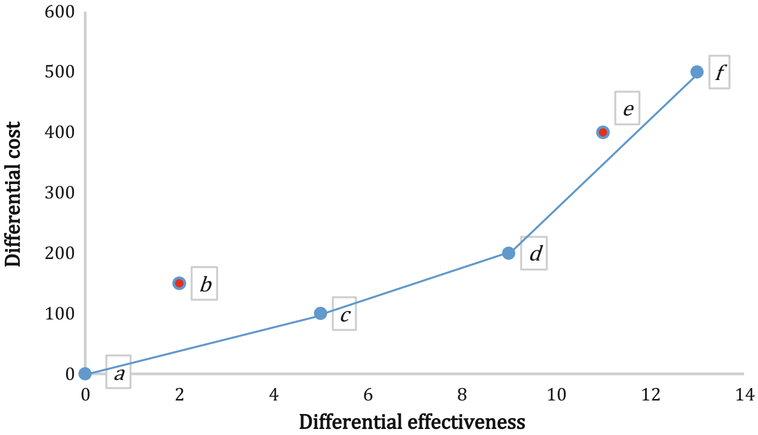


Fig. 10.12 Efficiency frontier: example 1

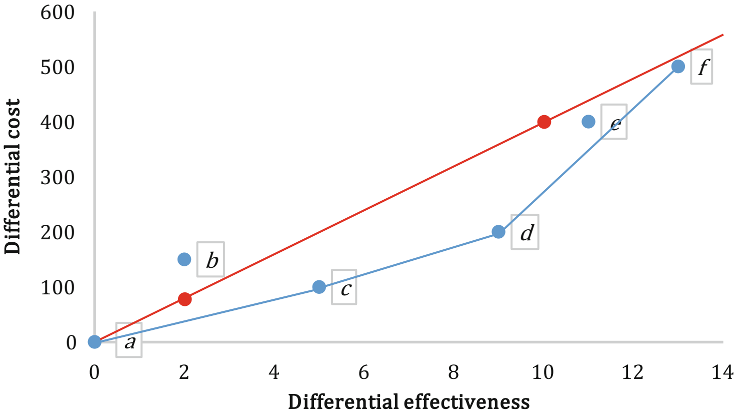


Fig. 10.13 Efficiency frontier and acceptable strategies: example 1

b is indeed more costly and less effective than strategy *c* while strategy *e* is dominated by all the linear combinations of strategies *d* and *f*. The frontier is by construction convex.

In order to complete the fourth step, one calculates the *ICER* from strategy *a* as the reference to the efficient alternative strategies *c*, *d* and *f* which gives respectively $20, $22.2 and $38.5 per unit of effectiveness. If *WTP* is for instance $40, then all three strategies are acceptable and one needs further criteria to draw out a preferred option. If *WTP* is for example $30, then strategy *f* is ruled out. Figure 10.13 illustrates that with a willingness to pay of $40 per unit of effectiveness.

Table 10.4 Evolution of *INB*s as a function of *WTP*: example 1

WTP	*INB* (*WTP*) of strategies				
	b	*c*	*d*	*e*	*f*
0	−150	−100	−200	−400	−500
10	−130	−50	−110	−290	−370
20	−110	0	−20	−180	−240
30	−90	50	70	−70	−110
40	−70	100	160	40	20
50	−50	150	250	150	150
60	−30	200	340	260	280
70	−10	250	430	370	410
80	10	300	520	480	540
90	30	350	610	590	670

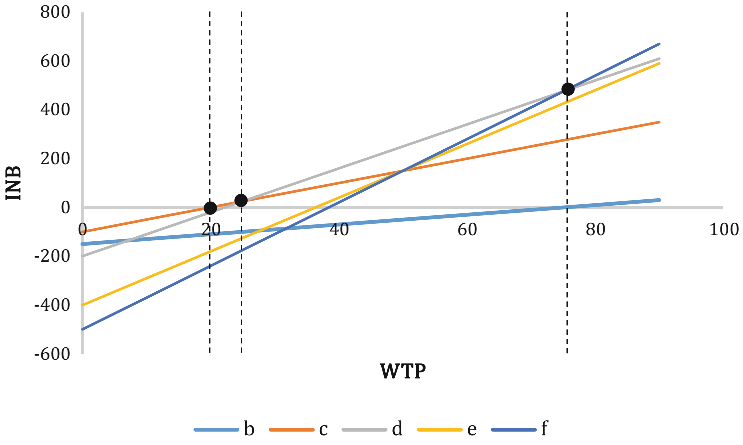


Fig. 10.14 Efficiency in the *INB* plane: example 1

Should a strategy yield 2 units of effectiveness, the decision-maker would be willing to pay \$80 at maximum. Should it be 10 units, the decision-maker would be willing to pay \$400 instead, and so on. This allows us to draw a continuous line below which all the represented strategies are potentially acceptable.

Another way of expressing the choice of strategies with respect to *WTP* builds on Fig. 10.10. Table 10.4 exemplifies the evolution of the *INB*s of the competing strategies for discrete values of *WTP*. One can draw the *INB* lines for strategy *b* to *f* with strategy *a* as the reference. Figure 10.14 shows the resulting changes in preferred option as *WTP* increases.

The analysis of efficiency in the *INB* plane confirms that strategy *b* is never the most preferred, neither is strategy *e* since their *INB* line is never the uppermost. The three vertical dotted lines on Figure 10.14 depict the choice facing the decision-maker when he or she is to select a project amongst the six proposed alternatives. In coherence with Table 10.3, below a *WTP* of $20 per unit of effectiveness gained, then the reference strategy *a* is preferred. For a *WTP* between $20 and $25, the choice should move to strategy *c*. Above $25 and until $75 per unit of effectiveness gained, strategy *d* should prevail. Strategy *f* is most effective but also very costly relatively to the others so that it is selected for high levels of consent only.

There is an alternative but equivalent way of building the efficiency frontier, with algorithms of elimination of SSD and SED strategies particularly suited for situations where a significant number of alternatives are competing. As already stressed, the efficiency frontier approach consists in comparing graphically the net benefit of available strategies in terms of effectiveness plotted on the *x*-axis, with the net costs of these strategies plotted on the *y*-axis. If a strategy is to the north west of the frontier, it is not cost-effective. For instance, in Fig. 10.15, six strategies *a*, *b*, ..., *f* are evaluated. Strategy *a* is used as reference policy. The orange line represents the efficiency frontier. Its construction is based on the evolution of the slope between the strategies, i.e. on *ICER* calculations. For instance, a move from strategy *a* to *b* implies a positive *ICER*, i.e. an additional investment of resources for each additional unit of effectiveness. A move from *b* to *c* generates an increase in the slope. In contrast, the *ICER* decreases when we move from *c* to *d* but it remains positive. This means that it is actually better to use *b* and *d*, as there would exist a combination of these strategies, denoted $c^{'}$, that would be less costly than *c*. By definition, strategy *c* is said to be subject to extended dominance and is therefore eliminated. Strategy *e* is also eliminated but for another reason. A move from *e* to *f* implies a negative *ICER*. This means that *e* is subject to simple dominance, i.e. this strategy generates higher costs than *f* but also lower benefits.

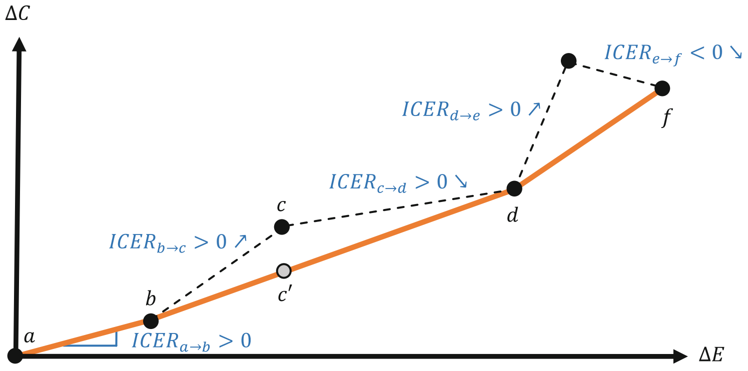


Fig. 10.15 Construction of the efficiency frontier

The construction of the efficiency frontier is thus based on two algorithms, one for the elimination of SSD alternatives, those that yield higher costs than the next more effective strategy (strategies are ordered by increasing effectiveness), and one for the elimination of SED strategies (those that yield a decrease in the *ICER*).

Figure 10.16 illustrates the methodology (example 3). SSD elimination rules out strategies such that $\Delta C \leq 0$ with respect to the next more effective strategy until all ΔC are strictly positive. In the example, the first step of SSD elimination identifies four strategies (namely *c*, *d*, *f* and *h*) that are SSD. The second step exhausts the possibilities of simple dominance with the exclusion of strategy *g* for which the cost difference with strategy *i* is null. All remaining strategies have a strictly positive cost difference with the next more effective strategy. The procedure goes on with the elimination of SED strategies. The algorithm eliminates strategies such that $\Delta ICER \leq 0$ with respect to the next more effective strategy until all $\Delta ICER$ are strictly positive. The first step excludes strategy *e* for the negative *ICER* difference with strategy *b*. The latter does not pass the second step when it is confronted with strategy *a*. Strategies *a*, *i* and *j* form the efficiency frontier (Fig. 10.17).

			Elimination of SSD strategies						Elimination of SED strategies								
			Step 1		Step 2		Step 3		Step 1			Step 2			Step 3		
	E_k	C_k		ΔC		ΔC		ΔC	*ICER* $S_{k-1} \to S_k$		$\Delta ICER$	*ICER* $S_{k-1} \to S_k$		$\Delta ICER$	*ICER* $S_{k-1} \to S_k$		$\Delta ICER$
a	10	10	S_1	4	S_1	4	S_1	4	S_1	NA	NA	S_1	NA	NA	S_1	NA	NA
b	15	14	S_2	26	S_2	22	S_2	22	S_2	0.80	0.08	S_2	0.8	−0.45			
c	20	40	S_3	−1													
d	30	39	S_4	−3													
e	40	36	S_5	16	S_3	4	S_3	4	S_3	0.88	−0.80						
f	50	52	S_6	−12													
g	70	40	S_7	20	S_4	0			S_4								
h	85	60	S_8	−20													
i	90	40	S_9	10	S_5	10	S_4	10	S_5	0.08	0.92	S_3	0.35	0.65	S_2	0.37	0.63
j	100	50	S_{10}	NA	S_6	NA	S_5	NA	S_6	1	NA	S_4	1	NA	S_3	1	NA

Fig. 10.16 Elimination of SSD and SED alternatives: example 3

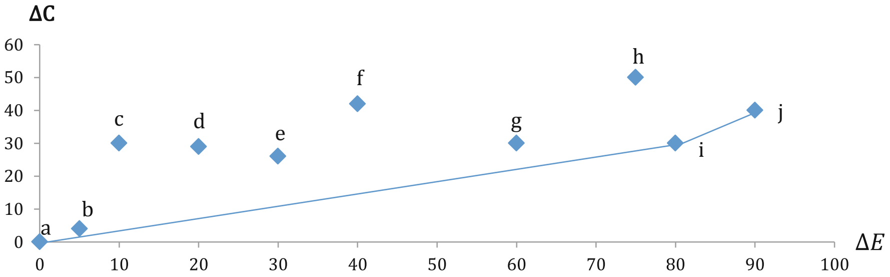


Fig. 10.17 Elimination of SSD and SED alternatives: example 3

As we have seen, the assessment of efficiency is ultimately based on aggregate cost and effectiveness data. However, the process by which such data is measured is far from straightforward. It requires conceiving and simulating a model that represents the evolution of the population targeted by the public project.

10.4 Decision Analytic Modeling

When comparing strategies, the ideal situation would be to rest on their true (population) mean costs and mean effects. In practice, information is usually obtained from sample means generated by a model simulating the evolution of a cohort under a given strategy. An often used framework is the Markov model, a simple and powerful tool. One of the most interesting properties of Markov modeling is that it allows reversible situations (you have a job once, lose it, then enter a training program and get a new one: the training policy has reversed your situation).

The first thing to define is the pertinent cohort, namely the subjects in the target population of the public project. All the relevant situations of those subjects should be listed in a finite number of Markov states m_i, $i = 1 \ldots M$. At one period, subjects are always located in one and only one Markov state. From one period to another, they may move from one Markov state to another, if that transition is allowed. The time horizon of the simulation is T. Periods are termed Markov cycles and denoted $t = 1 \ldots T$. Cycles are appropriate time increments, often years. Moving from state m_i to state m_j is expressed by transition probability $p_{ij} \geq 0$ with $\sum_{j=1}^{M} p_{ij} = 1$ for all $i = 1 \ldots M$.

A Markov process is characterized by $\mathbf{N}_0$, the initial allocation of subjects amongst the Markov states at $t = 0$:

$$\begin{array}{ccccccccc} & & m_1 & \cdots & m_i & \cdots & m_j & \cdots & m_M \\ \mathbf{N}_0 & = & [n_{01} & \cdots & n_{0i} & \cdots & n_{0j} & \cdots & n_{0M}] \end{array}$$

and by a transition matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c|ccccccc|} & m_1 & \cdots & m_i & \cdots & m_j & \cdots & m_M \\ \hline m_1 & p_{11} & \cdots & p_{1i} & \cdots & p_{1j} & \cdots & p_{1M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_i & p_{i1} & \cdots & p_{ii} & \cdots & p_{ij} & \cdots & p_{iM} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_j & p_{j1} & \cdots & p_{ji} & \cdots & p_{jj} & \cdots & p_{jM} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_M & p_{M1} & \cdots & p_{Mi} & \cdots & p_{Mj} & \cdots & p_{MM} \end{array}$$

For each row of this matrix we have:

$$p_{i1} + \ldots + p_{ii} + \ldots + p_{ij} + \ldots p_{iM} = 1, \forall i = 1 \ldots M$$

The transition matrix does not have to be symmetric (moving from state m_i to state m_j does not necessarily have the same probability as moving from state m_j to state m_i). For instance, it may be "easier" to move from regular school attendance to dropout than the opposite. Furthermore, if the matrix is upper triangular (all the entries below the main diagonal are zeros), pathways are not reversible and the Markov process is a one-way oriented decision tree. Last, if transitions are time-dependent, then the Markov process is non-stationary so that $p_{ij} = p_{ij}(t)$ and consequently $\mathbf{P} = \mathbf{P}_t$.

Let us consider the evolution of the cohort among the states. The state vector $\mathbf{N}_\mathrm{t}$ describes the number of subjects of the cohort who are present in each Markov state at a given cycle t. More specifically, $\mathbf{N}_t$ is a $[1, M]$ vector with elements resulting from the matrix product of $\mathbf{N}_{t-1}$ with the transition matrix $\mathbf{P}$:

$$\mathbf{N}_t = \mathbf{N}_{t-1} \times \mathbf{P}$$

The number of subjects at cycle t in state i is denoted $n_{t,i}$. We equivalently have:

$$\mathbf{N}_t = \begin{bmatrix} n_{t1} & \ldots & n_{tM} \end{bmatrix} = \begin{bmatrix} n_{t-1,1} & \ldots & n_{t-1,M} \end{bmatrix} \times \mathbf{P}$$

By definition of a matrix product, the elements of $\mathbf{N}_t$ are determined by $n_{t,i} = \sum_{j=1}^{M} n_{t-1,j} \times p_{ji}$. Examples of computation are provided below. Cycle after cycle, the state vectors yield what is called the Markov trace, i.e. a $[T, M]$ matrix that displays the evolution of the cohort among the states over the whole set of periods:

$$\mathbf{TRACE} = \begin{array}{c} \\ t=1 \\ \vdots \\ t=\tau \\ \vdots \\ t=T \end{array} \begin{array}{c} \begin{array}{ccccccc} m_1 & \ldots & m_i & \ldots & m_j & \ldots & m_M \end{array} \\ \begin{bmatrix} n_{11} & \ldots & n_{1i} & \ldots & n_{1j} & \ldots & n_{1M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{\tau 1} & \ldots & n_{\tau i} & \ldots & n_{\tau j} & \ldots & n_{\tau M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{T1} & \ldots & n_{Ti} & \ldots & n_{Tj} & \ldots & n_{TM} \end{bmatrix} \end{array} = \begin{bmatrix} \mathbf{N}_1 \\ \vdots \\ \mathbf{N}_\tau \\ \vdots \\ \mathbf{N}_T \end{bmatrix}$$

The Markov trace allows to determine an effectiveness measure, based on the number of subjects in each state over the whole time horizon. For instance, should m_1 be a state describing how many patients survived after a medical treatment, then the number of observations in m_1 would serve as a proxy for the desired outcome of the strategy. A discounted value can then be computed to better assess the present value of this outcome.

The computation of the total cost is based on a column vector of dimension $[M, 1]$ that describes the unit cost $uc_i \geq 0$ associated with the number of subjects in each state m_i, denoted $\mathbf{UC}$ afterwards:

$$\mathbf{UC} = \begin{array}{c} m_1 \\ \vdots \\ m_i \\ \vdots \\ m_j \\ \vdots \\ m_M \end{array} \left[\begin{array}{c} uc_1 \\ \vdots \\ uc_i \\ \vdots \\ uc_j \\ \vdots \\ uc_M \end{array}\right]$$

Since the total cost depends on the number of subjects in each state at each period, we need to compute a cost vector of dimension $[M, 1]$ defined as:

$$\mathbf{COST} = \mathbf{TRACE} \times \mathbf{UC}$$

The sum of the elements of **COST** gives the total policy cost. Again, a discounted value can be computed to assess its present value.

The following numerical example illustrates the basic functioning of Markov modeling. Consider a cohort of schoolchildren entering a 3-year educational program. Education authorities are concerned about demotivation and possible dropouts. Education analysts provide them with a scenario for the evolution of the cohort, synthesized by the state-transition Markov diagram of Fig. 10.18. The total number of subjects (here, schoolchildren) is 1000. Markov cycles are years. The time horizon of the simulation is $T = 3$. The set of Markov states is $\{m_1, m_2, m_3, m_4\}$. State m_1 comprehends pupils who attend school on a regular basis. State m_2 (respectively state m_3) concerns occasional (respectively frequent) school leavers. State m_4 deals with definitive dropouts. The education analysts' assumption is that the first three states are reversible while the fourth one is an absorbing state (those children would never go back to standard school). By definition state m_i is absorbing if $p_{ii} = 1$ and consequently $p_{i,j \neq i} = 0$. Let us assume for the moment that there

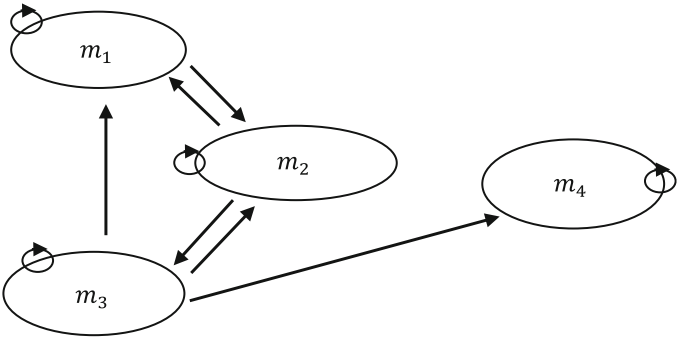


Fig. 10.18 Example of a state-transition Markov diagram

is no public intervention and denote a this strategy. The transition matrix provided by the education analysts is:

$$\mathbf{P}_a = \begin{bmatrix} p_{11} = 0.80 & p_{12} = 0.20 & p_{13} = 0.00 & p_{14} = 0.00 \\ p_{21} = 0.40 & p_{22} = 0.50 & p_{23} = 0.10 & p_{24} = 0.00 \\ p_{31} = 0.05 & p_{32} = 0.30 & p_{33} = 0.60 & p_{34} = 0.05 \\ p_{41} = 0.00 & p_{42} = 0.00 & p_{43} = 0.00 & p_{44} = 1.00 \end{bmatrix}$$

Matrix $\mathbf{P}_a$ can be based upon various information sources, depending on the actual context of the evaluation. They can stem from expert judgments, past experience, previous field experiments, or clinical trials in health. Transition probabilities can be a simple number or a complex combination of parameters leading to that number (for instance combining risk factors and socio-demographic characteristics).

The arrows on the state-transition Markov diagram of Fig. 10.18 describe the moves subjects can make in the model. For instance, a transition from state m_2 to state m_3 is allowed and the transition matrix quantifies it to $p_{23} = 0.10$, meaning that 10% of the subjects who were in state m_2 at cycle $t - 1$ will move to state m_3 at cycle t. Transition from state m_2 to state m_4 is not allowed which is quantified by $p_{24} = 0.00$. Circling arrows describe a strictly positive probability of remaining in the same Markov state from one cycle to the next. For instance, 50% of schoolchildren who were occasional school leavers during the previous period will remain so during the current period ($p_{22} = 0.50$). When absorbing states exist in the model, they only get incoming arrows and once subjects have reached such states, they can no longer get out of it. This is the case here of Markov state m_4, for which there is an incoming transition from m_3 but no way out since $p_{44} = 1$.

Let us assume that the initial allocation of subjects amongst the Markov states is $\mathbf{N}_0 = [1000; 0; 0; 0]$, i.e. all the students attend school on a regular basis (different initial allocations can of course be analyzed). Since transitions from matrix $\mathbf{P}_a$ are not time-dependent, the Markov process is stationary. In our case, it means that the passage of time is not supposed to affect the behavior of children (an alternative would be for example that as they get older, they also get more inclined to school leaving, in which case transition probabilities towards dropout states would relatively increase with time). What is the educational status of the cohort, year after year? Vectors $\mathbf{N}_t$ for $t = 1 \ldots 3$ provide the Markov trace of the process:

$$\mathbf{TRACE}_a = \begin{array}{cccccc} \text{Cycle} & m_1 & m_2 & m_3 & m_4 & \text{check} \\ t = 1 & 800 & 200 & 0 & 0 & 1{,}000 \\ t = 2 & 720 & 260 & 20 & 0 & 1{,}000 \\ t = 3 & 681 & 280 & 38 & 1 & 1{,}000 \end{array}$$

The Markov trace provides an annual account of the allocation of schoolchildren amongst the various Markov states describing their educational situation. Since n_{ti}

denotes the number of subjects at cycle t in state m_i, the elements of the state vectors $\mathbf{N}_t$ can be detailed as follows.

At cycle 1:

$$n_{11} = n_{01} \times p_{11} = 1000 \times 0.8 = 800$$

$$n_{12} = n_{01} \times p_{12} = 1000 \times 0.2 = 200$$

$$n_{13} = n_{01} \times p_{13} = 1000 \times 0.0 = 0$$

$$n_{14} = n_{01} \times p_{14} = 1000 \times 0.0 = 0$$

At cycle 2:

$$n_{21} = n_{11} \times p_{11} + n_{12} \times p_{21} + n_{13} \times p_{31}$$

$$= 800 \times 0.8 + 200 \times 0.4 + 0 \times 0.05 = 720$$

$$n_{22} = n_{11} \times p_{12} + n_{12} \times p_{22} + n_{13} \times p_{32}$$

$$= 800 \times 0.2 + 200 \times 0.5 + 0 \times 0.3 = 260$$

$$n_{23} = n_{12} \times p_{23} + n_{13} \times p_{33}$$

$$= 200 \times 0.1 + 0 \times 0.6 = 20$$

$$n_{24} = n_{13} \times p_{34} + n_{14} \times p_{44}$$

$$= 0 \times 0.05 + 0 \times 1 = 0$$

At cycle 3:

$$n_{31} = n_{21} \times p_{11} + n_{22} \times p_{21} + n_{23} \times p_{31}$$

$$= 720 \times 0.8 + 260 \times 0.4 + 20 \times 0.05 = 681$$

$$n_{32} = n_{21} \times p_{12} + n_{22} \times p_{22} + n_{23} \times p_{32}$$

$$= 720 \times 0.2 + 260 \times 0.5 + 20 \times 0.3 = 280$$

$$n_{33} = n_{22} \times p_{23} + n_{23} \times p_{33}$$

$$= 260 \times 0.1 + 20 \times 0.6 = 38$$

$$n_{34} = n_{23} \times p_{34} + n_{24} \times p_{44}$$

$$= 20 \times 0.05 + 0 \times 1 = 1$$

Facing the dropout problem exemplified above, the educational authorities may wish to put forward a prevention strategy, for instance by increasing support for schoolchildren who have been identified as occasional or frequent school leavers (those reaching states m_2 and m_3). That support strategy, labeled b hereafter, for instance results in the following modified transition matrix:

$$\mathbf{P}_b = \begin{bmatrix} p_{11} = 0.80 & p_{12} = 0.20 & p_{13} = 0.00 & p_{14} = 0.00 \\ p_{21} = 0.60 & p_{22} = 0.30 & p_{23} = 0.10 & p_{24} = 0.00 \\ p_{31} = 0.20 & p_{32} = 0.35 & p_{33} = 0.40 & p_{34} = 0.05 \\ p_{41} = 0.00 & p_{42} = 0.00 & p_{43} = 0.00 & p_{44} = 1.00 \end{bmatrix}$$

The support strategy is such that children who are occasional school leavers (state m_2) are taken care of for instance through interviews and subsequent follow-up. As a consequence, more of them move back (in probability) to state m_1 (60% instead of 40%), fewer stay in state m_2 (30% instead of 50%) which decreases the shift towards frequent dropout. Similarly, the children who are frequent school leavers (state m_3) move back more easily to state m_1 (20% instead of 5%) and state m_2 (35% instead of 30%), which reduces the probability of staying in state m_3 (40% instead of 60%). The result of the policy appears in the new Markov trace. One can thus check cycle after cycle the outcomes of the prevention strategy:

$$\mathbf{TRACE}_b = \begin{array}{cccccc} \text{Cycle} & m_1 & m_2 & m_3 & m_4 & \text{check} \\ t = 1 & 800 & 200 & 0 & 0 & 1000 \\ t = 2 & 760 & 220 & 20 & 0 & 1000 \\ t = 3 & 744 & 225 & 30 & 1 & 1000 \end{array}$$

In the example proposed here, the Markov trace shows that the dropout problem decreases when the support policy b is implemented. More children remain or come back to state m_1 while there are fewer dropouts at each cycle.

The next numerical example will comprehend both the cost and consequence aspects of a similar decision problem in education over a longer time horizon that will require discounting (in theory, the previous example should have used discounting too, but was mainly intended to provide an introduction to the mechanics of Markov modeling).

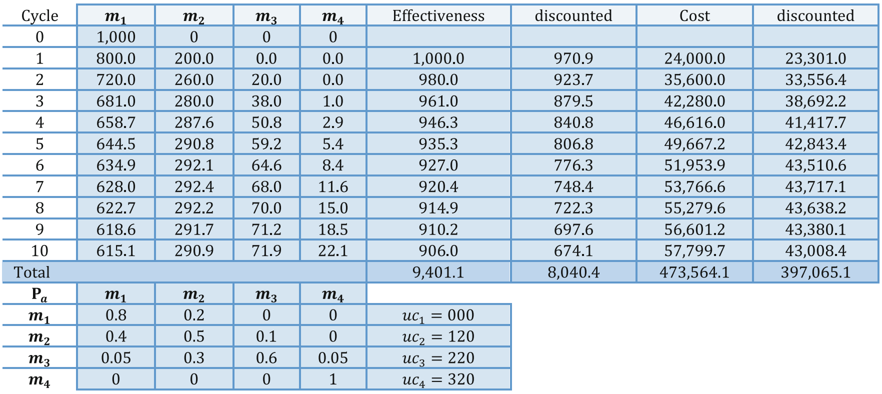
Assume that education authorities consider early school leaving over the whole horizon of mandatory school (for instance from the age of six to the age of sixteen), so that $T = 10$. That time span requires discounting and authorities choose a rate of 3%. Strategy a involves no particular prevention, strategies b, c and d each provide a different type of prevention. The Markov process still rests on Fig. 10.18 and is assumed again to be stationary. Each strategy is associated with a transition matrix and a cost vector. The cost perimeter is associated with Markov states m_2 , m_3 , m_4. This means that any subject entering (or staying in) those states triggers the corresponding costs. Effectiveness is defined by the headcount of schoolchildren in states m_1 and m_2. Of course, any other relevant measure of effectiveness can be used, provided that the model is equipped to deliver it. Starting with strategy a, matrix $\mathbf{P_a}$ is the same as in the previous numerical example; as to the cost matrix $\mathbf{UC_a}$, it is given by:

$$\mathbf{UC}_{\boldsymbol{a}} = \begin{bmatrix} uc_1 = 000 \\ uc_2 = 120 \\ uc_3 = 220 \\ uc_4 = 320 \end{bmatrix}$$

The Markov trace as well as measures of non-discounted and discounted cost and effectiveness are displayed in Fig. 10.19. The Markov trace is obtained by applying the transition matrix to each state vector. In Excel, state vectors $\mathbf{N}_t$ can be obtained with the command *MMULT*. This Excel function calculates the matrix product of two arrays. The format is $MMULT(array1, array2)$ where $array1$ and $array2$ are matrices. The number of columns in $array1$ is equal to the number of rows in $array2$. To input an array formula, one needs to (1) highlight the range of cells for the new matrix, (2) type the command $MMULT(array1, array2)$, and (3) press CTRL-SHIFT-Enter. The resulting matrix has the same number of rows as array1 and the same number of columns as array2.

Strategy b provides a first prevention policy that induces a new transition matrix as well as a new cost structure. Figure 10.20 displays them (with parameter changes indicated in blue) as well as the new Markov trace and its associated effectiveness and cost measures. Strategies c and d are similarly described in Figs. 10.21 and 10.22, changes with respect to strategy a indicated respectively with colors red and orange. The information generated by the Markov traces of the competing strategies allows first to build the efficiency frontier. Figure 10.23 shows how strategy b is SED. Strategies on the efficiency frontier are described in Fig. 10.24 and then characterized in Fig. 10.25.

Finally, Fig. 10.26 uses the *INB* approach to sort out the preferred strategies with respect to *WTP*. Strategy b is never the most preferred. In line with Fig. 10.24,



Cycle	m_1	m_2	m_3	m_4	Effectiveness	discounted	Cost	discounted
0	1,000	0	0	0				
1	800.0	200.0	0.0	0.0	1,000.0	970.9	24,000.0	23,301.0
2	720.0	260.0	20.0	0.0	980.0	923.7	35,600.0	33,556.4
3	681.0	280.0	38.0	1.0	961.0	879.5	42,280.0	38,692.2
4	658.7	287.6	50.8	2.9	946.3	840.8	46,616.0	41,417.7
5	644.5	290.8	59.2	5.4	935.3	806.8	49,667.2	42,843.4
6	634.9	292.1	64.6	8.4	927.0	776.3	51,953.9	43,510.6
7	628.0	292.4	68.0	11.6	920.4	748.4	53,766.6	43,717.1
8	622.7	292.2	70.0	15.0	914.9	722.3	55,279.6	43,638.2
9	618.6	291.7	71.2	18.5	910.2	697.6	56,601.2	43,380.1
10	615.1	290.9	71.9	22.1	906.0	674.1	57,799.7	43,008.4
Total					9,401.1	8,040.4	473,564.1	397,065.1

$\mathbf{P}_a$	m_1	m_2	m_3	m_4	
m_1	0.8	0.2	0	0	$uc_1 = 000$
m_2	0.4	0.5	0.1	0	$uc_2 = 120$
m_3	0.05	0.3	0.6	0.05	$uc_3 = 220$
m_4	0	0	0	1	$uc_4 = 320$

Fig. 10.19 Cost and effectiveness data: example 4 (strategy a)

Cycle	m_1	m_2	m_3	m_4	Effectiveness	discounted	Cost	discounted
0	1000	0	0	0				
1	800.0	200.0	0.0	0.0	1,000.0	970.9	32,800.0	31,844.7
2	760.0	220.0	20.0	0.0	980.0	923.7	42,480.0	40,041.5
3	744.0	225.0	30.0	1.0	969.0	886.8	47,120.0	43,121.5
4	736.2	226.8	34.5	2.5	963.0	855.6	49,785.2	44,233.5
5	731.9	227.4	36.5	4.2	959.3	827.5	51,579.3	44,492.8
6	729.3	227.4	37.3	6.0	956.6	801.2	52,982.6	44,372.1
7	727.3	227.1	37.7	7.9	954.4	776.0	54,209.6	44,077.4
8	725.6	226.8	37.8	9.8	952.4	751.9	55,356.5	43,698.9
9	724.1	226.4	37.8	11.7	950.5	728.5	56,466.3	43,276.7
10	722.7	226.0	37.8	13.6	948.7	705.9	57,558.3	42,828.8
Total					9,633.95	8,227.93	500,337.8	421,987.8

P_b	m_1	m_2	m_3	m_4	
m_1	0.8	0.2	0	0	$uc_1 = 000$
m_2	0.6	0.3	0.1	0	$uc_3 = 164$
m_3	0.2	0.35	0.4	0.05	$uc_4 = 320$
m_4	0	0	0	1	$uc_4 = 620$

Fig. 10.20 Cost and effectiveness data: example 4 (strategy *b*)

Cycle	m_1	m_2	m_3	m_4	Effectiveness	discounted	Cost	discounted
0	1,000	0	0	0				
1	800.0	200.0	0.0	0.0	1,000.0	970.9	40,000.0	38,835.0
2	780.0	200.0	20.0	0.0	980.0	923.7	47,200.0	44,490.5
3	772.0	206.0	21.0	1.0	978.0	895.0	49,400.0	45,208.0
4	770.2	206.1	21.7	2.1	976.3	867.4	50,326.0	44,714.0
5	769.1	206.1	21.7	3.1	975.2	841.2	51,031.1	44,019.9
6	768.2	205.9	21.7	4.2	974.1	815.8	51,684.7	43,285.2
7	767.4	205.7	21.7	5.3	973.0	791.2	52,328.2	42,547.7
8	766.5	205.4	21.7	6.4	972.0	767.3	52,969.0	41,814.2
9	765.7	205.2	21.6	7.5	970.9	744.1	53,608.7	41,086.6
10	764.9	205.0	21.6	8.5	969.8	721.7	54,247.5	40,365.3
Total					9,769.3	8,338.2	502,795.3	426,366.2

P_c	m_1	m_2	m_3	m_4	
m_1	0.8	0.2	0	0	$uc_1 = 000$
m_2	0.7	0.2	0.1	0	$uc_2 = 200$
m_3	0.4	0.5	0.05	0.05	$uc_3 = 360$
m_4	0	0	0	1	$uc_4 = 640$

Fig. 10.21 Cost and effectiveness data: example 4 (strategy *c*)

Cycle	m_1	m_2	m_3	m_4	Effectiveness	discounted	Cost	discounted
0	1,000	0	0	0				
1	800.0	200.0	0.0	0.0	1000.0	970.9	44000.0	42718.4
2	780.0	210.0	10.0	0.0	990.0	933.2	49700.0	46847.0
3	773.0	212.5	14.0	0.5	985.5	901.9	52000.0	47587.4
4	770.0	213.3	15.5	1.2	983.3	873.6	53205.3	47272.2
5	768.4	213.5	16.1	2.0	981.9	847.0	53995.3	46576.8
6	767.4	213.5	16.3	2.8	980.9	821.5	54626.2	45748.6
7	766.6	213.4	16.4	3.6	980.0	796.8	55196.1	44879.5
8	766.0	213.2	16.4	4.4	979.2	773.0	55742.3	44003.5
9	765.3	213.1	16.4	5.2	978.4	749.8	56279.2	43133.3
10	764.7	212.9	16.4	6.1	977.5	727.4	56812.2	42273.6
Total					9836.7	8395.1	531556.5	451040.2

P_d	m_1	m_2	m_3	m_4	
m_1	0.8	0.2	0	0	$\mathrm{uc}_1 = 000$
m_2	0.7	0.25	0.05	0	$\mathrm{uc}_2 = 220$
m_3	0.2	0.4	0.35	0.05	$\mathrm{uc}_3 = 350$
m_4	0	0	0	1	$\mathrm{uc}_4 = 700$

Fig. 10.22 Cost and effectiveness data: example 4 (strategy *d*)

Denomination of strategy	Notation of strategy	Effectiveness E_k	Cost ($) C_k	$ICER(S_{k-1} \rightarrow S_k)$
a	S_1	8,040.4	397,065.1	NA
b	S_2	8,227.9(< 8,338.2)	421,987.8	132.9(> 39.7)
c	S_3	8,338.2	426,366.2	39.7
d	S_4	8,395.1	451,040.2	434.1

Fig. 10.23 Elimination of SED strategies: example 4

Denomination of strategy	Notation of strategy	Effectiveness E_k	Cost ($) C_k	$ICER(S_{k-1} \rightarrow S_k)$
a	S_1	8,040.4	397,065.1	NA
c	S_2	8,338.2	426,366.2	98.4
d	S_3	8,395.1	451,040.2	434.1

Fig. 10.24 Strategies on the efficiency frontier: example 4

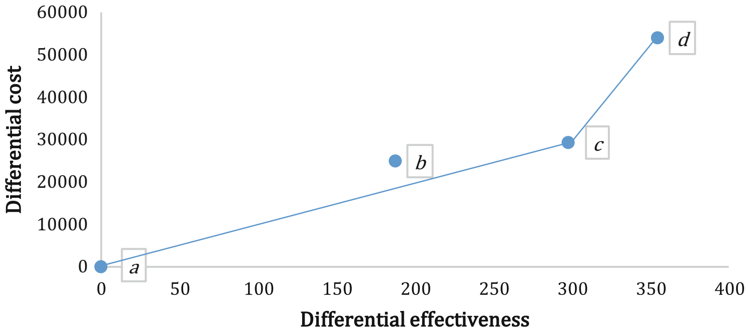


Fig. 10.25 Efficiency frontier: example 4

below a *WTP* of $98.4 per unit of effectiveness gained, the reference strategy *a* is preferred, so that no attempt at containing dropout is made. For a *WTP* between $98.4 and $434.4, then the prevention policy associated with strategy *c* should prevail. Above $434.4 per unit of effectiveness gained, strategy *d* should win through.

Note that for simplicity of exposition, the initial allocation of subjects amongst the states has been such that they were all in state m_1 so that $\mathbf{N}_0 = [1000; 0; 0; 0]$. This assumption can be easily relaxed if the context analysis has evidenced that subjects enter the model not only from m_1, but also from m_2 and m_3. Past or neighboring experience may for instance have shown that a relevant initial allocation would be $\mathbf{N}_0 = [700; 200; 100; 0]$. The model can then be rerun accordingly. The context of the analysis usually implies that subjects do not enter the model from an absorbing state. The new initial allocation of subjects uniformly applies to all the strategies.

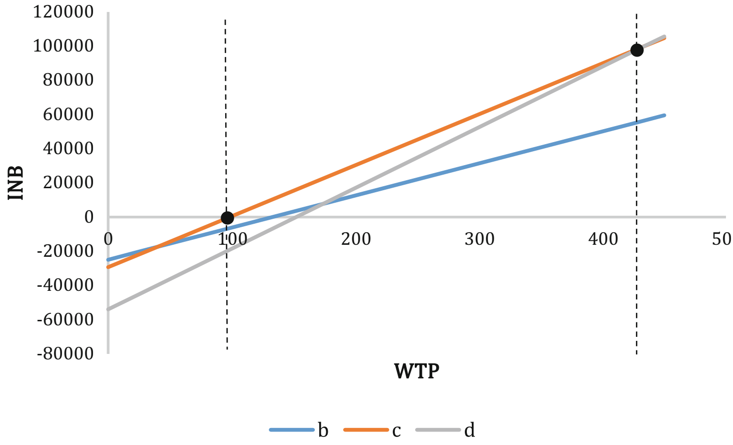


Fig. 10.26 Efficiency in the *INB* plane: example 4

10.5 Numerical Implementation in R-CRAN

In what follows, we introduce the main commands to be used in R-CRAN in order to implement a cost effectiveness analysis step by step. Figure 10.27 presents the command lines for deriving the Markov trace, as well as the cost and effectiveness data for strategy *a*, which was fully described in the previous section. The first stage is about loading the package *markovchain* using the command *library*. We also load the package *FinCal* which will be used to compute the discounted values of the cost and effectiveness measures. The second stage consists in creating a Markov chain object, labeled *P*, using command *new*. The option *states* allows the Markov states to be named as m_1, m_2, m_3 and m_4. Option *transitionMatrix* is used to specify the matrix of transition probabilities $\mathbf{P}_a$, which in our case is made of four columns and sixteen entries. Using command *plot* creates a graph displaying the Markov model. Since the position of the circles representing Markov states is randomly defined, command *plot* can be implemented several times until one gets a suitable graphic (for instance, that of Fig. 10.28).

In a third stage, we construct the Markov trace. Vector $\mathbf{N}_0$ is defined as $N0 = c$ $(1000, 0, 0, 0)$. To understand the functioning of the process, Fig. 10.27 first presents how one may derive the first state vector $\mathbf{N}_1$ by simply multiplying $N0$ by P. Similarly, the second vector $\mathbf{N}_2$ is obtained using $N0 * \mathrm{P}\,\hat{}\,2$ (similarly for $\mathbf{N}_3$). Then, a loop automatizes the process. This is done first by defining the time horizon $T = 10$, then by specifying the dimension of the Markov trace (i.e. T rows × 4 columns corresponding to the Markov states), and finally by developing the loop. The latter is made of an iterator t that goes from 1 to T and an iterable object

```
> library(markovchain)
> library(FinCal)
> M=4
> P=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
matrix(c(0.8,0.4,0.05,0,0.2,0.5,0.3,0,0,0.1,0.6,0,0,0,0.05,1),nrow=M))
> plot(P)

> N0=c(1000,0,0,0)
> N0*P
      m1  m2 m3 m4
[1,] 800 200  0  0
> N0*P^2
      m1  m2 m3 m4
[1,] 720 260 20  0
> N0*P^3
      m1  m2 m3 m4
[1,] 681 280 38  1

> T=10
> TRACE=matrix(NA,nrow=T,ncol=M)
> colnames(TRACE)=c("m1","m2","m3","m4")
> for(t in 1:T){TRACE[t,]=N0*P^t}

> a=data.frame(TRACE)
> a$PERIOD=1:T

> a$EFFECT=a$m1+a$m2
> UC=c(0,120,220,320)
> a$COST=TRACE%*%UC

> Disc.RATE=0.03
> a$DiscEFFECT=-pv.simple(Disc.RATE,a$PERIOD,a$EFFECT)
> a$DiscCOST=-pv.simple(Disc.RATE,a$PERIOD,a$COST)

> round(a)
    m1  m2 m3 m4 PERIOD EFFECT  COST DiscEFFECT DiscCOST
1  800 200  0  0      1   1000 24000        971    23301
2  720 260 20  0      2    980 35600        924    33556
3  681 280 38  1      3    961 42280        879    38692
4  659 288 51  3      4    946 46616        841    41418
5  645 291 59  5      5    935 49667        807    42843
6  635 292 65  8      6    927 51954        776    43511
7  628 292 68 12      7    920 53767        748    43717
8  623 292 70 15      8    915 55280        722    43638
9  619 292 71 19      9    910 56601        698    43380
10 615 291 72 22     10    906 57800        674    43008
```

Fig. 10.27 Cost and effectiveness data with R-CRAN: example 4

between brackets: for each of the values of t, a state vector is built and included into the *TRACE* matrix using command $TRACE[t,] = N0 * P \wedge t$.

The last stage consists in computing the effectiveness and cost vectors. For this purpose, it is preferable to work with a database instead of a matrix, since it allows additional variables to be easily included. In Fig. 10.27, we define a database for strategy a using command *data.frame*. The dollar sign notation is used to define each variable entering the dataset. The first variable that is created is *a$period* that will index each Markov cycle from 1 to T. This variable is also used afterwards to discount the cost and effectiveness measures. The second variable is our index of effectiveness (*a$EFFECT*) measured in this particular example as the sum of the subjects belonging to states m_1 and m_2. We specify the unit cost vector as $UC = c$

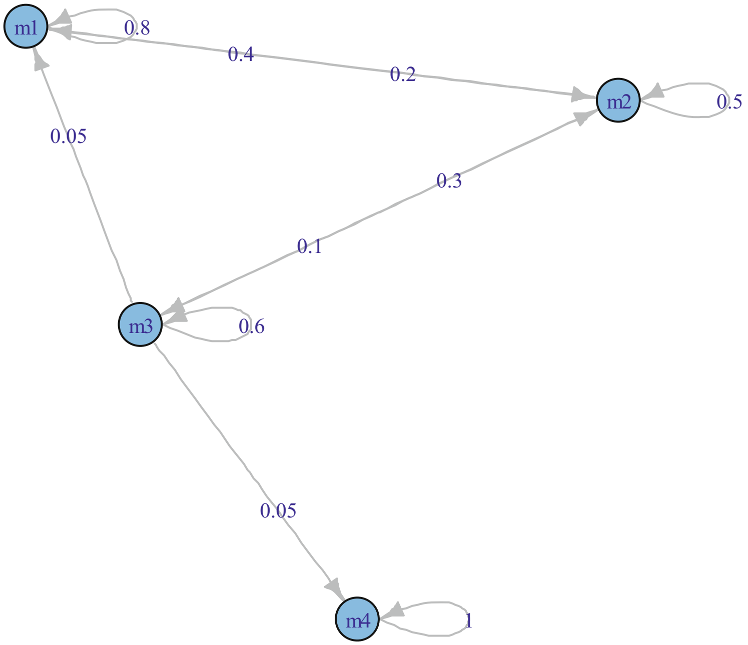


Fig. 10.28 State-transition Markov diagram with R-CRAN: example 4

(0,120,220,320). The final cost vector is obtained by multiplying the trace with *UC* to get *a*$*COST*. The multiplication sign has to be put between percentage signs (%) as we need here to implement a matrix multiplication. The $-pv\,.\,simple$ command is then used to compute the present value of each observation in *a*$*EFFECT* and *a*$*COST* by specifying both the discount rate (*Disc . RATE*) and the time period (*a*$*period*). The final database *a* is thus made of each relevant variable for the cost effectiveness analysis. Command *round* is used for presentation purpose and simply rounds the values of the dataset.

The program of Fig. 10.27 can be improved by creating a function that encompasses all the previous commands into a single one. The approach is then much faster, and allows the evaluator to examine a large set of competing strategies. In Fig. 10.29, we define a function labeled *cea* that depends on the transition matrix *P*, the unit cost vector *UC*, the time horizon *T*, the initial state vector *N*0, and the discount rate *Disc . RATE*. Between brackets are then specified the commands used to compute the trace, the cost and effectiveness vectors, as well as their discounted values. Command *print* finally specifies the final output of the function.

While *P* and *UC* are specific to each strategy, this is not the case for the time horizon, the initial vector and the discount rate. Those variables are common to all competing strategies and are thereby specified directly after the creation of function *cea*. In Fig. 10.29, we then define *P* and *UC* for each strategy. Command *cea* is used to create the relevant data for each strategy from *a* to *d*. Command *round* can be

```
> library(markovchain)
> library(FinCal)

> cea=function(P,UC,T,N0,Disc.RATE){
+ TRACE=matrix(NA,nrow=T,ncol=M)
+ TRACE[1,]=N0
+ colnames(TRACE)=c("m1","m2","m3","m4")
+ for(t in 1:T){TRACE[t,]=N0*P^t}
+ dat=data.frame(round(TRACE))
+ dat$PERIOD=1:T
+ dat$EFFECT=dat$m1+dat$m2
+ dat$discEFFECT=round(-pv.simple(Disc.RATE,dat$PERIOD,dat$EFFECT))
+ dat$COST=round(TRACE%*%UC)
+ dat$discCOST=round(-pv.simple(Disc.RATE,dat$PERIOD,dat$COST))
+ print(dat)}

> M=4
> T=10
> N0=c(1000,0,0,0)
> Disc.RATE=0.03

> #Strategy a
> Pa=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ matrix(c(0.8,0.4,0.05,0,0.2,0.5,0.3,0,0,0.1,0.6,0,0,0,0.05,1),
+ nrow=M))
> UCa=c(0,120,220,320)
> a=cea(Pa,UCa,T,N0,Disc.RATE)
    m1  m2 m3 m4 PERIOD EFFECT discEFFECT  COST discCOST
1  800 200  0  0      1   1000        971 24000    23301
2  720 260 20  0      2    980        924 35600    33556
3  681 280 38  1      3    961        879 42280    38692
...
> #Strategy b
> Pb=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ matrix(c(0.8,0.6,0.2,0,0.2,0.3,0.35,0,0,0.1,0.4,0,0,0,0.05,1),
+ nrow=M))
> UCb=c(0,164,320,620)
> b=cea(Pb,UCb,T,N0,Disc.RATE)
    m1  m2 m3 m4 PERIOD EFFECT discEFFECT  COST discCOST
1  800 200  0  0      1   1000        971 32800    31845
2  760 220 20  0      2    980        924 42480    40041
3  744 225 30  1      3    969        887 47120    43121
...
> #Strategy c
> Pc=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ matrix(c(0.8,0.7,0.4,0,0.2,0.2,0.5,0,0,0.1,0.05,0,0,0,0.05,1),
+ nrow=M))
> UCc=c(0,200,360,640)
> c=cea(Pc,UCc,T,N0,Disc.RATE)
    m1  m2 m3 m4 PERIOD EFFECT discEFFECT  COST discCOST
1  800 200  0  0      1   1000        971 40000    38835
2  780 200 20  0      2    980        924 47200    44491
3  772 206 21  1      3    978        895 49400    45208
...
> #Strategy d
> Pd=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ matrix(c(0.8,0.7,0.2,0,0.2,0.25,0.4,0,0,0.05,0.35,0,0,0,0.05,1),
+ nrow=M))
> UCd=c(0,220,350,700)
> d=cea(Pd,UCd,T,N0,Disc.RATE)
    m1  m2 m3 m4 PERIOD EFFECT discEFFECT  COST discCOST
1  800 200  0  0      1   1000        971 44000    42718
2  780 210 10  0      2    990        933 49700    46847
3  773 213 14  1      3    986        902 52000    47587
...
```

Fig. 10.29 Creating a cost effectiveness function in R-CRAN: example 4

```
> E=c(sum(a$discEFFECT),sum(b$discEFFECT),sum(c$discEFFECT),
+ sum(d$discEFFECT))
> C=c(sum(a$discCOST),sum(b$discCOST),sum(c$discCOST),sum(d$discCOST))

> # ICER k-1 to k
> delta.E=diff(E,1)
> delta.C=diff(C,1)
> ICER=delta.C/delta.E
> ICER
[1] 132.87862  39.68932 434.05073

> # ICER k-1 to k after elimination of SED strategy b (number 2)
> delta.E=diff(E[-2],1)
> delta.C=diff(C[-2],1)
> ICER=delta.C/delta.E
> ICER
[1]  98.36643 434.05073

> # Frontier
> Ea=sum(a$discEFFECT)
> delta.E=E-Ea
> Ca=sum(a$discCOST)
> delta.C=C-Ca
> plot(delta.E[-2], delta.C[-2],pch=c("a","c","d"), type="b",
+ xlab="Differential effectiveness",ylab="Differential cost")
> points(delta.E[2],delta.C[2],pch="b")

> # INB
> WTP=0:700
> INBb=WTP*delta.E[2]-delta.C[2]
> INBc=WTP*delta.E[3]-delta.C[3]
> INBd=WTP*delta.E[4]-delta.C[4]
> plot(INBb,type="l",col=2,ylim=c(-100000,200000), xlab="WTP",
+ ylab="INB")
> points(INBc,type="l",col=3
> points(INBd,type="l",col=4)
> abline(h=0)
> abline(v=ICER)
> legend("bottomright",c("Strategy b","Strategy c","Strategy d"),
+ lty=c(1,1,1), col=2:4)
```

Fig. 10.30 *ICER*, efficiency frontier and *INB* with R-CRAN: example 4

avoided if fully detailed results are preferred. Results can be compared to those of Figs. 10.19, 10.20, 10.21 and 10.22.

Figure 10.30 first displays the command for computing the *ICER*. It starts with the creation of two vectors E and C that are respectively made of the effectiveness and cost measures. For instance, *sum*(*a*$*discEFFECT*) is the sum of all the discounted values of effectiveness for strategy a, i.e. stands for E_a. In other words, E and C are each made of four elements: $E=(E_a, E_b, E_c, E_d)$ and $C=(C_a, C_b, C_c, C_d)$. As data is already ordered by increasing effectiveness ($E_a < E_b < E_c < E_d$), the next step consists in computing the first differences using the values of vector E and C and command *diff*. Two new vectors are created and denoted $delta\,.\,E$ and $delta\,.\,C$. They are used directly to compute the *ICER*. We eliminate strategy b (recall that it is subject to extended dominance) by excluding the second element of vectors E and C using commands $E[-2]$ and $C[-2]$, and compute again the first difference. Make sure that you do not use the *round* command when generating cost and effectiveness data vectors within the *cea*

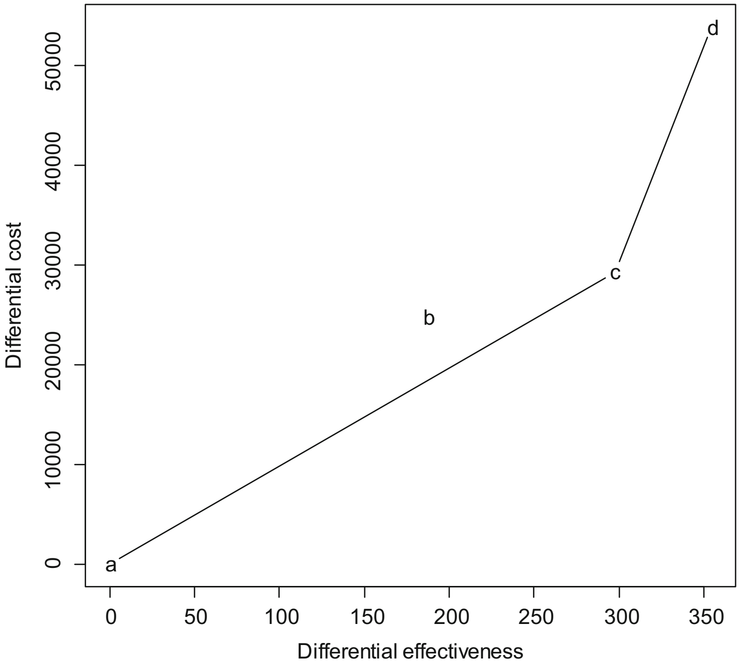


Fig. 10.31 Efficiency frontier with R-CRAN: example 4

function. Although it is quite useful for the presentation of results (as in Fig. 10.29), the use of rounded values precludes from obtaining exact outcomes.

The next step consists in plotting the efficiency frontier (Fig. 10.31). One needs to be careful as we need now to compute differences from each strategy to strategy *a* only, and not first differences as previously. To plot the frontier we first exclude strategy *b*. Then, to include separately strategy *b* on the graph we use command *points*. For both the *plot* and *points* functions we use command *pch* to define a vector of symbols. Command $type = ''b''$ used in the plot function allows to draw "both" a line and a symbol together. A description of graphic options is available by running command "?*plot . default*".

To implement the *INB* methodology (Fig. 10.32), one needs first to create a range of values for *WTP*. In Fig. 10.30, this range goes from 0 to 700. We then compute the *INB* for each value of *WTP* and for each strategy. The way to draw a graphic is similar to what has been done previously. Here, a range is defined for the *y*-axis in order to better represent each *INB*. We also specify the names of both the horizontal and vertical axes. The function *abline* is used to plot a horizontal line that goes through the origin ($h = 0$), as well as vertical lines ($v = ICER$) allowing to display the *WTP* for which strategies *c* and *d* are accepted. Last, a legend is created. The first entry relates to the position of the legend in the graphic. Option $lty = c(1, 1, 1)$ speficies the type of the line, "1" meaning continuous line. Option *col* defines the color of the lines, and *legend* their name.

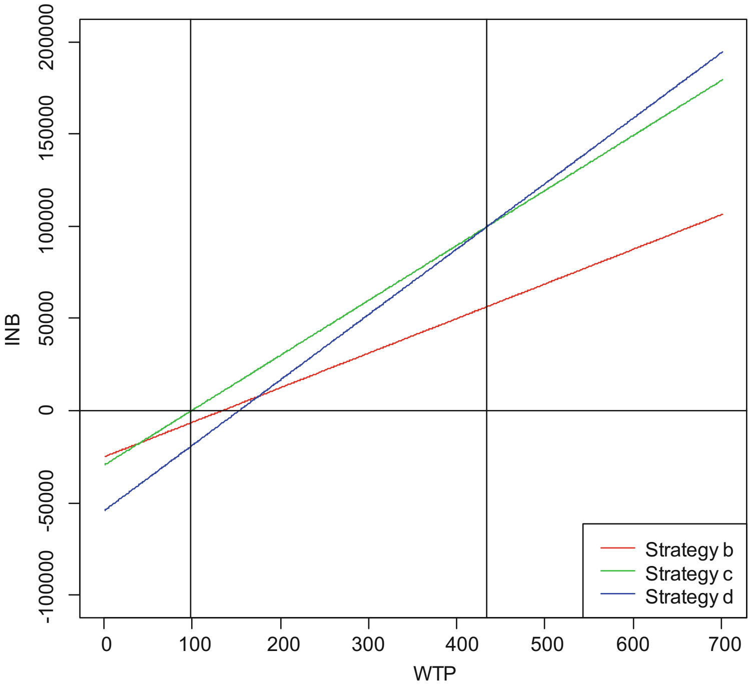


Fig. 10.32 Efficiency in the *INB* plane with R-CRAN: example 4

10.6 Extension to QALYs

A key feature of health programs is that patients can go through many health states (e.g., hospitalization for hip surgery, then convalescence period followed by progressive return to normal walk) that bring different levels of satisfaction. Accounting for these welfare changes can be essential to guide health-care resource allocation decisions. In this respect, a widely used measure of health effectiveness is that of quality adjusted life-years (QALYs). The idea is that the satisfaction of patients can be measured on a 0–1 scale which describes whether a health state is more desirable than another. As already stressed in a previous chapter (see Chap. 6), this measurement can be obtained from preference surveys in which standard gambles, time trade-offs or discrete choice experiments are used to create a ranking among health states. Health outcomes are translated into quality of life measures.

In practice, QALY increments or decrements are measured along discrete time intervals, for instance the cycles in a Markov model. Let us refer to example 4, strategy *a* where the four Markov states can be used to describe health states $[m_1, m_2, m_3, m_4]$. We make the assumption that quality of life deteriorates as individuals move from S_1 to S_2, S_2 to S_3, S_3 to S_4. Figure 10.33 provides the codes in R-CRAN. The first step is to upload packages *markovchain* and *FinCal*

```
> par(mar=c(4,4,2,1))
> library(markovchain)
> library(FinCal)
> M=4
> Pa=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ matrix(c(0.8,0.4,0.05,0,0.2,0.5,0.3,0,0,0.1,0.6,0,0,0,0.05,1),
+ nrow=M))
> N0=c(1000,0,0,0)
> T=50
> TRACE=matrix(NA,nrow=T,ncol=M)
> colnames(TRACE)=c("m1","m2","m3","m4")
> for(t in 1:T){TRACE[t,]=N0*Pa^t}
> a=data.frame(TRACE)
> a$PERIOD=1:T
> QALY=c(0.8,0.6,0.4,0.0)
> a$UTILITY=TRACE%*%QALY
> Disc.RATE=0.03
> a$DiscUTILITY=-pv.simple(Disc.RATE,a$PERIOD,a$UTILITY)
> plot(a$DiscUTILITY,xlab="Time increments",ylab="Discounted QALYs")
```

Fig. 10.33 QALY generation with R-CRAN

(for discounting) then create the Markov chain. The second step consists in generating the Markov trace and associating utility levels to it. At each time increment, here for instance a month, the initial population of 1000 subjects is reallocated among health states and their corresponding QALYs. In this numerical example, we put $[u(m_1) = 0.8, u(m_2) = 0.6, u(m_3) = 0.4, u(m_4) = 0.0]$. The discounted total utility reached by the cohort each month is represented in Fig. 10.34 until time horizon $T = 50$ months is reached.

Whenever effectiveness measures do not fully account for the condition of users or patients, QALYs appear as a way of apprehending it more accurately without any recourse to monetary valuation.

10.7 Uncertainty and Probabilistic Sensitivity Analysis

The exploration of uncertainty is of paramount importance in cost effectiveness analysis. The usual classification comes from health technology assessment guidelines and it includes methodological uncertainty, parameter uncertainty and structural uncertainty.

Methodological Uncertainty The vast majority of cost effectiveness analyses has to adhere to a reference case that prescribes the set of methods to be used. Inadequate compliance with guideline recommendations brings in discrepancies that undermine the scope and relevance of cost effectiveness outcomes. Methodological uncertainty comes from those failures. The first recommendation item is what is called the perspective of the model, namely: Who defines the effectiveness criterion, the cost perimeter? Second, the time horizon has to be carefully justified: too long, it dilutes costs and effects, particularly through discounting; too short, it may unduly truncate outcomes occurring too late in the evaluation process to be

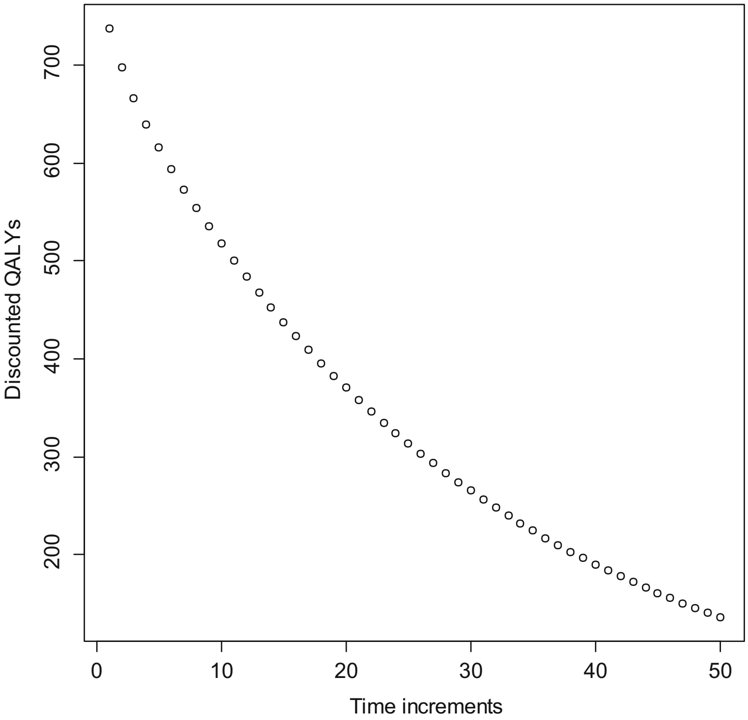


Fig. 10.34 QALY trace with R-CRAN

properly taken into account. Even in the absence of a guideline reference case, this is an important point. An example outside the field of health would be the choice of the right time horizon for an education program: should the horizon be limited to school achievements or should it reach beyond to catch the effects of the program on mid-term achievements on the job market? The third item is the choice of the discount rate. It influences the degree to which future outcomes are taken into account. Guidelines, when they exist, usually follow the recommendations of governments or supranational agencies.

The last two items are probably those who generate the more methodological uncertainty. The population of analysis (or population of interest, or of reference) should be targeted as precisely as possible in accordance with the goal defined during the analysis of the context of the program. Should it be too narrowly delineated, then relevant consequences would be missed. Should it be too broad, then outcomes would be blurred in too vast a population. Furthermore, in case of sampling, representativeness would be lessened. Finally, the choice of comparators (the competing strategies) is probably the main source of bias. If relevant comparators are left aside, the statistical part of the evaluation process may appear at first glance accurate, providing that it is properly carried out in mathematical terms. Nevertheless, it would be uninterpretable in terms of policy recommendations.

Parameter Uncertainty It relates to the estimation of the mean values of model inputs (e.g., unit costs, utility scores, mortality risk, recovery frequency, dropout rate, etc.). Parameter uncertainty first comes from the lack of justification of data sources. The quality of primary data, obtained from preexisting or homemade databases, must be openly assessed. Census or sample survey procedures should be fully described. The failure to meet those requirement may seriously impair the next steps of the evaluation process. The chapter dedicated to "Sampling and construction of variables" provides guidance in this respect. Equipped with a database of a given quality, a first way to deal with parameter uncertainty is to proceed to one-way deterministic sensitivity analysis. It examines how an outcome of interest (the *ICER* for instance) changes in response to variations in a single parameter, holding all other parameters constant. As an illustration, one may figure out an outcome associated linearly with a set of covariates, the one-way analysis consisting in calculating the partial derivative of the outcome with respect to the parameter of interest. One must be extremely cautious about the interpretation of such results. The underlying assumption of linearity undermines deterministic sensitivity analyses. Decision analytic modeling does not pre-specify any functional form relating model inputs and the ensuing outcomes. Models are usually complex enough to preclude reducibility to linear relations. For instance, an increase in effectiveness can come from fewer adverse events which breeds in turn smaller treatment costs while success in avoiding those negative health effects can be related to increased costs as more experienced staff are enrolled in the program. The ensuing reduced morbidity may in turn have longer term effects on patients' ability to bear subsequent care stages, increasing their utility, or allow them to be safely eligible to the next stage or line of treatment, etc. The main interest of deterministic sensitivity analysis is to check that the direction of change is consistent with common sense or prior belief as to that direction (e.g., a decrease in the price of an input should not increase the cost outcome). Since models are often complex, with many states and transitions organized in intricate patterns, deterministic sensitivity analysis can serve as debugging device, to check the internal coherence of the coding of the model. Parameter uncertainty is best dealt with through probabilistic sensitivity analysis and will be investigated in detail in the upcoming developments.

Structural Uncertainty It relates to the uncertainty around the constituting aspects of the model. The first problem that generates structural uncertainty is the possible omission of events by the evaluator when they frame the decision analytic model (in the case of a Markov model, there would be missing Markov states). Omission may be in relation to a lack of knowledge about the context or on the effects or consequences of the intervention. In terms of decision analysis under incomplete knowledge, this corresponds to the impossibility of an exhaustive classification of the states of nature associated with the intervention. For instance, in the case of a health program, the lack of knowledge of the natural history of the disease or the unfounded extrapolation of the effects of a drug beyond the time horizon of the associated clinical trial generate uncertainty that cannot be dealt with by using the standard statistical tools that address parameter uncertainty. The second issue associated with structural uncertainty is about the measurement of cost and above all of utility. The Chap. 6 has evidenced the difficulty to adequately assess individual preferences. The statistical handling of the uncertainty around utility parameters offers a partial answer but it gives it under the assumption that the utility measurement method is adequate, which assumption is surrounded by a statistically irreducible uncertainty.

Admittedly, the three types of uncertainty overlap to some extent. The standard classification described here nevertheless provides a useful framework and checklist to investigate the partial knowledge investigators usually face when they evaluate programs. Methodological uncertainty depends on the context of the analysis and will not be further studied. Structural uncertainty also very much depends on that context (e.g., utility formation and measurement is a fundamental question in health technology assessment that differs substantially from outcome measurement in education or social rehabilitation). The next developments will focus on the treatment of parameter uncertainty which all evaluation programs commonly face.

First, uncertainty in cost parameters is likely substantially to impact the project appraisal. This uncertainty can come from imprecision in measurement due to inadequate accounting structures, difficulty in using and comparing with cost data from other institutional context, absence of consensus on joint cost allocation methods, small sample size when gathering cost data, etc. An adequate distribution should rule out negative numbers since costs are counted positively, allow values greater than 1 as well as high values. Cost values often cluster around their mode, but particular situations are likely to generate outlying but nevertheless relevant observations. For instance, a given medical intervention may evidence costs that are homogenous amongst the vast majority of patients but nevertheless induce for a minority of them much longer stays in hospital requiring lengthy inpatient admission and care. Similarly, while most schoolchildren entail similar educational costs, some pupils with specific difficulties in their acquisition of knowledge or in their behavior may require special attention from the educational community. The counts of resource use (hospital days, hours of specialized teaching) are weighted by unit cost and sometimes, that count can be quite high. Cost data can thus be long-tailed (right skewed). Taking those constraints into account, both the Log-normal and Gamma distributions are suited to simulate cost parameters.

The Log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. In R-CRAN, the *rlnorm* command generates such random deviates:

$$rlnorm = rlnorm(obs, \mu, \sigma)$$

where *obs* stands for the number of observations randomly generated while μ is a location parameter (or log mean) and σ is a shape parameter (or log standard deviation). By definition, the expectation of the Log-normal distribution is $e^{\mu+\sigma^2/2}$ and the variance is $(e^{\sigma^2}-1)e^{\mu+\sigma^2/2}$. If a random variable X follows the Log-normal distribution, then ln$X$ has the normal distribution with mean μ and standard deviation σ. In other words, the location parameter is the mean of the data set after transformation by taking the logarithm, and the scale parameter is the standard deviation of the data set after transformation.

The method of moments can be used to estimate the parameters of the Log-normal distribution. The idea is to relate the population moments (i.e., equations of expectation and variance of the distribution) to the sample moments estimated from the sample. The equations are then solved for the shape parameters, using the sample moments in place of the (unknown) population moments. Assume for instance that we have some data about a cost parameter with sample mean $\bar{x}$ and standard error se. To fit a Log-normal distribution, we must have:

$$\bar{x} = e^{\mu+\sigma^2/2} \text{ and } se^2 = (e^{\sigma^2}-1)e^{\mu+\sigma^2/2}$$

We can invert these formulas to get μ and σ as functions of $\bar{x}$ and se:

$$\mu = \ln(\bar{x}) - \frac{1}{2}\ln\left((se/\bar{\mathrm{x}})^2+1\right)\text{and } \sigma = \sqrt{(se/\bar{x})^2+1}$$

For instance, if one were knowing that $\mu = 10$ and $se = 6$ we could easily use R-CRAN to fit the distribution in question. Note that the standard error se here denotes the standard deviation of the sampling distribution of the mean and not the sample standard deviation.

Figure 10.35 offers an example. First, using the above expressions, we define a unit cost $uc1$ with mean of 10 and standard error of 6. The first entry in the *rlnorm* command denotes the number of randomly generated observations (here 1,000,000), while the second and third entries stand for the μ and σ parameters, respectively. Figure 10.36a provides the related probability density function estimated with *plot*(*density*()). The bandwidth relates to the precision of the local estimations used to approximate the shape of the density curve. As can be seen, the distribution shifts to the right as the mean increases (Fig. 10.36b). Moreover, the lower the standard error, the less spread out the distribution (Fig. 10.36c).

The Gamma distribution is often preferred in health technology assessment. In R-CRAN, we have:

$$rgamma = rgamma(obs, \alpha, \beta)$$

where α is a shape parameter and β is a rate parameter. Both parameters are positive real numbers. Parameter α mainly determines the position of the density function. Higher values of α are for instance associated with a density function placed on the right of the x axis. Parameter β on the other hand has the effect of stretching or

```
> # Log-normal distribution
> xbar=10
> se=6
> uc1=rlnorm(1000000,meanlog=log(xbar)-0.5*log((se/xbar)^2+1),
+ sdlog=(log((se/xbar)^2+1))^0.5)
> plot((density(uc1)),main="23.1 Log-normal (mean=10,se=6)",
+ xlim=c(0,100))
> xbar=50
> se=6
> uc2=rlnorm(1000000,meanlog=log(xbar)-0.5*log((se/xbar)^2+1),
+ sdlog=(log((se/xbar)^2+1))^0.5)
> plot((density(uc2)),main="23.2 Log-normal (mean=50,se=6)",
+ xlim=c(0,100))
> xbar=50
> se=2
> uc3=rlnorm(1000000,meanlog=log(xbar)-0.5*log((se/xbar)^2+1),
+ sdlog=(log((se/xbar)^2+1))^0.5)
> plot((density(uc3)),main="23.3 Log-normal (mean=50,se=2)",
+ xlim=c(0,100))

> # Gamma distribution
> xbar=10
> se=6
> uc4=rgamma(1000000,shape=xbar^2/se^2,rate=xbar/se^2)
> plot((density(uc4)),main="23.4 Gamma (mean=10,se=6)",
+ xlim=c(0,100))
> xbar=50
> se=6
> uc5=rgamma(1000000,shape=xbar^2/se^2,rate=xbar/se^2)
> plot((density(uc5)),main="23.5 Gamma (mean=50,se=6)",
+ xlim=c(0,100))
> xbar=50
> se=2
> uc6=rgamma(1000000,shape=xbar^2/se^2,rate=xbar/se^2)
> plot((density(uc6)),main="23.6 Gamma (mean=50,se=2)",
+ xlim=c(0,100))
```

Fig. 10.35 Log-normal and gamma distributions in R-CRAN

compressing the range of the Gamma distribution. The higher β, the more spread out the distribution.

The expectation and variance of the gamma distribution are α/β and α/β^2, respectively. Thus, as previously, the shape and rate parameters can be calculated from a sample mean $\bar{x}$ and standard error se. Using the method of moments, we set:

$$\bar{x} = \frac{\alpha}{\beta} \text{ and } se^2 = \frac{\alpha}{\beta^2}$$

Solving for the values of α and β yields:

$$\alpha = \frac{\bar{x}^2}{se^2} \text{ and } \beta = \frac{\bar{x}}{se^2}$$

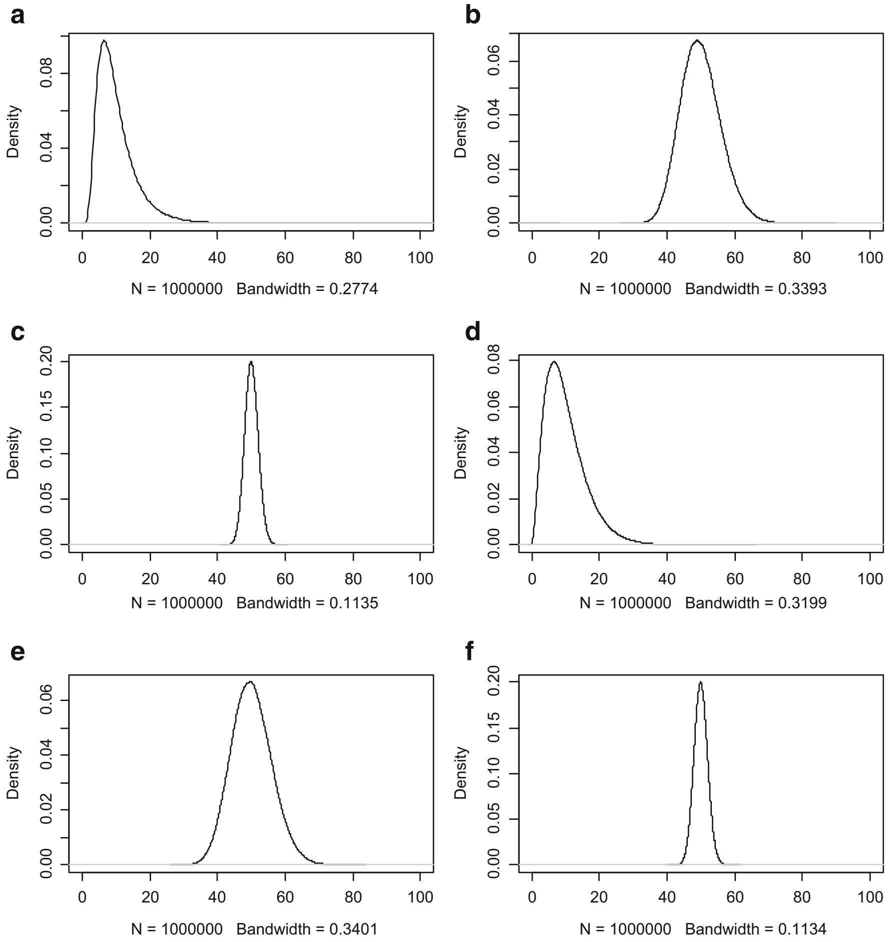


Fig. 10.36 Modeling cost parameters. (**a**) Log-normal (mean = 10, se = 6); (**b**) log-normal (mean = 50, se = 6); (**c**) log-normal (mean = 50, se = 2); (**d**) gamma (mean = 10, se = 6); (**e**) gamma (mean = 50, se = 6); and (**f**) gamma (mean = 50, se = 2)

The main difference between α and β comes from whether the mean is expressed in square in the numerator. An increase in the sample mean has thus a higher impact on α, and makes the distribution move to the right of the x axis.

Figures 10.35 and 10.36d, e, f simulate the gamma distribution. The first entry in the *rgamma* command denotes the number of randomly generated observations (here 1,000,000), while the second and third entries stand for the shape and rate parameters respectively. In Fig. 10.36d, the mean and standard error are set to 10 and 6 respectively. The density function moves to the right when the mean is set to a higher value (Fig. 10.36e). The lower the standard error, the less spread out the distribution (Fig. 10.36f). As can be observed, both the Log-normal and gamma distributions can be used effectively for describing positively skewed data.

Note that the Log-normal and gamma distributions are also frequently used to fit QALYs, especially when one needs to include the possibility of one or more health states with utility out of the [0,1] bounds. Specifically, utility parameter U may be negative if a health state is considered worse than death and thus allocated a value less than zero. The usual way to fit a distribution in that case is to focus on utility decrements $1-U$ so that values are positive only. The analysis moves from the utility scale to the disutility scale: distributions are censored at zero but they are unbounded above zero. The Log-normal and gamma distributions then become appropriate to model those weights.

Assigning a distribution on a probability parameter is challenging as one must account for the fact that measurements are constrained to lie between zero and one. Two distributions can be used which satisfy this property: the beta distribution and the Dirichlet distribution. While the beta distribution is well suited to model two-state transition probabilities, the Dirichlet distribution is more appropriate when faced with multiple-state environments. For instance, if one were using independent beta distributions for fitting the probabilities of several states of nature, one would face the risk of having a sum of parameters out of the [0, 1] bounds. The Dirichlet distribution overcomes this issue by generalizing the beta distribution.

In its simplest (non-generalized) form, the beta distribution is determined by two positive shape parameters, denoted α and β:

$$rbeta = rbeta(obs, \alpha, \beta)$$

Depending on those parameters, the beta distribution takes different shapes. When α and β have the same value, the distribution is symmetric. As their values increase, the distribution becomes more peaked. When α and β are different, the distribution is asymmetric.

To illustrate the beta distribution, assume that we deal with only two states of nature, say m_1 and m_2. Imagine that we have observed from a sample of size n that n_1 individuals are in state m_1 while n_2 are in state m_2. To fit the beta distribution based on this sample data, we simply need to set $\alpha = n_1$ and $\beta = n_2$. Another but equivalent way to specify α and β is to rely on the sample proportion p and sample error standard se. By definition, the beta distribution has an expectation of $\alpha/(\alpha+\beta)$ and a variance of $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$. Using the method of moments, we set:

$$p = \frac{\alpha}{\alpha+\beta} \text{ and } se^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Solving for α and β yields:

$$\alpha = p\left(\frac{p(1-p)}{se^2} - 1\right)\text{and } \beta = \alpha\frac{(1-p)}{p}$$

Assume for instance that we have 600 out of 1000 children who are occasional school leavers (state m_2) while the other 400 remain in state m_1 (they are regular attendants). The beta distribution representing the probability of school dropout can be defined as:

$$rbeta = rbeta(obs, 400, 600)$$

Figures 10.37 and 10.38a illustrate this case. If, on the other hand, only the sample proportion is known, e.g., p=0.4, and a confidence interval for a proportion has been estimated with $se = 0.05$, then the beta distribution is defined as:

$$rbeta = rbeta(obs, 38, 57)$$

Figure 10.37 describes the computation. As can be seen from Fig. 10.38b, the distribution is more spread out, yet the mean is still around 0.4.

The Dirichlet distribution is the multivariate generalization of the beta distribution to a larger set of states. It is particularly suitable for modeling a transition matrix:

$$rdirichlet = rdirichlet(obs, alpha)$$

where $alpha = (\alpha_1, \alpha_2, \ldots)$ is a vector or matrix of parameters. In R-CRAN, this function comes with the *mc*2*d* package which includes various distributions for Monte Carlo simulations.

Consider for instance the previous beta distribution $rbeta(obs,$400,600). In Fig. 10.38a, the command has been used to simulate 1,000,000 observations, e.g., 0.35, 0.41, 0.45, etc., which on average yields a proportion of being in state m_1 of 0.4. By construction, the probability of being in state m_2 is thus 1 minus those values, i.e. 0.65, 0.59, 0.55, etc. The Dirichlet distribution can save us some time by providing directly those proportions, as shown in Fig. 10.38c, d where both density functions are displayed (i.e. $\widehat{p} = 0.4$ and $1 - \widehat{p} = 0.6$ on average). To do so, in Fig. 10.37, we have specified *alpha* as a vector *c*(400,600).

More interestingly, the entry *alpha* in the *rdirichlet* command can take the form of a matrix. Coming back to example 4, let us assume that the transition matrix $\mathbf{P}_a$ of strategy a has been obtained empirically from samples of size 1000:

		***State* m_1**	***State* m_2**	***State* m_3**	***State* m_4**	***Total***
Sample 1 :	m_1	800	200	0	0	1000
Sample 2 :	m_2	400	500	100	0	1000
Sample 3 :	m_3	50	300	600	50	1000
Sample 4 :	m_4	0	0	0	1000	1000

```
> # Beta distribution
> p1=rbeta(1000000,400,600)
> plot((density(p1)),main="24.1 Beta (alpha=400,beta=600)",
+ xlim=c(0,1))
> mean(p1)
[1] 0.3999974

> mu=0.4
> se=0.05
> alpha=mu*(mu*(1-mu)/se^2-1)
> alpha
[1] 38
> beta=alpha*(1-mu)/mu
> beta
[1] 57
> p2=rbeta(1000000,alpha,beta)
> plot((density(p2)),main="24.2 Beta (p=0.4,se=0.05)",
+ xlim=c(0,1))
> mean(p1)
[1] 0.3999974

> # Dirichlet distribution / two-state transitions
> library(mc2d)
> alpha=c(400,600)
> p3=rdirichlet(1000,alpha)
> plot((density(p3[,1])),main="24.3 Dirichlet (400,600), p=0.4",
+ xlim=c(0,1))
> plot((density(p3[,2])),main="24.4 Dirichlet (400,600), p=0.6",
+ xlim=c(0,1))

> # Dirichlet distribution / multiple-state transitions
> alpha=matrix(c(800,400,50,0,200,500,300,0,0,100,600,0,0,0,50,1000),
+ nrow=4)
> alpha
     [,1] [,2] [,3] [,4]
[1,]  800  200    0    0
[2,]  400  500  100    0
[3,]   50  300  600   50
[4,]    0    0    0 1000
> rdirichlet(4,alpha)
           [,1]      [,2]      [,3]       [,4]
[1,] 0.78356459 0.2164354 0.0000000 0.00000000
[2,] 0.38298535 0.5165464 0.1004683 0.00000000
[3,] 0.04936275 0.3097588 0.5910538 0.04982458
[4,] 0.00000000 0.0000000 0.0000000 1.00000000
> rdirichlet(4,alpha)
           [,1]      [,2]      [,3]       [,4]
[1,] 0.84364772 0.1563523 0.0000000 0.00000000
[2,] 0.38188093 0.5114236 0.1066955 0.00000000
[3,] 0.05016427 0.2859346 0.6087972 0.05510398
[4,] 0.00000000 0.0000000 0.0000000 1.00000000
> rdirichlet(4,alpha)
           [,1]      [,2]       [,3]       [,4]
[1,] 0.80019536 0.1998046 0.00000000 0.00000000
[2,] 0.40538597 0.5112388 0.08337519 0.00000000
[3,] 0.04802104 0.3072435 0.59997891 0.04475652
[4,] 0.00000000 0.0000000 0.00000000 1.00000000
> #etc.
```

Fig. 10.37 Beta and Dirichlet distributions in R-CRAN

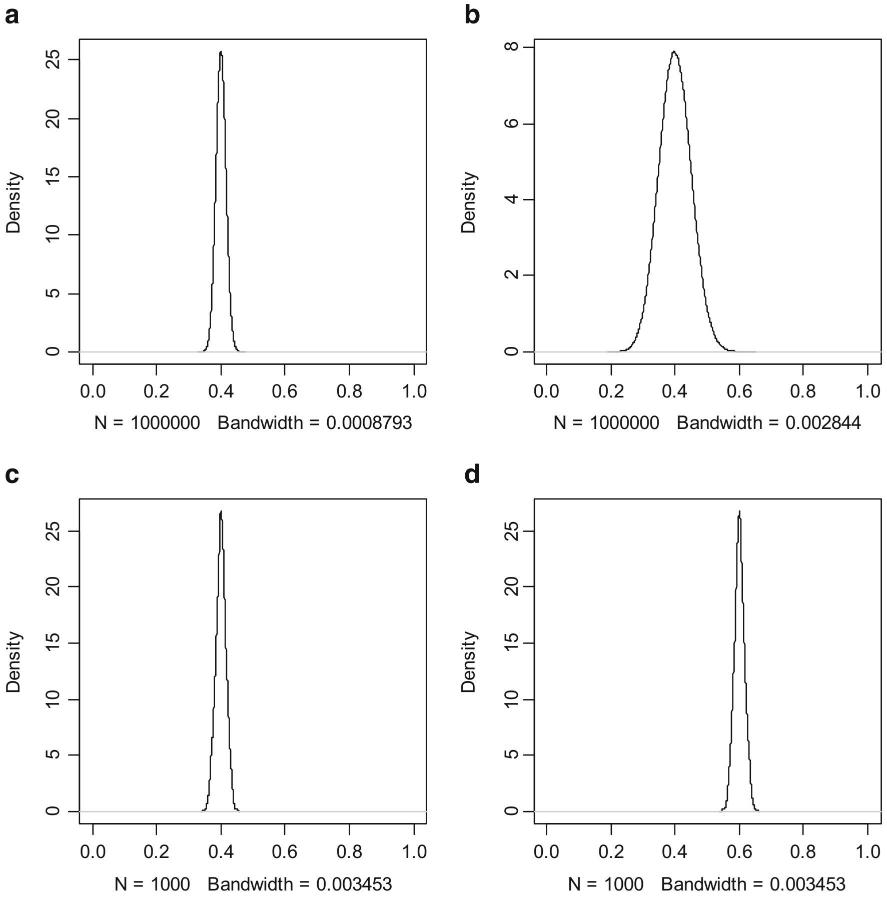


Fig. 10.38 Modeling transition probabilities. (**a**) Beta (alpha = 400, beta = 600); (**b**) beta (p = 0.4, se = 0.05); (**c**) Dirichlet (400, 600), p = 0.4; and (**d**) Dirichlet (400, 600), p = 0.6

For simplicity of exposition, the table is exactly the same as the Markov trace. However, in practice, it is unlikely that the sample sizes match the number of subjects in the Markov simulations. To fit the Dirichlet distribution, one simply needs to specify *alpha* using the values of the above matrix. In Fig. 10.37, we set:

$$alpha = matrix(c(800, 400, 50, 0, 200, 500, \ldots), nrow = 4)$$

Then, using $rdirichlet(4, alpha)$, we successively generate three random transition matrices. The approach can thereby be used in a Monte Carlo framework where each row of the transition matrix is assigned a random generator whose probabilities sum to 1.

Until now, parameters have been assumed to be independent. If there is a presumption that it is not the case, the Cholesky decomposition method can be used to assess uncertainty when two or more parameters are correlated. If we have access to the variance-covariance matrix, we can employ the Cholesky decomposition that provides correlated draws from a multivariate normal distribution. Chap. 9 offers an example of such a distribution.

A probabilistic sensitivity analysis (aka Monte Carlo simulations) assigns a probability distribution to all sensitive parameters. Figure 10.39 offers an illustration using example 4. The approach simulates $i = 1$ to $obs = 1000$ scenarios and examines the randomly generated parameters simultaneously in each loop. The parameters are thereby random in the sense that their value is subject to variations due to chance. For each strategy under examination, the analysis results in a range of cost and effectiveness measurements with their probabilities of occurrence.

More specifically, the Monte Carlo simulations of Fig. 10.39 start with uploading the *mc*2*d* package. Then the program sets the main parameters of the model. As previously, the focus is on $M = 4$ Markov states. The time horizon is $T = 10$ years. The initial allocation of subjects amongst the states is $N0 = c(1000, 0, 0, 0)$. The discount rate is set to $Disc.RATE = 0.03$. The number of strategies is $K = 4$. The second step consists in creating two matrices that will contain the randomly generated values of the cost and effectiveness measurements. The entry $obs = 1000$ defines the number of simulations. Matrices *SIM_E* and *SIM_C* are thus made of $K = 4$ columns and $obs = 1000$ rows.

The next step relates to the loop itself. A probability distribution is assigned to each of the parameters and the model is run a thousand of times to generate the probability distribution of the cost and effectiveness measures. Note that the number of loops can be set to a higher value, e.g., 10,000, in which case the software takes more time to generate the simulation outputs. For simplicity of exposition, the choice of the distributions in Fig. 10.39 is arbitrary. First, using the Dirichlet distribution, each strategy is assigned a transition matrix using the same approach as in Fig. 10.37 (i.e. we assume that each row of the transition matrix has been were obtained from a sample of size 1000). The unit costs are assigned the same values on average than those in Sects 10.5 and 10.6. For strategy a, the standard error is set to $\sqrt{0.5}$, while for strategy b, c and d, the standard error is set to $\sqrt{1.5}$, $\sqrt{2}$ and $\sqrt{3}$, respectively. Note that the *cea* function previously created in Sect. 10.6 is used to compute the effectiveness and cost measurements. The reader thus should be careful to run the program of Fig. 10.29 before implementing the program of Fig. 10.39. Last, the coding ends with two lines:

$$SIM\_E[i,] = c(sum(a\$discEFFECT), sum(b\$discEFFECT), \ldots)$$

and

$$SIM\_C[i,] = c(sum(a\$discCOST), sum(b\$discCOST), \ldots)$$

For each randomly generated data $i = 1 \ldots 1000$, the effectiveness and cost measurements are saved for subsequent analysis in the *SIM_E* and *SIM_C* matrices.

```
> library(mc2d)
> M=4
> T=10
> N0=c(1000,0,0,0)
> Disc.RATE=0.03
> K=4
> # Cost and effectiveness matrices
> obs=1000
> SIM_E=matrix(NA,nrow=obs,ncol=K)
> SIM_C=matrix(NA,nrow=obs,ncol=K)

> # Loop
> for(i in 1:obs){
+
+ #Strategy a
+ alpha_a=matrix(c(800,400,50,0,200,500,300,0,0,100,600,0,0,0,50,
+ 1000),nrow=M)
+ Pa=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ rdirichlet(4,alpha_a))
+ uc1=0
+ uc2=rgamma(1,shape=120^2/0.5,rate=120/0.5)
+ uc3=rgamma(1,shape=220^2/0.5,rate=220/0.5)
+ uc4=rgamma(1,shape=320^2/0.5,rate=320/0.5)
+ UCa=c(uc1,uc2,uc3,uc4)
+ a=cea(Pa,UCa,T,N0,Disc.RATE)
+
+ #Strategy b
+ alpha_b=matrix(c(800,600,200,0,200,300,350,0,0,100,400,0,0,0,50,
+ 1000),nrow=M)
+ Pb=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ rdirichlet(4,alpha_b))
+ uc1=0
+ uc2=rgamma(1,shape=164^2/1.5,rate=164/1.5)
+ uc3=rgamma(1,shape=320^2/1.5,rate=320/1.5)
+ uc4=rgamma(1,shape=620^2/1.5,rate=620/1.5)
+ UCb=c(uc1,uc2,uc3,uc4)
+ b=cea(Pb,UCb,T,N0,Disc.RATE)
+
+ #Strategy c
+ alpha_c=matrix(c(800,700,400,0,200,200,500,0,0,100,50,0,0,0,50,
+ 1000),nrow=M)
+ Pc=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ rdirichlet(4,alpha_c))
+ uc1=0
+ uc2=rgamma(1,shape=200^2/2,rate=200/2)
+ uc3=rgamma(1,shape=360^2/2,rate=360/2)
+ uc4=rgamma(1,shape=640^2/2,rate=640/2)
+ UCc=c(uc1,uc2,uc3,uc4)
+ c=cea(Pc,UCc,T,N0,Disc.RATE)
+
+ #Strategy d
+ alpha_d=matrix(c(800,700,200,0,200,250,400,0,0,50,350,0,0,0,50,
+ 1000),nrow=M)
+ Pd=new("markovchain",states=c("m1","m2","m3","m4"),transitionMatrix=
+ rdirichlet(4,alpha_d))
+ uc1=0
+ uc2=rgamma(1,shape=220^2/3,rate=220/3)
+ uc3=rgamma(1,shape=350^2/3,rate=350/3)
+ uc4=rgamma(1,shape=700^2/3,rate=700/3)
+ UCd=c(uc1,uc2,uc3,uc4)
+ d=cea(Pd,UCd,T,N0,Disc.RATE)
+
+ SIM_E[i,]=c(sum(a$discEFFECT),sum(b$discEFFECT),sum(c$discEFFECT),
+ sum(d$discEFFECT))
+ SIM_C[i,]=c(sum(a$discCOST),sum(b$discCOST),sum(c$discCOST),
+ sum(d$discCOST))
+ }
```

Fig. 10.39 Monte Carlo simulations of cost and effectiveness: example 4

```
> # Estimated distributions of effectiveness
> plot((density(SIM_E[,1])),main="25.1 Strategy a",xlim=c(7700,8500))
> plot((density(SIM_E[,2])),main="25.2 Strategy b",xlim=c(7700,8500))
> plot((density(SIM_E[,3])),main="25.3 Strategy c",xlim=c(7700,8500))
> plot((density(SIM_E[,4])),main="25.4 Strategy d",xlim=c(7700,8500))

> # Estimated distributions of cost
> plot((density(SIM_C[,1])),main="26.1 Strategy a",
+ xlim=c(300000,500000))
> plot((density(SIM_C[,2])),main="26.2 Strategy b",
+ xlim=c(300000,500000))
> plot((density(SIM_C[,3])),main="26.3 Strategy c",
+ xlim=c(300000,500000))
> plot((density(SIM_C[,4])),main="26.4 Strategy d",
+ xlim=c(300000,500000))
```

Fig. 10.40 Distributions of cost and effectiveness: example 4

Once the *SIM_E* and *SIM_C* matrices have been filled in with the randomly generated numbers, one can assess the relevance of the simulations with basic univariate analyses. For instance, Fig. 10.40 draws the estimated density probability functions of the cost and effectiveness measurements for each strategy (see Figs. 10.41 and 10.42). As previously, the *density* function is used to compute density estimates. Compared to the other strategies, it can be seen from Fig. 10.41 that strategy *a* has on average a lower effectiveness and a larger variance. Fig. 10.42 shows that costs and their variance increase as one moves from strategy *a* to strategy *b*, *c* and *d*. The next section aims to provide a more detailed comparison of those disparities.

10.8 Analyzing Simulation Outputs

In this section we review the three standard approaches to analyze Monte Carlo simulations in a cost effectiveness framework: the cost-effectiveness plane, the expected *INB* and the acceptability curves. All those approaches suppose a strategy against which all the others are compared. This reference can for instance be the current intervention or the one that is thought to be the less-cost effective. In what follows, pursuing with example 4, we will consider strategy *a* as this benchmark.

The cost-effectiveness plane method is described in Fig. 10.43. The idea is to compare each strategy with the reference by plotting the difference in effectiveness (ΔE) against the difference in cost (ΔC). For example, Figs. 10.44, 10.45 and 10.46 shows the previous 1000 Monte Carlo simulations in the $[\Delta E, \Delta C]$ plane using strategy *a* as the reference strategy. As already stated in Sect. 10.2, the slope of a line joining a point to the origin is the incremental cost-effectiveness ratio (*ICER*) which measures the incremental cost associated with one additional unit of effectiveness. The rectangle relates on the other hand to the 95% confidence intervals defining those distributions.

Figures 10.44, 10.45 and 10.46 have been created in R-CRAN using the matrices SIM_E and SIM_C of Fig. 10.39. Those matrices contains all the simulated values of cost and effectiveness from strategy *a* to strategy *d*. Figure 10.43 is coded as

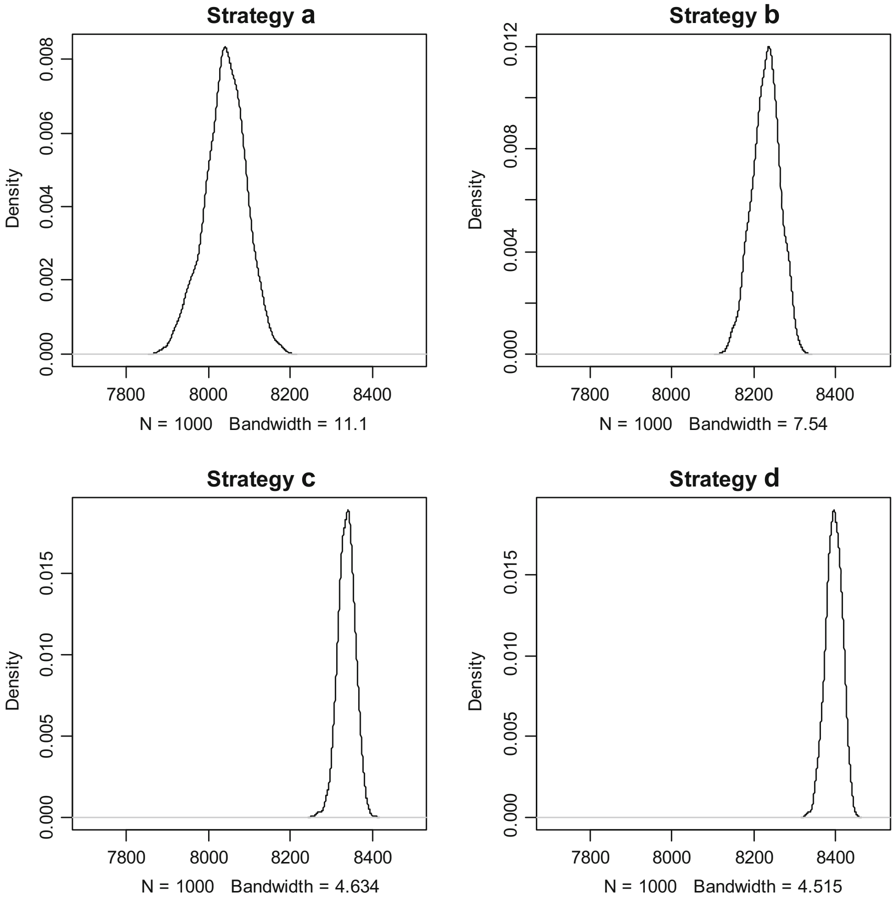


Fig. 10.41 Estimated distributions of effectiveness: example 4

follows. First, the differences in question are computed using strategy *a* (first colomn of *SIM_E* and *SIM_C*) as the reference. The *plot* command is then used to draw the graphs. For each strategy, the incremental cost (ΔC) is expressed as a function of the incremental effectiveness (ΔE). The *xlim* and *ylim* entry specify the range of the graph and is used to ensure comparability of graphs between strategies. The y- and x-axes are included using the $abline(v = 0)$ and $abline(h = 0)$ commands, respectively. The function *rect* draws a rectangle using the coordinates of two points. The method to compute those points is simple. The approach relies on the *quantile* function which is used here to compute the percentiles of the simulations vector. For instance, the function $quantile(delta.E[,2],.025)$ yields the lower bound of the 95% confidence interval of the 1000 observations related to strategy *b* and its differential effectiveness. In other words, 2.5% of the simulated observations lies below that value. Reciprocally, the function $quantile(delta.E$

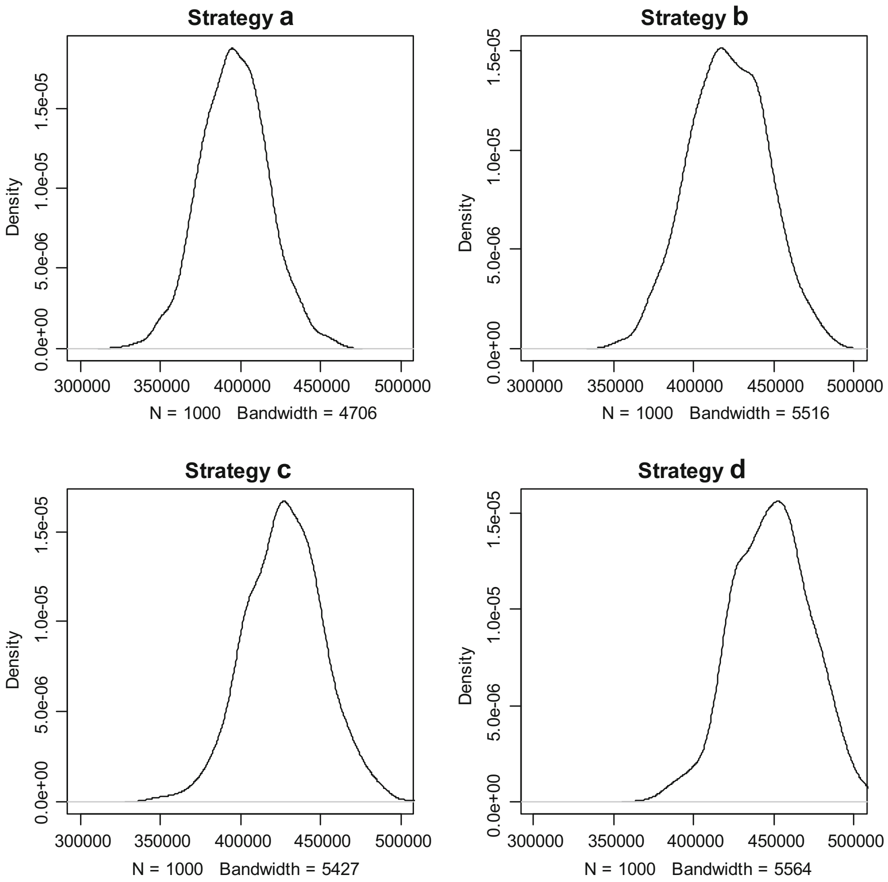


Fig. 10.42 Estimated distributions of cost: example 4

[, 2], .975) yield the upper bound of the 95% confidence interval with 97.5% of the observations lying above that limit. Using similar computations for costs, we are able to establish the coordinates of the rectangle whose height represents the 95% confidence interval on differential cost and width the 95% confidence interval on differential effectiveness.

The joint cost effectiveness density for strategies *b*, *c* and *d* with strategy *a* as the reference evidences a strong presumption that they all belong to the North East quadrant, with an increasing effectiveness from *b* to *c* and *d*. Cost structures are similar, with a small advantage to strategy *b* who offers more simulations in the South East quadrant where the strategy would be dominant. From Figs. 10.44 to 10.46 we cannot conclude that the costs of strategies are significantly different from those of strategy *a* (the rectangle crosses the horizontal axis). Yet, strategy *d* offers the greater number of simulations in the North East quadrant with the greater level of effectiveness.

```
> delta.E=SIM_E-SIM_E[,1]
> delta.C=SIM_C-SIM_C[,1]

> # Strategy b
> plot(delta.C[,2]~delta.E[,2],col=2,pch=21,
+ main="Strategy b",xlab="Differential effectiveness",
+ ylab="Differential cost",ylim=c(-100000,+200000),xlim=c(-200,+650))
> abline(v=0)
> abline(h=0)
> rect(quantile(delta.E[,2],.025),quantile(delta.C[,2],.025),
+ quantile(delta.E[,2],.975),quantile(delta.C[,2],.975),col="black",
+ density=10)

> # Strategy c
> plot(delta.C[,3]~delta.E[,3],col=3,pch=21,
+ main="Strategy c",xlab="Differential effectiveness",
+ ylab="Differential cost",ylim=c(-100000,+200000),xlim=c(-200,+650))
> abline(v=0)
> abline(h=0)
> rect(quantile(delta.E[,3],.025),quantile(delta.C[,3],.025),
+ quantile(delta.E[,3],.975),quantile(delta.C[,3],.975),col="black",
+ density=10)

> # Strategy d
> plot(delta.C[,4]~delta.E[,4],col=4,pch=21,
+ main="Strategy d",xlab="Differential effectiveness",
+ ylab="Differential cost",ylim=c(-100000,+200000),xlim=c(-200,+650))
> abline(v=0)
> abline(h=0)
> rect(quantile(delta.E[,4],.025),quantile(delta.C[,4],.025),
+ quantile(delta.E[,4],.975),quantile(delta.C[,4],.975),col="black",
+ density=10)
```

Fig. 10.43 Differential effectiveness and cost in R-CRAN: example 4

Although the previous approach offers an easy way of comparing strategies, it does suffer from a lack of generalization as the *ICER* is by construction to be compared with a single value of *WTP*. Moreover, the approach yields results that are difficult to interpret when the resulting confidence intervals are very large. To overcome those issues, one may rely instead on the expected incremental net benefit. The approach is depicted in Fig. 10.47 and consists in computing the usual *INB* indicator for a range of willingness to pay using the mean of the differential cost and effectiveness in the simulations. The codes are similar to those of Fig. 10.30 except that we now use averages. This yields Fig. 10.48 where strategy *b* is found to be never the most preferred strategy. Below a *WTP* of around \$100 per unit of effectiveness gained, the reference strategy *a* is preferred. For a *WTP* approximately between \$100 and \$440 strategy *c* is better. Then above \$440 per unit of effectiveness gained, strategy *d* should win through.

One inconvenient of the expected *INB* method is that it excludes risk considerations from the analysis since the variance in data is totally disregarded. In practice, the probability distributions used in the Monte Carlo framework are specified so that their expected value corresponds more or less to the parameters used in the deterministic analysis. Therefore, at best, the approach can be used as a robustness check that the simulations are able to reproduce the most likely scenario (see for instance the very similar Fig. 10.32).

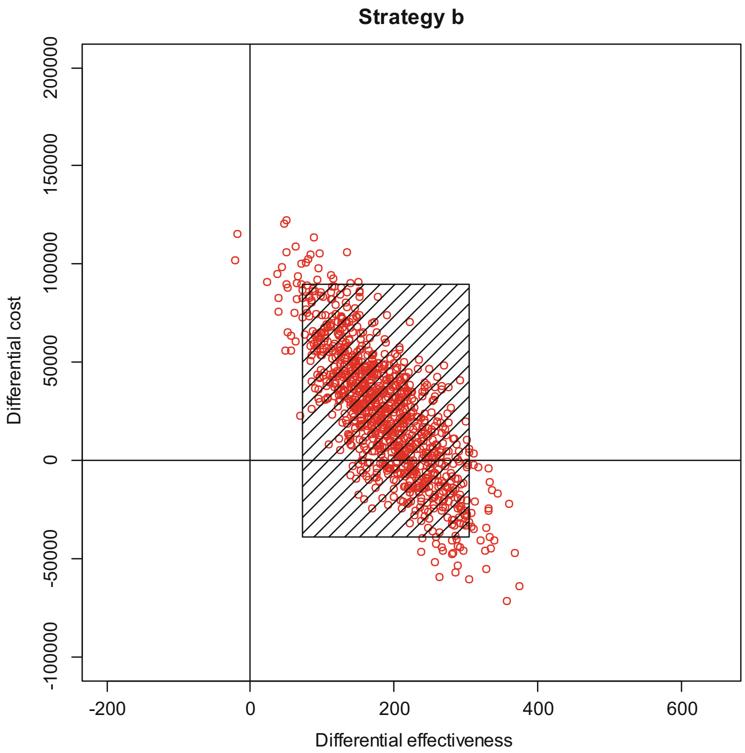


Fig. 10.44 Joint cost-effectiveness density for strategy *b*: example 4

A way to account for the variance in the data is to rely on cost-effectiveness acceptability curves. The principle is to compute the number of simulations in which each strategy is found to be most efficient and this, for each possible value of willingness to pay. Figure 10.49 provides the program in R-CRAN. The idea is again to analyze the simulated data (previously stored in *SIM_E* and *SIM_C*) and to run two loops: one for the change in the willingness to pay (loop 1), one for counting the number of times each strategy is optimal (loop 2).

Basically speaking, we are now dealing with four strategies, a range of *WTP* (from $1 to $700) and a set of 1000 simulations. Loop 1 starts with a *WTP* of $1 and computes an *INB* matrix for the whole set of simulations. Then loop 2 examines this *INB* matrix starting with the first set of simulations ($i = 1$). For each strategy, one must compute a dummy (*OPTa*, *OPTb*, *OPTc*, and *OPTd*) that specifies whether the strategy is optimal (value 1) or not (value 0) according to the *INB* criterion. The idea is to find the strategy that yields the highest *INB* given the willingness to pay. Loop 2 continues until all the data sets have been examined ($i = 2 \ldots obs$). Then the counter of Loop 1 is set to $2 and loop 2 restarts from $i = 1$. The process continues until *WTP*=700.

As can be understood, one needs to set a value 0 or 1 for each strategy, each *WTP* and each simulated data. To do so, Fig. 10.49 starts with the construction of four

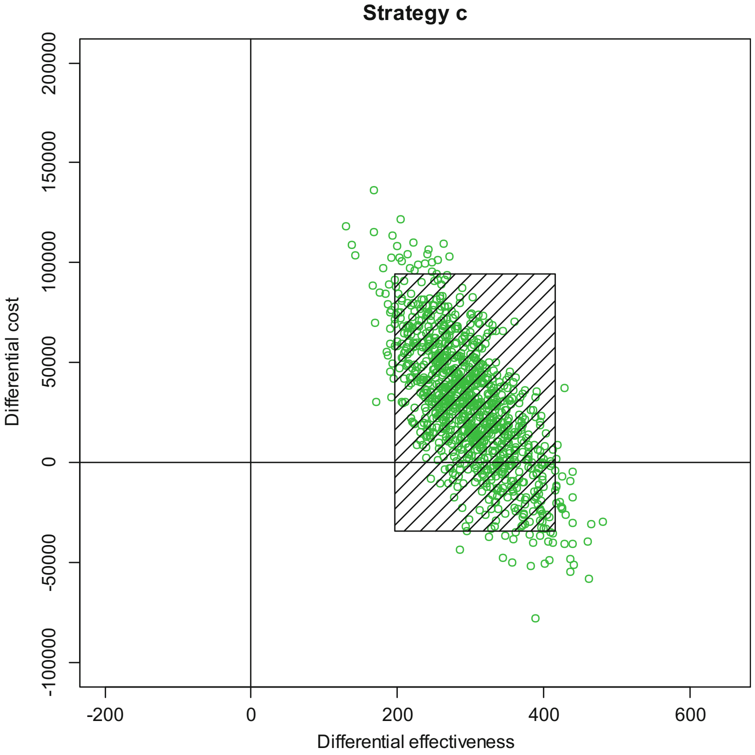


Fig. 10.45 Joint cost-effectiveness density for strategy *c*: example 4

matrices, one for each strategy, each made of as many rows as there are simulations, here, *obs* = 1000, and as many columns as they are values of *WTP*(here, from 1 to 700). For each *WTP*, loop 1 stores the incremental net benefits in a matrix called *MAT_INB*. Command *cbind* is used to combine the sequence of *INBs* by columns. Loop 2 then examines this matrix and specifies the dummies as follows:

$$OPTa[i, WTP] = (which.\max(MAT\_INB[i,]) == 1)^*1$$

$$OPTb[i, WTP] = (which.\max(MAT\_INB[i,]) == 2)^*1$$

$$OPTc[i, WTP] = (which.\max(MAT\_INB[i,]) == 3)^*1$$

$$OPTd[i, WTP] = (which.\max(MAT\_INB[i,]) == 4)^*1$$

Here *which*.max determines the location, i.e., index of the maximum in the numeric vector *MAT_INB*[*i*,]. The "==" command produces a type of value called a "logical", i.e. TRUE or FALSE, that specifies whether or not a condition (here, being equal to 1, 2, 3 or 4, referring to the four columns of matrix *MAT_INB*

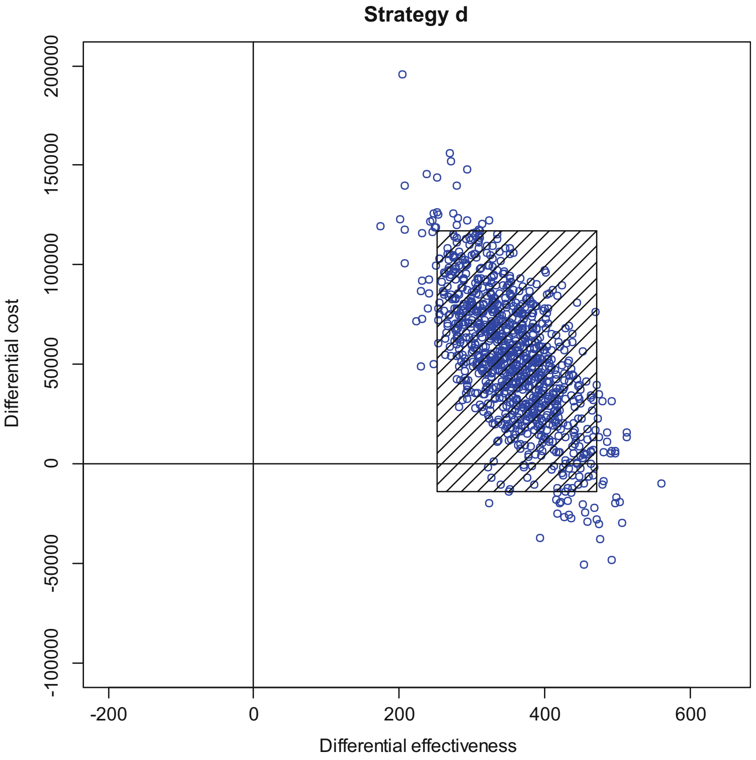


Fig. 10.46 Joint cost-effectiveness density for strategy *d*: example 4

```
> delta.E=SIM_E-SIM_E[,1]
> delta.C=SIM_C-SIM_C[,1]
>
> WTP=0:700
> INBb=WTP*mean(delta.E[,2])-mean(delta.C[,2])
> INBc=WTP*mean(delta.E[,3])-mean(delta.C[,3])
> INBd=WTP*mean(delta.E[,4])-mean(delta.C[,4])
> plot(INBb,type="l",col=2,ylim=c(-80000,160000),
+ xlab="WTP",ylab="Expected INB")
> points(INBc,type="l",col=3)
> points(INBd,type="l",col=4)
> abline(h=0)
> legend("top",c("Strategy b","Strategy c","Strategy d"),
+ lty=c(1,1,1), col=2:4)
```

Fig. 10.47 Expected incremental net benefit with R-CRAN: example 4

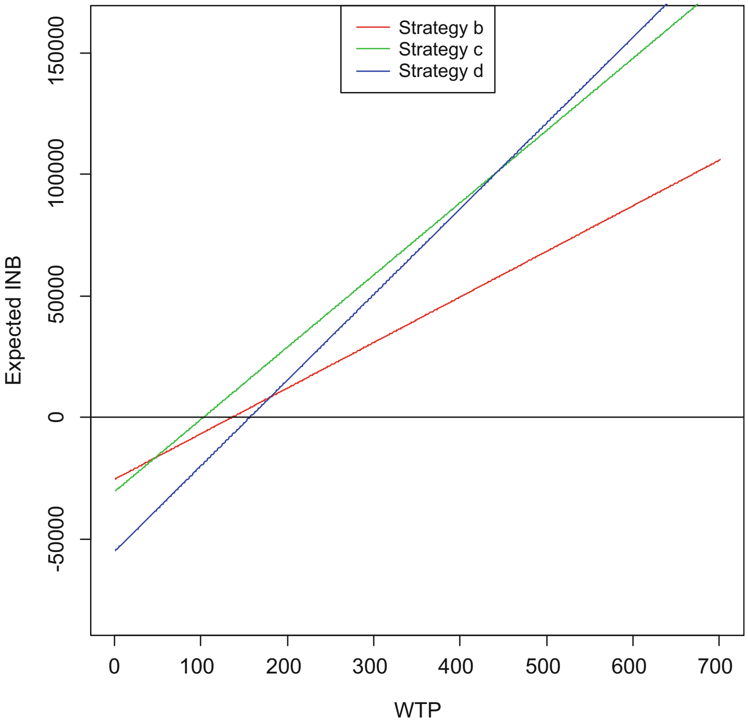


Fig. 10.48 Expected incremental net benefit: example 4

respectively associated with strategies a, b, c and d) is met. The " $*$ 1" command is used to transform the logical value TRUE or FALSE into a numerical value 1 or 0. For instance, if the index of the maximum is 1, it means that the *INB* is greater for the first strategy (namely strategy a) and a value of 1 is assigned to *OPTa* while zeros are assigned to other strategies (*OPTb*, *OPTc* and *OPTd*).

Once the loops are completed, one simply needs to compute for each *WTP* the number of times each strategy has been optimal. The command *colMeans* is used for this purpose. To illustrate, imagine for instance that *OPTa* has the following shape:

$$OPTa = \begin{bmatrix} WTP = \$1 & WTP = \$2 & WTP = \$3 & WTP = \$4 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here, the counter of Loop 1 goes from \$1 to \$4 while the counter of loop 2 goes from 1 to 6, i.e. we consider four values of *WTP* and 6 simulated data sets. For

```
> obs=1000
> MaxWTP=700
> OPTa=matrix(NA,nrow=obs,ncol=MaxWTP)
> OPTb=matrix(NA,nrow=obs,ncol=MaxWTP)
> OPTc=matrix(NA,nrow=obs,ncol=MaxWTP)
> OPTd=matrix(NA,nrow=obs,ncol=MaxWTP)

> delta.E=SIM_E-SIM_E[,1]
> delta.C=SIM_C-SIM_C[,1]

# Loop 1 (over ncol)
> for(WTP in 1:MaxWTP){
+ INBb=WTP*delta.E[,2]-delta.C[,2]
+ INBc=WTP*delta.E[,3]-delta.C[,3]
+ INBd=WTP*delta.E[,4]-delta.C[,4]
+ MAT_INB=cbind(0,INBb,INBc,INBd)

# Loop 2 (over obs)
+ for(i in 1:obs){
+ OPTa[i,WTP]=(which.max(MAT_INB[i,])==1)*1
+ OPTb[i,WTP]=(which.max(MAT_INB[i,])==2)*1
+ OPTc[i,WTP]=(which.max(MAT_INB[i,])==3)*1
+ OPTd[i,WTP]=(which.max(MAT_INB[i,])==4)*1
+ }
+ }

> WTP=1:MaxWTP
> plot(colMeans(OPTa)~WTP, type="l",col="1",ylim=c(0,1),
+ ylab="Proportion of simulation when strategy is optimal")
> points(colMeans(OPTb)~WTP, type="l",col="2")
> points(colMeans(OPTc)~WTP, type="l",col="3")
> points(colMeans(OPTd)~WTP, type="l",col="4")
> legend("top",c("Strategy a","Strategy b","Strategy c",
+ "Strategy d"),lty=c(1,1,1), col=1:4)
```

Fig. 10.49 Cost-effectiveness acceptability curves in R-CRAN: example 4

$WTP = \$1$, strategy *a* is found to be optimal four times. The mean is computed as 1 + 0 + 1 + 1 + 1 + 0/6 = 66.6%. In other words, for a willingness to pay of $1, strategy *a* is found to be optimal in two third of the simulations. Similarly, we can compute the mean for the next columns: 50% for $WTP = \$2$ and $3, 16.66% for $WTP = \$4$. Of course, in practice, the number of simulated data as well as the range of willingness to pay is much larger. By construction, if $OPTa[i = 1, WTP = 1] = 1$ then $OPTb[i = 1, WTP = 1]$ is equal to zero and so are *OPTc* and *OPTd*. This ensures that the vertical sum of acceptability curves is always equal to 100%.

Figure 10.50 provides the resulting graph. The results are in accordance with what has been previously observed. Below a *WTP* of around $100 per unit of effectiveness gained, the reference strategy *a* is preferred. For a *WTP* approximately between $100 and $440 strategy *c* is better. Then above $440 per unit of effectiveness gained, strategy *d* wins. The graph also has the advantage to provide information on the risk associated with each strategy. For a given *WTP*, the higher

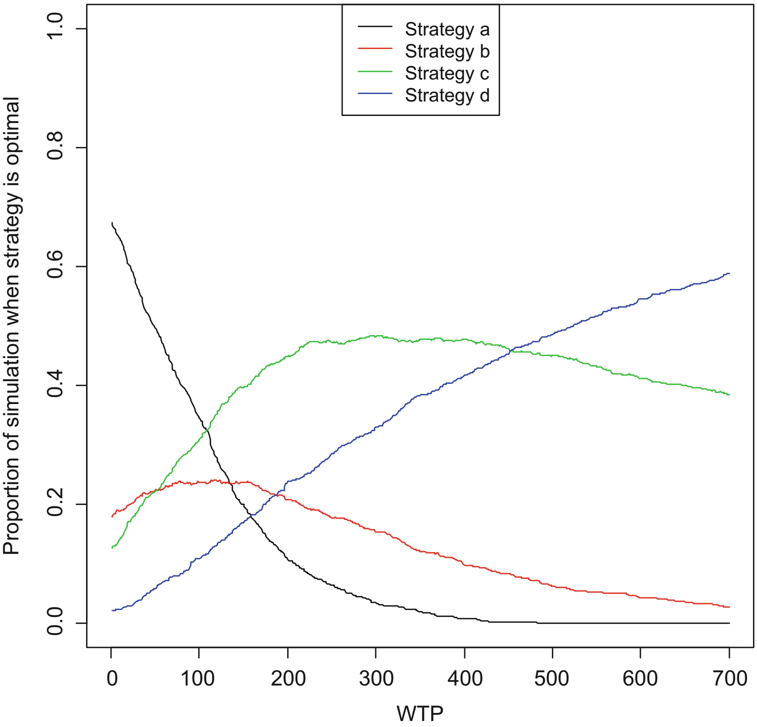


Fig. 10.50 Acceptability curves: example 4

the curve, the higher the confidence in the results. For instance, for small *WTP*, the black curve (strategy *a*) is above the other strategies, and is cost-effective with a probability of 40–60%. On the other hand, for large values of *WTP*, the blue curve (strategy *d*) is not so far from the green curve (strategy *c*) and strategy *d* progressively takes the advantage over all the other strategies. That advantage is confirmed for even greater values of WTP ($MaxWTP = 1500$) where strategy *d* evidences a strong degree of confidence that reaches 80% of simulations (Fig. 10.51).

Bibliographical Guideline

Cost effectiveness analysis has been developed initially as a decision tool for constrained optimization in operations research and management science (Briggs et al. 2006). It has then appeared to be particularly suited to public policies involving the maximization of a societal objective under a budget constraint. That constrained optimization would encompass two or more alternative options each with their own costs and consequences. In the 1990s, cost effectiveness analysis has developed in the field of education policies (Levin and McEwan 2000), in transport and environment (with Shoup 1973 as a precursor; a survey is

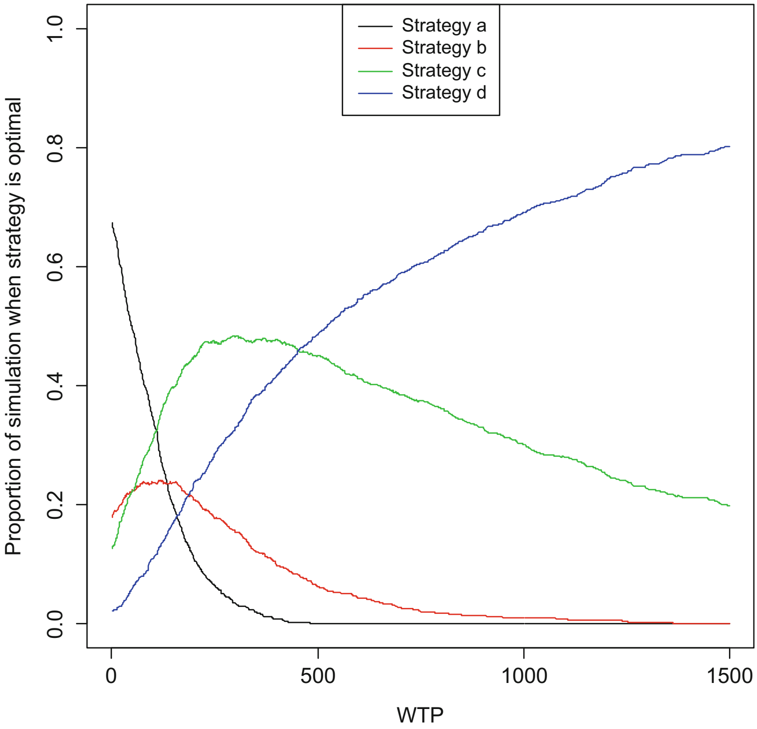


Fig. 10.51 Acceptability curves with higher *WTP*: example 4

provided by Kok et al. 2011) but mostly in public health were it has been extensively used to assess mass screening programs as well as access to social security reimbursement for innovative drugs or treatments (Gold et al. 1996; Drummond and McGuire 2001; Drummond et al. 2015). Furthermore, cost effectiveness analysis has become the official evaluation tool for several Health Technology Assessment national agencies including the United Kingdom, France, Australia, The Netherlands, Canada, etc. (Heintz et al. 2016 provide a systematic assessment of national practices in Europe).

Cost effectiveness analysis rests on cost effectiveness indicators that provide a synthetic measure of complex decision problems. The use of the differential effectiveness and differential cost mapping has been promoted by Black (1990). The incremental cost effectiveness method arises in the mid-1970s and has been systematized by Drummond et al. (1987) and Johannesson and Weinstein (1993). The decisional and statistical problems raised by the use of *ICER* have led to the introduction of the *INB* in the late 1990s (Stinnett and Mullahy, 1998). The efficiency frontier is a standard in management science; Laska et al. (2002) provide a systematic review of its properties. Willan (2011) reviews and illustrates methods for determining sample size requirements for cost effectiveness analysis in clinical trials.

Forerunners in the use of Markov models in decision analytic modeling are Sonnenberg and Beck (1993) and Briggs and Sculpher (1998). Alternatives to Markov processes are for instance decision trees or system dynamic models (Brennan et al. 2006 give a survey of the main DAM modeling tools). Markov models have been first used in this chapter in a deterministic setting. It is useful in order to get a broad picture of the policy question at stake. It also provides data for subsequent multicriteria decision analysis, especially with respect to the efficiency frontier. However, if a decision is to be taken on the basis of a cost effectiveness analysis, then the extension to a probabilistic setting is required. Uncertainty is systematically investigated by Barton et al. (2008), Claxton (2008), Briggs et al. (2012), Ghabri et al. (2016). A recent step by step approach to handling parameter uncertainty is provided by Edlin et al. (2015). Model biases in health technology assessment are explored by Raimond et al. (2014).

References

Barton, G., Briggs, A., & Fenwick, E. (2008). Optimal cost-effectiveness decisions: The role of the cost-effectiveness acceptability curve (CEAC), the cost-effectiveness acceptability frontier (CEAF), and the expected value of perfection information (EVPI). *Value in Health, 11*, 886–897.

Black, W. (1990). The CE plane: A graphic representation of cost-effectiveness. *Medical Decision Making, 10*, 212–214.

Brennan, A., Chick, S., & Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies. *Health Economics, 15*, 1295–1310.

Briggs, A., & Sculpher, M. (1998). An introduction to Markov modelling for economic evaluation. *PharmacoEconomics, 13*, 397–409.

Briggs, A., Claxton, K., & Sculpher, M. (2006). *Decision modelling for health economic evaluation*. Oxford: Oxford University Press.

Briggs, A., Weinstein, M., Fenwick, E., Karnon, J., Sculpher, M., & Paltiel, A. (2012). Model parameter estimation and uncertainty: A report of the ISPOR-SMDM modelling good research practices task force-6. *Value in Health, 15*, 835–842.

Claxton, K. (2008). Exploring uncertainty in cost-effectiveness analysis. *PharmacoEconomics, 26*, 781–798.

Drummond, M., & McGuire, A. (Eds.). (2001). *Economic evaluation in health care programs: Merging theory with practice*. Oxford: Oxford University Press.

Drummond, M., Stoddart, G., & Torrance, G. (1987). *Methods for the economic evaluation of health care programs*. Oxford: Oxford University Press.

Drummond, M., Sculpher, M., Claxton, K., Stoddart, G., & Torrance, G. (2015). *Methods for the economic evaluation of health care programs*. Oxford: Oxford University Press.

Edlin, R., McCabe, C., Hulme, C., Hall, P., & Wright, J. (2015). *Cost effectiveness modelling for health technology assessment*. Heidelberg: Springer.

Ghabri, S., Hamers, F., & Josselin, J.-M. (2016). Exploring uncertainty in economic evaluations of new drugs and medical devices: Lessons from the first review of manufacturers' submissions to the French National Authority for Health. *PharmacoEconomics, 34*, 617–624.

Gold, M., Siegel, J., Russell, L., & Weinstein, M. (Eds.). (1996). *Cost-effectiveness in health and medicine*. Oxford: Oxford University Press.

Heintz, E., Gerber-Grote, A., Ghabri, S., Hamers, F., Prevolnik Rupel, V., Slabe-Erker, R., et al. (2016). Is there a European view on health economic evaluation? Results from a synopsis of methodological guidelines used in the EUnetHTA partner countries. *PharmacoEconomics, 34*, 59–76.

Johannesson, M., & Weinstein, M. (1993). On the decision rules of cost-effectiveness analysis. *Journal of Health Economics, 12*, 459–467.

Kok, R., Annema, J., & van Wee, B. (2011). Cost-effectiveness of greenhouse gas mitigation in transport: A review of methodological approaches and their impact. *Energy Policy, 39*, 7776–7793.

Laska, E., Meisner, M., Siegel, C., & Wanderling, J. (2002). Statistical determination of cost-effectiveness frontier based on net health benefits. *Health Economics, 11*, 249–264.

Levin, H., & McEwan, P. (2000). *Cost-effectiveness analysis: Methods and applications*. New York: Sage.

Raimond, V., Josselin, J.-M., & Rochaix, L. (2014). HTA agencies facing model biases: The case of type-2 diabetes. *PharmacoEconomics, 32*, 825–839.

Shoup, D. (1973). Cost effectiveness of urban traffic law enforcement. *Journal of Transport Economics and Policy, 7*, 32–57.

Sonnenberg, F., & Beck, J. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making, 13*, 322–339.

Stinnett, A., & Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making, 18*, S68–S80.

Willan, A. (2011). Sample size determination for cost effectiveness trials. *PharmacoEconomics, 29*, 933–949.

Multi-criteria Decision Analysis 11

11.1 Key Concepts and Steps

Multiple criteria decision analysis (MCDA), also called multi-criteria analysis, is concerned with the analysis of multiple attribute environments and is devoted to the development of decision support tools to address complex decisions, especially where other methods fail to consider more than one outcome of interest. The approach is based on operational research and uses advanced analytical methods to evaluate a finite number of alternatives. These alternatives may be very broad. For example, in the context of public policy-making, MCDA can be used to compare the performance of different units (e.g., countries, municipalities, patients, students) based on their individual characteristics (wealth, education, health) or to select a particular policy option using a full range of social, environmental, technical, and financial indicators. A set of criteria is first established; then weights are assigned to reflect their relative importance. The analysis finally provides an ordered set of alternatives based on their overall performance.

In its simplest form (compensatory analysis), the idea behind MCDA is to simplify the decision-making process through the construction of a composite indicator, that is, a measurement of performance based on the aggregation of the different dimensions under examination. The approach is particularly useful for the comparison and ranking of countries or cities. The Human Development Index (HDI) offers an example. It is a measure of well-being which relates to three dimensions of human development: health, education, and income per capita:

$$HDI = \left(I_{\text{health}} \times I_{\text{education}} \times I_{\text{Income}}\right)^{\frac{1}{3}}$$

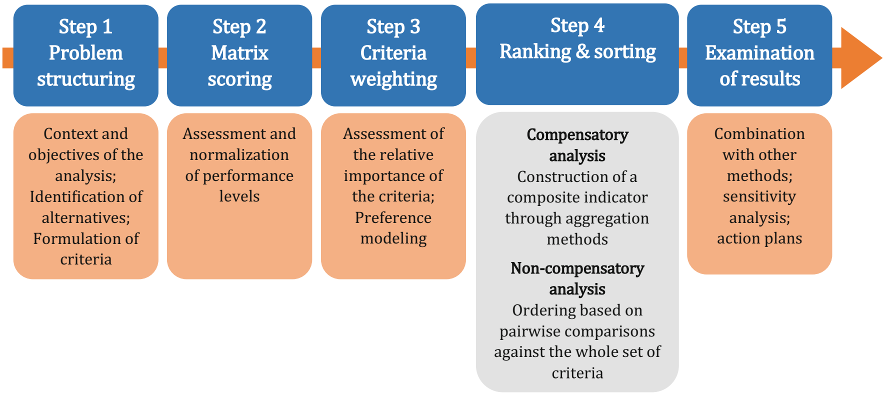
This index is an alternative to the traditional GDP per capita and can be used to follow improvement in human development over time or to compare development levels across countries. The method is said to be compensatory because a marginal decrease in the value of one dimension (e.g., income) can be compensated by a marginal increase in the value of another dimension (e.g., education). Composite

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_11

indicators are very attractive as they offer a simple way of ranking alternatives. The approach can also be used for comparing a range of public projects or action plans which are described in terms of their financial, social and environmental impacts. Policy options are then ordered from best to worst by means of the composite indicator.

In its most complex form (non-compensatory analysis), MCDA does not aim to aggregate the different criteria together but, instead, examines each dimension individually. In that context, tradeoffs among variables are of less importance. A bad performance may not be compensated for by a better score somewhere else. This type of analysis is generally dedicated to "sorting problems", where the aim is to assign alternatives to defined categories or to reduce the number of alternatives to be considered. For example, one of those non-compensatory methods, known as the outranking approach, relies on pairwise comparisons across the whole set of available criteria. The idea is to compute the number of times each alternative performs better than the others. Alternatives that outrank the others are considered the best, while those that are outranked are disregarded. The method is said to be non-compensatory because the final judgment is based on how many times each alternative differs positively from the others and not on the magnitude of that difference. In other words, an alternative i that performs slightly better on most criteria will rank higher than an alternative j that does much better for a smaller set of attributes.

The different steps of MCDA are illustrated in Fig. 11.1. Basically speaking, the approach starts with the identification of the problem (step 1): what are the objectives of the study or the questions to be answered? What is the set of variables to be analyzed? What are the alternatives under evaluation? This step is not to be neglected as it determines the way data is collected and the form each individual indicator will take. The approach then proceeds with gathering information about


Step 1 Problem structuring

Step 2 Matrix scoring

Step 3 Criteria weighting

Step 4 Ranking & sorting

Step 5 Examination of results

Context and objectives of the analysis; Identification of alternatives; Formulation of criteria

Assessment and normalization of performance levels

Assessment of the relative importance of the criteria; Preference modeling

Compensatory analysis
Construction of a composite indicator through aggregation methods

Non-compensatory analysis
Ordering based on pairwise comparisons against the whole set of criteria

Combination with other methods; sensitivity analysis; action plans

Fig. 11.1 Multi-criteria decision analysis

the performance of each alternative against the whole set of criteria (step 2). Data generally takes the form of a performance matrix:

	Criteria 1	**Criteria** 2	...	**Criteria** ***K***
Alternative 1	100	17,000	...	0.01
Alternative 2	160	20,000	...	0.03
⋮	⋮	⋮	⋮	⋮
Alternative ***n***	7	18,000	...	0.02

Each row of the matrix describes an evaluated unit/option and each column describes their performance against the criteria. Those criteria define the set of individual indicators under scrutiny. To make those indicators comparable, values in the performance matrix are generally normalized to be from 0 to 1, thereby constituting what is termed a score matrix.

Step 3 adds another element in the evaluation process. Numerical weights are assigned to criteria to better reflect their relative importance. In many cases, the elicitation process itself proves to be very helpful. The stakeholders can express and shape preferences in terms suited to the context. Weights and scores for each alternative are then combined to arrive at a ranking or sorting of alternatives (Step 4). The approach can be compensatory or non-compensatory. Should a compensatory analysis be implemented, the approach would rely on aggregation methods to build a composite indicator; for instance by multiplying or summing the individual indicators. Should a non-compensatory analysis be implemented, the approach would examine each dimension individually. Last, policy recommendations can be made on the basis of the MCDA model (step 5). A sensitivity analysis that examines the weights and scores is used to explore how changes in assumptions influence the results. Those successive steps are further detailed in the remaining of the chapter.

MCDA is very flexible as scores can be quantifiable in non-monetary terms (e.g., social or health effects) and be expressed in ordinal and numerical terms. It has applications in many different fields (human development, health, education, environment, etc.) and allows for the possibility of assessing performance over time. Often, it is considered as an efficient communication tool. Composite indicators are for instance very popular among government agencies as they can be used not only to assess performance, but also to capture the attention of the many actors involved in the decision-making process, to stimulate public policy debate, and to guide policy action. Because information is easier to interpret, stakeholders are offered a simple but general picture of the problem.

Despite its many qualities, MCDA is not the magic method that solves all the issues faced by the evaluator. The approach is often said to lack theoretical foundations and the analysis depends on so many factors, starting from step 1 (problem structuring) to step 5 (examination of results), that the final judgment can be far from relevant if those steps are not carefully addressed. While considered as an efficient decision aid tool, it may also offer a picture of the problem that is by far too simplistic. On top of those considerations, the weighting framework is generally based on human judgment and is thereby subjective (the aim of MCDA is to model the preferences of the decision-maker). Due to strategic choices or political dispute, the analysis can thereby be manipulated to support a particular outcome or to exclude another one. For those reasons, MCDA would only serve as a starting point for discussion, not to take the decision per se. A sound sensitivity analysis is also crucial to improve the quality of the analysis.

The outline of the chapter is as follows. Section 11.2 is about problem structuring, i.e. how to establish the decision context and how to generate the set of criteria for MCDA purpose. Sections 11.3 and 11.4 offer several methods for establishing relative preference scores and assigning weights to criteria. Section 11.5 explains how to combine those weights and scores to derive a composite indicator. Section 11.6 extends those methods to non-compensatory analysis. Last, Sect. 11.7 is about sensitivity analysis and how to interpret MCDA results.

11.2 Problem Structuring

The first step in a MCDA process is to define the problem faced by the policy-makers, the objective of the study and the persons who should be involved in the MCDA process. This task can be cognitively challenging since MCDA methods are generally used in the case of problems that are too complex to handle with traditional methods, involving multiple objectives and many conflicting views. Using reference sources such as budget requests, performance measurements and audit results may also render the analysis costly in terms of time and money spent. In that context, value trees (criteria hierarchy tree) are often seen as a useful tool to better understand the context of the decision.

Formally, a value tree can be divided in a set of three items: the goal (or main objective) of the study, a set of sub-objectives, and specific criteria or individual indicators. Figure 11.2 illustrates the approach using the Human Development Index. As can be seen, the HDI is not a comprehensive measure of well-being (for instance, inequalities are disregarded). It focuses on basic dimensions of human development, health, education, and income. Those dimensions are themselves measured in relation to specific individual indicators: life expectancy at birth, expected years of schooling (i.e. number of years of schooling that a child of school entrance age can expect to receive), mean years of schooling (i.e. average number of years of education received by people aged 25 and older) and gross national income per capita. The fundamental goal of building such a value tree is to help the evaluator organize all the information relevant to the analysis into a structure that can be easily apprehended.

A key issue in MCDA is to delineate the set of sub-objectives and their related measurement. Those dimensions can be very broad. By way of illustration, when evaluating a transport investment, MCDA can clarify the set of costs and benefits. Those can be expressed in terms of accessibility, safety, and environment. Other dimensions can also be included such as economic impact, future potential in terms

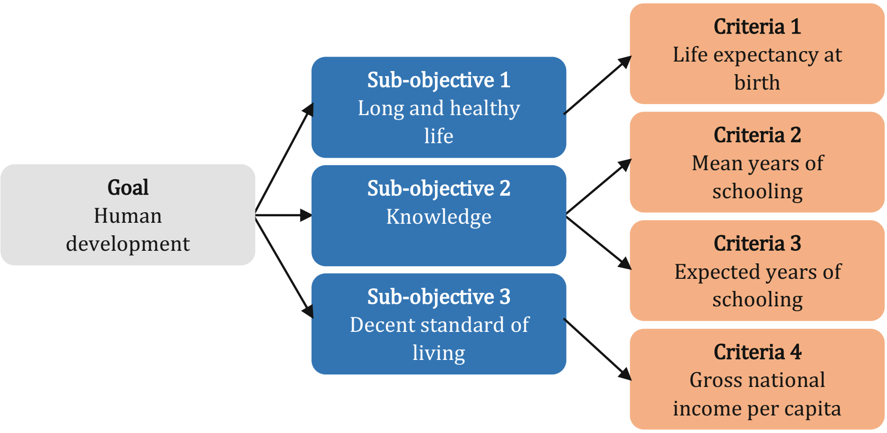


Fig. 11.2 Value tree: example of the Human Development Index

of supply, how the program integrates within the current environment, and deliverability. Those dimensions (or sub-objectives) may themselves be subdivided in specific indicators: improvement in landscape, noise, air quality, overall safety, etc. The literature distinguishes a set of requirements that are primordial in this respect. They can be summarized as follows:

1. **Completeness:** have all relevant criteria been considered? The selected criteria should account for all the important characteristics of the evaluated alternatives. Satisfying this condition is crucial as one does not want to miss an important item.
2. **Operationality:** can each alternative be assessed against each criterion? Data must be available and collected according to a well-defined scale of measurement.
3. **Non-redundancy:** are there unnecessary criteria? The MCDA process should avoid overlapping measures. Similar criteria should be eliminated or combined into a single indicator in order to facilitate the process of calculating criteria weights.
4. **Simplicity:** is the number of criteria excessive? The purpose of MCDA is to facilitate the evaluation process. The number of criteria should be kept as low as possible. A criterion can for instance be deleted if all the alternatives are likely to achieve a similar score when assessed against it.
5. **Independence:** are the criteria independent from policy choices? For example, high public spending (e.g., in education or health) cannot itself be a criterion. Public spending is a mean of achieving a goal, not the goal per se, and can only be considered as a cost.

When selecting the set of criteria a balance must be found between those five conditions. Note also that generating criteria is context dependent and will reflect the views of the different stakeholders in play: e.g., central and local authorities, interest groups, regulatory bodies, residents, scientific community, etc.

Several methods can be used to elicit the shape of the value tree. For instance, the set of sub-objectives can be identified using existing theories and policy statements, information from interest groups and government agencies, by observing the environment in which the study is conducted, from discussions with decision-makers, through a survey distributed to stakeholders, or via a panel of experts. In most cases, the organization of a focus group where relevant actors express their views and knowledge can be helpful. By definition, a focus group is a small group led through an open discussion by a moderator. In a MCDA context, the approach is used to learn about participants' opinions, e.g., issues and criteria they think are relevant, and to test assumptions. The group must be large enough (greater than 5) to be sufficiently representative but smaller enough (lower than 12) so that each participant has time to express their view. An agenda must be defined *ex ante*, with a specific timeline (from one to two hours), a set of open-ended questions (less than 10), and repeated sessions. A summary report is generally offered to the focus group after each session. The focus group represents also an opportunity to discuss the set of alternatives to be evaluated. The different views are then assembled to produce a final report and a graphical representation in the form of a value tree.

11.3 Assessing Performance Levels with Scoring

Once the set of criteria has been established, the next step is to assess how the alternatives perform with regard to each individual indicator. This process, also known as scoring, can take different forms. For instance, different groups (e.g., decision-makers, citizens, experts or stakeholders) can allocate the alternatives in question with a score in the [0,1] interval (direct rating). This is by far the easiest approach. Scores can also be obtained from a specially constructed information system or from already existing statistical sources. The resulting output is a performance matrix that relates each alternative to its performance with respect to the selected indicators.

Formally, a performance matrix (before scoring and weighting) is a table in which the rows describe the alternatives $i = 1 \ldots n$ (e.g., countries, policy options) and the columns x_k, $k = 1 \ldots K$, the individual indicators:

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ j \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & \cdots & x_k & \cdots & x_K \end{array} \\ \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{iK} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & \cdots & x_{jk} & \cdots & x_{jK} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{nK} \end{bmatrix} \end{array}$$

Table 11.1 Human development index for a set of countries: example 1

Country	Life expectancy at birth (years)	Expected years of schooling (years)	Mean years of schooling (years)	Gross national income per capita (2011 PPP $)	Human Development Index (HDI)
Norway	81.6	17.5	12.6	64,992	0.944
Denmark	80.2	18.7	12.7	44,025	0.923
Japan	83.5	15.3	11.5	36,927	0.891
Bulgaria	74.2	14.4	10.6	15,596	0.782
Egypt	71.1	13.5	6.6	10,512	0.690
Indonesia	68.9	13.0	7.6	9788	0.684
Cambodia	68.4	10.9	4.4	2949	0.555
Pakistan	66.2	7.8	4.7	4866	0.538
Haiti	62.8	8.7	4.9	1669	0.483
Niger	61.4	5.4	1.5	908	0.348

Data source: United Nations Development Program, 2014

Each element x_{ik} of matrix **P** represents the evaluation (or performance) of the i-th alternative against the k-th criterion. Those elements are generally expressed in quantitative terms, but can also take a qualitative form, e.g., a 5-star rating (★★★☆☆) or a specific color coding (red for high and green for low).

Table 11.1 offers an example of a performance matrix using the 2014 HDI. The table, based on data from the United Nations Development Program, shows the performance of a number of (randomly selected) countries in regard to the level they reach in terms of life expectancy, expected years of schooling for children of school-entering age, mean of years of schooling for adults aged 25 years and gross national income per capita. Last column offers the resulting HDI index. Its computation will be detailed later on. As can be observed, three of these criteria are measured in years and one in monetary (purchasing power parity PPP) terms. In addition, if the first three criteria are expressed in similar terms, their values are not directly comparable. For these reasons, MCDA usually requires transformation of raw measurements so that each individual indicator is normalized between 0 and 1 (or equivalently 100%).

Final scores are derived from partial value functions (marginal value functions). The aim of those functions, denoted $v_k = v_k(x_{ik})$ hereafter, is to translate the performance of the alternatives on a scale which allows direct comparisons among criteria. For each $k = 1 \ldots K$, they ideally satisfy three properties:

Property 1: alternative i is preferred to alternative j in terms of criterion k if and only if $v_k(x_{ik}) > v_k(x_{jk})$;

Property 2: indifference between i and j in terms of criterion k exists if and only if $v_k(x_{ik}) = v_k(x_{jk})$.

Property 3: $v_k(x_{\min,k})=0$ and $v_k(x_{\max,k})=1$, where $x_{\min,k}$ and $x_{\max,k}$ are the minimum and maximum level for criterion k, respectively.

Property 1 ensures that all scores are expressed in accordance with preferences and aims. It is preferable to have a high score (high v_k) than a low score (low v_k). Assume for instance that one aims to evaluate a range of public projects. Cost considerations are likely to belong to the set of relevant criteria. In that case, a high score must be associated with a low level of cost and, reciprocally, a low score must be associated with a high level of cost. Property 2 states that one is indifferent between two alternatives on the considered criterion if they have obtained the same score. Last, according to Property 3, those scores should lie between 0 and 1.

In accordance with property 3, it is conventional to assign a value using an interval scale. To do so, one needs to define two reference points. The scores are then computed in relation to those benchmarks. The approach, also known as unity-based normalization or Min-Max method, specifies a linear partial value function as follows:

$$v_k(x_{ik}) = \frac{x_{ik} - x_{\min,k}}{x_{\max,k} - x_{\min,k}}$$

The score is computed as the ratio of the difference between the raw value and the minimum value ($x_{ik} - x_{\min,k}$) to the difference between the maximum and minimum values ($x_{\max,k} - x_{\min,k}$). As illustrated in Fig. 11.3a, the approach is equivalent to rescaling the performance values using the equation of a line. For example, if one criterion has a minimum value of 2,000 and a maximum of 10,000, each score is computed as:

$$\text{Score} = \frac{\text{Performance} - 2{,}000}{10{,}000 - 2{,}000}$$

An alternative with a performance of 2,000 will obtain a score of 0 while an alternative with a performance of 10,000 will get a score of 1. For any alternative

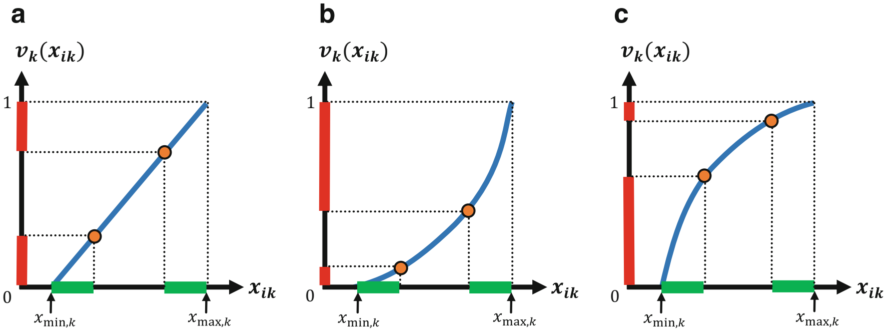


Fig. 11.3 Linear, convex and concave value functions. (**a**) Linear function, (**b**) Convex function, (**c**) Concave function

in between, say 3,000, we would obtain a score inside the [0,1] interval, here (3,000–2,000)/(8,000) = 0.125.

Note that when higher values in the performance matrix correspond to worse rather than better performance, for instance in the case of a pollution or a financial cost, then the linear value function must be modified as follows:

$$v_k(x_{ik}) = \frac{x_{\max,k} - x_{ik}}{x_{\max,k} - x_{\min,k}}$$

Consider for instance a measure of pollution whose values lie between 0 and 50. The value function becomes:

$$\text{Score} = \frac{50 - \text{Performance}}{50 - 0}$$

An alternative with a performance of 50 will get a score of 0 while an alternative with a performance of 0 will obtain a score of 1. For any alternative in between, say 25, we would have a score equal to (50–25)/(50) = 0.5.

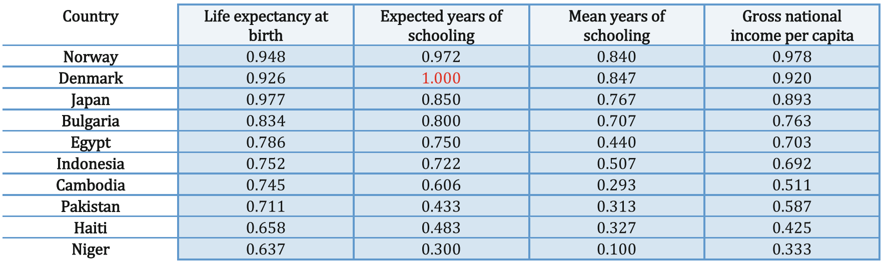
One interesting feature of partial value functions is that they can also be used to establish some preference over the performance levels. Mathematically speaking, partial value functions can be nonlinear. For example, preference over air quality can be marginally decreasing as people may not care about air quality improvements above a certain level. After an acceptable level of pollution, further marginal improvements are valued much less highly. In practice, for computational ease, it is simpler to rely on continuous functions as those illustrated in Fig. 11.3b, c. A convex function is more suitable when, for an equivalent variation (displayed in green in Fig. 11.3b), the decision-maker prefers an increase from high performance levels over an increase from low performance levels. On the other hand, a concave function is used when, for an equivalent variation, the decision-maker prefers an increase from low performance levels over an increase from high performance levels (Fig. 11.3c). In contrast, with a linear function, the decision-maker is indifferent with respect to the starting point (Fig. 11.3a). The value function is said to be neutral.

The minimum and maximum reference points can be established in many different ways. First and foremost, the scale can be global or local. By definition, a local scale is defined by the currently considered set of alternatives. In the sample, a score of 0 is assigned to the alternative with the worst performance and 1 is assigned to the alternative that best performs. Other alternatives will receive a score that lies in between those values. With global scaling on the other hand, the reference points are defined based on the worst and best possible performance using the whole range of conceivable values. Scores lie again in the [0,1] interval, but do not necessarily reach those extreme values. The choice between local and global scaling is mainly a matter of time and ease of use. Ideally, global scaling is chosen because the approach does not depend on the set of considered alternatives. It more easily accommodates new alternatives and allows for comparison through

Table 11.2 Reference points for the HDI: example 1

Country	Life expectancy at birth (years)	Expected years of schooling (years)	Mean years of schooling (years)	Gross national income per capita (2011 PPP $)
Minimum	20	0	0	100
Maximum	85	18	15	75,000

Data source: United Nations Development Program, 2014



Country	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income per capita
Norway	0.948	0.972	0.840	0.978
Denmark	0.926	1.000	0.847	0.920
Japan	0.977	0.850	0.767	0.893
Bulgaria	0.834	0.800	0.707	0.763
Egypt	0.786	0.750	0.440	0.703
Indonesia	0.752	0.722	0.507	0.692
Cambodia	0.745	0.606	0.293	0.511
Pakistan	0.711	0.433	0.313	0.587
Haiti	0.658	0.483	0.327	0.425
Niger	0.637	0.300	0.100	0.333

Fig. 11.4 Score matrix: example 1

time. Local scaling on the other hand permits an immediate estimation of scores and, most of all, does not require further human judgment.

The Human Development Index offers an example of global scaling. The reference points are presented in Table 11.2. The minima are fixed at 20 years for life expectancy, 0 years for expected years of schooling, 0 years for mean years of schooling, and $100 for GNI per capita. The maxima are respectively set to 85 years, 18 years, 15 years and $75,000. Those values are based on historical evidence and forecasts. Scores provided in Fig. 11.4 are computed in relation to those values. The different partial value functions can be described as follows:

$$v_k(x_{ik}) = \frac{x_{ik} - 20}{85 - 20} \quad (k = \text{Life expectancy})$$

$$v_k(x_{ik}) = \frac{\min\{x_{ik}, 18\} - 0}{18 - 0} \quad (k = \text{Expected years of schooling})$$

$$v_k(x_{ik}) = \frac{x_{ik} - 0}{15 - 0} \quad (k = \text{Mean years of schooling})$$

$$v_k(x_{ik}) = \frac{\ln x_{ik} - \ln 100}{\ln 75{,}000 - \ln 100} \quad (k = \text{GNI per capita})$$

Note that for the purpose of calculating the HDI value, expected years of schooling is capped at 18 years (see Denmark in Table 11.1 and Fig. 11.4), which equivalently means that the value function associated with that criterion is non-linear and concave. Similarly, the HDI uses the logarithm of income, i.e. a concave function, to reflect the diminishing importance of income with increasing GNI.

The score matrix is usually not the final product of the analysis. Yet, before any further calculations, a close examination of the scores can be informative and, to some extent, necessary. For instance, it can be interesting to check whether there are alternatives that are dominated by others. By definition, alternative i is said to dominate alternative j if i performs at least as well as j on all criteria and strictly better on at least one criterion. Reversely, an alternative is said to be non-dominated as long as it is not inferior to any other available alternative in all the considered criteria. Consider for instance Fig. 11.4. From this small sample of countries, it can be seen that Japan dominates all the alternatives but Norway and Denmark. Those alternatives are therefore those that best perform in the related sample. This result should be reflected in the composite indicator. When it comes to the evaluation of public projects, policy options that have proven to be dominated can be eliminated without any regret, unless those dominance relationships indicate that one or more relevant criteria are missing from the analysis.

11.4 Criteria Weighting

In most MCDA studies, weighting coefficients w_k, $k = 1 \ldots K$, lie between 0 and 1 and satisfy the following requirement:

$$\sum_{k=1}^{K} w_k = 1$$

Weights must reflect the relative importance of the criteria and, as such, be in accordance with the decision-maker's preferences. Assessing those weights is not straightforward. In most cases, weight assessment is based on successive iterations and involves the participation of stakeholders. Several methods can be used in this respect.

As exemplified with the HDI, the easiest way for establishing weights is to set $w_k = 1/K$ for all k. All criteria are judged to be of equal importance. While convenient at first, the approach may however induce redundancy by combining variables that are too similar. To overcome double-counting, criteria that prove to be highly correlated can be combined into a single indicator. This is for instance the case with the education index used to compute the 2014 HDI. We have:

$$I_{\text{Education}} = \frac{\text{Expected years of schooling} + \text{Mean years of schooling}}{2}$$

This index is defined as the arithmetic mean of the mean years of schooling indicator and the expected years of schooling indicator.

Another approach is direct subjective assessment through what is termed the analytic hierarchy process, also known as Saaty's method. The decision-maker evaluates the criteria using pairwise comparisons. Then, the results are assembled to compute weights. The method has proved to be very popular in MCDA. It is for instance described in "Multi-criteria analysis: a manual", a guide from the UK Government department for communities and local government in England. Specifically, the decision-maker must answer a set of questions of the form "how important is criterion k relative to criterion l?" The preference can be expressed verbally and converted on an ordinal scale as follows:

1 = Equally important;
3 = Moderately more important;
5 = Strongly more important;
7 = Very strongly more important;
9 = Overwhelmingly more important.

The results are then summarized in a judgment matrix:

$$\mathbf{A} = \begin{array}{cc} & \begin{array}{cccccc} x_1 & x_2 & \dots & x_k & \dots & x_K \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_K \end{array} & \begin{bmatrix} 1 & a_{12} & \dots & a_{1k} & \dots & a_{1K} \\ 1/\,a_{12} & 1 & \dots & a_{2k} & \dots & a_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/\,a_{1k} & 1/\,a_{2k} & \dots & 1 & \dots & a_{kK} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/\,a_{1K} & 1/\,a_{2K} & \dots & 1/\,a_{kK} & \dots & 1 \end{bmatrix} \end{array}$$

where each a_{kl} represents the pairwise comparison rating for criterion k against l. The decision-maker is assumed to be consistent in his/her preference. For example, if k is felt to be "moderately more important" than l, then the value 1/3 is assigned to l against k. Because of this reciprocity, the method requires only $K(K-1)/2$ pairwise comparisons.

One popular method for deriving weights is the geometric mean method. First, the geometric mean of each row of $\mathbf{A}$ is computed as:

$$\pi_1 = (1 \times a_{12} \times \ldots \times a_{1k} \times \ldots \times a_{1K})^{1/K}$$

$$\pi_2 = (1/a_{12} \times 1 \times \ldots \times a_{2k} \times \ldots \times a_{2K})^{1/K}$$

$$\ldots$$

$$\pi_k = (1/a_{1k} \times 1/a_{2k} \times \ldots \times 1 \times \ldots \times a_{kK})^{1/K}$$

$$\ldots$$

$$\pi_K = (1/a_{1K} \times 1/a_{2K} \times \ldots \times 1/a_{kK} \times \ldots \times 1)^{1/K}$$

Second, the sum $\pi_1 + \pi_2 + \ldots \pi_k + \ldots + \pi_K$ of those geometric means is calculated. Last, weights are obtained through normalization by dividing each of the geometric means by their total:

$$w_k = \frac{\pi_k}{\sum_{k=1}^{K} \pi_k}$$

The approach offers a simple way of assessing preferences, especially if the decision-maker finds direct rating rather difficult. Greater weights are given to criteria which are considered to be more important. Yet, the method can be costly in terms of time as it relies on a large number of pairwise comparisons.

To illustrate Saaty's method, let us assume that we have the following judgment matrix:

$$A = \begin{bmatrix} 1 & 1/5 & 1/9 \\ 5 & 1 & 1/3 \\ 9 & 3 & 1 \end{bmatrix}$$

Criterion 2 is judged to be "strongly more important" than criterion 1. Criterion 3 is "overwhelmingly more important" than criterion 1 and "moderately more important" than criterion 2. The geometric means are computed as follows:

$$\pi_1 = (1 \times 1/5 \times 1/9)^{1/3} \approx 0.281$$

$$\pi_2 = (5 \times 1 \times 1/3)^{1/3} \approx 1.186$$

$$\pi_2 = (9 \times 3 \times 1)^{1/3} = 3.000$$

The total of those geometric means is $\pi_1 + \pi_2 + \pi_3 \approx 4.467$. Finally, the weights are obtained by dividing the geometric means by their total:

$$w_1 \approx 0.281/4.467 \approx 0.063$$

$$w_2 \approx 1.185/4.467 \approx 0.265$$

$$w_3 \approx 3.000/4.467 \approx 0.672$$

As can be seen, the weights sum to one, the highest weight being allocated to criterion 3, while the lowest is allocated to criterion 1.

Another possibility for computing weights is through use of a regression model. The approach is suitable when the main outcome of interest is directly measurable, but at high cost, which prevents the measure to be used repeatedly over time. Consider for instance a value tree where the goal is to get a measure of y (e.g., welfare) and that only a set of normalized individual indicators $x_1, x_2, \ldots, x_K$ are available (e.g., per capita income, level of inequality, unemployment rate, mortality rate, etc.). A survey can be implemented in period t to get a value of outcome y (e.g., through a scale from 1 to 10 where people indicate how satisfied they are with their life). Then, using ordinary least squares, the following regression model can be estimated to assess the link between y and the individual indicators $x_1, x_2, \ldots, x_K$:

$$y_i = w_1 x_{1i} + w_2 x_{2i} + \ldots + w_K x_{Ki} + \epsilon_i$$

In that context, $w_1, \ldots w_K$ are the coefficients to be estimated and ϵ_i is the error term, i.e. unobserved factors that affect the dependent variable. Once estimated, the model offers a way to predict the value of y using the individual indicators. The approach does not require assumptions about the weights and rely only on statistical evidence. The method can also be used to validate a set of already chosen weights. One inconvenience, however, is that negative weights can be assigned. The approach also relies on statistical expertise as a set of critical assumptions must be verified in order to apply the method. For instance, the econometric approach requires a large sample and appropriate checks to provide accurate results. Among other things, individual indicators must be uncorrelated to avoid multicollinearity problems.

Last, the weights may reflect the quality of the data. Individual indicators that prove to be statistically unreliable (e.g., due to sampling error, missing values) can be assigned lower weights. It is also possible to rely on experts or citizens' opinion, e.g., through survey or focus group methods, to determine the weights. Computed as such, they would better reflect the importance of the criteria from the society's point of view and not only from the viewpoint of the decision-maker. Finally, to avoid double counting, a sound principal component analysis, which provides a multidimensional analysis of the context, can be used to characterize the different correlations that are in play in the data.

11.5 Construction of a Composite Indicator

Once the weights have been determined, one must proceed with the evaluation of the alternatives per se. A possible approach is to rely on aggregation methods (aka American school) to build a single measure of performance. Based on the framework of multi-attribute utility theory, a value function is constructed so as to express overall preference. The idea is to combine all the partial scores and to assemble them into a global score. The approach is compensatory because a poor performance in some indicators can be compensated by sufficiently high values in other indicators.

Formally, let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ characterize the performance of alternative i against the whole set of criteria. Aggregation is implemented via a value function $V(\mathbf{x}_i) = V(x_{i1}, x_{i2}, \ldots, x_{iK})$ which assigns a global score to each alternative i $(i = 1 \ldots n)$ using the whole set of criteria x_{ik} $(k = 1 \ldots K)$. The value function ideally satisfies three properties:

Property 1. Alternative i is preferred to alternative j against all criteria if and only if $V(\mathbf{x}_i) > V(\mathbf{x}_j)$;

Property 2. Indifference between i and j against all criteria exists if and only if $V(\mathbf{x}_i) = V(\mathbf{x}_j)$.

Property 3. Preferences and indifferences are transitive: for any alternatives h, i and j, if alternative h is preferred to i, and alternative i is preferred to j, then alternative h is preferred to j, and similarly for indifference.

Properties 1 and 2 imply that preferences are complete: for any alternatives i and j, either one is strictly preferred to the other or there is indifference between them.

The purpose of the value function is to construct a preference order of the alternatives that is consistent with the decision-maker's viewpoint, the final aim being to compile all individual indicators into a single composite indicator. The decision-maker's preference can be modeled in many ways depending on the aggregation technique. The simplest form is the additive model, which sums the weighted and normalized individual indicators:

$$V(\mathbf{x}_i) = \sum_{k=1}^{K} w_k \times v_k(x_{ik})$$

Another widespread form is the geometric model:

$$V(\mathbf{x}_i) = \prod_{k=1}^{K} [v_k(x_{ik})]^{w_k}$$

In both cases we have:

$$\sum_{k=1}^{K} w_k = 1 \text{ and } 0 \leq w_k \leq 1 \text{ for all } k$$

One advantage of the additive model is that it allows the assessment of the marginal contribution of each variable separately. An increase in the partial value function $v_k(x_{ik})$ by one unit yields an increase in the value function $V(\mathbf{x}_i)$ by w_k units. Moreover, it is easy to compute a "marginal rate of substitution" (MRS) for a pair of criteria. For criteria k and l, consider a simultaneous variation of their respective partial value functions, keeping all other partial value functions constant. If the value function level is to remain unchanged, then the total derivative of V is 0:

$$dV = w_k dv_k + w_l dv_l = 0$$

Which yields:

$$MRS_{v_k \to v_l} = \frac{dv_k}{dv_l} = -\frac{w_l}{w_k}$$

By definition, this rate represents how much a score for one criterion must increase to compensate a decrease by one percent in a score on another criterion. By construction, this rate is independent of the values of the $K-2$ other indicators.

One inconvenient of the additive model is the full compensability it implies. A bad score for one criterion (e.g., an environmental index) can be easily compensated for by a good score on another (e.g., a growth index). The geometric model overcomes this issue. For instance, when a bad score is obtained, e.g., a score close to zero, the value function approaches zero as well. In that context, alternatives with low scores in some individual indicators are more likely to rank last under a geometric aggregation. To illustrate those differences let us consider again the HDI. In 2010, the HDI methodology has changed from arithmetic to geometric aggregation. The reason behind this change was that poor performance in any dimension was not reflected in the additive model. The geometric model on the other hand allocates a low value for countries with uneven development across dimensions. Table 11.3 offers an illustration using the sample of countries from example 1. Scores for individual indicators are derived from Fig. 11.4. A stated previously, the education index stands for the arithmetic mean of the mean years of schooling indicator and the expected years of schooling indicator. For Japan, the geometric model is obtained as follows:

$$V\left(\mathbf{x}_{\mathbf{Japan}}\right) = (0.977)^{1/3} \times (0.808)^{1/3} \times (0.893)^{1/3} \approx 0.890$$

The additive version of the HDI (which is not applied anymore) yields instead:

$$V\left(\mathbf{x}_{\mathbf{Japan}}\right) = \frac{1}{3}0.977 + \frac{1}{3}0.808 + \frac{1}{3}0.893 \approx 0.893$$

As can be observed from Table 11.3, no significant difference is observed between the two models although discrepancies increase for countries with lower values of

Table 11.3 Aggregation of partial value functions: example 1

Country	Health index	Education index	Income index	HDI Geometric model	HDI Additive model
Norway	0.948	0.906	0.978	0.944	0.944
Denmark	0.926	0.923	0.920	0.923	0.923
Japan	0.977	0.808	0.893	0.890	0.893
Bulgaria	0.834	0.753	0.763	0.783	0.783
Egypt	0.786	0.595	0.703	0.690	0.695
Indonesia	0.752	0.614	0.692	0.684	0.686
Cambodia	0.745	0.449	0.511	0.555	0.568
Pakistan	0.711	0.373	0.587	0.538	0.557
Haiti	0.658	0.405	0.425	0.484	0.496
Niger	0.637	0.200	0.333	0.349	0.390

HDI. For a larger set of countries, however, the resulting ranking can be very different.

When faced with a high number of alternatives, it is possible to summarize information by allocating the alternatives to different categories based on the value they obtain with respect to the composite indicator. The categories can for instance be established on the basis of a rating scale (1, 2, 3. . .) or a qualitative scale (e.g., fully achieved, partly achieved, etc.). Cut-off points can be specified using information about how the composite indicator is distributed. For instance, in the 2015 Human Development Report, those cut-off points are derived from the quartiles of distributions: a country is classified in the "very high development" group if its HDI is in percentiles 76–100%, in the "high development" group if its HDI is in the percentiles 51–75%, and so on.

11.6 Non-Compensatory Analysis

When using the previous aggregation methods, results can be highly sensitive to changes in scores and the way individual indicators are constructed and possibly traded off. Some composite indicators can be in favor of one alternative while other value functions are in favor of another. To overcome this issue one may rely on a non-compensatory analysis. Non-compensatory models (aka French school) rely on pairwise comparisons of alternatives with respect to each individual indicator. The approach is particularly suitable for solving sorting problems. Among the most popular outranking methods are the ELECTRE methods (ELimination Et Choix Traduisant la REalité), also known as the "elimination and choice translating reality" methods. Outranking relations among alternatives are examined based on two measurements: a concordance index and a discordance index. Several versions of the approach exist: ELECTRE I, II, III, Tri. Other methods are also available: e.g., PROMETHEE I and II. For the sake of simplicity, this section focuses on ELECTRE I only.

First a concordance set of criteria is defined for each pair of alternatives i and j. This set is denoted $\Omega\ (i,j)$ hereafter. It is the set of all $k \in \{1, \ldots, K\}$ for which alternative i is preferred to alternative j:

$$\Omega(i,j) = \left\{k|\ v_k(x_{ik}) \geq v_k\left(x_{jk}\right)\right\}, \text{for } i \neq j$$

Let $\Omega^C(i,j)$ denote the complement of the concordance set. We have:

$$\Omega^C(i,j) = \left\{k|v_k(x_{ik}) < v_k\left(x_{jk}\right)\right\}, \text{for } i \neq j$$

Hereafter, we will refer to this set as the discordance set. It denotes the set of criteria for which alternative i is worse than alternative j.

The concordance index between alternatives i and j is defined as the weighted measure of the number of criteria for which alternative i is preferred to alternative j. The calculation of this index is based on the concordance set and is defined as:

$$c(i,j) = \frac{\sum_{k \in \Omega(i,j)} w_k}{\sum_{k=1}^{K} w_k}$$

This index lies in the [0,1] interval and is measured as the normalized sum of all weights for which alternative i scores at least as highly as alternative j. If we assume that the weights $w_1 \ldots w_k$ have been normalized *a priori* (their sum is equal to 1), then the concordance index becomes:

$$c(i,j) = \sum_{k \in \Omega(i,\mathrm{j})} w_k$$

This index offers a measure of how far we are from dominance. A value of one indicates that alternative i dominates alternative j (alternative i always yields a score at least as high as alternative j). A value of zero indicates that alternative j strictly dominates i (alternative i always yields a lower score). Any value between 0 and 1 indicates non-dominance.

The different values of the concordance indices are included in a concordance matrix of dimension $n \times n$:

$$\mathbf{C} = \begin{array}{c|ccccccc|} & 1 & \ldots & i & \ldots & j & \ldots & n \\ \hline 1 & \mathrm{NA} & \ldots & c(1,i) & \ldots & c(1,j) & \ldots & c(1,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ i & c(i,1) & \ldots & \mathrm{NA} & \ldots & c(i,j) & \ldots & c(i,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ j & c(j,1) & \ldots & c(j,i) & \ldots & \mathrm{NA} & \ldots & c(j,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n & c(n,1) & \ldots & c(n,i) & \ldots & c(n,j) & \ldots & \mathrm{NA} \end{array}$$

Those values represent the (weighted) number of times each alternative wins against another. Because they are meaningless, the diagonal elements of **C** are not available (an alternative is not compared to itself).

The discordance index between alternatives i and j measures the maximum observed difference in scores for which alternative j is preferred to alternative i:

$$d(i,j) = \frac{1}{\delta} \max_{k \in \Omega^C(i,j)} v_k\left(x_{jk}\right) - v_k(x_{ik})$$

where δ is the range of the scores, i.e. maximum score minus minimum score over all criteria $1 \ldots K$. The discordance index lies between 0 and 1. By definition, it assigns a value of zero when the discordance set Ω^C is empty, i.e. when alternative i dominates alternative j. Overall, this yields what is termed a discordance matrix:

$$\mathbf{D} = \begin{array}{c|ccccccc|} & 1 & \dots & i & \dots & j & \dots & n \\ \hline 1 & \text{NA} & \dots & d(1,i) & \dots & d(1,j) & \dots & d(1,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ i & d(i,1) & \dots & \text{NA} & \dots & d(i,j) & \dots & d(i,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ j & d(j,1) & \dots & d(j,i) & \dots & \text{NA} & \dots & d(j,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n & d(n,1) & \dots & d(n,i) & \dots & d(n,j) & \dots & \text{NA} \end{array}$$

Somehow, those values establish the largest opportunity cost or "regret" (i.e. worst score) incurred from having selected alternative i instead of alternative j.

Basically speaking, one would like the concordance index to be maximized and the discordance index to be minimized. For this purpose, preferences are modeled using binary outranking relations S whose meaning is "at least as good as":

$$iSj \Leftrightarrow c(i,j) \geq \tilde{c} \text{ and } d(i,j) \leq \tilde{d}$$

where $\tilde{c}$ and $\tilde{d}$ are chosen by the decision-maker in the [0,1] range. For an outranking relation to hold true, both concordance and discordance indices should lie in a given range of value. If iSj is verified, then we say that "alternative i outranks alternative j". An outranking relation is not necessarily complete or transitive. Four situations actually may occur when alternative i is compared to alternative j, $i \neq j$:

***iSj* and not *jSi*:** alternative i is strictly preferred to alternative j.
***jSi* and not *iSj*:** alternative j is strictly preferred to alternative i.
***iSj* and *jSi*:** alternative i is indifferent to alternative j
Not *iSj* and not *jSi*: alternative i is incomparable to alternative j

The final solution to the decision problem is given by the set of non-outranked alternatives. The higher the concordance threshold $\tilde{c}$ and the lower the discordance threshold $\tilde{d}$, the lower is the number of alternatives that are outranked and, thus, the less severe is the comparison.

To illustrate the approach, let us consider the matrix of scores $v_k(x_{ik})$ presented in Table 11.4 where $n = 5$ alternatives are evaluated against $K = 4$ criteria. Last row yields the weights w_k associated with each dimension. Based on the score matrix, we can compare alternative 1 versus alternative 2. As can be observed, the concordance set for this pair of alternatives is made of criteria x_1, x_2 and x_4. Alternative 1 indeed performs at least better for those criteria. By summing the related weights we obtain:

$$c(1,2) = 0.6 + 0.2 + 0.1 = 0.9$$

In a similar manner, the concordance set for a comparison of alternative 2 versus alternative 1 is $\{x_2, x_3\}$. We obtain:

Table 11.4 Score matrix for example 2

	Criterion $\boldsymbol{x}_1$	Criterion $\boldsymbol{x}_2$	Criterion $\boldsymbol{x}_3$	Criterion $\boldsymbol{x}_4$
Alternative 1	0.8	0.5	0.1	0.5
Alternative 2	0.4	0.5	0.6	0.4
Alternative 3	0.2	0.9	0.7	0.1
Alternative 4	0.5	0.4	0.2	0.9
Alternative 5	0.2	0.3	0.1	0.3
Weights	**0.6**	**0.2**	**0.1**	**0.1**

$$c(2,1) = 0.2 + 0.1 = 0.3$$

Applying a comparable reasoning for the whole set of alternatives, we obtain the following concordance matrix:

$$\mathbf{C} = \begin{bmatrix} \mathrm{NA} & 0.9 & 0.7 & 0.8 & 1 \\ 0.3 & \mathrm{NA} & 0.7 & 0.3 & 1 \\ 0.3 & 0.3 & \mathrm{NA} & 0.3 & 0.9 \\ 0.2 & 0.7 & 0.7 & \mathrm{NA} & 1 \\ 0.1 & 0 & 0.7 & 0 & \mathrm{NA} \end{bmatrix}$$

From the 1's in the matrix, we can conclude that alternatives 1, 2 and 4 dominate alternative 5. From the 0's we can also conclude that only alternatives 2 and 4 strictly dominate alternative 5.

The discordance set for a comparison of alternative 1 against alternative 2 is made of criterion x_3 only. For this criterion we observe a difference in score equal to 0.6–0.1 = 0.5. The range $\delta = 0.8$ is defined by the difference between the maximum value (0.9) and the minimum value (0.1) observed in the score matrix (Table 11.4). The discordance index for this pair of alternatives is thus:

$$d(1,2) = \frac{0.5}{0.8} = 0.625$$

Reciprocally, the discordance set when alternative 2 is compared to alternative 1 is $\{x_1, x_4\}$. We have:

$$d(2,1) = \frac{\max\{0.8 - 0.4, 0.5 - 0.4\}}{0.8} = \frac{0.4}{0.8} = 0.500$$

After completion of the comparisons, we obtain:

$$\mathbf{D} = \begin{bmatrix} \text{NA} & 0.625 & 0.750 & 0.500 & 0.000 \\ 0.500 & \text{NA} & 0.500 & 0.625 & 0.000 \\ 0.750 & 0.375 & \text{NA} & 1.000 & 0.250 \\ 0.375 & 0.500 & 0.625 & \text{NA} & 0.000 \\ 0.750 & 0.625 & 0.750 & 0.750 & \text{NA} \end{bmatrix}$$

Assume now that the thresholds are defined as follows: $\tilde{c} = 0.6$ and $\tilde{d} = 0.5$. The results can be summarized in the form of a matrix filled with 1 and 0 depending on whether the conditions $c(i,j) \geq \tilde{c}$ and $d(i,j) \leq \tilde{d}$ do hold (1 = TRUE, 0 = FALSE). Using information from matrices **C** and **D**, we obtain the following result matrix:

$$\mathbf{R} = \begin{bmatrix} \text{NA} & 0 & 0 & 1 & 1 \\ 0 & \text{NA} & 1 & 0 & 1 \\ 0 & 0 & \text{NA} & 0 & 1 \\ 0 & 1 & 0 & \text{NA} & 1 \\ 0 & 0 & 0 & 0 & \text{NA} \end{bmatrix}$$

When both conditions are satisfied ($c(i,j) \geq \tilde{c}$ and $d(i,j) \leq \tilde{d}$), there is no reason to eliminate alternative *i* when assessed against alternative *j*. Considering the relative performance of the criteria, it performs better a sufficiently high number of times and, when this is not the case, the difference in scores is sufficiently low. For instance, when comparing alternative 1 versus alternative 4, the concordance index is 0.8 and the discordance index is 0.5. This means that alternative 1 performs sufficiently well against alternative 4, i.e. alternative 1 outranks alternative 4 ("1" is displayed in red in the result matrix). In this particular case, it can be seen that the reverse is not true. Alternative 4 does not outrank alternative 1 ("0" is displayed in red in the result matrix). Alternative 1 is thus strictly preferred to alternative 4.

The set of best alternatives can be identified using a graph similar to that of Fig. 11.5. Each circle represents an alternative and each arrow specifies whether an alternative (start point) outranks another (end point). More specifically, the ELECTRE I method implies searching the set of non-outranked alternatives. This set defines a "kernel" which contains the solutions to the decision problem. By definition a kernel Γ associated with an outranking relation *S* satisfies two properties:

Property 1. For any alternative *j* outside Γ there exists an alternative *i* in Γ such that *iSj* (stability property).

Property 2. Whatever the alternatives *i* and *j* inside Γ, we have neither *iSj* nor *jSi* (absorption property).

Property 1 states that any alternative outside the kernel must be outranked by an alternative inside the kernel. Property 2 states that an alternative inside the kernel cannot be outranked by another alternative inside the kernel. As we shall see, there can be several kernels or no kernel at all.

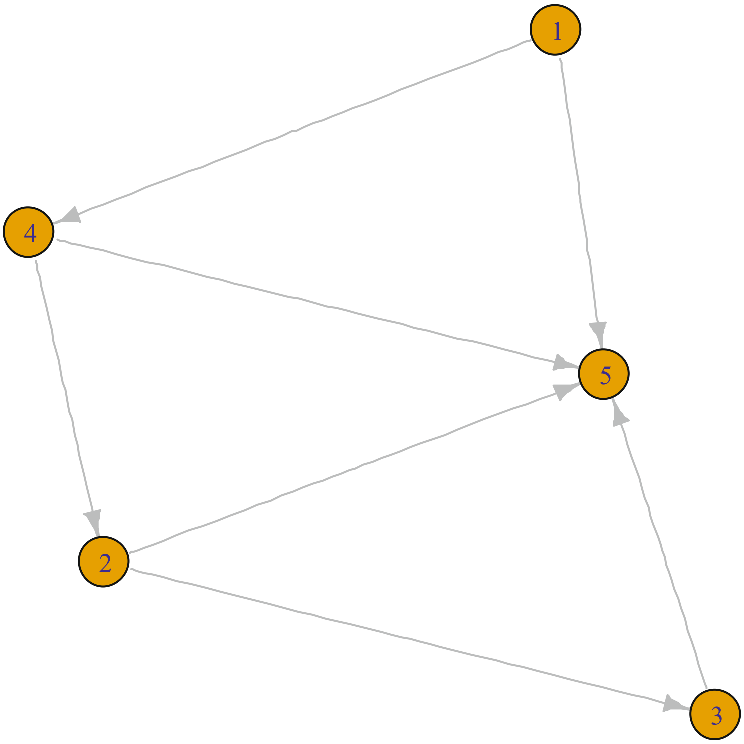


Fig. 11.5 Graph of S when $\tilde{c} = 0.6$ and $\tilde{d} = 0.5$: example 2

The following algorithm can be used to find the set of possible solutions: (1) place in the kernel all the alternatives that are not outranked; (2) examine whether there are alternatives outside the kernel that are not outranked by at least one alternative in the kernel. If yes, include them in the kernel and reiterate from 1, otherwise stop. Let us consider Fig. 11.5. Alternative 1 is not outranked and must be included in the kernel (property 1). Yet, alternatives 2 and 3 are not outranked by alternative 1. They must be included in the kernel (property 1). Since alternative 3 is outranked by alternative 2, it cannot be included in the same kernel as alternative 2 (property 2). We finally have two kernels: $\{1, 2\}$ and $\{3\}$. Alternatives outside those kernels can be dropped as they are not fundamentally better compared to the alternatives in the kernels.

One inconvenient of the approach is that the choice of the thresholds is subjective (as is the choice of the weights). Results can thereby be manipulated. For instance, when the thresholds are set to $\tilde{c} = 0.7$ and $\tilde{d} = 0.125$, the approach yields a different result:

$$\mathbf{R} = \begin{bmatrix} \text{NA} & 0 & 0 & 0 & 1 \\ 0 & \text{NA} & 0 & 0 & 1 \\ 0 & 0 & \text{NA} & 0 & 0 \\ 0 & 0 & 0 & \text{NA} & 1 \\ 0 & 0 & 0 & 0 & \text{NA} \end{bmatrix}$$

The graph associated with this matrix is provided in Fig. 11.6. The kernel is made of a larger set of alternatives: $\{1, 2, 3, 4\}$. When the thresholds are set to more severe values, respectively $\tilde{c} = 0.5$ and $\tilde{d} = 0.75,$ we obtain instead:

$$\mathbf{R} = \begin{bmatrix} \text{NA} & 1 & 1 & 1 & 1 \\ 0 & \text{NA} & 1 & 0 & 1 \\ 0 & 0 & \text{NA} & 0 & 1 \\ 0 & 1 & 1 & \text{NA} & 1 \\ 0 & 0 & 1 & 0 & \text{NA} \end{bmatrix}$$

In this case, as shown in Fig. 11.7, the kernel is characterized by a single element: alternative 1.

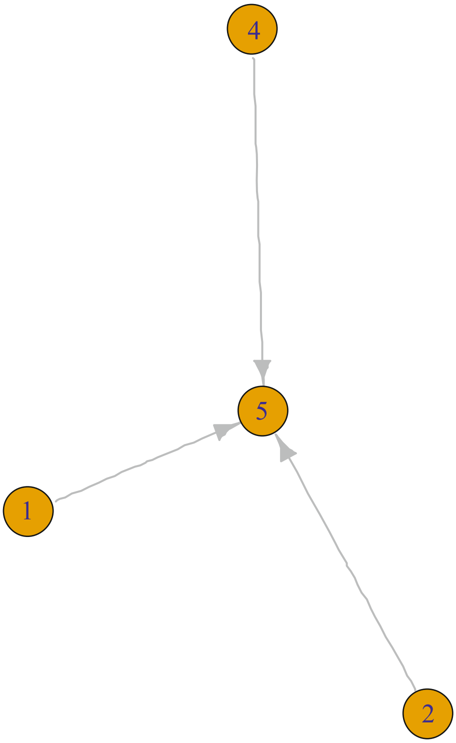


Fig. 11.6 Graph of S when $\tilde{c} = 0.7$ and $\tilde{d} = 0.125$: example 2

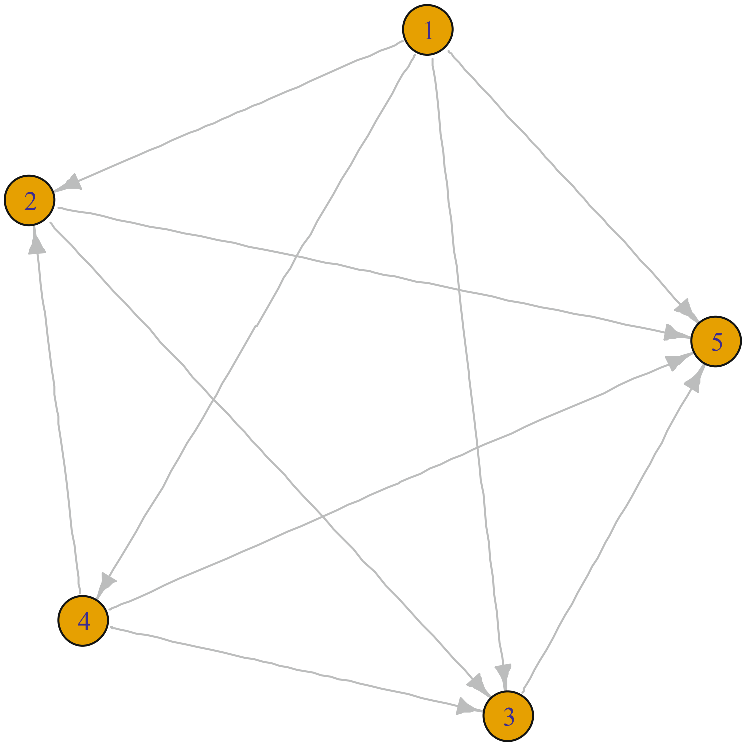


Fig. 11.7 Graph of S when $\tilde{c} = 0.5$ and $\tilde{d} = 0.75$: example 2

Note that we obtain more or less similar results using a compensatory approach. Applying an additive model on the score matrix of Table 11.4, we have:

Alternative 1: $0.8 \times 0.6 + 0.5 \times 0.2 + 0.1 \times 0.1 + 0.5 \times 0.1 = 0.64$
Alternative 2: $0.4 \times 0.6 + 0.5 \times 0.2 + 0.6 \times 0.1 + 0.4 \times 0.1 = 0.44$
Alternative 3: $0.2 \times 0.6 + 0.9 \times 0.2 + 0.7 \times 0.1 + 0.1 \times 0.1 = 0.38$
Alternative 4: $0.5 \times 0.6 + 0.4 \times 0.2 + 0.2 \times 0.1 + 0.9 \times 0.1 = 0.49$
Alternative 5: $0.2 \times 0.6 + 0.3 \times 0.2 + 0.1 \times 0.1 + 0.3 \times 0.1 = 0.22$

Alternative 1 would be ranked first, then alternative 4, alternative 2, alternative 3 and alternative 5. The ELECTRE I method, however, has the advantage of pointing out particularities. For instance, we can see that alternative 1 (0.64) obtains a higher global score than alternative 3 (0.38). Yet, in Fig. 11.5, alternative 3 is not outranked by alternative 1. The reason behind this result is that alternative 3 performs much better on criteria x_3 than alternative 1. The ELECTRE approach can thus be used as a preliminary screening process in order to select a set of promising alternatives but also to avoid large opportunity costs.

```
> library(OutrankingTools)
>
> scoreMatrix=cbind(
+ c(0.8,0.4,0.2,0.5,0.2),
+ c(0.5,0.5,0.9,0.4,0.3),
+ c(0.1,0.6,0.7,0.2,0.1),
+ c(0.5,0.4,0.1,0.9,0.3))
>
> alternatives=c("1","2","3","4","5")
> criteria=c("x1","x2","x3","x4")
> w=c(0.6,0.2,0.1,0.1)
> direction=c("max","max","max","max")
>
> par(mar=c(0,0,0,0))
> Electre_1(scoreMatrix,alternatives,criteria,w,direction,
+ concordance_threshold=0.6,discordance_threshold=0.5)[2:3]

$`Concordance Matrix`
    1   2   3   4   5
1 1.0 0.9 0.7 0.8 1.0
2 0.3 1.0 0.7 0.3 1.0
3 0.3 0.3 1.0 0.3 0.9
4 0.2 0.7 0.7 1.0 1.0
5 0.1 0.0 0.7 0.0 1.0

$`Discordance Matrix`
      1     2     3     4    5
1 0.000 0.625 0.750 0.500 0.00
2 0.500 0.000 0.500 0.625 0.00
3 0.750 0.375 0.000 1.000 0.25
4 0.375 0.500 0.625 0.000 0.00
5 0.750 0.625 0.750 0.750 0.00
```

Fig. 11.8 Outranking relations with R-CRAN: example 2

The graphs of Figs. 11.5, 11.6, and 11.7 have been produced in R-CRAN. Figure 11.8 provides the codes for Fig. 11.5. The program starts with the *library* command which loads the package *OutrankingTools*. This package offers several functions to process outranking ELECTRE methods. The program then continues with the creation of the score matrix using information from Table 11.4. Each vector represents a column of the score matrix. Four vectors are then created to specify (1) the names of each option c("1", "2", "3", "4", "5"), (2) the names of the criteria c("x1", "x2", "x3", "x4"), (3) the weights c(0.6,0.2,0.1,0.1), and (4) the direction of each indicator c(" max ", " max ", " max ", " max "), i.e. whether the criteria have to be minimized or maximized. Function *par* is used to modify the margins of the box using a numerical vector of the form c(bottom,left,top,right) that gives the number of lines of margin to be specified on the four sides of the plot. Last, the command *Electre*_1 from the package *OutrankingTools* draws the graph of Fig. 11.5. Both entries, *concordance_threshold* = 0.6 and *discordance_threshold* = 0.5, specify the thresholds to be used in the analysis. Including [2 : 3] at the end of the function allows the concordance and discordance matrices to be displayed.

11.7 Examination of Results

The last step of MCDA is the examination of results. The size of alternatives, whether they are countries, cities, schools, hospitals or policy options, may play an important role in this respect. To illustrate, let us examine again the HDI. As already stated, the aim of this composite indicator is to provide a measurement of human development. The approach is very simple by nature as it relies only on three dimensions. Yet, those dimensions represent averages: income per capita, expected and mean years of schooling, number of years a newborn infant could expect to live. If one were not using averages but values expressed in level instead, largest countries would necessarily be ranked first because they would reach higher performance in all dimensions. The HDI would be meaningless and the measurement would be related to the population size only. To avoid those effects, individual indicators should be adjusted so as to control for the size of alternatives (e.g., using variables per capita, per GDP, per dollar spent). However, addressing this issue is not as easy as it might look. It is usually dependent on the objective and context of the analysis.

Consider for instance a decision-maker who must choose among several policy options (e.g., public transportation modes). For each option, the dimensions of the problem are divided into a financial cost c and several benefit measures $b_1, \ldots b_K$, each of them being a specific indicator. With MCDA, there are three ways of comparing the options in question. First, one may consider the cost c as a criterion along with the various benefit measures. Second, one may divide each individual indicator $b_1, \ldots b_k$ by c and focus on benefits per dollar spent. Third, one may examine $b_1, \ldots b_k$ regardless of the cost and build an effectiveness indicator which is compared to the spending level c. Each approach has its pros and cons.

The first approach is commonly used in MCDA. To compute the weights, the decision-maker must be able to assess the importance of the project cost with respect to each benefit. The recommendation in the literature is to ask the decision-maker: “How important is cost relative to effectiveness?” The answer is likely to depend on budget issues (if affordability is not an issue per se, a zero weight would be assigned to costs). The process can therefore be highly sensitive and so can be the conclusion itself. Moreover, depending on the selected weights, the analysis may be biased toward smaller projects. Assume for instance that we are comparing two transportation projects according to their cost (c) and how they improve both traffic (b_1) and air quality (b_2). The performance matrix is the following:

	b_1	b_2	c
Alternative 1	3,000	400	\$10,000
Alternative 2	5,000	600	\$20,000

Alternative 2 is more costly but induces larger benefits at the same time. Thus, assigning a high weight on cost will be detrimental to this alternative, not only because it is more costly, but also because the weights must add up to 1 (the weights on b_1 and b_2 will be correspondingly lower).

The second approach which consists in dividing all individual benefits by the project cost can be misleading. Rescaling b_1 and b_2 (using respectively 10,000 and 1,000 as upper bounds and 0 as lower bound) yields:

	b_1	b_2
Alternative 1	0.3	0.4
Alternative 2	0.5	0.6

By dividing b_1 and b_2 by c (expressed in thousand dollars), we obtain:

	b_1/c	b_2/c
Alternative 1	0.030	0.04
Alternative 2	0.025	0.03

Those values relate to performance per thousand of dollars spent. In other words, projects are now compared based on how good they are at achieving each sub-objective. Alternative 2 is strictly dominated by alternative 1 and thereby eliminated. One would reach a similar conclusion using a composite indicator. For example, assuming equal weights:

$$\text{Alternative 1}: \frac{0.030 + 0.04}{2} = 3.50\%$$

$$\text{Alternative 2}: \frac{0.025 + 0.03}{2} = 2.75\%$$

Yet, such an analysis focuses on one aspect of the problem only. Alternative 2 is indeed costly but it is also the alternative that best performs with respect to b_1 and b_2. The decision-maker may actually decide to dedicate extra money to reach this higher level of effectiveness. The approach which is described above completely disregards this possibility.

The third approach consists in treating the cost as an extra variable and computing a hybrid benefit-cost ratio, denoted *HBCR* hereafter. First, as previously, normalized scores are computed:

	b_1	b_2
Alternative 1	0.3	0.4
Alternative 2	0.5	0.6

Then weights are assigned to each criterion and an effectiveness indicator is calculated. Assuming equal weights and using the additive model, alternative 1 could be for instance assigned a final score of $(0.3 + 0.4)/2 = 35\%$, while alternative 2 would obtain $(0.5 + 0.6)/2 = 55\%$. Those scores offer a global measure of effectiveness. The hybrid benefit-cost ratio is then computed as:

$$HBCR = \frac{\text{Effectiveness}}{\text{Cost}}$$

This approach is for instance described in “Principles and Guidelines for Economic Appraisal of Transport Investment and Initiatives” (Government of New South Wales). The aim is to be as close as possible to the framework of a cost benefit analysis. Using amounts expressed in thousand dollars, we get:

$$HBCR(1) = \frac{35\%}{10} = 3.50\%$$

$$HBCR(2) = \frac{55\%}{20} = 2.75\%$$

Projects can then be ranked based on their *HBCR*. As can be deduced, the approach is mathematically equivalent to the second one. Yet, it is more flexible as it allows various weightings. Costs can be compared against benefits in many other ways. The decision-maker may wish to assess whether it is advantageous to allocate extra resources to reach a higher level of effectiveness. A cost-effectiveness analysis would then be implemented to select the most efficient options.

MCDA provides a way of structuring complex decisions, and helps the decision-maker assess the relative importance of the criteria. Yet, the method is often subjective and sometimes complex. Weights might not be transferable from one decision context to another. They may also differ among stakeholder groups. Data about future outcomes may rely on imprecise forecasting methods. The previous discussion also points out the importance of the decision method. To avoid inaccurate decisions, a sound MCDA analysis generally ends with a sensitivity analysis. The purpose is to check whether the solutions obtained are robust to changes not only in the performance matrix, but also in the value functions, the weights and the aggregation method. The approach is very similar to that used in financial analysis and cost benefit analysis.

First, the sensitivity analysis can be performed by changing each important parameter individually through a well-specified range. Consider for instance the weight on cost in the following model:

$$\text{V} = w_c c + w_1 b_1 + w_2 b_2 + \ldots + w_K b_K$$

We must account for the fact that the sum of weights is equal to one.

$$(w_1 + w_2 + \ldots + w_K) = 1 - w_c$$

The impact of a change in w_c is therefore established as follows:

$$\mathrm{V} = w_c c + (1 - w_c)\left[\frac{w_1}{\sum_k w_k} b_1 + \frac{w_2}{\sum_k w_k} b_2 + \ldots + \frac{w_K}{\sum_k w_k} b_K\right]$$

To illustrate, let us consider the score matrix of example 2 (Table 11.4). The weights are $w_1 = 0.6$, $w_2 = 0.2$, $w_3 = 0.1$ and $w_4 = 0.1$. Assume now that we would like to examine how sensitive the results are to the weight on x_1. Since $w_2 + w_3 + w_4 = 0.4$, we have:

$$V = w_1(x_1) + (1 - w_1)\left(\frac{0.2}{0.4} x_2 + \frac{0.1}{0.4} x_3 + \frac{0.1}{0.4} x_4\right)$$

Replacing the x's with their true value for each alternative:

Alternative 1: $w_1 \times 0.8 + (1 - w_1)\left(\frac{0.2}{0.4} 0.5 + \frac{0.1}{0.4} 0.1 + \frac{0.1}{0.4} 0.5\right)$
Alternative 2: $w_1 \times 0.4 + (1 - w_1)\left(\frac{0.2}{0.4} 0.5 + \frac{0.1}{0.4} 0.6 + \frac{0.1}{0.4} 0.4\right)$
Alternative 3: $w_1 \times 0.2 + (1 - w_1)\left(\frac{0.2}{0.4} 0.9 + \frac{0.1}{0.4} 0.7 + \frac{0.1}{0.4} 0.1\right)$
Alternative 4: $w_1 \times 0.5 + (1 - w_1)\left(\frac{0.2}{0.4} 0.4 + \frac{0.1}{0.4} 0.2 + \frac{0.1}{0.4} 0.9\right)$
Alternative 5: $w_1 \times 0.2 + (1 - w_1)\left(\frac{0.2}{0.4} 0.3 + \frac{0.1}{0.4} 0.1 + \frac{0.1}{0.4} 0.3\right)$

In other words:

Alternative 1: $w_1 \times 0.8 + (1 - w_1) \times 0.4$
Alternative 2: $w_1 \times 0.4 + (1 - w_1) \times 0.5$
Alternative 3: $w_1 \times 0.2 + (1 - w_1) \times 0.65$
Alternative 4: $w_1 \times 0.5 + (1 - w_1) \times 0.475$
Alternative 5: $w_1 \times 0.2 + (1 - w_1) \times 0.25$

By varying w_1 between 0 and 1 we obtain the sensitivity plot of Fig. 11.10, which has been established in R-CRAN.

Figure 11.9 provide the codes to construct Fig. 11.10. First, a vector $w1$ is created which takes values between 1% and 100%. This vector is used to create the value functions $V1, \ldots, V5$. Function *plot* then creates the graph and draws $V1$. Entries *xlab* and *ylab* specify the label on the x axis and y axis, respectively; $ylim = c(0, 1)$ sets up the limits for the vertical axis; $type = l$ gives the type of plot desired, here a line; $col = 1$ specifies the color of that line, here black. Function *points* adds the other value functions to the graph. Last, a legend is included using *legend*() which specifies the names to be used ("alternative 1", "alternative 2" and so on), as well as the type of plot desired and the color for each curve.

Several conclusions may be drawn from Fig. 11.10. In particular, alternative 3 (in green) and alternative 1 (in black) appear at the top of the graph. While alternative 1 is ranked first for high value of w_1, alternative 3 is preferred for low values of w_1. The intersection of those lines is obtained as follows:

```
> w1=(1:100)/100
> w1
  [1] 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12
 [13] 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24
 [25] 0.25 0.26 0.27 0.28 0.29 0.30 0.31 0.32 0.33 0.34 0.35 0.36
 [37] 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48
 [49] 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.60
 [61] 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72
 [73] 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84
 [85] 0.85 0.86 0.87 0.88 0.89 0.90 0.91 0.92 0.93 0.94 0.95 0.96
 [97] 0.97 0.98 0.99 1.00

> V1=w1*0.8+(1-w1)*0.4
> V2=w1*0.4+(1-w1)*0.5
> V3=w1*0.2+(1-w1)*0.65
> V4=w1*0.5+(1-w1)*0.475
> V5=w1*0.2+(1-w1)*0.25
>
> plot(V1~w1,xlab="Weight w1",ylab="Value function",type="l",
+ ylim=c(0,1),col=1)
> points(V2~w1,type="l",col=2)
> points(V3~w1,type="l",col=3)
> points(V4~w1,type="l",col=4)
> points(V5~w1,type="l",col=5)
>
> legend("topleft",legend=c("Alternative 1","Alternative 2",
+ "Alternative 3","Alternative 4","Alternative 5"),lty=1,col=1:5)
```

Fig. 11.9 Sensitivity analysis with R-CRAN: example 2

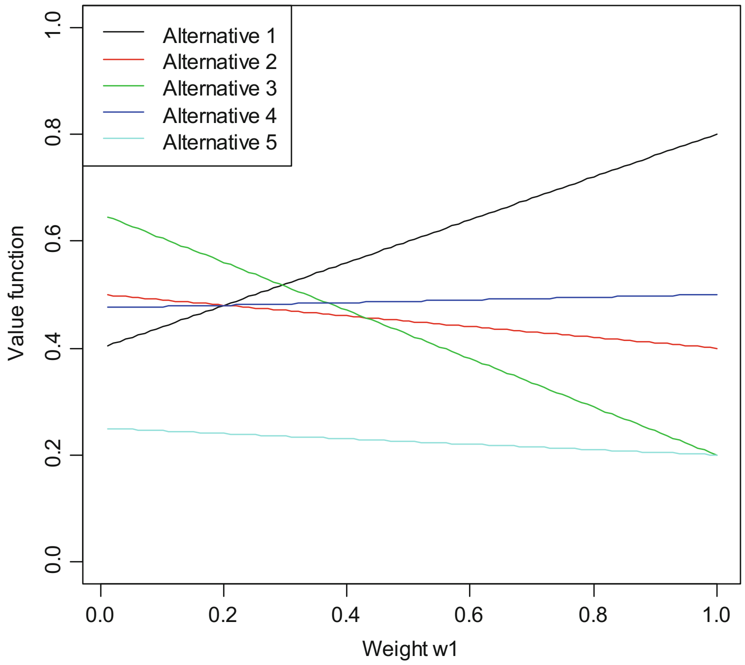


Fig. 11.10 Sensitivity plot for example 2

$$w_1 \times 0.8 + (1 - w_1\ ) \times 0.4 = w_1 \times 0.2 + (1 - w_1) \times 0.65$$

That is to say:

$$w_1^* \approx 0.29$$

If the weight on cost is thought to be fundamentally greater than this level, then the decision goes toward alternative 1. The analysis can go further by examining the scores themselves, e.g., by considering a set of possible scenarios or, when possible, implementing Monte-Carlo simulations.

Bibliographical Guideline

The debate between using a compensatory approach or a non-compensatory approach is often said to originate in the social science literature, and more particularly in the seminal works of Borda (1784) and Condorcet (1785). The "Handbook on constructing composite indicators" from the OECD offers a description of those works. Basically speaking, Borda and Condorcet were arguing about the best voting rule for selecting a particular candidate from a set of politicians. If several individuals or voters participate in the decision, how can we translate the diverse views regarding the election outcome into a group or societal choice? While Borda was in favor of the compensatory approach (modeled through what is now termed the Borda count), Condorcet was on the other hand in favor of the non-compensatory approach. For Condorcet, the best voting rule is to elect the candidate that would win by majority rule in all pairings against the other candidates. In MCDA, the approach has been extended by Roy and several coauthors (the French school) to a family of methods known as ELECTRE (see Roy 1968; Roy and Berthier 1973; Roy and Hugonnard 1982; Roy and Bouyssou 1993 among others). Replacing "voters" by "criteria", the best decision rule would be to select the alternative that best performs in all pairwise comparisons. The American school is on the other hand represented by Saaty (1980), a professor of statistics and operations research who developed the Analytic Hierarchy Process (AHP) method.

It should be stressed that the aim of this chapter was to describe a set of techniques which are commonly in play in MCDA. The chapter does not provide an exhaustive review of all techniques but, instead, offers a synthesized view of the MCDA approach. To go further, the reader may rely on several textbooks that present an introduction to MCDA followed by more detailed chapters about the methods and/or software used in this field (see, among others, Hobbs and Meier 2000; Pomerol and Barba-Romero 2000; French et al. 2009; Ishizaka and Nemery 2013). Additional textbooks (e.g., Beroggi 1991; Beinat 1997) provide an introduction to the main analytic concepts in MCDA. The reader can also rely on Melese et al. (2015) who offer a discussion of MCDA methods in the context of cost benefit analysis.

Last, several guides that present the best practice in a policy-making context are available online. We may name in particular "Tools for Composite Indicators

Building" prepared for the EU (Nardo et al. 2005), the "Handbook on Constructing Composite Indicators" by the OECD, "Multi-criteria analysis: a manual" from the UK Government department for communities and local government in England, and "Principles and Guidelines for Economic Appraisal of Transport Investment and Initiatives" from the New South Wales Government. Health technology assessment agencies also show a growing interest in MCDA methods (Thokaka et al. 2016).

References

Beinat, E. (1997). *Value functions for environmental management*. Berlin: Springer.

Beroggi, G. (1991). *Decision modeling in policy management: An introduction to the analytic concepts*. Heidelberg: Springer.

de Borda, J. C. (1784). Mémoire sur les Elections au Scrutin. In *Histoire de L'Academie Royale des Sciences*.

de Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la probabilité des voix*. Paris: De l'imprimerie royale.

Department for Communities and Local Government. (2009). Multi-criteria analysis: A manual.

French, S., Maule, J., & Papamichail, N. (2009). *Decision behaviour, analysis and support*. University of Manchester, Cambridge University Press.

Hobbs, B. F., & Meier, P. (2000). *Energy decisions and the environment: A guide to the use of multicriteria methods*. International Series in Operations Research & Management Science.

Ishizaka, A., & Nemery, P. (2013). *Multi-criteria decision analysis: Methods and software*. New York: Wiley.

Melese, F., Richter A. and Solomon B. (2015). Military cost-benefit analysis: Theory and practice. In *Studies in defence and peace economics*. Routledge.

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). *Tools for composite indicators building*. Prepared for the EU Commission.

New South Wales Government. (2013). *Principles and guidelines for economic appraisal of transport investment and initiatives*.

OECD. (2008). *The handbook on constructing composite indicators: Methodology and user guide*.

Pomerol, J.-C., & Barba-Romero, S. (2000). *Multicriterion decision in management principles and practice*. Heidelberg: Springer.

Roy, B. (1968). *Classement et choix en présence de points de vue multiples (la méthode Electre)*. Revue française d'automatique, d'informatique et de recherche opérationnelle 8.

Roy, B., & Berthier, P. (1973). *La méthode ELECRE II*. Rapport technique. Note de travail, l42. METRA, Direction Scientifique.

Roy, B., & Bouyssou, D. (1993). *Aide multicritère à la décision: méthodes et cas*. Paris: Économica.

Roy, B., & Hugonnard, J. C. (1982). Ranking of suburban line extension projects on the Paris Metro System by a multicriteria method. Transportation Research A16.

Saaty, T. (1980). *The analytical hierarchy process*. New York: Wiley.

Thokala, P., Devlin, N., Marsh, K., Baltussen, R., Boysen, M., Kalo, Z., et al. (2016). Multiple criteria decision analysis for health care decision making—An introduction: Report 1 of the ISPOR MCDA emerging good practices task force. *Value in Health, 19*, 1–13.

Part III

Ex post Evaluation

Project Follow-Up by Benchmarking 12

12.1 Cost Comparisons to a Reference

The implementation of large programs is usually devolved to several decision making units often located in different places on the territory concerned by the policy. Those are often facilities that operate independently in their geographical area for the provision of the outcome planned by the program. Typical examples include schools, hospitals, prisons, social centers, fire departments, which are in charge of carrying out a mission as defined by the selected strategy. How those facilities perform is a rather important question since there is no straightforward way of measuring the relative efficiency of the facilities involved in the implementation of the project. They may face different constraints, various demand settings and may have chosen different organizational patterns.

Benchmarking is a follow-up evaluation tool that compares the cost structure of facilities with that of a given reference, the benchmark or yardstick. What is assessed is not a policy per se, but the facilities in charge of implementing it. Benchmarking should be applicable to any public service operating within a multiple-input multiple-output setting and equipped with a cost accounting system. The method is particularly relevant for services dedicated to a variety of target groups. For instance, in education, pupils may come from differentiated social backgrounds and require different learning and caring methods. Cost comparisons that would not take into account differences in these demand motives would miss essential information and surely distort assessments.

The first step in benchmarking is to highlight and delineate the effects of the demand structure on the cost of the assessed facility (often called the case-mix effect). The set of services that is supplied must be clearly identified in the accounting system, with lists of users for each service. Offering those services implies combinations of inputs that vary from one service to the other. For instance, in the case of a fire department, fire suppression does not require the same vector of inputs as a rescue mission. For this reason, as it also determines the quantity of inputs used, the demand for a set of services plays a determinant role in explaining

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_12

Table 12.1 The benchmarking methodology through a simple example

	School A		School B	
	Average cost per year	Share of students	Average cost per year	Share of students
Degree in physics	\$50,000	40%	\$45,000	70%
Degree in literature	\$30,000	60%	\$25,000	30%
		100%		**100%**

the average cost of a facility. A sound analysis must therefore identify and isolate the effect of the demand structure on cost. Another example is that of a social care facility which is located in an area where social distress is in relative terms much higher than usual.

The purpose of benchmarking is to explore how inputs are translated into outputs. A facility is considered more efficient than another when it maximizes the level of outputs for a given set of inputs, or when it minimizes the cost of inputs required to produce a given level of outputs. Among others, inputs include labor, equipment, energy, maintenance and administration costs. They are the resources used by the facility to produce and deliver the service in a given time period. Outputs on the other hand are defined as the quantity of goods or services produced. In the context of a public program, outputs are often measured by the number of users that benefit from the services in question.

Consider two schools A and B that offer two types of degrees, one in physics and one in literature. One would like to assess which facility achieves its task in the most efficient way. Table 12.1 provides the average cost per student incurred in supplying those services. The cost per year is much higher in physics than in literature in both schools. This can be explained for instance by the equipment required in physics for operating the classes. In the meantime, the demand structure is found to differ from one school to the other. The majority of students in school A are registered in literature while the majority of students in school B are registered in physics. A rough comparison of their costs would be misleading. For school A, the total average cost is defined as:

$$\bar{c}(A) = \$50000 \times 40\% + \$30000 \times 60\% = 38000$$

For school B, we have:

$$\bar{c}(B) = \$45000 \times 70\% + \$25000 \times 30\% = 39000$$

We can compute the following cost ratio:

$$\frac{\bar{c}(A)}{\bar{c}(B)} = \frac{38000}{39000} = 0.974$$

School A is thus cost saving by $1-0.974=2.6\%$. Yet, this simple comparison does not account for the fact that the distribution of users is in favor of school A. In this school, most of the students are registered in literature, a much less costly degree. From Table 12.1, school B is actually more efficient as the cost per student is lower both in physics (\$45,000 < \$50,000) and literature (\$25,000 < \$30,000).

The basic tenet of the benchmarking methodology is to isolate the effect of the demand structure:

$$\text{Total effect} = (\text{Demand effect}) \times (\text{Production effect})$$

This is done by applying the demand structure of the assessed facility to another facility chosen as the reference for the evaluation. How would the benchmark perform if it were to face conditions similar to that of the evaluated unit? In our numerical example, assume that school B is chosen as the benchmark. The approach consists in applying the demand structure of A to the costs of B:

$$\bar{c}(B|A) = \$45000 \times 40\% + \$25000 \times 60\% = 33000$$

The cost ratio can be rewritten as:

$$\frac{\bar{c}(A)}{\bar{c}(B)} = \frac{\bar{c}(B|A)}{\bar{c}(B)} \times \frac{\bar{c}(A)}{\bar{c}(B|A)} = \frac{33000}{39000} \times \frac{38000}{33000} = 0.85 \times 1.15$$

The first ratio relates the adjusted cost of B to the true cost of B; only the demand structure is different. In this respect, we conclude that the distribution of students in school A generates an extra saving of $1-0.85=15\%$. The second ratio compares the cost of A to the adjusted cost of B, i.e. once the demand structure has been controlled for. We conclude that the production structure of A generates an extra cost of 15%.

There remains to be check whether the extra cost observed in school A is due to price considerations (for instance, do the teachers in school A have a higher salary?) or to the allocation of inputs among services (does school A use more teachers?). Formally, the production effect can be decomposed into two elements:

$$\text{Production effect} = (\text{Price effect}) \times (\text{Quantity effect})$$

Differences will now concern the input combination for each demand motive, as well as the price of those inputs. Quantity and price effects will thus complete the comparison of the assessed facility to its benchmark. In this respect, two approaches can be implemented. The first one is service-oriented: in the case of the previous example, the services provided are the degrees that students enroll in. The accounting system is organized by service, the use of inputs being detailed for each of them successively. With the second approach, the accounting system is organized by input and the allocation of their use among the various services provided by the facility (for instance, how many square meters of classrooms are respectively allocated to the two degrees).

In practice, benchmarking requires adequate and precise information on demand and cost. That can have been planned ex ante, as a requirement before the implementation of the project: facilities in charge of enacting it would have to demonstrate that they do have, or will have at the time the project is started, a cost accounting structure suitable for follow-up by benchmarking. Information on inputs must thus be detailed and reliable. There is also a need for balance between aggregation and disaggregation of data to be provided by the accounting system. Too disaggregated, results may be intricate and uninterpretable; too aggregated, they may provide an inaccurate picture of underlying but hidden problems.

There are three usual shortcomings associated with benchmarking. First, there can be doubts about the quality of the accounting system and of data reporting. While inaccuracies can be due to a lack of technical skill, one cannot also rule out strategic misreporting due to cheating behavior. Second, benchmarking is not equipped to assess the value of the services delivered by the facility. Only the input combination is judged. In the case of our previous example, nothing can be said with this method about the quality of the degrees offered by schools A and B. This remark should be kept in mind as benchmarking can be used in further policy-making. Benchmarking is neither about cost effectiveness nor about Pareto-optimality as in cost benefit analysis. The approach does not relate the costs to the degree to which objectives are achieved or to the satisfaction the users derive from the services under evaluation. Third, the benchmark is usually an average of reference facilities and may itself fall under the previous two shortcomings. Benchmarking should then not be used as a “punishment and reward” device, especially in a non-profit framework, with the exception of obviously outlying decision-making units. It should rather help facilities in their learning process of good practices.

Finally, when the project is implemented through a single facility, or when the decision-maker wants to focus on a specific facility amongst several, benchmarking can also be used in a self-evaluation dynamic process. The assessed facility is compared to itself with respect to a previous period of activity so as to learn about how its practices have evolved over time.

The chapter is organized as follows. Section 12.2 considers the cost accounting framework that is required for cost comparisons between the assessed facility and the benchmark. Section 12.3 examines the effects of the demand structure and of the production structure on cost. Section 12.4 decomposes the price and quantity components of the production effect, building on a service-oriented cost accounting. Section 12.5 offers an alternative input-oriented decomposition. Last, Sect. 12.6 explains how the method can be used in a performance improvement process for the assessed facilities.

12.2 Cost Accounting Framework

The first step of the benchmarking process is to make sure that the facility under assessment, denoted F hereafter, has an adequate cost accounting system. As illustrated in Fig. 12.1, different quantities of input may be used depending on the type and number of services that are provided by the facility. The number of users may also vary from one service to the other, which may in return affect the quantity of inputs employed. Analytical accounts must be able to reflect the organization of service provision for the various types of demand faced by the facility. The time horizon is usually the budget year as it is defined by the accounting rules set by current public regulations.

Formally, let $s = 1 \ldots S$ denote the different services supplied by facility F. Those can be for instance degrees delivered by a university, categories of emergency services provided by a fire station, types of care (to the young, the elderly, single mothers, etc.) delivered by a social service unit, disease related groups in hospitals. For each service s, the accounting system should assemble a list of users $1, \ldots, n_s$ where n_s denotes the total number of users in service s. On the supply side, one must list all production factors (the inputs $k = 1 \ldots K$), their unit price p^k and the quantities of input used $q^k_{s,i}$ by each user i in service s. Input prices can be derived by directly using information about the prices charged to the facility (external pricing). More complex structures like hospitals can also resort to internal pricing, in which case the internal price is the ratio between the observed total cost of input k and the total quantity of k used in the facility.

To sum up, the cost accounting system should provide information about price and quantity for all inputs, with prices common to all services, and quantities varying from one service to another, and, for a given service, from one user to

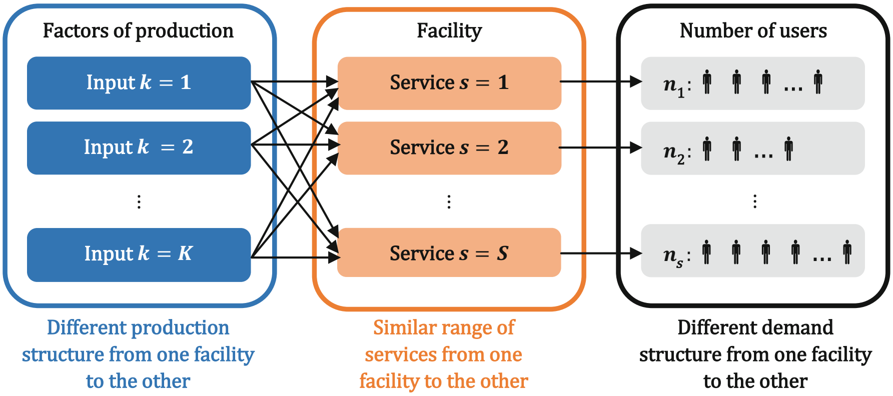

Factors of production
Input k = 1
Input k = 2
⋮
Input k = K
Facility
Service s = 1
Service s = 2
⋮
Service s = S
Number of users
n1: …
n2: …
⋮
nS: …
Different production structure from one facility to the other
Similar range of services from one facility to the other
Different demand structure from one facility to the other

Fig. 12.1 The cost accounting system

another. Equipped with this analytical accounting, the facility can calculate the average quantity of factor k used by service s:

$$\bar{q}_s^k = \frac{\sum_{i=1}^{n_s} q_{s,i}^k}{n_s} \qquad k = 1 \ldots K$$

Consider for instance that service s is the obstetrics care unit in a given hospital. The list of inputs may include umbilical cord clamps, latex gloves, sterile pads, etc. Assume that this care unit receives three patients who use the following combinations of inputs:

Inpatient 1: 1 umbilical cord clamp, 2 latex gloves, 3 sterile pads
Inpatient 2: 1 umbilical cord clamp, 6 latex gloves, 7 sterile pads
Inpatient 3: 1 umbilical cord clamp, 4 latex gloves, 5 sterile pads

The average quantity of clamps is (1 + 1 + 1)/3 = 1, the average number of latex gloves is (2 + 6 + 4)/3 = 4, and the average number of sterile pads is (3 + 7 + 5)/3 = 5.

Using information about input prices, we are able to record the average cost of service s:

$$\bar{c}_s = \sum_{k=1}^{K} \bar{q}_s^k \times p^k$$

Coming back to our example, if the price of one clamp is \$0.05, \$0.03 for one glove and \$0.04 for one sterile pad, the average cost in the obstetrics care unit is computed as:

$$1 \times \$0.05 + 4 \times \$0.03 + 5 \times \$0.04 = \$0.37$$

To compute the total average cost of facility F (total cost per user), one must use information about how the users are distributed among the services. Let $N = n_1 + n_2 + \ldots + n_S$ denote the total number of users. For each service s, we can compute a relative frequency $f_s = n_s/N$ which represents the share of users who have been using service s. This distribution expresses the demand structure of the facility (also known as “case-mix” in health). The average cost of facility F is then defined as a weighted sum of the average costs per service:

$$\bar{c} = \sum_{s=1}^{S} \bar{c}_s \times f_s$$

To illustrate, Table 12.2 provides the analytical accounting of a facility that uses three inputs in order to provide four types of services. As mentioned earlier, the price of inputs is common to all demand motives while quantities are demand-

Table 12.2 Cost accounting of facility F

	Average quantity of input $\bar{q}_s^k$	Price of input p^k	Average cost $\bar{c}_s$	Number of users n_s	Relative frequency f_s
Service 1			$\bar{c}_1 = \$8885$	$n_1 = 5200$	$f_1 = 0.257$
Input 1	12	$290			
Input 2	25	$55			
Input 3	31	$130			
Service 2			$\bar{c}_2 = \$3270$	$n_2 = 2500$	$f_2 = 0.124$
Input 1	4	$290			
Input 2	10	$55			
Input 3	12	$130			
Service 3			$\bar{c}_3 = \$4820$	$n_3 = 6000$	$f_3 = 0.297$
Input 1	6	$290			
Input 2	30	$55			
Input 3	11	$130			
Service 4			$\bar{c}_4 = \$6525$	$n_4 = 6500$	$f_4 = 0.322$
Input 1	6	$290			
Input 2	35	$55			
Input 3	22	$130			
				$\boldsymbol{N = 20200}$	**100%**

specific. Each service generates an average cost. For instance, for service 1 we have:

$$\bar{c}_1 = 12 \times \$290 + 25 \times \$55 + 31 \times \$130 = \$8885$$

Last column of Table 12.2 provides the distribution of users among the services. Using this information, the average cost of facility F is computed as:

$$\bar{c}(F) = \$8885 \times 25.7\% + \ldots + \$6525 \times 32.2\% = \$6223.24$$

The relative frequencies are used to weight the cost of each service. Note that the numbers have been rounded for convenience (25.7% should actually be replaced with 5200/20,200 and 32.2% with 6500/20,200 for finding the result above).

12.3 Effects of Demand Structure and Production Structure on Cost

The cost comparison is undertaken with respect to a benchmark that can be, depending on the institutional context, national, regional or whatever is relevant for the background of the assessment. This benchmark, denoted B hereafter, is usually built on a sample of "representative" facilities equipped with a full accounting system and adequate reporting, and for which data is available. The assessed facility and the benchmark must have similar accounting structures.

Table 12.3 shows the data for the benchmark used to assess the performance of facility F. The interpretation of that table is similar to that of Table 12.2. The average cost of facility B is computed as:

$$\bar{c}(B) = \$7500 \times 22.7\% + \ldots + \$6680 \times 33.2\% = \$5461.23$$

A rudimentary cost comparison of the assessed facility F and the benchmark B is given by the cost ratio:

Table 12.3 Cost accounting of benchmark B

	Average quantity of input $\bar{q}_s^k$	Price of input p^k	Average cost $\bar{c}_s$	Number of users n_s	Relative frequency f_s
Service 1			\$7500	4800	0.227
Input 1	10	\$300			
Input 2	30	\$50			
Input 3	25	\$120			
Service 2			\$3180	3000	0.142
Input 1	3	\$300			
Input 2	12	\$50			
Input 3	14	\$120			
Service 3			\$3640	6300	0.299
Input 1	4	\$300			
Input 2	20	\$50			
Input 3	12	\$120			
Service 4			\$6680	7000	0.332
Input 1	8	\$300			
Input 2	40	\$50			
Input 3	19	\$120			
				$N = 21{,}100$	**100%**

$$\frac{\bar{c}(F)}{\bar{c}(B)} = \frac{\$6223.24}{\$5461.23} = 1.140$$

The assessed facility *F* evinces an extra cost of 14%, but this result is too aggregate to provide an accurate assessment of the cost structure. One needs to distinguish between the effects of the production structure and those resulting from a different distribution of users. For instance, from the supply side, we have:

Service 1: cost greater in *F* ($8885) than in *B* ($7500);
Service 2: cost greater in *F* ($3270) than in *B* ($3180);
Service 3: cost greater in *F* ($4820) than in *B* ($3640);
Service 4: cost lower in *F* ($6525) than in *B* ($6680).

While *B* dominates *F* on the first three services, *F* dominates *B* for service 4. In other words, for services 1 to 3, we can say that *B* is more cost-efficient than *F*: for a same number of users, facility *B* incurs lower costs for those services. Please note that by "more cost-efficient", we do not mean of better quality. The definition of "efficiency" is thus narrower than in cost effectiveness analysis where the output under examination relates to some effectiveness measure (e.g., lower mortality rate or higher success at school).

From Tables 12.2 and 12.3, differences also exist with respect to the demand structure:

Service 1: relative frequency higher in *F* (0.257) than in *B* (0.227);
Service 2: relative frequency lower in *F* (0.124) than in *B* (0.142);
Service 3: relative frequency lower in *F* (0.297) than in *B* (0.299);
Service 4: relative frequency lower in *F* (0.322) than in *B* (0.332).

The users of facility *F* are more concentrated in service 1, relatively speaking. This generates an extra cost for this facility as this service is relatively costly compared to facility *B*.

Radar charts can be used to compare the production and demand structures of the facilities under examination. Figure 12.2 provides an example. The first chart offers a comparison of the costs of facility *F* (displayed in blue) with those of facility *B* (displayed in red). The second chart uses information about relative frequencies and compares the demand structures. Each service $s = 1$ to 4 is displayed on a separate axis. For simplicity of exposition, and to make the charts readable, the origin of each axis corresponds to the minimum value observed in facilities *F* and *B*, while the end of the axis corresponds to the maximum value. For a given service, if *F* is placed at the extremity of the axis, while *B* is placed at the origin, it means that the value for *F* is higher than the value for *B*. The approach is thus only qualitative as it does not account for the magnitude of the observed difference.

Radar charts are very useful when one must examine a significant number of services. Let us first examine Fig. 12.2a. For service $s = 1$, the origin of the axis relates to the cost of facility *B* (i.e. $7500) while the end of the axis stands for the

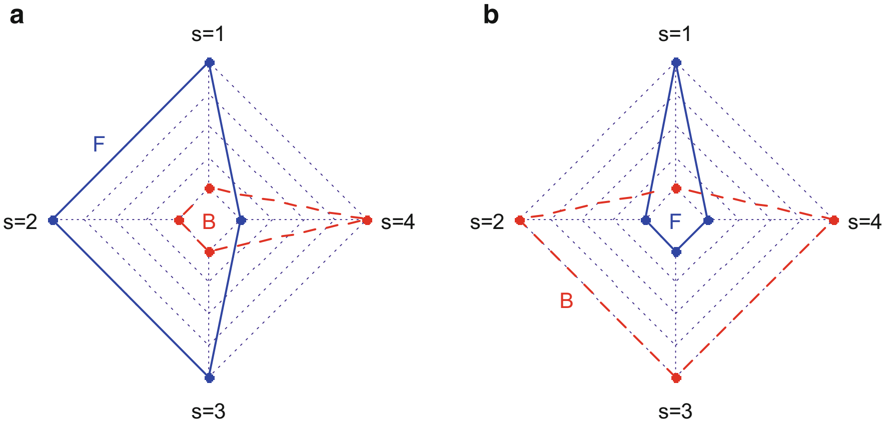


Fig. 12.2 Qualitative comparison of facilities F and B. (**a**) Production structure, (**b**) Demand structure

cost of facility F (i.e. $8885). In a similar manner, minimum and maximum values have been used to construct the axes of services $s = 2$ to 4. When for a given service F is located at the tip of the corresponding radar axis, while B is located at its origin, it means that facility F is less cost-efficient than B. In Fig. 12.2b, data about relative frequencies have been used to draw the axes. If we consider service 2 for instance, B faces a relatively higher demand than F. The figures can thus be used to visualize the dominance relationships both in terms of production (Fig. 12.2a) and demand (Fig. 12.2b). For instance, it can be easily seen that the demand in facility F is relatively oriented toward service 1 despite the fact that F is more cost-efficient in supplying service 4.

Figure 12.3 provides the code to be used in R-CRAN to create Fig. 12.2. Command *par* first specifies the margins of the box plot using a vector of the form $c(bottom, left, top, right)$ that gives the number of lines of margin on the four sides of the plot. The term $mfrow = c(1, 2)$ specifies the number of graphs to be drawn, i.e. one row made of two box plots arranged in column. Variables cF, cB, fF, and fB denote the cost and frequency vectors of facility F and facility B, respectively. The values are entered manually using information from Tables 12.2 and 12.3. The $c()$ function is used to combine those values into a vector. A variable *label* is created to name the axes of the radar chart. Two databases are produced, namely D and E. Database D combines the cost vectors cF and cB (through the *rbind* command). The function *radarchart* (from the package *fmsb*) is then used to produce the radar chart of Fig. 12.2a. The entry $maxmin = FALSE$ states that the maximum and minimum values for each axis will be calculated as actual maximum and minimum of the data. The entry *title* gives the title of the graph, *vlabel* specifies the names for variables, *plwd* defines a vector of line widths for plot data, and *pcol* yields the color of the lines. In a similar manner, database E combines the frequency

```
> par(mar=c(1,1,1,1),mfrow = c(1,2))

> cF=c(8885,3270,4820,6525)
> cB=c(7500,3180,3640,6680)
> fF=c(0.257,0.124,0.297,0.322)
> fB=c(0.227,0.142,0.299,0.332)

> label=c("s=1","s=2","s=3","s=4")

> D=data.frame(rbind(cF,cB))
> library(fmsb)
> radarchart(D,maxmin=FALSE,title="2.1 Production structure",
+ vlabel=label,plwd=c(2,2),pcol=c("blue","red"))
> text(0,0,"B",col="red",lwd=2)
> text(-0.7,0.5,"F",col="blue",lwd=2)
>
> D=data.frame(rbind(fF,fB))
> radarchart(D,maxmin=FALSE,title="2.2 Demand structure",
+ vlabel=label,plwd=c(2,2),pcol=c("blue","red"))
> text(0,0,"F",col="blue",lwd=2)
> text(-0.7,-0.5,"B",col="red",lwd=2)
```

Fig. 12.3 Radar charts in R-CRAN for benchmarking purpose

vectors *fF* and *fB* and is used to produce Fig. 12.2b. Last, the *text* command includes a legend ("F" or "B").

In benchmarking, it is crucial to uncover the underlying structural effects explaining average cost differences. Basically speaking, the allocation of users amongst services does influence cost. One must therefore isolate the effect of the demand structure. This is done by decomposing the cost ratio as follows:

$$\frac{\bar{c}(F)}{\bar{c}(B)} = \frac{\sum_s \bar{c}_s(F) f_s(F)}{\sum_s \bar{c}_s(B) f_s(B)} = \underbrace{\frac{\sum_s \bar{c}_s(B) f_s(F)}{\sum_s \bar{c}_s(B) f_s(B)}}_{\text{Demand effect}} \times \underbrace{\frac{\sum_s \bar{c}_s(F) f_s(F)}{\sum_s \bar{c}_s(B) f_s(F)}}_{\text{Production effect}}$$

The ratio $\sum_s \bar{c}_s(B) f_s(F) / \sum_s \bar{c}_s(B) f_s(B)$ specifies the extent to which the demand structure of F is responsible for the extra cost (or for the cost saving). The ratio $\sum_s \bar{c}_s(F) f_s(F) / \sum_s \bar{c}_s(B) f_s(F)$ on the other hand compares the cost of facility F with that of facility B using the demand structure of F. It measures the extra cost (or extra saving) generated by the production structure of F.

Let $\bar{c}(B|F) = \sum_s \bar{c}_s(B) f_s(F)$ express the cost of B with the demand structure of F. The cost ratio can be rewritten as:

$$\frac{\bar{c}(F)}{\bar{c}(B)} = \underbrace{\frac{\bar{c}(B|F)}{\bar{c}(B)}}_{\text{Demand effect}} \times \underbrace{\frac{\bar{c}(F)}{\bar{c}(B|F)}}_{\text{Production effect}}$$

In the numerical example, we have:

Table 12.4 Cost accounting of facility G

	Average quantity of input $\bar{q}_s^k$	Price of input p^k	Average cost $\bar{c}_s$	Number of users n_s	Relative frequency f_s
Service 1			$7005	5600	0.281
Input 1	11	$280			
Input 2	22	$40			
Input 3	29	$105			
Service 2			$2770	2500	0.126
Input 1	5	$280			
Input 2	8	$40			
Input 3	10	$105			
Service 3			$3725	3800	0.191
Input 1	5	$280			
Input 2	24	$40			
Input 3	13	$105			
Service 4			$5265	8000	0.402
Input 1	4	$280			
Input 2	38	$40			
Input 3	25	$105			
				$\boldsymbol{N = 19{,}900}$	**100%**

$$\bar{c}(B|F) = \$7500 \times 25.7\% + \ldots + \$6680 \times 32.2\% = 5554.95$$

The cost ratio becomes:

$$\frac{\bar{c}(F)}{\bar{c}(B)} = \frac{\$5554.95}{\$5461.23} \times \frac{\$6223.24}{\$5554.95} = 1.017 \times 1.120$$

In the case of facility F, the demand structure only accounts for 1.7% in the extra cost. The positive cost differential is mostly driven by the use of inputs. The ratio $\bar{c}(F)/\bar{c}(B|F)$ shows that the production structure of F is indeed cost increasing and it accounts for 12% of the extra cost.

Another case is that of a facility, say G, which is cost saving. Table 12.4 provides the data. The average cost is computed as:

$$\bar{c}(G) = \$7005 \times 28.1\% + \cdots + \$5265 \times 40.2\% = \$5147.14$$

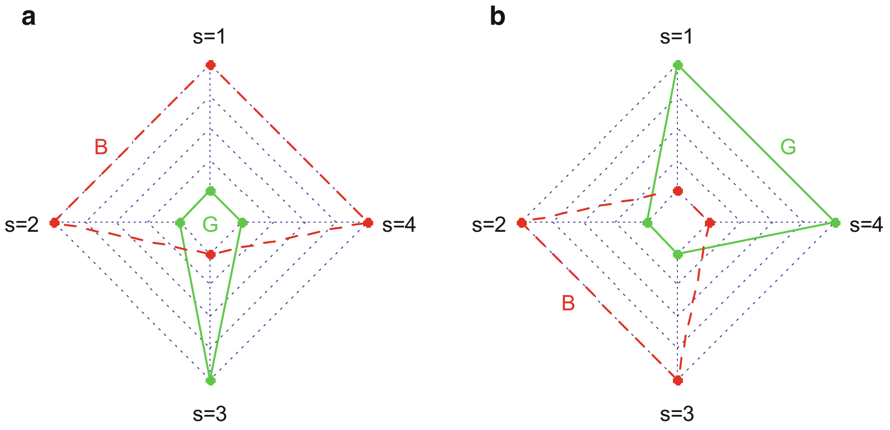


Fig. 12.4 Qualitative comparison of facilities G and B. (**a**) Production structure, (**b**) Demand structure

From Fig. 12.4a, we can see that facility G is more cost-efficient than B for all services but $s = 3$. In the meantime, Fig. 12.4b shows that facility G is more specialized than facility B in services 1 and 4. Overall, the cost ratio is:

$$\frac{\bar{c}(G)}{\bar{c}(B)} = \frac{\$5147.14}{\$5461.23} = 0.942$$

Facility G is thus cost saving by 5.8%. The cost of B with the demand structure of G is:

$$\bar{c}(B|G) = \$7500 \times 28.1\% + \ldots + \$6680 \times 40.2\% = \$5890.55$$

The cost ratio is decomposed as:

$$\frac{\bar{c}(G)}{\bar{c}(B)} = \frac{\$5890.55}{\$5461.23} \times \frac{\$5147.14}{\$5890.55} = 1.079 \times 0.874$$

In the case of facility G, the demand structure implies an extra cost of 7.9% which is more than compensated by a favorable production structure which is cost saving by 12.6%.

Finally, Table 12.5 illustrates a pure demand structure effect. As can be deduced from Fig. 12.5a, facility H has the same cost structure as the benchmark both in terms of price and quantity. We thus have $\bar{c}_s(H) = \bar{c}_s(B)$ for all $s = 1 \ldots 4$. Facility H differs from its benchmark only by the way the users are allocated amongst services (see Fig. 12.5b). The average cost of H is computed as:

Table 12.5 Cost accounting of facility H

	Average quantity of input $\bar{q}_s^k$	Price of input p^k	Average cost $\bar{c}_s$	Number of users n_s	Relative frequency f_s
Service 1			\$7500	3000	0.142
Input 1	10	\$300			
Input 2	30	\$50			
Input 3	25	\$120			
Service 2			\$3180	4000	0.190
Input 1	3	\$300			
Input 2	12	\$50			
Input 3	14	\$120			
Service 3			\$3640	8000	0.379
Input 1	4	\$300			
Input 2	20	\$50			
Input 3	12	\$120			
Service 4			\$6680	6100	0.289
Input 1	8	\$300			
Input 2	40	\$50			
Input 3	19	\$120			
				$N = $ **21,100**	**100%**

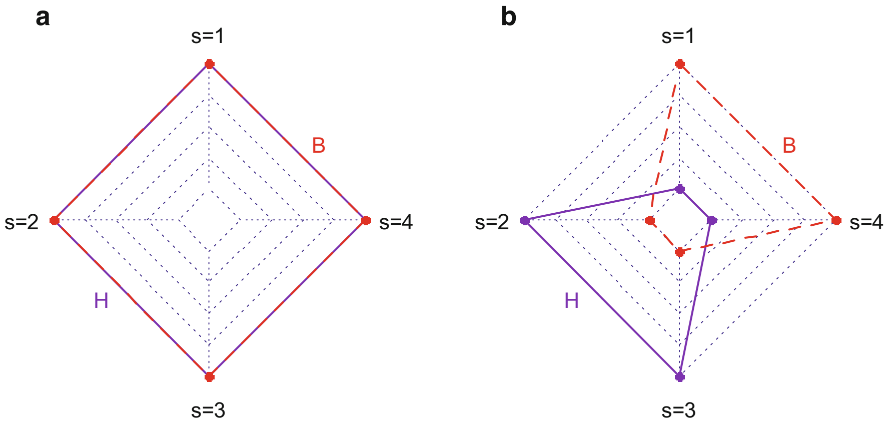


Fig. 12.5 Qualitative comparison of facilities *H* and *B*. (**a**) Production structure, (**b**) Demand structure

$$\bar{c}(H) = \$7500 \times 14.2\% + \ldots + \$6680 \times 28.9\% = \$4980.47$$

The cost ratio is:

$$\frac{\bar{c}(H)}{\bar{c}(B)} = \frac{\$4980.47}{\$5461.23} = 0.912$$

Given that facility H differs from B with respect to the distribution of users only, we can conclude that the demand structure in facility H is cost saving by 8.8%. More specifically, the cost of B with the demand structure of H is:

$$\bar{c}(B|H) = \$7500 \times 14.2\% + \ldots + \$6680 \times 28.9\% = \$4980.47 = \bar{c}(H)$$

The cost ratio can thus be decomposed as:

$$\frac{\bar{c}(H)}{\bar{c}(B)} = \frac{\$4980.47}{\$5461.23} \times \frac{\$4980.47}{\$4980.47} = 0.912 \times 1.000$$

The production structure effect is by construction non-existent. The demand structure of facility H puts it in a relatively favorable position compared to the benchmark.

Having identified the influence of the demand structure and of the production structure on cost, the benchmarking analysis moves on to measuring the effects of input prices and quantities on cost. To do so, one must have neutralized the demand effect by applying the demand structure of the assessed facility to the benchmark.

12.4 Production Structure Effect: Service-Oriented Approach

Cost differences do not only come from demand dissimilarities, they also depend on the combination of inputs and their prices. The next step is thus to decompose the production effect $\bar{c}(F)/\bar{c}(B|F)$ into a price effect and a quantity effect. We need to control for the distribution of users, but also for price and quantity differences. Formally, the production effect can be expressed as:

$$\frac{\bar{c}(F)}{\bar{c}(B|F)} = \frac{\sum_s \bar{c}_s(F) f_s(F)}{\sum_s \bar{c}_s(B) f_s(F)} = \frac{\sum_s \left[\sum_k \bar{q}_s^k(F)\, p^k(F)\right] f_s(F)}{\sum_s \left[\sum_k \bar{q}_s^k(B) p^k(B)\right] f_s(F)}$$

Using this expression we can distinguish between a price effect, concerning $p^k(F)$ and $p^k(B)$ and expressing financial and managerial choices, and a quantity effect, involving quantities of factors $\bar{q}_s^k(F)$ and $\bar{q}_s^k(B)$ and expressing productive choices.

Price and quantity effects can be isolated using Laspeyres and Paasche indices. The approach has been originally conceived to assess the changes in the price level between two periods using indexes such as the consumer price index. In our context, the aim is to compare two facilities. To illustrate, consider two facilities X and Y which supply one service only. There are K inputs for this purpose. Under this framework, one does not need to control for the demand structure since all users benefit from the same service and the focus is on average costs only. The production effect is directly obtained from:

$$\frac{\bar{c}(X)}{\bar{c}(Y)} = \frac{\sum_k \bar{q}^k(X) p^k(X)}{\sum_k \bar{q}^k(Y) p^k(Y)}$$

We would like to know how much each item in the ratio contributes to the observed difference between X and Y. Facility Y is taken as the origin (the equivalent of the base year for a price index) while facility X is the arrival (the equivalent of the current year). A price and quantity index is obtained using the following Fisher decomposition:

$$\frac{\bar{c}(X)}{\bar{c}(Y)} = \begin{pmatrix} \text{Fisher price} \\ \text{index} \end{pmatrix} \times \begin{pmatrix} \text{Fisher quantity} \\ \text{index} \end{pmatrix}$$

with

$$\begin{matrix} \text{Fisher price} \\ \text{index} \end{matrix} = \left[\frac{\sum_k \bar{q}^k(Y)\, p^k(X)}{\sum_k \bar{q}^k(Y)\, p^k(Y)} \times \frac{\sum_k \bar{q}^k(X)\, p^k(X)}{\sum_k \bar{q}^k(X)\, p^k(Y)} \right]^{1/2}$$

and

$$\begin{matrix} \text{Fisher quantity} \\ \text{index} \end{matrix} = \left[\frac{\sum_k \bar{q}^k(X)\, p^k(X)}{\sum_k \bar{q}^k(Y)\, p^k(X)} \times \frac{\sum_k \bar{q}^k(X)\, p^k(Y)}{\sum_k \bar{q}^k(Y)\, p^k(Y)} \right]^{1/2}$$

The first ratio of the Fisher price index is a Laspeyres price index, the second one is a Paasche price index. The first ratio of the Fisher quantity index is a Paasche quantity index while the second one is a Laspeyres quantity index.

The Fisher price index specifies the extent to which the price structure of X is responsible for the extra-cost (or for the cost saving) relatively to Y. For each ratio composing this index, the quantities in the numerator and denominator are equal, only the prices change. The Fisher quantity index on the other hand measures the extent to which quantities contribute to the observed difference in costs. For each ratio composing this index, the prices in the numerator and denominator are equal, only the quantities are changing.

More generally, by applying this approach to facilities F and B and still controlling for the demand structure, we obtain:

$$\begin{matrix} \text{Fisher price} \\ \text{index} \end{matrix} = \left[\frac{\sum_s [\sum_k \bar{q}_s^k(B)\, p^k(F)] f_s(F)}{\bar{c}(B|F)} \times \frac{\bar{c}(F)}{\sum_s [\sum_k \bar{q}_s^k(F) p^k(B)] f_s(F)} \right]^{1/2}$$

	$f_s(F)$	$\bar{q}_s^k(B)$	$p^k(F)$	$\left[\sum_k \bar{q}_s^k(B)\, p^k(F)\right] f_s(F)$	$\bar{q}_s^k(F)$	$p^k(B)$	$\left[\sum_k \bar{q}_s^k(F)\, p^k(B)\right] f_s(F)$
Service 1	0.257			$7800×0.257			$8570×0.257
Input 1		10	$290		12	$300	
Input 2		30	$55		25	$50	
Input 3		25	$130		31	$120	
Service 2	0.124			$3350×0.124			$3140×0.124
Input 1		3	$290		4	$300	
Input 2		12	$55		10	$50	
Input 3		14	$130		12	$120	
Service 3	0.297			$3820×0.297			$4620×0.297
Input 1		4	$290		6	$300	
Input 2		20	$55		30	$50	
Input 3		12	$130		11	$120	
Service 4	0.322			$6990×0.322			$6190×0.322
Input 1		8	$290		6	$300	
Input 2		40	$55		35	$50	
Input 3		19	$130		22	$120	
				5806.44			**5958.86**

Fig. 12.6 Price and quantity effects for facility *F*

and

$$\text{Fisher quantity index} = \left[\frac{\bar{c}(F)}{\sum_s\left[\sum_k \bar{q}_s^k(B)\, p^k(F)\right] f_s(F)} \times \frac{\sum_s\left[\sum_k \bar{q}_s^k(F) p^k(B)\right] f_s(F)}{\bar{c}(B|F)}\right]^{1/2}$$

Benchmark B with the demand structure of F is taken as the origin while the assessed structure F is the arrival.

Figure 12.6 gives details of calculations. Recall that in the case of F, the production structure accounts for 12% in the extra cost with respect to the benchmark:

$$\frac{\bar{c}(\mathrm{F})}{\bar{c}(B|F)} = \frac{\$6223.24}{\$5554.95} = 1.120$$

The Fisher decomposition requires the intermediate calculations of $\sum_k \bar{q}_s^{\,k}(B)\; p^k(F)$ (the cost of service s in F when this facility uses the same input quantities as B) and of $\sum_k \bar{q}_s^{\,k}(F) p^k(B)$ (the cost of service s in F when this facility faces the same prices as B). Using the total row of Fig. 12.6, the Fisher decomposition is specified as follows:

$$\underbrace{\frac{\bar{c}(F)}{\bar{c}(B|F)}}_{1.120} = \underbrace{\left[\frac{\$5806.44}{\$5554.95} \times \frac{\$6223.24}{\$5958.86}\right]^{1/2}}_{\text{Price effect} = 1.045} \times \underbrace{\left[\frac{\$6223.24}{\$5806.44} \times \frac{\$5958.86}{\$5554.95}\right]^{1/2}}_{\text{Quantity effect} = 1.072}$$

Input prices are responsible for the extra cost to the extent of 4.5% while quantities used explain 7.2% of the cost differential, keeping in mind that the demand structure effect has already been controlled for.

By using the same procedure for facility G, one obtains the following results (for the sake of simplicity, calculations are not detailed):

$$\underbrace{\frac{\bar{c}(G)}{\bar{c}(B|G)}}_{0.874} = \underbrace{\left[\frac{\$5167.79}{\$5890.55} \times \frac{\$5147.14}{\$5872.66}\right]^{1/2}}_{\text{Price effect} = 0.877} \times \underbrace{\left[\frac{\$5147.14}{\$5167.79} \times \frac{\$5872.66}{\$5890.55}\right]^{1/2}}_{\text{Quantity effect} = 0.996}$$

Facility G is cost saving by 12.6%. Intermediate calculations yield $\sum_k \bar{q}_s^k(B)\, p^k(G) = \$5167.79$ and $\sum_k \bar{q}_s^k(G)\, p^k(B) = \$5872.66$. Both the price effect (0.877) and the quantity effect (0.996) contribute to the cost advantage but the price effect is the most prominent. For facility H, we have instead:

$$\underbrace{\frac{\bar{c}(H)}{\bar{c}(B|H)}}_{1} = \underbrace{\left[\frac{\$4980.47}{\$4980.47} \times \frac{\$4980.47}{\$4980.47}\right]^{1/2}}_{\text{Price effect} = 1} \times \underbrace{\left[\frac{\$4980.47}{\$4980.47} \times \frac{\$4980.47}{\$4980.47}\right]^{1/2}}_{\text{Quantity effect} = 1}$$

As this facility shares the same cost structure as facility B, both in terms of price and quantity, there are by construction no price and quantity effects.

12.5 Production Structure Effect: Input-Oriented Approach

With the service-oriented approach to the production structure effect, the emphasis is put on the demand motives and the data is gathered within a framework that fits that purpose. This is usually the case for instance with hospitals where demand is allocated to homogenous disease related groups (the services in our presentation) and where the case-mix (the allocation of demand amongst services) is a crucial health management feature. The stakeholders of the public project may also wish to get alternative or complementary information based on the role of inputs in the formation of the production structure effect. A simple reorganization of the data allows it.

Figure 12.7 offers an example of such a reorganization of data for facility F and facility B. The data comes originally from Tables 12.2 and 12.3. The rows are now divided into three inputs, which are in turn divided in four services. For facility F, the frequency column $f_s(F)$ and the quantity column $\bar{q}_s^k(F)$ can be used together to compute the average quantity of input k used in the whole facility:

	$f_s(F)$	$\bar{q}_s^k(F)$	$\bar{q}^k(F)$	$p^k(F)$	$\bar{q}_s^k(B)$	$\bar{q}^k(B\|F)$	$p^k(B)$
Input 1			7.30	$290		6.71	$300
Service 1	0.257	12			10		
Service 2	0.124	4			3		
Service 3	0.297	6			4		
Service 4	0.322	6			8		
Input 2			27.85	$55		28.02	$50
Service 1	0.257	25			30		
Service 2	0.124	10			12		
Service 3	0.297	30			20		
Service 4	0.322	35			40		
Input 3			19.81	$130		17.85	$120
Service 1	0.257	31			25		
Service 2	0.124	12			14		
Service 3	0.297	11			12		
Service 4	0.322	22			19		

Fig. 12.7 Data reorganization for facilities *F* and *B*

$$\bar{q}^k(F) = \sum_{s=1}^{S} \bar{q}_s^k(F) \times f_s(F)$$

For instance, for facility *F*, the average quantity of input 1 is computed as:

$$\bar{q}^1(F) = 12 \times 25.7\% + 4 \times 12.4\% + 6 \times 29.7\% + 6 \times 32.2\% = 7.30$$

Therefore, on average, facility *F* uses 7.30 units of input 1 per user.

For the benchmark facility *B*, the approach is quite similar except that we need to control for the demand structure by using the frequency distribution of facility *F*:

$$\bar{q}^k(B|F) = \sum_{s=1}^{S} \bar{q}_s^k(B) \times f_s(F)$$

For instance, the average quantity of input 1 used by facility *B* when this facility faces the same demand structure as facility *F* is computed as:

$$\bar{q}^1(B|F) = 10 \times 25.7\% + 3 \times 12.4\% + 4 \times 29.7\% + 8 \times 32.2\% = 6.71$$

On average, were facility *B* facing the demand structure of facility *F*, it would use 6.71 units of input 1 per user.

Using information from Fig. 12.7, the input-oriented comparison of facility *F* with its benchmark *B* is:

Input 1: price lower in *F* ($290) than in *B* ($300);
Input 2: price greater in *F* ($55) than in *B* ($50);
Input 3: price greater in *F* ($130) than in *B* ($120).

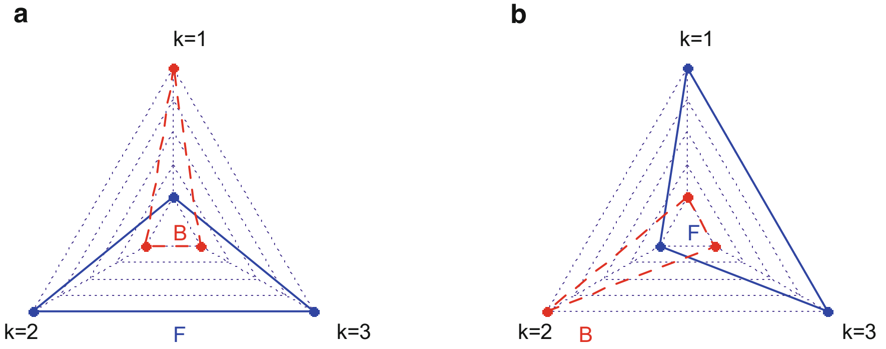


Fig. 12.8 Qualitative comparison of prices and quantities in *F* and *B*. (**a**) Price structure, (**b**) Quantity structure

Similarly, for the quantities (and controlling for the demand structure), we have:

Input 1: quantity higher in *F* (7.30) than in *B* (6.71);
Input 2: quantity lower in *F* (27.85) than in *B* (28.02);
Input 3: quantity higher in *F* (19.81) than in *B* (17.85).

Figure 12.8 illuminates the dominance relationships. We can see that facility *F* is twice dominated by *B* with respect to prices (inputs 2 and 3) and quantities (input 1 and 3).

The codes to produce Fig. 12.8 are provided in Fig. 12.9. The approach is quite similar to what has been done previously (see Fig. 12.3). First the command *par* specifies the margins of the box plot and the number of graphs to be drawn. Variables *pF*, *pB*, *qF*, and *qB* denote the price and quantity vectors of facility *F* and facility *B*, respectively. Two databases are again created, one that combines the price vectors, and one that combines the quantity vectors. The function *radarchart* then plots the graph. Last, the *text* function includes a legend.

This new organization of data can be used to decompose the production effect into price and quantity effects. The production effect can be written as:

$$\frac{\bar{c}(F)}{\bar{c}(B|F)} = \frac{\sum_k \bar{q}^k(F)p^k(F)}{\sum_k \bar{q}^k(B|F)p^k(B)}$$

which can be decomposed into:

$$\text{Fisher price index} = \left[\frac{\sum_k \bar{q}^k(B|F)p^k(F)}{\bar{c}(B|F)} \times \frac{\bar{c}(F)}{\sum_k \bar{q}^k(F)p^k(B)}\right]^{1/2}$$

```
> par(mar=c(0,0,4,0),mfrow = c(1,2))

> pF=c(290,55,130)
> pB=c(300,50,120)
> qF=c(7.30,27.85,19.81)
> qB=c(6.71,28.02,17.85)

> label=c("k=1","k=2","k=3")

> D=data.frame(rbind(pF,pB))
> library(fmsb)
> radarchart(D,maxmin=FALSE,title="5.1 Price structure",
+ vlabel=label,plwd=c(2,2),pcol=c("blue","red"))
> text(0,0,"B",col="red",lwd=2)
> text(0,-0.62,"F",col="blue",lwd=2)

> D=data.frame(rbind(qF,qB))
> radarchart(D,maxmin=FALSE,title="5.2 Quantity structure",
+ vlabel=label,plwd=c(2,2),pcol=c("blue","red"))
> text(0,0,"F",col="blue",lwd=2)
> text(-0.67,-0.62,"B",col="red",lwd=2)
```

Fig. 12.9 Comparison of prices and quantities in R-CRAN

and

$$\begin{array}{c}\text{Fisher quantity}\\ \text{index}\end{array} = \left[\frac{\bar{c}(F)}{\sum_k \bar{q}^k(B|F)p^k(F)} \times \frac{\sum_k \bar{q}^k(F)p^k(B)}{\bar{c}(B|F)}\right]^{1/2}$$

Using information from Fig. 12.7, we have:

$$\bar{c}(F) = 7.30 \times \$290 + 27.85 \times \$55 + 19.81 \times \$130 = 6223.24$$

$$\bar{c}(B|F) = 6.71 \times \$300 + 28.02 \times \$50 + 17.85 \times \$120 = 5554.95$$

$$\sum\nolimits_k \bar{q}^k(B|F)p^k(F) = 6.71 \times \$290 + 28.02 \times \$55 + 17.85 \times \$130 = 5806.44$$

$$\sum\nolimits_k \bar{q}^k(F)p^k(B) = 7.30 \times \$300 + 27.85 \times \$50 + 19.81 \times \$120 = 5958.86$$

Thus, as previously, we obtain the following decomposition:

$$\underbrace{\frac{\bar{c}(\mathrm{F})}{\bar{c}(B|F)}}_{1.120} = \underbrace{\left[\frac{\$5806.44}{\$5554.95} \times \frac{\$6223.24}{\$5958.86}\right]^{1/2}}_{\text{Price effect} = 1.045} \times \underbrace{\left[\frac{\$6223.24}{\$5806.44} \times \frac{\$5958.86}{\$5554.95}\right]^{1/2}}_{\text{Quantity effect} = 1.072}$$

However, we now specifically point out the consequences of differences in input management between facility F and facility B. From Figs. 12.7 and 12.8, we know that the prices of inputs 2 and 3 induce extra costs and are responsible for the price effect while the quantities of inputs 1 and 3 explain the quantity effect. If facility

F wants to improve its performance, those items would appear as a priority for a reorganization of production. Of course, a similar approach can be used to assess the performance of facilities *G* and *H*.

12.6 Ranking Through Benchmarking

Benchmarking is a tool that can be used to motivate operations improvement or to help a decision-maker understand where the performance falls in comparison to others. The approach can also promote emulation among facilities. Table 12.6 offers an example of how the results of the analysis can be summarized. Using information gathered from both the demand analysis (Sect. 12.3) and the production analysis (Sects. 12.4 and 12.5), we are able to offer a clear picture of where the facilities best perform.

First, by examining the total cost effect $\bar{c}(X)/\bar{c}(B)$ for each facility $X \in \{F, G, H\}$, we can see that facility *H* performs better than any other facility. Compared to the benchmark, this facility generates an extra saving of 8.8%. As stressed in Sect. 12.2, this effect can be decomposed into a demand effect (−8.8%) and a production effect (0%). The extra saving thus appears to be generated by the distribution of users only. How facility *H* responds to the demand can be a source of inspiration to the other facilities, provided that it is ethically acceptable to reduce access to some of the services or to encourage the use of the most efficient ones. Since most public programs are dedicated to the well-being of their users, this may not be agreeable. This caveat applies to the whole ranking process.

Second, by examining the second column of Table 12.6, we can also see that facility *F* is ranked last. The decomposition of the total effect tells us that an improvement should not only concern the organization of supply (production effect: +12%) but also that the demand pattern is cost-increasing (+17%).

Table 12.6 Summary of the benchmarking analysis[a]

	Cost ratio	Decomposition of total effect		Decomposition of production effect	
Facility	**Total effect** Comparison of the facility's average cost with the benchmark	**Demand effect** Impact of the distribution of users among services	**Production effect** Impact of the organization of services	**Price effect** Impact of financial and managerial choices	**Quantity effect** Impact of the production policy
B	0%	0%	0%	0%	0%
F	+14%	+17%	+12%	+4.5%	+7.2%
G	–5.8%	+7.9%	–12.6%	–12.3%	–0.04%
H	–8.8%	–8.8%	0%	0%	0%

[a] Figures indicate extra costs (+) or extra saving (−) compared to facility ***B*** (in percent)

The last two columns of Table 12.6 decompose the production effect into a price effect and a quantity effect. It can be seen for instance that facility G yields the best performance in terms of productive structure. Compared to facility B, this facility is able to generate cost saving of -12.3% due its managerial policy and of –0.04% due its input-use arrangements. In this respect, facility G should serve as a reference in any service improvement work.

Bibliographical Guideline

The health sector has been a forerunner (Hansen and Zwanziger 1996) in the comparison of costs amongst facilities providing a common service to various types of demand. Many national health services advocate such cost comparisons. Hospitals have a complex cost structure with, depending on national analytical account techniques, up to one hundred inputs. They also face numerous types of services that generate as many outputs: the corresponding disease-related groups may number around two thousand. Benchmarking is a simple tool for comparison to a reference, usually an "average" hospital which can prove to be somewhat elusive and difficult to define (Llewellyn and Northcott 2005). It is also usually a part of a more general performance assessment framework, like the one launched in 2003 by the World Health Organization Regional Office in Europe (Veillard et al. 2005) and evaluated in 2013 (Veillard et al. 2013). Agarwal et al. (2016) provide a survey of quality of management practices of public hospitals in the Australian healthcare system, as well as comparisons with practices in the United States of America, the United Kingdom, Sweden, France, Germany, Italy and Canada.

More generally, in the 1990s, the Organization for Economic Cooperation and Development has promoted developments in public sector benchmarking (OECD 1997) through the Public Management Service (PUMA). Afterwards, a famous exercise in benchmarking has been the PISA test (OECD 2013).

Our own presentation of benchmarking has mainly focused on the technical side of cost comparisons in presence of heterogeneous demand structures amongst the assessed facilities aiming at implementing a public program. The range of applications can be quite large, keeping in mind that the quality of the data involved in the evaluation process is an essential factor of its relevance.

References

Agarwal, R., Green, R., Agarwal, N., & Randhawa, K. (2016). Benchmarking management practices in Australian public healthcare. *Journal of Health Organization and Management, 30*, 31–56.

Hansen, K., & Zwanziger, J. (1996). Marginal costs in general acute care hospitals: A comparison among California, New York and Canada. *Health Economics, 5*, 195–216.

Llewellyn, S., & Northcott, D. (2005). The average hospital. *Accounting, Organizations and Society, 30*, 555–583.

OECD. (1997). *International benchmarking. Experiences from OECD countries.*

OECD. (2013). *International benchmarking for school improvement. OECD tests for schools (based on PISA).*

Veillard, J., Champagne, F., Klazinga, N., Kazandjian, V., Arah, O., & Guisset, A. (2005). A performance assessment framework for hospitals: The WHO regional office for Europe PATH project. *International Journal for Quality in Health Care, 17*, 487–496.

Veillard, J., Schiotz, M., Guisset, A., Brown, A., & Klazinga, N. (2013). The PATH project in eight European countries: An evaluation. *International Journal of Health Care Quality Insurance, 26*, 703–713.

Randomized Controlled Experiments 13

13.1 From Clinical Trials to Field Experiments

Policy questions are mostly about the impact or causal effect of projects on the life of the citizens who would be exposed to them. Does free access to public schools shift children away from child labor? To what extent are communicable diseases (e.g., infections) driven out by primary health care delivered in local dispensaries? Is voluntary paid work in prisons an adequate tool for preparing the reinsertion of inmates? Answering those questions can be tricky as several confounding factors may also affect the outcome variable in question:

$$\text{Change in outcome} = \text{Impact of intervention} + \text{Confounding factors}$$

Confounding factors can be of two kinds. On the one hand, individuals may differ with respect to their personal characteristics, e.g., their productivity, motivation, health condition and, in return, these characteristics are likely to affect the outcome of interest. Observed changes can also be induced by various events and shocks occurring during the period of observation, that modify the environment of the program. As can be easily understood, the presence of these confounding factors makes the identification of causal effects rather difficult. One solution to avoid potential bias is to run a randomized controlled experiment. The basic tenet of the method is to assign the subjects to different groups using randomization, such that they share similar characteristics on average from one group to the other.

Take an example of "before and after" evaluation. A new drug is delivered to 1,000 patients suffering from type 2 diabetes over an observation period of one year during which a clinical endpoint is scrutinized, for instance, the baseline risk of hypoglycemia. At the end of the trial, an investigator observes that the baseline risk has decreased on average, compared to the situation at the time of patients' inclusion. Can we conclude that this new drug leads to a decrease in hypoglycemia? The first problem is that there are many factors that also affect the outcome of interest. Those factors can be of several kinds, e.g., environmental (diabetes is more

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_13

prevalent in poorer people), behavioral (physical activity is a real asset in fighting the disease), or clinical (comorbidities may affect the endpoint). If these factors have changed since the program was introduced, then the results of the study are likely to be biased.

The second problem is a statistical phenomenon known as the "regression toward the mean" effect. Patients' selection is usually based on the value of a biological (e.g., glycemic control) or clinical (e.g., pain intensity) pretest. Unfortunately, those pretests can be subject to random errors and erroneously select healthy individuals. Once the intervention has taken place, a posttest can yield misleading conclusions for this particular reason: a significant impact can be pointed out while the effect is due to chance only. Consider for instance a game involving a large number of (fair) coins. Suppose that the pretest consists simply in tossing the coins and recording the number of tails. As might be expected, the test is likely to give 50% of heads (H) and 50% of tails (T). Now assume that the tails represent the subjects who receive treatment:

	Selected group	**Non selected group**
Pretest	$T, T, T, T, T, etc.$	$H, H, H, H, H, etc.$

A posttest (new flipping) would inevitably yield bias conclusions as we are likely to obtain again 50% of tails and 50% of heads:

	Selected group
Posttest	$T, T, T, \ldots, H, H, H$

In that case, a simple after-versus-before comparison would yield naive conclusions. Despite the absence of any treatment, the posttest would make us erroneously conclude that the selected coins have now a lower probability of tails.

The temptation is great to compare one exposed individual over time and estimate how he or she fared "before and after" so as to identify variations (positive or negative) in their individual well-being. As we have seen, however, such comparisons can be misleading. An alternative is to contrast participants to those who were not exposed to the program. Yet, can we so easily compare those two types of individuals? The answer is no. Using treated versus non-treated comparisons can be misleading as well. Individuals necessarily differ in their personal characteristics. For all those reasons, and whatever the policy question that is tackled, no individual impact on the path of an exposed individual can be obtained. Experimentation can however assess the average impact of a program on those who benefited from it (were "exposed to it") by comparing them to those who did not. The first set of individuals is generally referred to as the treatment group while the second set is the comparison group.

Taken as a whole, experimental designs are based on the idea that, in the absence of the intervention, individuals from the treatment group would have had an outcome (level of literacy in an educational program, glycemic control in the diabetes example) similar on average to that of the comparison group. One thus needs to find a relevant and adequate comparison group that resembles on average the treatment group in everything but the fact of receiving the intervention. In practice, this proves to be an issue as well. Estimating the impact of a program at the average level does necessarily mean that the confounding factors are held constant in comparisons. In many cases, confounding factors can influence the results at the group level, hence the need to find proper criteria for allocating individuals to the groups.

Consider for instance a federal government that wishes to experiment an educational policy in a subset of States. Low-income families would get free access to a new and special educational policy aimed at fighting illiteracy. The policy question associated with that treatment group is: how would participants have fared without the program? Concurrently, the comparison group could be the low-income families from a subset of other States within the Federation: how would have they fared with the program? However, mere absence of exposure to the treatment is a poor criterion for a comparison since any difference between the outcomes of the exposed and non-exposed can be attributed to both the program and differences due for instance to different systems of social assistance from one State to the other, prevalence of poverty, gender discrimination, tolerance for child labor, etc.

Differences between the treatment and comparison groups may also come from self-selection if the program is not mandatory. Subjects select themselves into a group, causing a biased sample. For instance, children who register to a non-compulsory educational program can belong to the most motivated ones. Another example is when one compares smokers with non-smokers to assess the effect of smoking. This can be hazardous since those who smoke may behave differently with respect to other items as well (e.g., alcohol and food consumption). Differences in subject adherence to a mandatory treatment can also be observed. A treatment can be painful or demanding, involve out-of-pocket or transportation costs that could be an obstacle for low-income patients, etc. On top of these considerations, selection errors may occur when wrong or not fully adequate criteria are used to identify the population of interest (coverage errors). Concurrently, the sample under study may not accurately represent the population in question (sampling errors, measurement errors, etc.). Overall, those selection biases blur the results of the experiment and may prevent isolating the average treatment effect.

By definition, a selection bias occurs when the subjects in a study differ from the treatment group to the comparison group or when they do not perfectly represent the larger population from which they are drawn. If there are important differences, the results of the experiment may be biased. It is thus crucial to identify, isolate or at least control for selection biases. A natural tool for doing so is randomization. Broadly speaking, a randomization process allocates participants to the experiment in such a way that they have the same chance of being assigned to either the treatment group or the comparison group. The groups should thereby share similar characteristics on average, in which case the comparison group is also termed a “control group”. Randomization is thus a constitutive part of experiments. Unlike observational studies, confounding factors are controlled for *ex ante*, which allows direct causal inference.

Randomized controlled experiments are often considered as the most rigorous way of testing a causal relationship. A distinction is however made between a “clinical trial” and a “field experiment”. Clinical trials evaluate biomedical or health-related outcomes with the aim of assessing the effectiveness of a treatment. A field experiment generalizes the approach to any type of intervention. As the name suggests, it is a form of investigation performed outside the laboratory in which units of observation (e.g., individuals, municipalities, etc.) are randomly assigned to treatment and control groups. This type of investigation has been booming recently, with variations in protocol that evidence that there is still need for harmonization in their designs. Their implementation may encounter many difficulties which, most of the time, are dependent on the context in which they are used.

The experimental methodology is highly indebted to health technology assessment (evaluation of properties, effects, and/or impacts of drugs, medical devices, screening, vaccination, etc.). For this reason, our presentation of experiments mostly rests on the clinical trial methodology. Medical experiments or clinical trials are at the very heart of research in medicine. They are a necessary condition for the development of any new drug or medical device and often a legal requirement if they are to be authorized by medical authorities and further reimbursed by mutual funds or social security. For instance, applications for marketing authorization for human medicines in the European Economic Area must have been carried out in accordance with the requirements set by the European Medicines Agency. Because clinical trials have since long a mature and robust framework, they should serve as the “gold standard” or benchmark when using their methodology outside the field of medical research, even if admittedly that methodology should adapt to the characteristics and circumstances of the so-evaluated public policies.

Medical research can be subcategorized into a set of four phases, from early investigation to extensive experimentation. To begin with, what is termed the “phase I study” attempts to estimate the pharmacokinetics and pharmacodynamics of drugs and they are tested on healthy volunteers or patients already under existing standard treatment. Phase I is mostly observational and it aims at assessing tolerability, toxicity and drug activity, usually through dose escalation. Escalation often obeys a Bayesian design, with the investigator’s prior estimate of toxicity/activity updated and assigned to the next participants. What are termed “phase II studies” concentrate instead on the biological effects of the drug and they may involve controls (concurrent control groups, not necessarily randomized, historical group data, and individual history of patients in the treatment group). It is a kind of individual before-after evaluation with as strict as possible a control of environmental factors. By providing information and data on the response rate, namely the biological activity of the drug on patients, they prepare the ground for “phase III studies”. Phase III trials are the randomized controlled experiment per se. They should be able to provide relevant outcomes (biomarkers such as blood pressure, progression free survival, recovery rate, etc.) for both exposed and non-exposed groups. Longer-term evaluation, or “phase IV studies”, can also be conducted after the approval of the drug or medical device by the regulatory agency in charge of its endorsement. They involve the real-life follow-up of a larger cohort of patients using the authorized drug. The aim is progressively to build a retrospective knowledge of the disease treatment and of the side-effects of the drug as it is used in real-life monitoring.

Randomized clinical trials (phase III) are the most accomplished phase of medical research. Experiments yield average policy effects contrasting the outcome in the treatment group to the outcome in the control group. The first aim of an experiment is thus to assess the average impact of the public project through the comparison of the respective means. The trial also provides indicators of policy effects or outcomes such as event occurrence (e.g., number of medical complications in a screening program), relative risk or odds ratios (e.g., risk of re-offense or school dropout). Those indicators are built over the whole time span of the experiment, without indication of the timing of events. To go farther, event analysis through survival curves is motivated by the fact that usual comparisons of means or proportions with chi-squares or equivalent statistics raise problems when the length and timing of individual observations differ among participants. Seemingly comparable final outcomes may hide significant differences in the patterns of evolution from one group to another. The Mantel-Haenszel test for conditional independence proves to be very useful in such cases.

Note that the practical operation of randomization can be tricky and it depends both on the assessed program and on the context in which it is to be enacted. Naturally, randomization is best controlled and structured in health technology assessment. It nevertheless remains that the very notion of experiment may raise concern as it may have considerable bearing on the individual welfare of exposed and non-exposed patients. This is why the ethics of experimentation are a constitutive part of that evaluation tool and should not be neglected. When randomization is definitely not a feasible protocol for organizational or ethical reasons, quasi-experimental techniques such as difference-in-differences, propensity score matching, regression discontinuity design or instrumental variable estimation can prove quite powerful substitutes. The last chapter of this book details those procedures.

The remainder of the chapter is organized as follows. Section 13.2 introduces methods for the random allocation of subjects to the treatment and control groups and also deals with the ethics of randomization. Section 13.3 presents the usual statistical tests to assess the significance of a treatment. Section 13.4 explains how to compute the statistical power of the test, i.e. the probability of detecting a predefined clinical significance. Section 13.5 is about sample size calculation and how to determinate the minimum number of subjects to enroll in a study to achieve a given statistical power. Section 13.6 offers additional but illuminative indicators of the effect of the evaluated policy intervention. Section 13.7 proposes an extended experimental framework where the survival patterns of the treated and control groups can be compared through Kaplan-Meier curves. Last, Sect. 13.8 explains how to perform the Mantel-Haenszel test in order to compare those curves.

13.2 Random Allocation of Subjects

A trial is basically speaking a comparative study between two (or more) groups of subjects (patients, target population of a public policy, etc.) that are chosen from a population of reference. Figure 13.1 offers an illustration. This population of reference is defined through general clinical criteria which have to be carefully discussed and evaluated (e.g., young adults suffering from diabetes). It is also the task of the investigator to select a sample that is as representative as possible of this population of reference. The theoretical sample size (i.e. the number of patients or experimental units required for the trial) will partly depend on the size of the population of reference, and partly on the precision the investigator would like to achieve (statistical criteria). The time and budget constraints associated with data collection are also determinant factors. Several sample size calculators and sampling procedures exist. In this respect, the reader can refer to Sect. 13.5.

A particularity of clinical trials is that they generally set guidelines about who can participate. The construction of a sample is thus often dependent on the context of the experiment. Inclusion and exclusion criteria are designed with the help of clinical experts and statisticians in order to select the most suitable participants. Inclusion criteria can for instance be related to age, gender, current medications, etc. Exclusion criteria may concern pre-existing medical conditions, administrative matters, or exclude subjects that are for instance too ill to participate in the experiment. The focus is not about sample representativeness only (quality of subject selection), but also on operationality and whether the experiment will suffer from additional complications (e.g., complex enrolment process, protocol violation, censoring). As a matter of fact, if the inclusion and exclusion criteria are too strict, the final sample can be non-representative of the population of reference (see Fig. 13.1).

The aim of the experiment must also be clearly defined. An experiment is usually designed to test one or several hypotheses and explore a particular set of data. If the investigator intends to collect socio-economic data or have evaluative questionnaires filled in by patients (for instance quality of life inquiry), then this must be conceived and planned from the beginning. Once the experiment is on its way, the protocol should not be altered anymore. An exception is when the

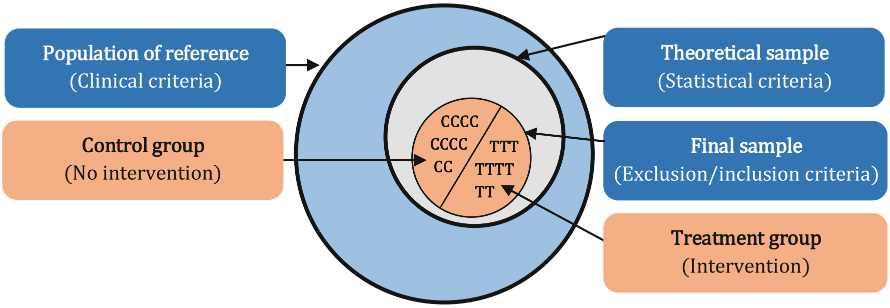


Fig. 13.1 From the population of reference to randomized groups

experiment must be terminated, e.g., in case of full and early success, conversely because of harmful effects of the drug endangering the life of patients (mortality exposure), or because of effects strongly compromising their health condition (morbidity exposure).

Once a subset of subjects has been selected, the next step is to randomly assign them to different groups. In practice, because of limited space, time constraint or for safety reasons, the number of groups, subgroups or blocks, can be greater than two. Yet, the standard practice is to allocate equal numbers of patients to treatment and control blocks. For this reason, and without loss of generality, we will assume in the remaining of the chapter that there are two groups only, namely the control group and the treatment group, also referred to as "control arm" and "treatment arm" in health. A control trial is about how the treatment group (exposed to the treatment or the policy) fares in comparison with the control group (not exposed to the treatment or the policy). A randomized control trial is such that the assignment to the treatment group or to the control group follows a random process ensuring an equal likelihood of being assigned to either group. It warrants comparability between the two groups. Covariates like prognostic factors, personal and environmental characteristics of the patients should on average have the same magnitude and direction in both groups (even though groups can never be perfectly balanced for all relevant covariates).

The randomization process must ensure that the allocation of participants between the two groups (exposed and non-exposed) is not influenced by the investigators or the participants, on the ground of explicit criteria (e.g., a bad prognosis may encourage the patient to ask for an innovative treatment or may deter the investigators from testing a treatment) or of implicit criteria (e.g. empathy or conversely suspicion of bad observance). It is thus important to define *ex ante* the degree of information of the parties involved, namely the investigators and the participants. In this respect, three cases may arise. First, the experiment can be un-blinded or open, in which case the allocation of participants to the control and treatment arms of the trial is common knowledge. This is often the case with surgical procedures. The open design is susceptible to biases like the strategic reporting of adverse events or psychological effects associated with being treated or not. Second, single-blinded trials are such that only the investigators are aware of which arm of the trial, control or treatment, each participant is allocated to. Biases lie mostly with the investigators, in their reporting and particularly in their advice to patients. For instance, they may provide compensatory guidance for patients assigned to the control group, which may affect the comparison among groups. Finally, double-blind allocation is recommended especially in trials of drugs: they minimize preconceptions and biased reporting for both the investigators and the participants. Neither the medical staff nor the patient know which group the patient belongs to. This is meant to neutralize psychological effects associated with the administration of the new drug (e.g., enthusiasm on the part of the doctor, optimism on the part of the patient, better adherence associated with greater expectations, etc.) and the possibly opposite psychological effects for those patients assigned to the control group who do not benefit from the assumed innovation.

Randomization can follow a fixed allocation procedure or be adaptive. In the first case, group assignment obeys a pre-specified probability that remains constant throughout the experiment. In the second case, the allocation probability changes as the experiment progresses. In what follows, we will consider only examples of fixed allocation procedures. Before that, we should first rule out a method, the "assignment by alternating sequences", which is sometimes used but should not be. The procedure goes as follows. The first patient stepping in is randomly allocated to one group (e.g., treatment group T), the second one is allocated to the other group (here control group C), the third one to the first group again, and so on, so that the assignment follows a sequence $TCTCTC\ldots$ From a purely statistical point of view, each subject has the same probability of being assigned to the treatment group. Yet, under this procedure, the person in charge of the assignment knows the next assignment and, for this reason, could influence which subjects are allocated to which group. Even for a double blind trial, the entire sequence is known as long as one subject of the chain has been identified.

Among the randomized fixed allocation procedures, "simple randomization" is by far the easiest method. A uniform random algorithm generates a number in interval $[0, 1]$. For a cut-point p, the subject or participant draws a number $x \in [0, 1]$. If $x \leq p$ (respectively $x > p$) then he or she is allocated to the treatment group (respectively to the control group). Most of the time, the investigator chooses an equal assignment of patients among groups so that $p = 50\%$. The approach is somewhat equivalent to the tossing of an unbiased coin. For instance, if for a patient the coin turns up tail, then he or she is allocated to the treatment group T. If it turns up head, allocation is to the control group C. With $p = 50\%$, a long run assignment with steady patients' enrollment will lead approximately to groups of equal size. The procedure is based on the Law of Large Numbers. This latter suggests that, given a sufficiently large number of subjects, the share of subjects assigned to each group should converge toward a predictable average proportion, here p. This result however no longer holds when the sample size is small. In that case, one may face the risk of imbalances in group size and average pre-trial characteristics.

In order to control for potentially serious imbalances in group sizes, one may use "blocked randomization". Blocked randomization encompasses several algorithms. One of the simplest designs randomly assigns participants to blocks of a given even size, for instance $s = 4$, with the constraint of containing two Ts (individuals allocated to treatment) and two Cs (individuals allocated to control). For a given block and a list of enrolled participants $i = 1 \ldots 4$, there exists a total of six combinations of group assignment:

$$TTCC, TCTC, TCCT, CCTT, CTTC, CTCT$$

One combination is selected at random. For instance, if the third assignment is selected:

$$\begin{array}{cccc} T & C & C & T \\ i=1 & i=2 & i=3 & i=4 \end{array}$$

Then, the procedure moves on to the next block. If the trial size is not a multiple of the block size, then the last block is incomplete which may cause imbalance in this block. Yet, this imbalance is necessarily small and, by construction, the two groups have approximately an equal size. In the case of a double blind trial, the T and C denominations are hidden behind anonymous group names A and B for instance.

Another frequent source of imbalance relates to a large variance of individual risk factors in relation to their individual characteristics (e.g., the smoking history of the patient). If this is likely to be the case, then participants can be stratified with respect to prognostic factors relevant to the clinical frame of the experiment. Stratification may occur at the inclusion: for instance, it can be appropriate to consider first the two subsets of men and women separately and then randomize them to the control and treatment arms so that both groups contain an equal number of males and females. When imbalances involve several prognostic factors likely to influence the outcome of the trial, then stratification can take place within the randomization process itself. The first step consists in identifying those prognostic factors that will be used as stratification criteria. The investigator should avoid selecting too many of them in order to preserve the operationality of the assignment procedure. Table 13.1 shows an example with three prognostic factors (age, sex and body mass index (BMI) excess) of early diabetes. Each factor gets a level classification. The choice of the levels is based on clinical considerations except for obvious factors like gender. Age has two levels (intervals $[15, 19]$ and $[20, 24]$). The BMI excess factor has three levels (low, moderate and high). This leaves us with twelve $(2 \times 2 \times 3)$ strata. The second step consists in the allocation of subjects among the treatment and control groups per se. If we keep a block size $s = 4$, we can proceed to blocked randomization within strata as in Table 13.1. From the second column, note that the strata can be of different size (in accordance with the sample and population characteristics). The number of blocks in each stratum is thus different. When the number of subjects in a stratum is not a multiple of the block size, then the last block is incomplete.

Randomized controlled trials may suffer from two difficulties: noncompliance (or no-shows) and missing outcomes. For this reason, the standard protocol for inclusion is the intention to treat (ITT) procedure: all participants are randomized and the associated encountered events should be accounted for throughout the study. Admittedly, it may happen, during the experiment, that a number of patients do not fit the inclusion criteria (although they seemed to at the time of inclusion). Similarly, patients may not completely comply with the protocol or may unintendedly fail to follow it closely. However, excluding them would severely bias the experiment since such behaviors or events are likely also to occur if the treatment is finally selected and implemented in real life.

ITT experiments involve treatment and control groups that may not be of constant size as participants may move from one group to another. Those

Table 13.1 Stratified randomization

Strata	Number of subjects	Age	Gender	Excess BMI	Number of blocks	Group assignment (example)
1	78	[15; 19]	M	Low	19 of size 4 + 1 of size 2	TTCC, CTCT, . . ., CC
2	93	[15; 19]	M	Moderate	23 of size 4+1 of size 1	CTCT, CTTC, . . ., T
3	123	[15; 19]	M	High	30 of size 4+1 of size 3	CTCT, TTCC, . . ., CTC
4	65	[15; 19]	F	Low	16 of size 4+1 of size 1	TCCT, CTCT, . . ., T
5	47	[15; 19]	F	Moderate	11 of size 4+1 of size 3	TCTC, TCTC, . . ., TCC
6	89	[15; 19]	F	High	22 of size 4+1 of size 1	TCTC, TTCC, . . .,C
7	210	[20; 24]	M	Low	52 of size 4+1 of size 2	TTCC, CTTC, . . .,CC
8	198	[20; 24]	M	Moderate	49 of size 4+1 of size 2	CCTT, CTTC, . . .,TC
9	234	[20; 24]	M	High	58 of size 4+1 of size 2	CTCT, TTCC, . . .,CT
10	301	[20; 24]	F	Low	75 of size 4+1 of size 1	CTTC, CTTC, . . .,T
11	214	[20; 24]	F	Moderate	53 of size 4+1 of size 2	TCCT, TTCC, . . .,TT
12	101	[20; 24]	F	High	25 of size 4+1 of size 1	TTCC, TCTC, . . ., C

participants are labeled “crossovers”. In clinical trials, it is customary to designate a patient who goes from control to treatment as a “drop-in”; conversely, a patient switching from treatment to control is a “drop-out”. This type of experiments is often opposed to so-called “per protocol analyses” (PPA). PPA removes lost or non-adherent patients from the statistical and clinical analysis which leads to serious concerns about the validity of the subsequent conclusions as to the efficacy of the treatment. The denomination itself is misleading as it wrongly conveys the idea that the PPA would be the preferred one as it would “stick to the protocol” whereas it does not.

Last, randomized control trials cannot but raise ethical questions since they deal with matters of personal integrity and welfare, particularly in the field of health. When experimenting a new drug within a randomized control trial, about one-half of the patients do not have the opportunity to benefit from it although the investigator expects it to bring health improvement. Conversely, if for instance the control group receives a standard treatment where side effects are reasonably mastered, the treatment group may face harmful unknowns, unexpected side effects that were not detected during the phase II of the trial. Ethical issues are present not only in health but also in development, education, etc. Think for instance of the case of a policy towards disadvantaged children who would be offered safe shelter and balanced diet in a specially designed educational facility while children in the control group would be left to their condition.

Depriving participants from treatment may raise ethical concerns that cannot be eluded in the methodological design of any experiment, be it in health or other fields of intervention. They can be classified into foundational, operational and reporting concerns. Admittedly, the following remarks do not contend to exhaust the subject, and they will concentrate on foundational concerns since operational and reporting problems are largely context-dependent.

Foundational concerns primarily relate to what is labeled "clinical equipoise", namely the uncertainty surrounding the benefits and harms of a new drug. Clinical research is fundamentally meant to move from total ignorance ("We don't know what we don't know") to recognized ignorance ("We know what we don't know"). What we do not know is the differential benefit/harm ratio of a suggested new treatment or intervention vis-à-vis the existing ones. That ignorance can be explored through phase I and phase II clinical trials, and then reduced through phase III randomized clinical trials. All these phases contribute to provide answers with respect to what we know about the relative merits of the new drug or treatment. Yet, they raise ethical concerns. From phase I to phase III, participants who are exposed to treatment run the risk of unexpected or expected (but with unknown frequency) harmful events. Thus, even if it seems that the research question requires a clinical trial, one should always keep in mind the question of individual integrity. On top of these considerations, participants from the control group face a potential loss of opportunity since by construction they do not benefit from the treatment (but conversely are not exposed to its adverse events). This is why informed consent is required from participants in both arms of the trial under the close scrutiny of (now always mandatory) ethics committees.

When balancing a health research program with the individual welfare of participants, one should always have as a primary objective a Pareto improvement. Namely, we search for a net increase in the welfare of as many participants as possible (reasonably those from the treatment group in phase III if the intervention is clinically relevant, volunteers or patients in phases I and II, keeping in mind that early experiment phases are by essence risky and potentially harmful). Concurrently, the "do no harm" maxim should apply so that no one is made worth by his or her participation in the trial. Due to the inherent risks of experimentation, the design of the trial should contribute to minimize individual welfare losses so as to tend to a Pareto-improving situation.

13.3 Statistical Significance of a Treatment Effect

Once the participants have been assigned to their respective groups, the subjects in the control group are either treated with a placebo (placebo controlled trials) or with the standard treatment against which the new treatment is assessed. In both cases, the fundamental point is that the result of the allocation of a subject to a group is by nature unpredictable: randomization precludes selection biases. It nevertheless remains that the balance between the two groups, in terms of risk or prognosis factors, may not always be ensured. However, the comparability of the covariates for the two groups increases with the trial size. Randomization generates groups of similar (or better, equal) size with similar entry characteristics on average. The differentiated care paths of the two groups should then illuminate the relevance of introducing the new drug.

How can we measure the treatment effect? A simple numerical example illustrates the answer. Consider a public health obesity policy that imposes regular physical activity to the treatment group while the control group is not constrained to attend such a program. The health indicator is the average overweight (in kilograms), the reference being the standard body mass index. Table 13.2 provides the data for a sample of 60 participants randomly assigned to the treatment group and to the control group. The randomization procedure has ensured equal size of the groups. Patients allocated to the control group are coded "0" while patients receiving the treatment are coded "1". A quick glance at the data shows that patients exposed to treatment usually have a lower overweight at the end of the experiment.

In our example, what can be measured is the average treatment effect (*ATE*), calculated as the average outcome in the treatment group (T) minus the average outcome for the non-exposed patients of the control group (C). Let $\bar{x}^T$ denote the mean outcome in the treatment group and $\bar{x}^C$ the mean outcome in the control group, with respective individual outcomes x_i^T and x_i^C. We have:

$$ATE = \frac{1}{n_T}\sum\nolimits_{i=1}^{n_T} x_i^T - \frac{1}{n_C}\sum\nolimits_{i=1}^{n_C} x_i^C = \bar{x}^T - \bar{x}^C$$

where n_C and n_T denote the group sizes, respectively. In the case of example 1:

$$ATE = \frac{1134}{30} - \frac{1328}{30} \approx 37.8 - 44.3 \approx -6.5$$

At first glance, the program seems to be effective since on average, patients from the treatment group reduce their overweight by 6.5 kg. Yet, those results are dependent on the selected sample. Statistical methods must be used to make inference about the population of reference. To do so, we compute the standard error (*se*) of the average treatment effect:

$$se(ATE) = \sqrt{\frac{(s^T)^2}{n^T} + \frac{(s^C)^2}{n^C}}$$

where s^T and s^C are the sample standard deviations of the treatment and control groups, respectively:

Table 13.2 Raw results of an experiment: example 1

Patient	Overweight (kg)	Group (1=Treatment; 0=Control)
1	43	1
2	56	0
3	28	1
4	33	0
5	49	1
6	49	0
7	21	1
8	38	1
9	29	0
10	43	1
11	47	0
12	37	1
13	36	1
14	33	1
15	55	0
16	49	1
17	51	1
18	42	0
19	36	0
20	35	1
21	34	0
22	60	0
23	58	0
24	42	1
25	37	0
26	44	1
27	49	0
28	53	0
29	52	0
30	30	0
31	43	1
32	27	1
33	31	1
34	30	0
35	33	1
36	36	0
37	49	0
38	40	1
39	43	0
40	47	0
41	34	1
42	38	1
43	53	0

(continued)

Table 13.2 (continued)

Patient	Overweight (kg)	Group (1=Treatment; 0=Control)
44	48	1
45	53	1
46	59	0
47	30	0
48	34	0
49	26	1
50	49	0
51	34	1
52	58	0
53	39	1
54	45	1
55	43	0
56	30	1
57	27	1
58	34	0
59	43	0
60	37	1

$$s^j = \sqrt{\frac{1}{n^j - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad j = T, C$$

For instance, in our sample we have $s^T \approx 8.10$, $s^C \approx 10.00$, we therefore find:

$$se(ATE) = \sqrt{\frac{(8.10)^2}{30} + \frac{(10.00)^2}{30}} \approx 2.35$$

It is an estimate of the variability of the ATE. As we will see below, this value can be used to check the significance of the average treatment effect.

Descriptive statistics are useful on their own, as they provide a description of the sample, but yet cannot be used to generalize the results to the population of reference. In contrast, inferential statistics allows hypotheses to be tested and can be used to determine if observed differences between groups are real or occur simply by chance. First, one must make an assumption about the population parameters, denoted μ^T and μ^C hereafter, in the treatment and control groups, respectively:

$$H_0 : \mu^T = \mu^C \text{ } (ATE \text{ is not significant})$$

This assumption is referred to as the null hypothesis H_0. It assumes that the observed difference is due to chance only. Second, one needs to define the alternative hypothesis H_1:

$$H_1 : \mu^T \neq \mu^C \ \ (ATE \text{ is significant})$$

The alternative hypothesis states that the observed difference is the result of some non-random cause. A statistical test is then implemented to determine whether there is enough evidence to reject the null hypothesis.

The test is usually referred to as a "two-sample *t*-test for equal means". Formally, the test statistic is defined as a *t*-score:

$$t^* = \frac{ATE}{se(ATE)} = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{(s^T)^2}{n^T} + \frac{(s^C)^2}{n^C}}}$$

A Student *t*-distribution approximates the way this statistic is spread. A common way to test a hypothesis is to rely on a 95% confidence level or, reciprocally, a 5% significance level. This confidence level defines an acceptance region which contains the values of the test statistic for which we fail to reject the null hypothesis. Figure 13.2a provides an illustration. The shape of the Student distribution depends on a specified number of degrees of freedom (for instance, 55.6 in Fig. 13.2). This number approximates the "true" sample size by taking into account the number of observations required for computational purposes. As the number of degrees of freedom decreases, the tails of the distribution become larger, and so does the acceptance region. On the other hand, as the number of degrees of freedom increases, the *t*-distribution approaches the normal distribution with mean 0 and variance 1.

In Fig. 13.2a, to achieve a significance level of 5%, the absolute value of the test statistic t^* must be outside the region of acceptance. Because the Student distribution is symmetrical, one usually compares the test statistic t^* to a single critical value which delimits the acceptance region. Let $t_{\alpha/2}(df)$ denote this value, where α denotes the significance level, and *df* stands for the number of degrees of freedom. For the two-sample *t*-test, the number of degrees of freedom is given by:

$$df = \frac{\left(\frac{(s^T)^2}{n^T} + \frac{(s^C)^2}{n^C}\right)^2}{\frac{\left((s^T)^2/n^T\right)^2}{n^T-1} + \frac{\left((s^C)^2/n^C\right)^2}{n^C-1}}$$

If the test statistic t^* is higher in absolute value than the critical value $t_{\alpha/2}(df)$, we reject the null hypothesis: the average treatment effect is significant. Reciprocally, facing a non-significant difference in means, the analyst can conclude that there is absence of proof of a significant effect of the treatment, but that does not prove the absence of an effect. Hence the saying: "absence of proof is not proof of absence".

Note that the previous test is defined as a two-tailed (or two-sided) test. It is also possible to perform what is termed a one-tailed (or one-sided) test. We could test for instance:

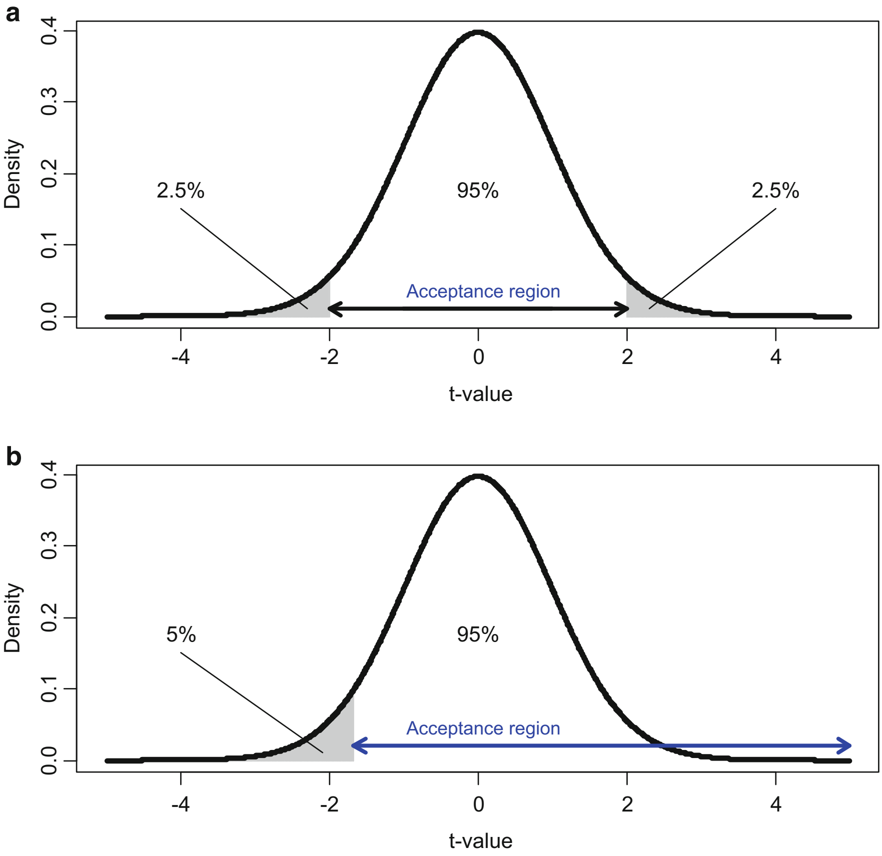


Fig. 13.2 The Student distribution ($df = 55.6$). (**a**) Two-tailed test, (**b**) one-tailed test

$$H_0 : \mu^T \geq \mu^C \ (ATE \text{ is not significant})$$

$$H_1 : \mu^T < \mu^C \ (ATE \text{ is strictly negative})$$

Here, we are interested in one side of the Student distribution only, as illustrated in Fig. 13.2b. In that case, we implicitly assume that the average treatment effect cannot be positive. The critical value is now denoted $t_\alpha(df)$ and, by construction, the acceptance region is smaller than previously. Hence, one-tailed tests make it easier to reject the null hypothesis and to detect an effect. If one cares about not missing an effect, the approach can be appropriate. However, choosing a one-tailed test for the purpose of attaining significance can be suspicious. For instance, choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is scientifically questionable. With a one-tailed test, one also makes an assumption about the direction of the relationship and completely disregards the possibility of a relationship in the other direction.

Coming back to example 1, a two-sample *t*-test can be performed to test whether the overweight for the treatment group is on average significantly different than the overweight in the control group. The null hypothesis ($H_0 : \mu^T = \mu^C$) states that the treatment is not more effective than standard practice in controlling overweight: there is no difference between means μ_T and μ_C in population. On the other hand, the alternative hypothesis ($H_1 : \mu^T \neq \mu^C$) states that there is a difference. If a significant difference is found, then the result from the sample ($ATE = -6.5\text{kg}$) can be generalized to the population of reference.

The test statistic is computed as follows:

$$t^* = \frac{ATE}{se(ATE)} \approx \frac{-6.5}{2.35} \approx -2.8$$

The number of degrees of freedom is:

$$df = \frac{\left(\frac{8.10^2}{30} + \frac{10.00^2}{30}\right)^2}{\frac{\left(8.10^2/30\right)^2}{30-1} + \frac{\left(10.00^2/30\right)^2}{30-1}} \approx 55.6$$

By definition, we have $t_{2.5\%}(55.6)$=2.004. This value can for instance be obtained in Excel using the *TINV*(5%, 55.6) command. The test statistic t^* is thus higher in absolute value than the critical value. The null hypothesis is rejected: the two means are significantly different. Provided that the intervention is cost effective and its budget burden is sustainable, then it can be implemented in the target population.

Figure 13.3 displays the R-CRAN codes of the two-sample *t*-test. The database (saved as a *.csv* file on disc *C*:) is uploaded using the command $read.table$ and, for simplicity of exposition, saved under a new name *D*. The two groups are distinguished using the entries $[D\$Group == 1]$ and $[D\$Group == 0]$ so as to get their summary statistics. This provides the respective means and sample standard deviation which can for instance be used to compute the *t*-score and the number of degrees of freedom. The test is implemented using the command $t.test$. We have 55.6 degrees of freedom and the test statistic is found to be 2.7535 in absolute value. The critical value is obtained with the *qt*(0.975,55.6) command. As previously, we conclude that there is a significant overweight reduction when patients benefit from the intervention instead of standard practices. This is alternately confirmed by a *p*-value of 0.0079 lower than 5%. The *p*-value gives the level of significance for which one would be indifferent between rejecting and not rejecting H_0. In other words, if the *p*-value is less than the significance level α=5%, the null hypothesis is rejected. On the other hand, if the *p*-value is greater, then the null hypothesis is not rejected.

Until now, we have been considering tests of equality of means and consequently have been using a Student test. The chi-square test of independence should

```
> D=read.table("C://mydataTrial.csv",head=TRUE,sep=";")
> head(D)
  Patient Overweight Group
1       1         43     1
2       2         56     0
3       3         28     1
4       4         33     0
5       5         49     1
6       6         49     0
> # Summary statistics
> summary(D$Overweight[D$Group==1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21.0    33.0    37.5    37.8    43.0    53.0
> summary(D$Overweight[D$Group==0])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   34.50   45.00   44.27   52.75   60.00
> sd(D$Overweight[D$Group==1])
[1] 8.095976
> sd(D$Overweight[D$Group==0])
[1] 9.996321
> # Two-sample t-test
> t.test(Overweight~Group,D)

        Welch Two Sample t-test

data:  Overweight by Group
t = 2.7535, df = 55.6, p-value = 0.007951
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.761199 11.172135
sample estimates:
mean in group 0 mean in group 1
       44.26667        37.80000
> qt(0.975,55.6)
[1] 2.003559
```

Fig. 13.3 Comparing treatment and control groups in R-CRAN: example 1

be used when one wants to test for equality of proportions between two samples (e.g., two baseline risks). The first step is to create a two-way table. Consider for instance a randomized controlled trial that involves a single period of time. The two arms of the trial consist of a set of schoolchildren in a standard educational system (control group) and a set of schoolchildren hosted in a specialized care center (treatment group). The policy context is to avoid adverse events, i.e. reoffenders in the target population of children from disadvantaged backgrounds. Allocation to the two groups has been randomized. We assume that no individuals are lost or censored during the experiment: they are all present from the beginning to the end of the trial and they are either reoffenders or not.

After one year of experiment, the count result is summed up as in Table 13.3(a). The size of the treatment and control groups is the same ($n^T = n^C = 200$). Over the whole sample (treatment plus control), the number of subjects with event (140) is lower than the number of subjects without event (260). The research question is whether there is a difference between the treatment group and the control group and whether the observed difference is statistically significant. For each group, the baseline risks are computed as follows:

$$p^T = \frac{60}{200} = 30\% \text{ and } p^C = \frac{80}{200} = 40\%$$

Table 13.3 Observed and expected frequencies: example 2

	Participants WITH adverse event during trial	Participants WITHOUT adverse event during trial	Total
(a) Observed frequencies			
Treatment group	60	140	**200**
Control group	80	120	**200**
Total	**140**	**260**	**400**
(b) Expected frequencies			
Treatment group	(140×200)/400=70	(260×200)/400=130	**200**
Control group	(140×200)/400=70	(260×200)/400=130	**200**
Total	**140**	**260**	**400**

The resulting difference is $30\,\% \, - 40\,\% \, = \, - 10\%$. As expected, the intervention generates a decrease in the number of adverse events. It remains to be seen whether this result can be inferred to the population of reference. One needs to implement a chi-square test of independence.

The test hypotheses are specified as:

$$H_0 : \pi^{\mathrm{T}} = \pi^{\mathrm{C}}$$

$$H_1 : \pi^{\mathrm{T}} \neq \pi^{\mathrm{C}}$$

where π^T and π^C denote the corresponding population parameters. The next step is to compute the test statistic:

$$\chi^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency in cell i, E_i is the expected frequency in cell i and C is the total number of cells. The critical value is obtained from a chi-square distribution. The number of degrees of freedom is given by:

$$df = (\text{number of row categories} - 1) \times (\text{number of column categories} - 1)$$

Although the alternative hypothesis is two-sided, the chi-square test is a one-tailed test. If the test statistic is higher than the critical value $\chi^2_\alpha(df)$, with $\alpha = 5\%$, we reject the null hypothesis H_0 of equality of proportions.

The chi-square test is based on a comparison between the observed frequencies, expressed in the two-way table, and the expected frequencies that would be observed under the null hypothesis. In our example, if the probability of an adverse effect were independent from the assignment to the control group or the treatment group, the joint frequencies would be those of Table 13.3(b), i.e. we would obtain a similar distribution of events. From the previous formula, the test statistic is computed as:

$$\chi^2 = \underbrace{\frac{(60-70)^2}{70} + \frac{(80-70)^2}{70}}_{\text{First column}}$$
$$+ \underbrace{\frac{(140-130)^2}{130} + \frac{(120-130)^2}{130}}_{\text{Second column}} \approx 4.40$$

The number of degrees of freedom is:

$$df = (2-1) \times (2-1) = 1$$

The critical values can been generated from Excel using the function *CHIINV* (5%, 1). We find:

$$\chi^2_{5\%}(1) = 3.841$$

The test statistic is greater than the critical value. We thereby reject the null hypothesis.

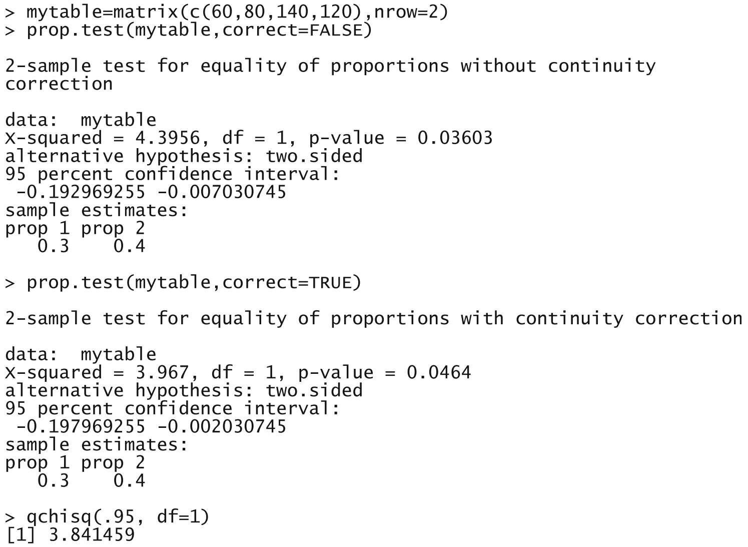
Note that for small data counts (in particular when one cell of the table has a count smaller than 5), Yates' continuity correction can be used as an approximation in the analysis of 2×2 tables. In that case, half the sample size is subtracted from all frequency differences:

$$\chi^2_{\text{Yates}} = \frac{1}{n} \sum_{i=1}^{C} \frac{\left(|O_i - E_i| - \frac{n}{2}\right)^2}{E_i}$$

In our example, the test statistic is computed as:

$$\chi^2_{\text{Yates}} = \frac{1}{400} \underbrace{\frac{\left(10 - \frac{400}{2}\right)^2}{70} + \frac{\left(10 - \frac{400}{2}\right)^2}{70}}_{\text{First column}}$$
$$+ \underbrace{\frac{\left(10 - \frac{400}{2}\right)^2}{130} + \frac{\left(10 - \frac{400}{2}\right)^2}{130}}_{\text{Second column}} \approx 3.96$$

Again, the test statistic is greater than the critical value, though now much closer to it. We still reject the null hypothesis. Yates' correction is however sometimes criticized as it tends to reduce the chi-squared value and, for this reason, may fail to reject the null hypothesis when it should be rejected (a so-called type II error as we shall see later).



```
> mytable=matrix(c(60,80,140,120),nrow=2)
> prop.test(mytable,correct=FALSE)

2-sample test for equality of proportions without continuity
correction

data:  mytable
X-squared = 4.3956, df = 1, p-value = 0.03603
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.192969255 -0.007030745
sample estimates:
prop 1 prop 2
   0.3    0.4

> prop.test(mytable,correct=TRUE)

2-sample test for equality of proportions with continuity correction

data:  mytable
X-squared = 3.967, df = 1, p-value = 0.0464
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.197969255 -0.002030745
sample estimates:
prop 1 prop 2
   0.3    0.4

> qchisq(.95, df=1)
[1] 3.841459
```

Fig. 13.4 Testing for equality of proportions with R-CRAN: example 2

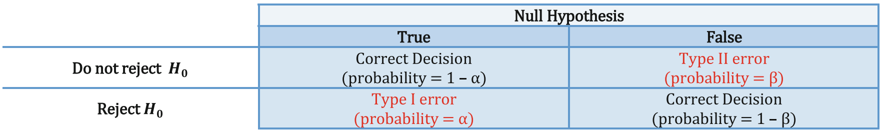
Figure 13.4 performs the test in R-CRAN using the function *chisq.test*. A two-way table is created using the *matrix* command, which is used along with *prop.test* to perform the test. The entry *correct* indicates whether Yates' continuity correction should be applied. As can be seen, the conclusion of the tests is the same. The value of the chi-square statistic is significant (the *p*-value is lower than 5%). Last, *qchisq*(.95, *df* = 1) offers the critical value. Note that the command yields the 95% confidence interval estimate of the difference between the proportions. Note also that no assumption is made about the direction of the effect (two-sided test). Yet, from the confidence interval, we have confirmation that the proportion of adverse events in the treatment group is lower than that in the control group.

13.4 Clinical Significance and Statistical Power

When performing a hypothesis test, two types of errors are possible: type I and type II, also referred to as "alpha" and "beta" errors. A type I error occurs when the null hypothesis is true and one erroneously rejects it. This type of error is related to the significance level α, which denotes the probability of making this type of error:

$$\alpha = \Pr\{\text{Type I error}\} = \Pr\{\text{reject } \mathrm{H}_0 \text{ when } \mathrm{H}_0 \text{ is true}\}$$

For instance, a significance level of 5% means that one is willing to accept a 5% chance that one is wrong when one rejects H_0. To decrease this risk, one must use a lower value of α, i.e., enlarge the region where H_0 is accepted. Yet, using a larger region of acceptance also means that one is less likely to detect a true difference, a



	Null Hypothesis	
	True	False
Do not reject H_0	Correct Decision (probability = 1 – α)	Type II error (probability = β)
Reject H_0	Type I error (probability = α)	Correct Decision (probability = 1 – β)

Fig. 13.5 Summary of type I and II errors

so-called type II error: the null hypothesis is false and one actually fails to reject it. The probability of making this type of error is usually denoted β:

$$\beta = \Pr\{\text{Type II error}\} = \Pr\{\text{reject } \mathrm{H}_1 \text{ when } \mathrm{H}_1 \text{ is true}\}$$

Often the literature refers to $1-\beta$ to denote the power of a statistical test. It is the probability that a Type II error is not committed.

Figure 13.5 illustrates the four possible cases. Type I errors are mostly about the statistical quality of the test, namely the control of random fluctuations in sampling. On the other hand, Type II errors concern the scientific quality of the test or "clinical significance", i.e. the identification of the magnitude of the effect of the intervention. Which type of error is more damageable? In most cases, as they do not want to see an effect where there is not, statisticians focus on the type I error and, to reduce this type of risk, choose a small level of significance α for implementing a test. However, when evaluating a medical treatment, missing an effect can also be detrimental to the patients as they may incur a loss of opportunity, hence the importance of accounting for type II errors.

Consider a public program aimed at minimizing the occurrence of an adverse event in a target population. It may be a new drug that diminishes side effects for the patients treated for a disease, or a new organization of institutional care for young offenders that would decrease the risk of re-offense. The sample statistics for the treatment and control groups are denoted p^T and p^C, while the (unobserved) population parameters are denoted π^T and π^C, respectively. Ideally, we expect that the treatment reduces the probability of adverse event, i.e., we would like $\pi^T < \pi^C$. The problem is that we only observe a sample difference $(p^T - p^C)$, which itself depends on the subjects that have been selected. In the case of a drug for instance, the adverse event for a given patient is triggered by his or her own clinical characteristics and the natural history of the pathology. Young offenders have their own personal history and personality characteristics which, coupled with triggering events, would lead them to re-offense. We may thus face type I and type II errors when testing $H_0 : \pi^T = \pi^C$ versus $H_1 : \pi^T \neq \pi^C$.

Imagine for instance an experiment where the sample gives $p^T = 7\%$ and $p^C = 15\%$. At first glance, the treatment group seems less likely to face the adverse event than the control group. In this case, a type I error would be the rejection of the null hypothesis whereas, in population, the two probabilities are equal (e.g., $\pi^T = \pi^C = 10\%$). The variation in probabilities observed at the sample level may be due to sampling error. The usual way for reducing this type of risk is to define an acceptance region so that the probability of observing such an error is sufficiently small. That small risk of error is usually kept to $\alpha = 5\%$. If the type I error is considered by the decision-maker as a risk that the society should not take, then the edge value should be lowered to 1%.

Let us now move on to the reverse situation and let us consider for instance an experiment where the sample gives $p^T = p^C = 10\%$. At first glance, the treatment group seems to face the same frequency of adverse events as the control group. In this case, a type II error would be the acceptance of the null hypothesis whereas, in population, the two probabilities would differ (say $\pi^T = 5\%$ and $\pi^C = 15\%$). Should this experiment be used for policy-making, the inability to reject the null hypothesis would imply that the policy associated with the intervention will not be implemented, leading to a loss of opportunity for the population of reference who would otherwise have benefited from the treatment. This loss of opportunity is an important feature of public policies since, by giving up the implementation of a program, the decision-maker misses the possibility of improving the welfare of the initially targeted population.

Often, for simplicity of exposition, the computation of the type II risk is presented in the context of a one-sample *t*-test. This type of test is used to determine whether the mean of a group differs from a specified value. For instance, we may want to know whether a treatment group differs or not from a general population. Imagine that we expect the mean of the treatment group μ^T to be lower than the mean μ_0 of the general population (e.g., we expect a reduction in the level of adverse effects). The test hypotheses can be defined as:

$$H_0 : \mu^T = \mu_0$$

$$H_1 : \mu^T < \mu_0$$

The population mean μ_0 is not always known, but can be hypothesized, e.g., from the observation of a control group. Similarly, we only know the sample mean for the treatment group, denoted $\bar{x}^T$ hereafter. In this case, a type I error means to erroneously conclude that $\mu^T < \mu_0$ while there is not effect. On the other hand, a type II error means to erroneously conclude that $\mu^T = \mu_0$ while there is an effect.

At this stage, to compute the type II error, we need to understand that, depending on the selected sample, our conclusions can be different. To account for that uncertainty, we need to assume a distribution for $\bar{x}^T$. Imagine for instance that the null hypothesis is true: we have $\mu^T = \mu_0$. Yet, we do not know μ^T and only have an approximation of it, that is $\bar{x}^T$. The problem is that in one given sample, we may find that the mean $\bar{x}^T$ is lower than μ_0, in another sample it can be higher, and so on. Given this uncertainty, we need to specify a confidence interval which is likely to include the unknown population parameter μ^T. If μ_0 belongs to this interval, then we do not reject H_0. For a 5%, significance level, this means that one is willing to accept a 5% chance that one is wrong when one rejects H_0. In practice, a convenient way to address this question is to normalize the problem and to rely on a standard normal distribution or a Student distribution to implement the test. Basically speaking, a test statistic is computed and compared to a critical value which defines

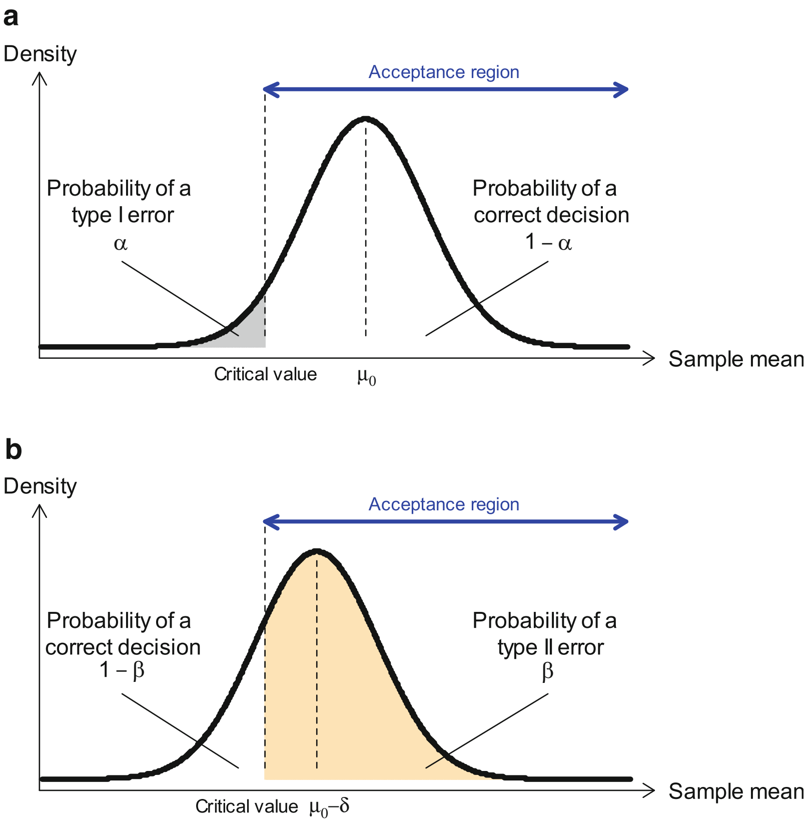


Fig. 13.6 Type I and type II errors. (**a**) Under the null hypothesis, (**b**) Under the alternative hypothesis

the acceptance region. The tricky issue is that the distribution in question also depends on which hypothesis is true, i.e. whether it is H_0 or H_1.

Under the null hypothesis, the probability of facing a type I error is defined as:

$$\alpha = \Pr\{\text{reject H}_0 \text{ when H}_0 \text{ is true}\} = \Pr\left\{\bar{x}^T < \text{critical value when H}_0 \text{ is true}\right\}$$

Figure 13.6a illustrates this case in the mean-density plane. The null hypothesis is assumed to be true and, for this reason, the distribution of the sample mean has a mean μ_0. The critical value is chosen so that the probability that $\bar{x}^T$ falls in the acceptance region is $1 - \alpha$ (confidence level). To implement the test, one generally focuses on a well-known distribution such as a t-distribution. Instead of focusing on $\bar{x}^T$, the focus is on the following statistic:

$$t^{*}=\frac{\bar{x}^{T}-\mu_{0}}{s^{T}/\sqrt{n^{T}}}$$

which denotes the standardized sample mean. The latter is computed as the difference between the sample mean $\bar{x}^T$ and the hypothesized mean μ_0 relative to the standard error of the mean $se=s^T/\sqrt{n^T}$, where s^T denotes the sample standard deviation and n^T the sample size. Under H_0, this statistic follows a Student distribution. The risk of a type I error is thus defined as:

$$\alpha=\Pr\left\{t^{*}<t_{\alpha}(df)\mid t^{*}\sim t\left(n^{T}-1\right)\right\}$$

where $t_\alpha(df)$ denotes the standardized critical value, with $df=n^T-1$. In practice, for this kind of test, it is best to use a Student distribution, instead of a normal distribution, whenever the population standard deviation is unknown. If the sample statistic is found to be lower than the critical value, then we reject the null hypothesis. This equivalently means that μ_0 does not belong to the acceptance region, i.e. that it differs from the treatment group mean.

Under the alternative hypothesis, the test statistic t^* follows a non-central distribution. Assume that $\delta>0$ denotes the true (unobserved) difference between the population means so that the treatment group has now a population mean equal to $\mu^T=\mu_0-\delta=\mu^*$. The type II error is given by:

$$\beta=\Pr\{\text{reject H}_1\text{ when H}_1\text{ is true}\}=\Pr\left\{\bar{x}^{T}>\text{critical value when H}_1\text{ is true}\right\}$$

This case is illustrated in Fig. 13.6b where a small difference δ is considered. The alternative hypothesis is assumed to be true and, for this reason, the distribution of the sample mean has now a mean equal to $\mu_0-\delta$. Compared to Fig. 13.6a, the distribution of the sample mean shifts slightly to the left, which generates a high probability of a type II error. As can be noticed, the probability β that $\bar{x}^T$ falls in the acceptance region is determined by the previously chosen critical value. Type I and Type II errors are therefore interrelated: as one increases, the other decreases. Moreover, the larger is the difference δ, the lower is the risk of a type II error and the higher is the power of the test.

Through standardization, we get:

$$\beta=\Pr\{t^{*}>t_{\alpha}(df)\mid H_{1}\}$$

Under H_1, we know that:

$$\frac{\bar{x}^{T}-\mu^{*}}{s^{T}/\sqrt{n^{T}}}\sim t\left(n^{T}-1\right)$$

Moreover, we have:

$$t^* = \frac{\bar{x}^T - \mu_0}{s^T/\sqrt{n^T}} = \frac{\bar{x}^T - \mu_0 + \mu^* - \mu^*}{s^T/\sqrt{n^T}} = \frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} - \frac{\delta}{s^T/\sqrt{n^T}}$$

The risk of a type II error can thus be expressed as:

$$\beta = \Pr\left\{\frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} - \frac{\delta}{s^T/\sqrt{n^T}} > t_\alpha(df) \;\middle|\; \frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} \sim t\left(n^T - 1\right)\right\}$$

Equivalently, we have:

$$1 - \beta = \Pr\left\{\frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} - \frac{\delta}{s^T/\sqrt{n^T}} < t_\alpha(df) \;\middle|\; \frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} \sim t\left(n^T - 1\right)\right\}$$

Thus, to calculate the power of a test, one must first compute the non-centrality parameter $\sqrt{n^T}\,\delta/s^T$. The non-central t-distribution is a generalization of the usual t-distribution. It describes the distribution of a test statistic when the null hypothesis is false. Figure 13.7 offers an illustration with different values of the non-centrality parameter. If the parameter is zero, the distribution is identical to a distribution in the central family. If the non-centrality parameter is nonzero, then the distribution shifts either to the left or to the right. Note also that the tails of a non-central distribution are larger from those of the central distribution.

Consider example 1 (see Table 13.2). The indicator of effect is the overweight and the treatment is meant to help patients decrease that overweight. For individuals exposed to the policy, the average sample overweight is $\bar{x}^T = 37.8$ kg. From the control group, we may infer the average overweight among the non-exposed: $\mu_0 = 44.7$ kg. The policy objective is to reach $\mu^T < \mu_0$. Under the null hypothesis, the test statistic is computed as:

$$t^* = \frac{\bar{x}^T - \mu_0}{s^T/\sqrt{n^T}} = \frac{37.8 - 44.7}{8.10/\sqrt{30}} \approx -4.66$$

The number of degrees of freedom is $n^T - 1 = 29$. If we arbitrarily choose a risk of a type I error equal to 5%, we obtain a critical value equal to $t_{5\%}(29) \approx -1.70$. The test statistic t^* is thus greater in absolute value than the critical value. At a 5% risk level, the null hypothesis is rejected: the effect is considered as significant. Figure 13.8 offers the program in R-CRAN. Note that the p-value ($3.188e-05$) is lower than that of the two-sample t-test (p-value = 0.0007951 in Sect. 13.3, Fig. 13.3), which equivalently means that the support for the effect is larger with the one sample t-test. There are two reasons for this: (1) the one sample t-test is one-sided only and (2) we assume that μ_0 is perfectly known (the uncertainty inherent to the control group is not taken into account here).

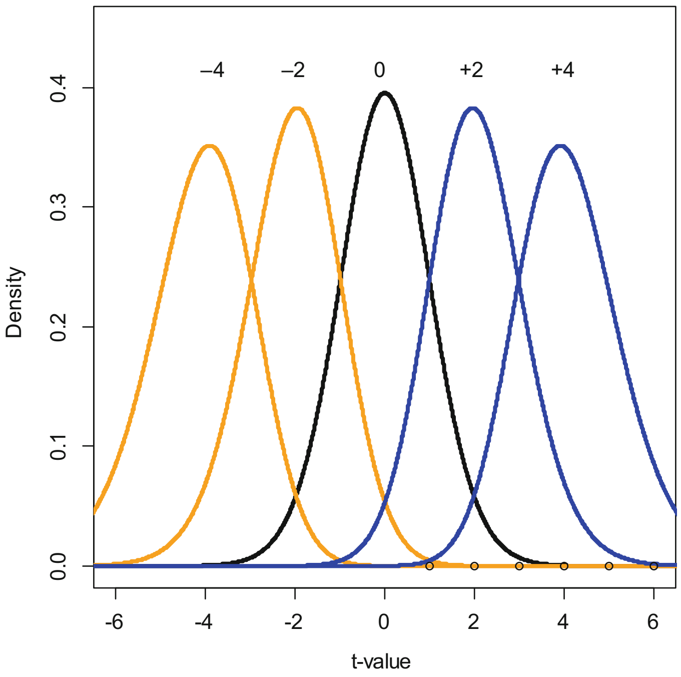


Fig. 13.7 Non-central Student distributions (df=29)

Beyond the statistical risk represented by type I errors, what is the scientific risk incurred if H_0 cannot be rejected, i.e. if we cannot prove the existence of the effect from the experiment although it presumably exists in population (or at least we would like to demonstrate it)? To answer that question, one must know the effect size that is scientifically required to significantly conclude that the impact of the treatment does matter. This kind of information is obtained from experts and it stands as a parameter for the evaluator, even though it can be varied in agreement with the experts in order to check the (univariate) sensitivity of results to changes in its value. For instance, let us assume that the weight decrease must reach $\delta = 3$kg to be clinically significant. The power of the test is obtained from:

$$1 - \beta = \Pr\left\{ \frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} - \frac{3}{8.10/\sqrt{30}} < -1.70 \,\middle|\, \frac{\bar{x}^T - \mu^*}{s^T/\sqrt{n^T}} \sim t\left(n^T - 1\right) \right\}$$

It is equal to the probability that a non-central t-distributed random variable with 29 degrees of freedom and non-centrality parameter 2.03 is lower than -1.70. To find this value, one can rely on the *pt* command in R-CRAN which gives the related distribution function for a set of non-centrality parameters. In Fig. 13.8, we use this command and find a statistical power equal to 63.19%. An equivalent way is to rely

```
> D=read.table("C://mydataTrial.csv",head=TRUE,sep=";")
> n=30
> sT=sd(D$Overweight[D$Group==1])
>
> # One sample t-test
> t.test(D$Overweight[D$Group==1],mu=44.7,alternative="less")

        One Sample t-test

data:  D$Overweight[D$Group == 1]
t = -4.6681, df = 29, p-value = 3.188e-05
alternative hypothesis: true mean is less than 44.7
95 percent confidence interval:
     -Inf 40.31151
sample estimates:
mean of x
     37.8

> # Critical value
> qt(0.05,29)
[1] -1.699127
>
> # Power of the test 1
> delta=3
> stand.delta=n^0.5*delta/sT
> pt(qt(0.05,29),29,ncp=-stand.delta)
[1] 0.6319631
>
> # Power of the test 2
> library(pwr)
> d=delta/sT
> pwr.t.test(d=-d,n=n,sig.level=0.05,type="one.sample",
+ alternative="less")

     One-sample t test power calculation

              n = 30
              d = -0.3705545
      sig.level = 0.05
          power = 0.6319631
    alternative = less
```

Fig. 13.8 Computing the power of a test with R-CRAN

on the *pwr.t.test* function from the package *pwr*. In that case the effect size is defined as:

$$d = \frac{\pm\delta}{s^T}$$

In our example, we have $d = 3/8.10 \approx 0.37$ and this effect is supposed to be negative. In general, for any type of *t*-test or for testing difference between proportions, d values of 0.2, 0.5, and 0.8 (in absolute value) represent small, medium, and large effect sizes, respectively. Note also that a power above 80% is usually considered as satisfactory so that in this numerical example (we found 0.6319631) one cannot reasonably conclude that the proposed health policy is worth implementing. Last, our analysis of effectiveness does not presume of the efficiency of the policy since we do not have any information about the differential cost if the health program replaced standard practices.

13.5 Sample Size Calculations

Controlling for the beta risk implies that the test must be sufficiently powerful, i.e. capable of providing (in probability) a significant difference in outcomes if the treatment is indeed more effective. The power of the test depends on several parameters, among which the effect size d, the sample size n^T, and the significance level α. Table 13.4 offers an example for different sets of parameters. The base case corresponds to example 1 (Sect. 13.4) with $\alpha = 5\%$, $n = 30$, $\delta = 3$ and $d \approx 0.37$. As shown previously, the statistical power is 63.2%. Figure 13.9 computes the different scenarios in R-CRAN. Note that if one adds $\$power$ to the end of the $pwr.t.test$ command, then only the statistical power will be displayed.

As can be observed from Table 13.4, the power of the test increases with α. Type I and type II errors are antagonistic. A smaller significance level α implies less test power, conversely a greater α risk-taking increases the odds of implementing the policy and hence minimize losses of opportunity. The antagonism between type I and type II risks is however mitigated by the fact that there is a well establish consensus to use a type I risk of $\alpha = 5\%$. Thus, in practice, the investigator exogenously sets α to 5% and computes the power of the test from the remaining parameters.

Second, the required effect size d appears as a quite sensitive parameter. The power of the test increases with the non-centrality parameter. When the scientific indicator of effect size implies a treatment mean value too close to that of the control group, the experiment loses most of its power. On the other hand, a large required effect size may prove too ambitious and difficult to reach.

Last, the power of the test increases with n. Sample size has the obvious and intuitive effect: the greater the sample size, the more powerful is the test. Limits to sample size are nevertheless often practical. Running an experiment is costly and, usually, the higher the sample size, the higher its cost. Randomization also requires the formation of two groups and, for particular diseases, finding and following up patients may not be so easy, hence the concentration of efforts on the inclusion of a smaller group. In the meanwhile, budgets are seldom limitlessly extendable. It is thus natural to seek for the minimum sample size required to obtain a pre-defined statistical power. A test power analysis can be performed for this particular purpose.

The statistical justification of the number of inclusions is a rather difficult topic but nevertheless important. Sample size calculation is always an estimate surrounded by a lot of uncertainty since the parameters involved are themselves estimates. The examples below are mostly illustrative and cannot encompass all cases since sample size formulas depend on the hypothesis test that has been selected and the type of clinical trial that has been implemented.

In R-CRAN, the pwr package provides several functions for sample size calculation. Let us consider again the base case of Table 13.4 with $\alpha = 5\%$, $d \approx -0.37$, $n = 30$ and $1 - \beta = 63.2\%$. In R-CRAN, when one of the parameters is not specified in the $pwr.t.test$ function, then that parameter is determined from the others. We can therefore specify the statistical power one would like to achieve and compute

Table 13.4 Sensitivity of test power to parameter variation: example 2

	Probability of a Type I error α			Sample size n^T			Required effect size δ, d		
	–	Base-case	+	–	Base-case	+	–	Base-case	+
Value	0.01	0.05	0.1	10	30	40	$\delta = 1$ $d \approx 0.12$	$\delta = 3$ $d \approx 0.37$	$\delta = 5$ $d \approx 0.62$
Power	34.8%	63.2%	76.4%	28.8%	63.2%	74.5%	16.3%	63.2%	95.1%

```
> library(pwr)
> # Base-case
> d=delta/s
> pwr.t.test(d=-d,n=n,sig.level=0.05,type="one.sample",
+ alternative="less")$power
[1] 0.6319631
> # Variation of parameters
> pwr.t.test(d=-d,n=n,sig.level=0.01,type="one.sample",
+ alternative="less")$power
[1] 0.3479142
> pwr.t.test(d=-d,n=n,sig.level=0.1,type="one.sample",
+ alternative="less")$power
[1] 0.7639277
> pwr.t.test(d=-d,n=10,sig.level=0.05,type="one.sample",
+ alternative="less")$power
[1] 0.2875263
> pwr.t.test(d=-d,n=40,sig.level=0.05,type="one.sample",
+ alternative="less")$power
[1] 0.7446563
> d=1/s
> pwr.t.test(d=-d,n=n,sig.level=0.05,type="one.sample",
+ alternative="less")$power
[1] 0.1625534
> d=5/s
> pwr.t.test(d=-d,n=n,sig.level=0.05,type="one.sample",
+ alternative="less")$power
[1] 0.9512743
```

Fig. 13.9 Sensitivity of statistical power with R-CRAN: example 2

```
> library(pwr)
> pwr.t.test(d=-0.37,power=0.632,sig.level=0.05,type="one.sample",
+ alt="less")$n
[1] 30.08852
> pwr.t.test(d=-0.37,power=0.80,sig.level=0.05,type="one.sample",
+ alt ="less")$n
[1] 46.54421
> pwr.t.test(d=-0.5,power=0.80,sig.level=0.05,type="two.sample",
+ alt ="two.sided")$n
[1] 63.76561
> pwr.2p.test(h = 0.5,power = 0.80,sig.level = 0.05,
+ alt = c("two.sided"))$n
[1] 62.79088
```

Fig. 13.10 Sample size calculations with R-CRAN

the required sample size. Figure 13.10 offers several examples. For instance, when specifying:

$$d = -0.37, power = 0.63, sig.level = 0.05, type = \text{"}one.sample\text{"}, alt = \text{"}less\text{"}$$

and indicating $\$n$ to get the ensuing sample size, one approximately finds a sample size of 30. If one sets the power of the test to the 80% level, the required sample size becomes 47.

The *pwr.t.test* command can also be used to find the sample size for a two-sided two-sample t-test (see Sect. 13.3) given a medium effect size ($d = 0.5$) and a power of 80%. The entries of the *pwr.t.test* function must be specified as:

$$d = 0.5, power = 0.80, sig.level = 0.05, type = "two.sample", alt = "two.sided"$$

Since the test is two-sided, the sign of the effect size does not matter anymore (we can set $d = 0.5$ or -0.5 indifferently). From Fig. 13.10, we obtain approximately a sample size of $n = 64$. One should be careful here as n now denotes the number of subjects in each group, i.e. control and treatment (the total number of subjects is 128). Using the $pwr.2p.test$ function, R-CRAN can also calculate the sample size required to test for equality of proportions. For instance, with an effect size equal to $h = 0.5$:

$$h = 0.5, power = 0.80, sig.level = 0.05, type = "two.sample", alt = "two.sided"$$

We obtain a sample size of 63 subjects in each group.

13.6 Indicators of Policy Effects

When the experiment reaches its end, the investigator can summarize and interpret information through various indicators. Often the approach relies on counting the number of successes and failures in each group. The outcome of interest is thereby categorical. The most popular indicators include the absolute risk reduction (*ARR*), the relative risk ratio (*RR*), the odds ratio (*OR*) and the number needed to treat (*NNT*). They are successively presented below. Since their computation depends on the sample under scrutiny, a margin of error $\pm e$ is generally calculated. It consists in defining a confidence interval which is likely to contain the true population measurement:

$$e = \text{critical value} \times se$$

The level of the critical value depends on the confidence level chosen by the evaluator (often, a 95% confidence level) and the probability distribution assumed behind the statistic. The standard error *se* is the standard deviation of the sampling distribution.

The simplest way of accounting for the treatment effectiveness is to rely on the absolute risk reduction, also called risk difference, which simply measures the size of a difference between two treatments. Let $p^T$and p^C denote the baseline risks, i.e. the probability of an event in the treatment group and the control group respectively. The absolute risk reduction is defined as:

$$ARR = p^T - p^C$$

It represents the proportion of subjects who are spared the adverse outcome as a result of having received the intervention. The standard error (se) is specified as:

Table 13.5 Relative risk: example 2

	Participants WITH event during trial	Participants WITHOUT event during trial	Participants at risk prior to trial	Baseline risk
Treatment group	$a = 60$	$b = 140$	$a + b = 200$	$a/(a + b) = 0.30$
Control group	$c{=}80$	$d = 120$	$c + d = 200$	$c/(c + d) = 0.40$
	$a + c = 140$	$b + d{=}260$	$n = a + b + c + d$ $=400$	$RR = \frac{a/(a+b)}{c/(c+d)} = 0.75$
	Total number of events during trial	Total number of unaffected participants during trial	Total of participants at risk prior to trial	Relative risk ***RR***

$$se(ARR) = \sqrt{\frac{p^T(1 - p^T)}{n^T} + \frac{p^C\ (1 - p^C)}{n^C}}$$

where n^T and n^C denotes the size of the control group and the treatment group, respectively. The critical value is the appropriate $z_{\alpha/2}$ value from the standard normal distribution for the desired confidence level. For instance, for a 5% significance level, the critical value is $z_{5\,\%\,/2}{=}1.96$. The corresponding confidence interval is:

$$ARR \pm 1.96 \times se$$

To illustrate, consider again example 2 (Table 13.3). The policy context is to avoid adverse events. Table 13.5 replicates the results. The baseline risk indicates the frequency of events evidenced from the two arms of the sample. In the treatment group, it amounts to $p^T = a/(a + b) = 0.30$. It is $p^C = c/(c + d) = 0.40$ in the control group. The group size are $n^T = a + b = 200$ and $n^C = c + d = 200$, respectively. The standard error is computed as:

$$se(ARR) = \sqrt{\frac{0.30(1 - 0.30)}{200} + \frac{0.40\ (1 - 0.40)}{200}} \approx 0.047$$

The confidence interval is given by $-0.10 \pm 1.96 \times 0.047 \approx [-0.193, -0.007]$ which corresponds to the confidence interval obtained with the $prop\,.\,test$ function in Fig. 13.4 (without Yates' correction for continuity).

Another possibility to assess the effectiveness of an intervention is to calculate the relative risk. It is defined as the ratio of the probability of event occurrence in the treatment group to the probability of the event occurring in the control group:

$$RR = \frac{\text{Baseline risk in the treatment group}}{\text{Baseline risk in the control group}} = \frac{p^T}{p^C}$$

If the policy under scrutiny is such that $RR \leq 1$, its implementation implies a relative improvement. If $RR > 1$ then the event we wish to avert is unfortunately more likely to occur in the treatment group than in the control group. This means that the assessed project should not be generalized from the trial sample to the whole population of interest. Last, if $RR = 1$ then the two arms experience the same outcome.

The calculation of the confidence interval rests on the assumption that the distribution of the logarithm ln(RR) is approximately normal with standard error:

$$se(\ln RR) = \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}$$

For a 5% significance level, the critical value is $z_{5\,\%/2}$=1.96. The corresponding confidence interval is expressed as $\ln(RR) \pm 1.96 \times se$.

For instance, in Table 13.5, the relative risk RR is defined as:

$$RR = \frac{p^T}{p^C} = \frac{0.30}{0.40} = 0.75$$

The risk for the treatment group is 75% that of the control group. There is a relative improvement that can also be expressed as the relative risk reduction RRR:

$$RRR = 1 - \mathrm{RR} = 25\%$$

If the policy measure is to be adopted, then the risk of adverse event (e.g., re-offense in example 2) should decrease by 25%. This statement should however be verified through the calculation of a confidence interval. We have $se = \sqrt{0.0192} \approx 0.138$, the confidence interval amounts to:

$$\exp(\ln(0.75) \pm 0.138 \times 1.96) = [0.57, 0.98]$$

The switch value $RR = 1$ is not included in the confidence interval, which confirms the previous statement.

Note that the ARR test and the RR test can sometimes yield different conclusions. The reason behind such a phenomenon is that the indicators do not measure the same thing. Imagine for instance that we have the following baseline risks: $p^T = 2\%$ and $p^C = 7\%$. Then, the $ARR = -0.05$ is relatively small while the RR amounts to 0.28. A small difference in proportions can thus lead to a high RR. The proportion of subjects getting the disease does not differ substantially from the control group to the treatment group (which is common in studying low incidence rates). Yet, relatively speaking, this difference is quite sensible. Reciprocally, when the

Table 13.6 Odds ratio: example 2

	Participants WITH event during trial	Participants WITHOUT event during trial	Participants at risk prior to trial	Odds
Treatment group	$a = 60$	$b = 140$	$a + b = 200$	$a/b = 0.43$
Control group	$c=80$	$d = 120$	$c + d = 200$	$c/d = 0.67$
	$a + c = 140$	$b + d=260$	$n = a + b + c + d$ $=400$	$OR = \frac{a/b}{c/d} = 0.64$
	Total number of events during trial	Total number of unaffected subjects during trial	Total of participants at risk prior to trial	Odds ratio *OR*

difference in proportions is large (e.g., $ARR = 90\,\% - 70\,\% = 20\%$), this does not mean that the RR is necessarily low (here, $RR = 77\%$).

Another popular measure of exposure to an adverse event is the odds ratio:

$$OR = \frac{\text{Odds for the treatment group}}{\text{Odds for the control group}} = \frac{a/b}{c/d}$$

Odds for the treatment group are the ratio of the count of event exposures (a) to the count of non-exposed participants (b). A similar ratio is calculated for the control group (c/d). If $OR = 1$ then both groups have the same odds. An OR less than 1 means that the treatment group is less likely to experience the adverse event. Table 13.6 exemplifies it. The policy implies a relative improvement since $OR = 0.64 < 1$. The sampling distribution of the odds ratio is approximately normally distributed on the natural log scale. The standard error is:

$$se(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

For a 5% significance level, the critical value is $z_{5\,\%/2}$=1.96. The corresponding confidence interval is thus expressed as $\ln(OR) \pm 1.96 \times se$. In our example, the standard error amounts to $se(\ln OR) = 0.211$, approximately. The confidence interval is expressed as:

$$\exp(\ln(0.64) \pm 0.211 \times 1.96) = [0.42, 0.97]$$

Again, the resulting confidence interval gives support to the program.

In practice, the relative risk ratio and odds ratio indicators can yield different results. In our example, the odds ratio is much more favorable to the policy since for the treatment group the risk of event occurrence is only 64% of that of the control group, compared to 75% in the case of the relative risk ratio (Table 13.5). This discrepancy takes place when the event has a high incidence (high probability of event in both groups). Conversely, if event occurrence is low, the two indicators converge. Table 13.7 illustrates this phenomenon. The ratio a/c is similar to that of Table 13.5 and, for this reason, the RR remains equal to 0.75. The odds ratio on the other hand increases to 74%. To systematize this result, assume groups of equal size and rewrite the relative risk and odds ratios as:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a}{c}$$

$$OR = \frac{a/b}{c/d} = \frac{a}{c} \times \frac{d}{b}$$

When the incidence of the event is very low compared to sample size, then $d/b \rightarrow 1^-$ and the odds ratio gets very close to the relative risk ratio.

Another indicator provided by the count framework is the number needed to treat (NNT, hereafter). In the case of an adverse event, it is the average number of subjects who should be treated to avoid the occurrence of one event. Formally:

$$NNT = \frac{1}{[a/(a+b)] - [c/(c+d)]} = \frac{1}{p^T - p^C} = \frac{1}{ARR}$$

It is defined as the inverse of the absolute risk reduction. For instance, in example 2, we have:

$$NNT = \frac{1}{0.30 - 0.40} = \frac{1}{-0.10} = -10$$

According to the ARR criterion, ten events on average are avoided every 100 treated subjects. Equivalently, with the NNT, we conclude that one event is avoided every 10 treated subjects. The confidence interval is directly obtained from the ARR:

$$\left[\frac{1}{-0.007}, \frac{1}{-0.193}\right] \approx [-142, -5]$$

Note that the NNT should not be mistaken with the baseline risk $p^T = 30\%$. It does not mean that by treating 10 subjects, one of them only will avoid the event. Instead, it is the average number of subjects who need to be treated to prevent one additional adverse event.

When $p^T = 0\%$ (the treatment works in every case), while $p^C = 100\%$ (every subject in the control group face the adverse event), then the number needed to treat is $NNT = 1/(0.0 - 1.0) = -1$. If $p^T = 20\%$ (an event is observed 2 times out of 10 in the treatment group) and $p^C = 40\%$ (an event is observed 4 times out of 10 in the control group), then the number needed to treat is $NNT = 1/(0.2 - 0.4) = -20$. On

Table 13.7 Converging relative risk and odds ratios: example 3

	Participants WITH event during trial	Participants WITHOUT event during trial	Participants at risk prior to trial	Baseline and relative risk	Odds ratio
Treatment group	$a=6$	$b=194$	$a+b=200$	$a/(a+b)=0.03$	$a/b=0.031$
Control group	$c=8$	$d=192$	$c+d=200$	$c/(c+d)=0.04$	$c/d=0.042$
	$a+c=14$	$b+d=386$	$n=400$	$RR=\frac{a/(a+b)}{c/(c+d)}=0.75$	$OR=\frac{a/b}{c/d}=0.74$

average, one event is avoided every 20 treated subjects. The higher is the *NNT*, the less effective is the treatment. If the treatment group and the control group have the same size ($a + b = c + d$), the *NNT* is expressed as:

$$NNT = \frac{(a+b)(c+d)}{[a/(c+d)] - [c/(a+b)]} = \frac{(a+b)^2}{[a/(a+b)] - [c/(a+b)]} = \frac{a+b}{a-c}$$

In this case, the *NNT* is the ratio of the group size to the reduction in the number of events allowed by the intervention.

13.7 Survival Analysis with Censoring: The Kaplan-Meier Approach

How long are you going to keep your job? Will that be influenced by the labor policy of your state or region? How long would you remain at school before you dropout, (if ever you do)? How long will your treatment keep you alive? Would you live longer if treated with another drug? Answering those questions can be difficult as the subjects may face a different evolution of their situation through time. Adverse events can be observed all along the experiment, from the beginning until the end of the period of observation. For instance, the same endpoint result for two groups can hide very different outlines of evolution. Imagine a medical experiment with a long time horizon where none of the participants have survived. Yet, the subjects in the treatment group may have survived longer thanks to the intervention. It is thus quite important to catch the survival profiles. For this reason, we must extend the analysis to a framework were the timing of event occurrence is explicitly accounted for.

This new framework involves the comparison of a control group versus the treatment group over several time periods. Two-way tables are built for each period, and then compared in order to check whether the intervention provides significant improvement. In the affirmative, the experiment can be extended to larger populations or other jurisdictions. For instance, depending on the context, adverse events can be job loss, training disruption in labor policy, school dropout in education, re-offense after prison, side-effect of a drug, etc. In those cases, the effectiveness of a program is not only about the number of successes and failures, but also about when they are observed.

Consider an experiment where all the subjects are followed until they face an adverse event (e.g., progression of the disease). The Kaplan-Meier survival analysis first requires to set the time frame of individual exposure to treatment (time horizon *H*, time period *t* for event count: months, years, etc.). In practice, the design should also carefully describe the censoring conditions. Censoring occurs when some participants do not experience the event before trial termination or when administrative termination takes place because of the budget constraint, of early benefits or harmful effects, etc. Data arrangement should also take into consideration the fact that some participants can get lost to follow-up without experiencing the event.

Following the previous structure (see Sect. 13.6), let us first define the two-way table observed in period t as follows:

Period t	**With event**	**Without event**	**Total**
Treatment group	a_t	b_t	$a_t + b_t$
Control group	c_t	d_t	$c_t + d_t$

For each group and each period, a conditional probability of surviving at any particular time period is calculated as:

$$(1 - p_t) = \frac{\text{Number of subjects at risk} - \text{Number of adverse events}}{\text{Number of subjects at risk}}$$

In other words, for each time interval t, $(1 - p_t)$ is defined as the number of subjects who did not experience the adverse event divided by the number of subjects at risk. This amount to compute the following probabilities for the treatment and control groups:

$$\left(1 - p_t^T\right) = \frac{(a_t + b_t) - a_t}{a_t + b_t} = \frac{b_t}{a_t + b_t}, \qquad \left(1 - p_t^C\right) = \frac{(c_t + d_t) - c_t}{c_t + d_t} = \frac{d_t}{c_t + d_t}$$

Those probabilities are conditional because they measure the probability of no event in period t given that no event has occurred in period $(t - 1)$. For instance, in the case of a medical treatment, they denote the probability of surviving the t^{th} period given that the participant has survived the previous time intervals.

The survival rate in period t is calculated by multiplying all the conditional probabilities preceding and including that time period:

$$S_t^j = \prod_{\tau=1}^{t} \left(1 - p_\tau^j\right), \quad j = T, C$$

It is the percentage of subjects in group j who did not experience the adverse event by the end of period t. Note that the Kaplan-Meier approach has the advantage to exclude the number of censored subjects from the calculation. This is crucial. In many cases, we cannot rely on a simpler computation based on the number of remaining subjects in each group. The reason is that the censored subjects would be counted as subjects who did not face the adverse event and one would obtain instead an overestimation of the survival rate (lower probability of adverse effects). This is why the Kaplan-Meier approach should be used anytime censoring is prevalent.

To get the intuition of the Kaplan-Meier survival analysis, consider the following example. The follow-up horizon is two years. The initial cohort in the treatment group and in the control group each have 40 participants. At $t=1$, we have the two-way table:

$t=1$	**With event**	**Without event**	**Total**
Treatment group	$a_1 = 1$	$b_1 = 39$	40
Control group	$c_1 = 7$	$d_1 = 33$	40

The conditional probabilities of surviving period $t=1$ are computed as $\left(1-p_1^T\right) = 39/40 = 97.5\%$ and $\left(1-p_1^C\right) = 33/40 = 82.5\%$, respectively. In the first period, these probabilities directly yield the survival rates S_1^T and S_t^C. Now, assume that at $t=2$ one participant from the treatment group and one from the control group are censored. We have:

$t=2$	**With event**	**Without event**	**Total**
Treatment group	$a_2 = 2$	$b_2 = 36$	38
Control group	$c_2 = 4$	$d_2 = 28$	32

Because one individual has been censored, the number of subjects at risk in the treatment group at $t=2$ (i.e. 38) does not match the observed number of subjects without event at the end of period 1 (i.e. 39). Equivalently, in the control group, one individual has been left out ($32 \neq 33$). The conditional probabilities are computed as $\left(1-p_2^T\right) = 36/38 \approx 94.7\%$ and $\left(1-p_2^C\right) = 28/32 \approx 87.5\%$, respectively. Survival probabilities at the end of period $t=2$ are thus:

$$S_2^T = \left(1-p_1^T\right) \times \left(1-p_2^T\right) = \frac{39}{40} \times \frac{36}{38} \approx 92.4\%$$

and

$$S_2^C = \left(1-p_1^C\right) \times \left(1-p_2^C\right) = \frac{33}{40} \times \frac{28}{32} \approx 72.2\%$$

Consider now Fig. 13.11 which extends the previous illustration. For the treatment group (Fig. 13.11a) the survival curve is built as follows. The observation period consists of 6 consecutive time intervals (e.g., years) indexed t for an initial cohort of 40 individuals. During the first period, 1 subject is exposed to the event and 1 is lost or censored. That leaves $40-1-1=38$ subjects for the next period during which 2 participants face the event and 1 is lost or censored. The third period thus begins with $38-2-1=35$ subjects, etc. At the end of the experiment (administrative termination at the end of period 6), 9 subjects have "survived" and they are by convention allocated to the set of censored participants. To draw the corresponding survival curve, one needs first to calculate the conditional survival probabilities $\left(1-p_t^T\right)$ period after period, then multiply them to compute the survival rates. A similar approach is used for the control group (see Fig. 13.11b).

a

Period	Subjects at risk (at the beginning of period t)	Subjects with event (during period t)	Participants lost or censored (during period t)	Conditional survival probability	Estimated survival step function
t	$a_t + b_t$	a_t		$1 - p_t^T$	S_t^T
1	40	1	1	0.975	0.975
2	38	2	1	0.947	0.924
3	35	5	2	0.857	0.792
4	28	7	0	0.750	0.594
5	21	6	2	0.714	0.424
6	13	4	9	0.692	0.294

b

t	$c_t + d_t$	c_t		$1 - p_t^C$	S_t^C
1	40	7	1	0.825	0.825
2	32	4	0	0.875	0.722
3	28	7	2	0.750	0.541
4	19	7	0	0.632	0.342
5	12	8	1	0.333	0.114
6	3	3	0	0.000	0.000

Fig. 13.11 Survival step functions: example 4. (**a**) Treatment group. (**b**) Control group

The survival step functions for the treatment and control groups are depicted in Fig. 13.12. The code for plotting the curves is provided in Fig. 13.13. First, the survival probabilities S_t^T and S_t^C are entered manually. The related curves are then drawn with the *stepfun* function. This function is used because the usual way of plotting survival curves is through step curves, each segment representing survival at the end of the current time period. A legend is then included with the usual command. Note that R-CRAN also offers a package named *survival*, with several tools for analyzing data at the individual level with heterogeneous time intervals.

As can be understood, survival rates can be very helpful to assess the effect of a program over time. They can be used to inform a decision-maker about survival probabilities after a number a periods. A median survival time can also be provided. For instance, in Fig. 13.11, the estimated probability of surviving 2 years is 72.2% without treatment and 92.4% with treatment. The median survival time without treatment is approximately 3 years and is between 4 and 5 years with treatment. There remains to check whether these differences are statistically significant.

13.8 Mantel-Haenszel Test for Conditional Independence

As can be deduced from the previous section, what we would like to measure is the area that lies between two survival curves. The larger is that difference, the larger is the effect of the treatment over the whole time period. To do so, the usual approach is to rely on a Mantel-Haenszel test. As previously, this test is used when we have successive data from two-way tables at different time periods $t = 1 \ldots H$:

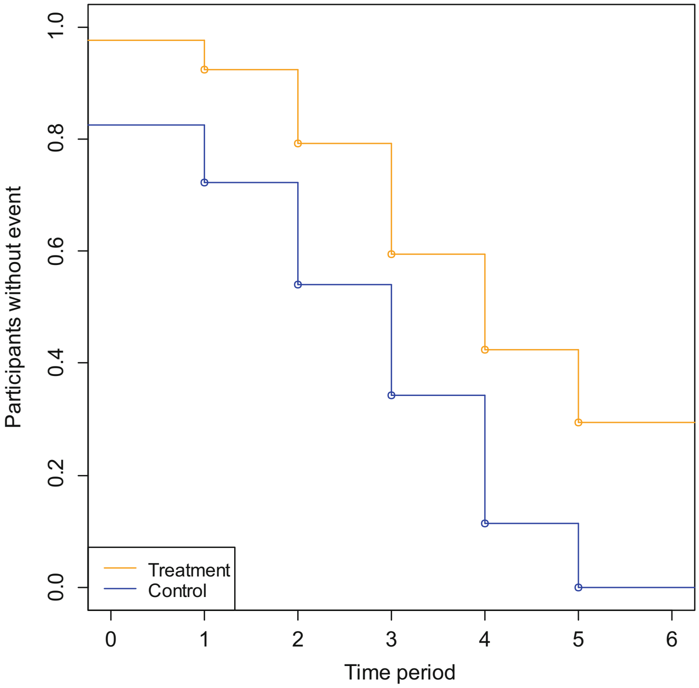


Fig. 13.12 Kaplan-Meier survival step function: example 4

```
# Survival curves
>
> par(mar=c(4,4,2,2))
> ST=c(0.975,0.924,0.792,0.594,0.424,0.294)
> SC=c(0.825,0.722,0.541,0.342,0.114,0.000)
>
> plot(stepfun(1:5,ST),main="",ylim=c(0,1),xlab="Time period",
+ ylab="Participants without event",col="Orange")
> lines(stepfun(1:5,SC),col="Blue")
> legend("bottomleft",legend=c("Treatment","Control"),
+  col=c("orange","blue"),lty=c(1,1))
```

Fig. 13.13 Plotting survival curves in R-CRAN

Period t	**With event**	**Without event**
Treatment group	a_t	b_t
Control group	c_t	d_t

where $n_t = (a_t + b_t + c_t + d_t)$ represents the total number of subjects. Under this framework, the Mantel-Haenszel test first defines what is termed a common odds ratio:

$$\text{Common OR} = \frac{\sum_{t=1}^{H} a_t d_t / n_t}{\sum_{t=1}^{H} b_t c_t / n_t}$$

It is a way of summarizing how big the differences are between the two groups. The null hypothesis of the test is that the population common odds ratio is equal to 1. The alternative hypothesis is that it is different from unity. In other words, the Mantel-Haenszel test checks whether there is no consistent difference in the two-way tables over the whole period of study.

Under the Mantel-Haenszel framework, the test statistic is expressed as:

$$\chi^2_{MH} = \frac{\left(\sum_{t=1}^{H} a_t - E(a_t) \right)^2}{\sum_{t=1}^{H} s^2(a_t)}$$

where $E(a_t)$ denotes what is termed the expected number of events in the treatment group and $s^2(a_t)$ the variance of the observed number of events in the treatment group, respectively:

$$E(a_t) = (a_t + b_t) \times \frac{a_t + c_t}{n_t}$$

$$s^2(a_t) = \frac{(a_t + c_t)(b_t + d_t)(a_t + b_t)(c_t + d_t)}{{n_t}^2 (n_t - 1)}$$

The Mantel-Haenszel statistic has a chi-square distribution with one degree of freedom. At the 5% confidence level, significant difference between two groups is reached when $\chi^2_{\mathrm{MH}} > 3.84$. At the 1% confidence level, the critical value increases to 6.63.

Let us now illustrate the Mantel-Haenszel test with the data from example 4. Figure 13.14 reproduces the analysis in R-CRAN. Technically speaking, it does not matter how the two-way tables are arranged as any of the four values of the two-way tables can be used as a_t. However, to fit the Kaplan-Meier framework, the database is arranged so as to correspond to the following order: *a*, *c*, *b*, *d*. The command *read.table* reads the file and allows to check the values entered in the .*csv* file. Using the command *array*, the whole set of two-way tables is generated. To do so, the initial dataframe *D* must be specified as a list of numbers. Then the command *mantelhaen.test* implements the test. Note that by specifying *correct=TRUE* in the *mantelhaen.test* command, R-CRAN performs the Mantel-Haenszel test with a continuity correction. In addition to testing the null hypothesis, the Mantel-Haenszel test also produces an estimate of the common odds ratio.

In our example, the common odds ratio is found to be 0.313 approximately, so that the treatment group is substantially less likely to experience the adverse event than the control group. The test statistic is $\chi^2_{\mathrm{MH}} = 13.242 > 6.63$. The survival patterns of the treatment and control groups are thus significantly different at the

```
> D=read.table("C://mydataMH.csv",head=TRUE,sep=";")
> D
  a c  b  d
1 1 7 39 33
2 2 4 36 28
3 5 7 30 21
4 7 7 21 12
5 6 8 15  4
6 4 3  9  0
> # Mantel-Haenszel test
> list=as.numeric(as.list(data.matrix(t(D))))
> mytables=array(list,dim=c(2,2,6), dimnames=list(
+ Group= c("Treatment","Control"), Outcome=c("With event","Without"),
+ Time=1:6))
> mytables
, , Time = 1
         Outcome
Group      With event Without
  Treatment         1      39
  Control           7      33
, , Time = 2

         Outcome
Group      With event Without
  Treatment         2      36
  Control           4      28
...
> mantelhaen.test(mytables,correct=FALSE)

        Mantel-Haenszel chi-squared test without continuity correction

data:  mytables
Mantel-Haenszel X-squared = 13.242, df = 1, p-value = 0.0002737
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.1644801 0.5940887
sample estimates:
common odds ratio
        0.3125952
> mantelhaen.test(mytables,correct=TRUE)

        Mantel-Haenszel chi-squared test with continuity correction

data:  mytables
Mantel-Haenszel X-squared = 12.15, df = 1, p-value = 0.000491
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.1644801 0.5940887
sample estimates:
common odds ratio
        0.3125952
```

Fig. 13.14 Mantel-Haenszel test with R-CRAN: example 4

1% significance level. For small samples, a continuity correction can be used, in which case the numerator of the test statistic becomes:

$$\left(\left|\sum_{t=1}^{H}[a_t - E(a_t)]\right| - 0.5\right)^2$$

It reduces the test statistic to 12.15 without changing the conclusion in our case.

Bibliographical Guideline

According to Friedman et al. (2010), clinical trials and the idea of comparison groups date back to the eighteen's century, with studies of scurvy in the English Navy. Randomization was introduced in agricultural research in the 1930s and modern clinical trials in medicine began in the 1960s. Friedman et al. (2010)

provide a comprehensive and very useful survey of all the sequences of the clinical trial methodology. A valuable reference for the ethics in clinical trials is Freedman (1987). More specifically, chi-square tests and the event count framework have been initially developed by Cochran (1954). Non parametric estimations of survival curves were initiated by Kaplan and Meier (1958) and by Cutler and Ederer (1958). Survival curves comparisons originate in Mantel and Haenszel (1969).

There is also a long history of randomized trials in social science and program evaluation. The Rand health insurance experiment in the mid-1970s is still a benchmark (Aron-Dine et al. 2013). Since then, randomized trials have been applied to study a variety of research areas. Among others, we may name labour economics (see for instance Crépon et al. 2014 for a recent application), education economics (e.g., Attanasio et al. 2012) and development economics (e.g., Duflo et al., 2015). The question of randomization in social science, in particular in development programs, is addressed by Duflo et al. (2007) and Deaton (2010).

Randomized experiments offer an original evidence-based approach to evaluate public programs. This import of the clinical trial methodology is, however, not without pitfalls. Concerns about the use (or misuse) of randomized control trials have been recently raised by Deaton and Cartwright (2016). Conjointly, Favereau (2016) provides a critical assessment of the analogy between medicine and program evaluation, especially in the field of economic policies fighting poverty. The latter cannot not meet all the requirements of the clinical trial methodology, which could weaken their evaluative power.

Note also that despite all the efforts made to have similar treatment and control groups, a few differences may pertain. One reason is non-compliance which may finally result in a selection bias. Other possible reasons are difficulties in the design itself. For example, one needs to ensure that the group size is sufficiently large, that the participants give their consent, or that the experiment benefits from the jurisdiction's support. In many cases, those items can be a serious concern. An easy way to detect a selection bias is to rely on simple comparison of means and proportions using the exogenous variables in the database (e.g., gender, age, etc.). Those simple descriptive statistics can be accompanied with a one-way *t*-test and a chi-square test of independence to assess the significance of those differences (see Sect. 13.3). If a bias is observed, then one needs to control for it *ex post*. As already stressed in the introduction, the usual way of doing it is to rely on quasi-experimental techniques such as difference-in-differences, propensity score matching, regression discontinuity design or instrumental variable estimation. The next chapter offers a description of those substitutes.

References

Attanasio, O., Meghir, C., & Santiago, A. (2012). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA. *Review of Economic Studies, 79*, 37–66.

Cochran, W. (1954). Some methods for strengthening the common chi-square tests. *Biometrics, 10*, 417–451.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2014). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics, 128*, 531–580.

Cutler, S., & Ederer, F. (1958). Maximum utilization of the lifetable method in analyzing survival. *Journal of Chronic Diseases, 8*, 699–712.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature, 48*, 424–455.

Deaton, A., & Cartwright, N. (2016). *Understanding and misunderstanding randomized control trials*. National Bureau of Economic Research w22595.

Duflo, E., Glennester, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. In P. Schultz & J. Strauss (Eds.), *Handbook of development economics* (vol. 4, pp. 3895–3962).

Duflo, E., Dupas, P., & Kremer, M. (2015). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review, 105*, 2757–2797.

Favereau, J. (2016). On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine. *Journal of Economic Methodology, 23*, 203–222.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine, 317*, 141–145.

Friedman, L., Furberg, C., & DeMets, D. (2010). *Fundamentals of clinical trials*. Heidelberg: Springer.

Kaplan, E., & Meier, P. (1958). Non parametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Mantel, N., & Haenszel, W. (1969). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Quasi-experiments 14

14.1 The Rationale for Counterfactual Analysis

Impact evaluation assesses the degree to which changes in a specific outcome or variable of interest as measured by a pre-specified set of indicators can be attributed to a program rather than to other factors. Such evaluation generally requires a counterfactual analysis to assess what the outcome would have looked like in the absence of the intervention. The main issue is that the counterfactual cannot be observed individually (the same unit cannot be exposed and unexposed at the same time) which means that one cannot directly calculate any individual-level causal effect. Instead, the counterfactual must be approximated with reference to a comparison group. Broadly speaking, one needs to compare a group that received the intervention, the “treatment group”, against a similar group, the “comparison group”, which did not receive the intervention. The observed difference in mean outcome between the treatment group and the comparison group can then be inferred to be caused by the intervention. What is observed in the comparison group serves as the counterfactual of what would have happened in the absence of the intervention.

Two types of methods can be used to generate the counterfactual: randomized controlled experiments and quasi-experiments. Both approaches rely on the estimation of the average causal effect in a population. In the first case, the treatment group and the comparison group (also termed “control group” in this case) are selected randomly from the same population. Similarly, quasi-experimental evaluation estimates the causal impact of an intervention, the difference being that it does not randomly assign the units between the treatment group and the comparison group. Hence, a key issue with quasi-experimental methods is to find a proper comparison group that resembles the treatment group in everything but the fact of receiving the intervention. The term “comparison group” differs from the narrower term “control group” in that the former is not necessarily selected randomly from the same population as program participants.

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_14

Table 14.1 Treated and non-treated group in a two-period setting

	Before intervention $P=0$	After intervention $P=1$	Δ after/before
Non-treated group $S=0$	$\bar{y}^{00}$	$\bar{y}^{01}$	
Treated group $S=1$	$\bar{y}^{10}$	$\bar{y}^{11}$	$\bar{y}^{11}-\bar{y}^{10}$
Δ treated/non-treated		$\bar{y}^{11}-\bar{y}^{01}$	

To illustrate the many difficulties that occur with a quasi-experimental design, let us consider a four-outcome setting by distinguishing the units that are exposed to the program from those who are not, and whether the outcomes are observed before or after the intervention. More specifically, let $\bar{y}^{SP}$ denote the average outcome in each case, with S being equal to one if the group was selected to receive treatment (and zero otherwise) and P denoting the time period ($P=0$ before the intervention and $P=1$ after). The four possible cases are depicted in Table 14.1. How can we assess the impact of the intervention using these four possible outcomes? Answering that question is not straightforward since a variation in the variable of interest, when measured by single differences, can always be thought of as the sum of two elements: the true average effect of the intervention (E hereafter) and biases due to the quasi-experimental design itself.

First, should we compare the outcome observed for the treated units after and before they have been exposed to the intervention, the analysis would suffer from an omitted-variable bias. The within-subjects estimate of the treatment effect, which measures difference over time, is given by:

$$\Delta^{S=1}_{\text{after/before}}=\bar{y}^{11}-\bar{y}^{10}=E+\text{omitted variable bias}$$

The fundamental problem here is that the observed change through time could be due to the true effect E of the intervention but also due to other changes occurring over time during the same period. The only case in which after-versus-before comparisons are relevant is when no other factor could plausibly have caused any observed change in outcome.

Second, should we concentrate on the period after the intervention, and compare the group that has been exposed to the intervention with the group that has not, then the analysis could suffer from a selection bias. The between-subjects estimate of the treatment effect, which measures the difference between the treated and non-treated groups, is as follows:

$$\Delta^{P=1}_{\text{treated/non treated}}=\bar{y}^{11}-\bar{y}^{01}=E+\text{selection bias}$$

A selection bias appears when the comparison group is drawn from a different population than the treatment group. The differences that are observed between the treated and non-treated groups could have been generated by the selection process itself and not necessarily caused by the intervention.

Assume for instance that one aims to evaluate the effect of a tutoring program for children at risk of school failure. It consists in lessons that prepare students to pass the examinations of the second semester. Only students who volunteered attend these sessions. To establish the impact of tutoring we may try to compare the average marks $\bar{y}^{10}$ of those who participated in the program before they were exposed to the intervention (e.g., first semester) with the marks $\bar{y}^{11}$ they obtained after the intervention (second semester). Imagine now that the teachers of the first semester were more prone to give good marks than the teachers from the second semester. Approximating the impact of the intervention using after-versus-before comparison would yield an underestimation of the effect. Should we focus on the second semester only, we could compare the marks $\bar{y}^{11}$ of those who benefited from the training sessions with the marks $\bar{y}^{01}$ of those who did not. However, the evaluation may be affected in this case by a selection bias. For instance, those who have decided to participate in the tutoring program may also be those who are the most motivated, and not necessarily those at risk of failing exams. Using treated versus non-treated comparisons would yield in this case an overestimation of the impact.

Basically, the ideal way to eliminate a selection bias is to randomly select the units who belong to the non-treated and treated groups. However, implementing randomized controlled experiments is not always feasible given the many legal, ethical, logistical, and political constraints that may be associated with it. Another problem with randomization is that other biases may appear due to the sampling process itself, but also to the fact that the experimental design may demotivate those who have been randomized out, or generate noncompliance among those who have been randomized in. In those instances, the alternative we are left with is quasi-experiment, i.e. not to allocate participants randomly.

The key feature of quasi-experimental evaluation is that one needs to identify a comparison group among the non-treated units that is as similar as possible to the treatment group in terms of pre-intervention characteristics. This comparison group is supposed to capture the counterfactual, i.e. what would have been the outcome if the intervention had not been implemented. The average treatment effect is then given by:

$$E = \bar{y}^{11} - \bar{y}^{c}$$

where $\bar{y}^{c}$ denotes the counterfactual outcome. While a quasi-experimental design aims to establish impact in a relevant manner, it does not relate the extent of the effect to the cost of the intervention. Instead, the challenge is to prove causality by using an adequate identification strategy. An identification strategy is the manner in which one uses observational data to approximate a randomized experiment. In practice, the counterfactual $\bar{y}^{c}$ is approximated using quasi-experimental methods such as difference-in-differences, regression discontinuity design, propensity score matching, or instrumental variable estimation. The chapter provides a review of these statistical tools.

14.2 Difference-in-Differences

Difference-in-differences, also known as double differencing, is by far the simplest method to estimate the impact of an intervention, especially as it does not necessarily require a large set of data. The method consists in comparing the changes in outcome over time between treatment and comparison groups. Unlike single differences (within- and between-subjects estimates), the approach considers both the time period P and the selection process S to estimate the impact. Using the setting developed in the previous section, we have:

$$\widehat{E} = \Delta^{\mathrm{P=1}}_{\mathrm{treated/non\ treated}} - \Delta^{\mathrm{P=0}}_{\mathrm{treated/non\ treated}} = \left(\bar{y}^{11} - \bar{y}^{01}\right) - \left(\bar{y}^{10} - \bar{y}^{00}\right)$$

The method thus requires that outcome data be available for treated and non-treated units, both before and after the intervention. The assumption underlying the identification strategy is that the selection bias is constant through time:

$$\text{selection bias} \approx \Delta^{\mathrm{P=0}}_{\mathrm{treated/non\ treated}}$$

In other words, the approach aims to eliminate any potential difference between the treated and comparison groups by using information from the pre-intervention period.

An alternative but equivalent way to explain the difference-in-differences approach is to consider the change observed over time among the treated and non-treated units. As stated in the previous section, this difference cannot be interpreted as the impact of the intervention, because other factors might have caused the observed variation. However, one plausible way to take this dynamics into account is to use the change in outcome observed over time among the non-treated units:

$$\widehat{E} = \Delta^{S=1}_{\mathrm{after/before}} - \Delta^{S=0}_{\mathrm{after/before}} = \left(\bar{y}^{11} - \bar{y}^{10}\right) - \left(\bar{y}^{01} - \bar{y}^{00}\right)$$

The assumption underlying the identification strategy is that the trend of the treated group in the absence of the intervention would have been the same as that of the non-treated group:

$$\text{omitted variable bias} \approx \Delta^{S=0}_{\mathrm{after/before}}$$

This assumption is also known as the parallel-trend assumption.

In practice, it is common to present the result of a difference-in-differences evaluation using a framework similar to that of Table 14.2, where single and double

Table 14.2 Double difference calculations

	Before intervention $P=0$	After intervention $P=1$	Δ after/before
Non-treated group $S=0$	$\bar{y}^{00}$	$\bar{y}^{01}$	$\bar{y}^{01}-\bar{y}^{00}$
Treated group $S=1$	$\bar{y}^{10}$	$\bar{y}^{11}$	$\bar{y}^{11}-\bar{y}^{10}$
Δ treated/non-treated	$\bar{y}^{10}-\bar{y}^{00}$	$\bar{y}^{11}-\bar{y}^{01}$	$\bar{y}^{11}+\bar{y}^{00}-\bar{y}^{01}-\bar{y}^{10}$

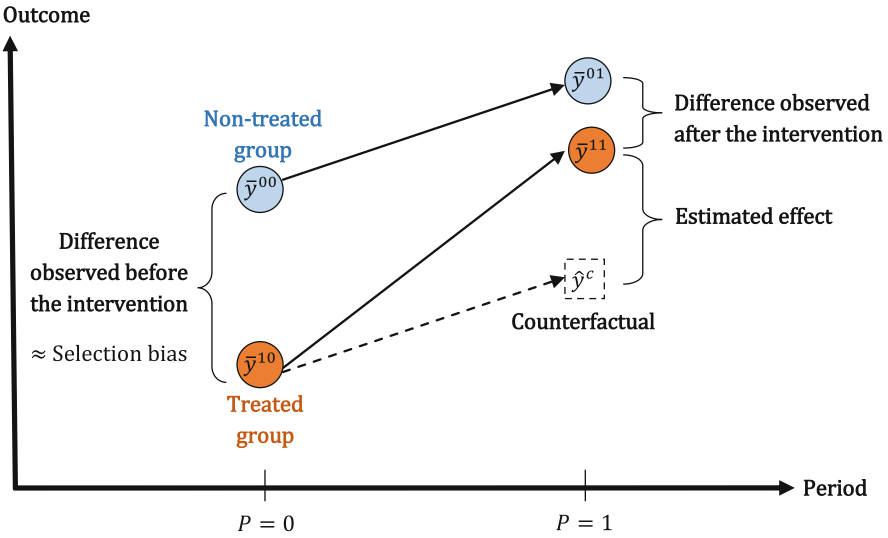


Fig. 14.1 The difference-in-differences approach

differences are elicited. The treated-versus-non-treated and after-versus-before approaches yield the same result:

$$\widehat{E}=\bar{y}^{11}+\bar{y}^{00}-\bar{y}^{01}-\bar{y}^{10}=\bar{y}^{11}-\left(\bar{y}^{01}+\bar{y}^{10}-\bar{y}^{00}\right)$$

The counterfactual is thus approximated by:

$$\widehat{y}^{c}=\left(\bar{y}^{01}+\bar{y}^{10}-\bar{y}^{00}\right)$$

In these expressions, the hat symbol over variable y means that the calculated quantity is only an estimate of the counterfactual $\bar{y}_c$, and so is the observed impact $\widehat{E}$ of the intervention.

Figure 14.1 illustrates the approach. The treatment group ($S=1$) is represented in orange while the non-treated group ($S=0$) is represented in blue. The outcome variable is measured both before and after the intervention takes place. After the intervention ($P = 1$), the difference observed between the treated group and the non-treated group does not reveal the true effect of the intervention. A difference was already observed before the intervention at $P = 0$. The difference-in-differences approach controls for this selection bias by subtracting the difference observed between the two groups before the intervention from the difference observed after the intervention. In other words, we assume that without any intervention, the trend of the treated group would have been similar to that of the non-treated. Graphically, this is equivalent to drawing a line parallel to the trend observed among non-treated units, but starting where the treated units are at $P = 0$. The dotted line yields the counterfactual $\widehat{y}_c$, which is depicted by a dotted square at $P = 1$.

The results can also be reproduced through an econometric analysis. The model requires the use of a database with information on treated and non-treated units, both before and after the intervention. Formally, for each unit i and time period t, the outcome y_{it} can be modeled via the following equation:

$$y_{it} = \alpha + \beta S_i + \gamma P_t + \delta(S_i \times P_t) + \epsilon_{it}$$

where S_i is the group variable, P_t is the dummy that controls for the timing of the treatment, the coefficients α, β, γ and δ are the parameters to be estimated and ε_{it} is an error term which contains all the factors the model omits. The product $S_i \times P_t$ is an interaction term that represents the treatment variable, i.e. whether an individual from group $S = 1$ received treatment in period $P = 1$.

The estimated counterpart of the equation can be written as:

$$y_{it} = \widehat{\alpha} + \widehat{\beta} S_i + \widehat{\gamma} P_t + \widehat{\delta}(S_i \times P_t) + \widehat{\epsilon}_{it}$$

Ordinary least squares (OLS) regressions are such that the mean of residuals is exactly equal to zero. Thus, on average, we have:

$$\bar{y}^{SP} = \widehat{\alpha} + \widehat{\beta} S + \widehat{\gamma} P + \widehat{\delta}(S \times P)$$

We thereby obtain the four possible outcomes of Table 14.2. If $S = P = 0$, we have $\bar{y}^{00} = \widehat{\alpha}$. Similarly, if $S = 1$ and $P = 0$, we get $\bar{y}^{10} = \widehat{\alpha} + \widehat{\beta}$. If $S = 0$ and $P = 1$, we obtain $\bar{y}^{01} = \widehat{\alpha} + \widehat{\gamma}$. Last, if $S = P = 1$, then we have $\bar{y}^{11} = \widehat{\alpha} + \widehat{\beta} + \widehat{\gamma} + \widehat{\delta}$. Under this setting, the single differences between subjects are given by:

$$\Delta^{\text{P=0}}_{\text{treated/non treated}} = \bar{y}^{10} - \bar{y}^{00} = \widehat{\beta}$$

$$\Delta^{\text{P=1}}_{\text{treated/non treated}} = \bar{y}^{11} - \bar{y}^{01} = \widehat{\beta} + \widehat{\delta}$$

The estimated impact of the intervention is:

$$\widehat{E} = \Delta^{\text{P=1}}_{\text{treated/non treated}} - \Delta^{\text{P=0}}_{\text{treated/non treated}} = \widehat{\delta}$$

The regression approach entails a loss in terms of simplicity, but has the advantage to provide a level of significance with respect to the estimated impact. Moreover, to go beyond the assumption that other covariates do not change across time, the regression model can be extended by including additional variables that may affect the outcome of interest.

Let us exemplify the approach through a simple application (example 1). Imagine that we would like to estimate the effects of a community-based health program on newborn mortality. Assume that this program provides primary care through the use of nurse teams intervening at the city level (e.g., counseling and prevention). The data is provided in Table 14.3. Our dataset comprises 20 jurisdictions, among which the nine municipalities numbered from 12 to 20 were selected for the program ($S = 1$). We also have information about the pre-intervention period ($P = 0$). The mortality rate is expressed per thousand of newborns. The last column provides information about the gross domestic product (GDP) per capita in each city, for both periods. By definition, Table 14.3 forms a panel database, as each column contains observations over two periods for each unit.

Table 14.4 provides descriptive statistics for each period $P \in \{0, 1\}$ and both groups $S \in \{0, 1\}$. The program appears to have been implemented in municipalities that were poorer in terms of GDP and had worse mortality rates, which creates a selection bias. Table 14.5 provides a more detailed view of the evolution of the groups. The first frame of interpretation is that of single differences. Non-treated municipalities evidence a decrease (-1) in their average mortality rate that by construction cannot be attributed to the policy, but to some yet unspecified variables. The treated municipalities show a relatively larger decrease (-3.22) that must nevertheless be related to the initial gap (6.71) between the two groups, as well as to yet unspecified variables. This gap still remains but decreases to 4.49 with the intervention. To control for the initial differences between the two groups and for their evolution over time, the difference-in-differences approach provides a second and more accurate evaluation of the effect of the intervention:

$$\widehat{E} = \bar{y}^{11} + \bar{y}^{00} - \bar{y}^{01} - \bar{y}^{10} = 17.22 + 13.73 - 12.73 - 20.44 = -2.22$$

The actual impact of the health program on the treated group is in fact -2.22, less than the observed reduction by -3.22.

A possible extension of the method is to relax the parallel-trend assumption by including additional variables. The previous interpretation assumes a similar pattern among the treated and non-treated units. We can go beyond this statement through the use of Ordinary Least Squares (OLS) regression. For instance, according to Table 14.4, the increase in per capita GDP observed for the cities not covered by the program might have influenced their mortality rate. However, the increase in GDP observed among the treated units was much smaller, suggesting that their mortality rate would not have followed a similar track in the absence of the intervention. This could mean that we previously underestimated the effect of

Table 14.3 Database for example 1

Municipality	S	P	Mortality rate	Municipal GDP
1	0	0	15	9865
1	0	1	14	10,608
2	0	0	16	8698
2	0	1	15	9692
3	0	0	17	9520
3	0	1	16	9820
4	0	0	15	8542
4	0	1	15	9876
5	0	0	20	6200
5	0	1	19	7023
6	0	0	12	12,698
6	0	1	11	13,466
7	0	0	12	13,569
7	0	1	11	14,569
8	0	0	16	7231
8	0	1	15	8965
9	0	0	10	10,236
9	0	1	8	11,598
10	0	0	8	12,589
10	0	1	7	13,569
11	0	0	10	13,202
11	0	1	9	14,598
12	1	0	19	7566
12	1	1	15	7727
13	1	0	22	5640
13	1	1	18	5964
14	1	0	20	6720
14	1	1	17	7023
15	1	0	20	6560
15	1	1	17	6780
16	1	0	22	5201
16	1	1	19	5469
17	1	0	21	5678
17	1	1	18	6521
18	1	0	19	7021
18	1	1	16	7243
19	1	0	21	5023
19	1	1	18	6038
20	1	0	20	6541
20	1	1	17	6456

Table 14.4 Summary statistics for example 1

	S	P	Mortality rate	Municipal GDP
Non-treated group				
Before	0	0	13.73	10213.64
After	0	1	12.73	11253.09
Treated group				
Before	1	0	20.44	6216.67
After	1	1	17.22	6580.11

Table 14.5 Basic difference-in-differences: example 1

	Before intervention $P=0$	After intervention $P=1$	Δ after/before
Non-treated group $S=0$	$\bar{y}^{00}=13.73$	$\bar{y}^{01}=12.73$	$\bar{y}^{01}-\bar{y}^{00}=-1$
Treated group $S=1$	$\bar{y}^{10}=20.44$	$\bar{y}^{11}=17.22$	$\bar{y}^{11}-\bar{y}^{10}=-3.22$
Δ treated/non-treated	$\bar{y}^{10}-\bar{y}^{00}=6.71$	$\bar{y}^{11}-\bar{y}^{01}=4.49$	$\widehat{E}=-2.22$

the intervention. We may relax the parallel-trend assumption by including GDP per capita as an additional variable in the analysis. Figure 14.2 provides the coding to be used in R-CRAN (our programs only display outputs directly used in the analysis). Command *read.table* is used to upload the database in R-CRAN using the path *C://mydataDID.csv* (that denotes the location of the file), saved afterwards under the name "D". The file format is *.csv*, with ";" as a separator, and can be created with Excel. The command *head* displays the first rows of the dataset.

The first regression (*reg*1) consists in reproducing the double-difference results using OLS (command *lm*). By default, in R-CRAN, one needs to specify only the interaction term $S*P$ in the *lm* command; the software will automatically include S and P with the interaction term. As expected, we obtain the same result as previously. The constant (*intercept*) amounts to 13.73, which is the average outcome $\bar{y}^{00}$ for the non-treated group before the intervention (see Table 14.5). The second coefficient stands for the single difference between the treated and non-treated for the pre-intervention period $\left(\bar{y}^{10}-\bar{y}^{00}=20.44-13.73=6.71\right)$. The third and here non-significant coefficient represents the decrease in the mortality rate observed for the municipalities not covered by the program $\left(\bar{y}^{01}-\bar{y}^{00}=12.73-13.73=-1\right)$. Last, the coefficient on the interaction term yields the effect of the intervention (−2.22), which is also found to be non-significant. The second regression (*reg*2) extends the model through the inclusion of the GDP per capita among the covariates. The coefficient on the interaction term is much higher, as anticipated; it is significant and amounts to −3.07. Command *confint* yields the 95% confidence interval for the interaction term, namely [−5.05; −1.10]. The implementation of the program is therefore associated with a significant reduction in mortality.

```
> D=read.table("C://mydataDID.csv",head=TRUE,sep=";")
> reg1=lm(Mortality~S*P,D)
> summary(reg1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7273     0.8618  15.928  < 2e-16 ***
S             6.7172     1.2848   5.228 7.47e-06 ***
P            -1.0000     1.2188  -0.820    0.417
S:P          -2.2222     1.8169  -1.223    0.229
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> reg2=lm(Mortality~GDP+S*P,D)
> summary(reg2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.5344612  1.4188103  18.702  < 2e-16 ***
GDP         -0.0012539  0.0001314  -9.543 2.84e-11 ***
S            1.7052502  0.8643609   1.973  0.05645 .
P            0.3034036  0.6654613   0.456  0.65126
S:P         -3.0698918  0.9749484  -3.149  0.00335 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> confint(reg2, "S:P")
        2.5 %    97.5 %
S:P -5.049142 -1.090641
```

Fig. 14.2 Difference-in-differences with R-CRAN: example 1

14.3 Propensity Score Matching

The idea behind matching is to select and pair units that would be identical in everything but the fact of receiving the intervention. Several matching algorithms exist. One difficulty they share is that the units may differ in more than one variable, which yields a problem known as the curse of dimensionality. To overcome that, the propensity score matching method estimates a model of the probability of participating in the treatment using a set of observed characteristics (overt bias), and then uses the fitted values of this model to match the units. It thus allows the multidimensional problem to be reduced to a single dimension: that of the propensity score. If the score is accurately computed, the outcome observed for the comparison group should provide a satisfactory counterfactual.

Figure 14.3 illustrates the approach. The main focus is on the post-intervention period, although the scores are often estimated based on pre-intervention characteristics. The orange dots represent the treated units ($S = 1$), while the blue ones represent those units that did not receive the intervention ($S = 0$). The counterfactual is approximated using the units that could have been selected in theory (with similar propensity scores), but were not. Matched units are indicated with two squares connected by a dotted line. By matching on the propensity score, we are able to construct two comparable groups. The difference in mean outcome between these groups yields the estimated impact of the intervention. As can be seen, one condition for using the method is the existence of a sufficient overlap between the propensity scores. This is known as the common support condition. For example, in

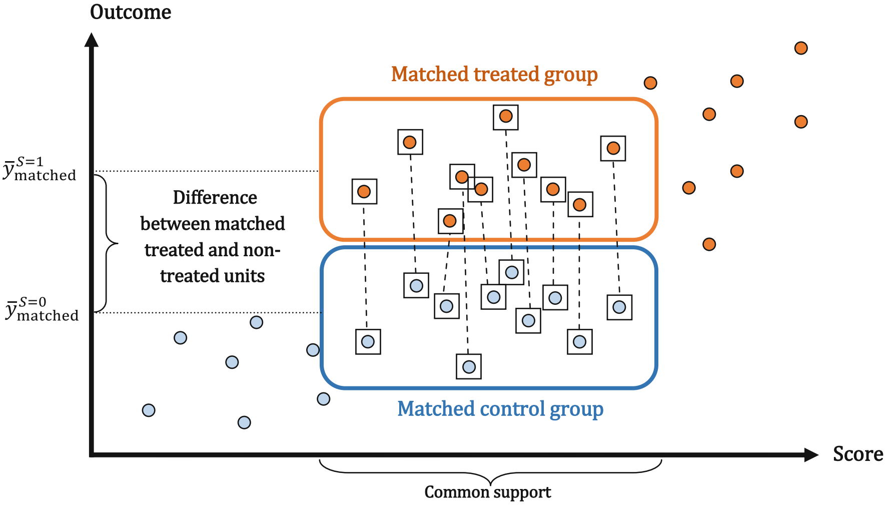

Outcome
Matched treated group
$\bar{y}^{S=1}_{\text{matched}}$
Difference between matched treated and non-treated units
$\bar{y}^{S=0}_{\text{matched}}$
Matched control group
Score
Common support

Fig. 14.3 The propensity score matching approach

Fig. 14.3, those units with very low and very high propensity scores, respectively to the left and to the right, are excluded from the analysis.

Propensity scores are derived from a qualitative response regression that estimates the probability of a unit's exposure to the intervention, conditional on a set of observable characteristics that may affect participation in the program. For instance, a logit model can be used and be specified as:

$$\ln\left(\frac{S_i}{1-S_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K\, x_{Ki}$$

where i stands for unit i, S_i specifies whether unit i belongs to the treatment group, the x's represent the individual characteristics, and the β's are the coefficients to be estimated. Once estimated, the model yields the propensity score, defined as the estimated probability $\widehat{S}_i$ that unit i receives treatment, given a vector of observed covariates.

An important problem with respect to propensity score matching is to identify the x variables to be included in the model. In general, any variable that is thought to simultaneously influence the exposure S and the outcome variable y should be included. One should not use variables observed after the intervention, as they could themselves be influenced by the intervention. Thus, a crucial issue is the availability of characteristics observed before the intervention takes place.

Once the model has been estimated, the treated units are matched to the non-treated units that are most similar in terms of their propensity score $\widehat{S}$. The two most common methods are nearest neighbor matching and caliper matching. With nearest neighbor matching, each unit in the treatment group is matched with a unit from the control group that is closest in terms of propensity score. The second

Table 14.6 Data for example 2

Unit	Score	Outcome	Average outcome
	Control group		26.67
1	0.2	10	
2	0.3	30	
3	0.4	40	
	Treatment group		56.25
4	0.3	45	
5	0.4	50	
6	0.5	60	
7	0.6	70	

method uses a standardized distance which is acceptable for any match. This tolerance level is imposed on the propensity scores and observations which are outside of the caliper are dropped. As a matter of comparison, caliper matching generally gives more accurate results, as nearest neighbor matching may link units with very different scores if no closer match is available.

The average treatment effect is estimated by computing the difference in means between the two groups:

$$\widehat{E} = \bar{y}^{S=1}_{\text{matched}} - \bar{y}^{S=0}_{\text{matched}}$$

Three measures of $\widehat{E}$ exist depending on whether the focus is on the sample average treatment effect for the treated group (ATT), the sample average treatment effect for the control group (ATC), or the sample average treatment effect (ATE). The following example 2 illustrates the difference (Table 14.6). The non-treated group consists of three units (controls 1, 2 and 3), while the treated group comprises four units (treated 4, 5, 6 and 7). Each unit displays a score value (in this illustrative case, unidimensional and exogenous) and a related outcome. Figure 14.4 shows the corresponding scatter plot, with blue dots for the control group and orange dots for the treated group. Without matching, the difference in outcome means between the two groups is:

$$\widehat{E}_{\text{unmatched}} = \bar{y}^{S=1}_{\text{unmatched}} - \bar{y}^{S=0}_{\text{unmatched}} = 56.25 - 26.67 = 29.58$$

The matching principle takes into account the fact that a number of units are relatively distant from the others and that it would be more relevant to compare the outcomes of units with neighboring locations by giving more weights to central observations.

As was mentioned before, the nearest neighbor matching may link units in three different ways. With the ATT method, the focus is on the treated units (4, 5, 6 and 7) and how they differ from their non-treated counterparts (controls 1, 2 and 3). Figure 14.5 displays the corresponding matching. All four orange dots are matched. The analysis excludes unit 1 as it differs too much from any potential counterpart.

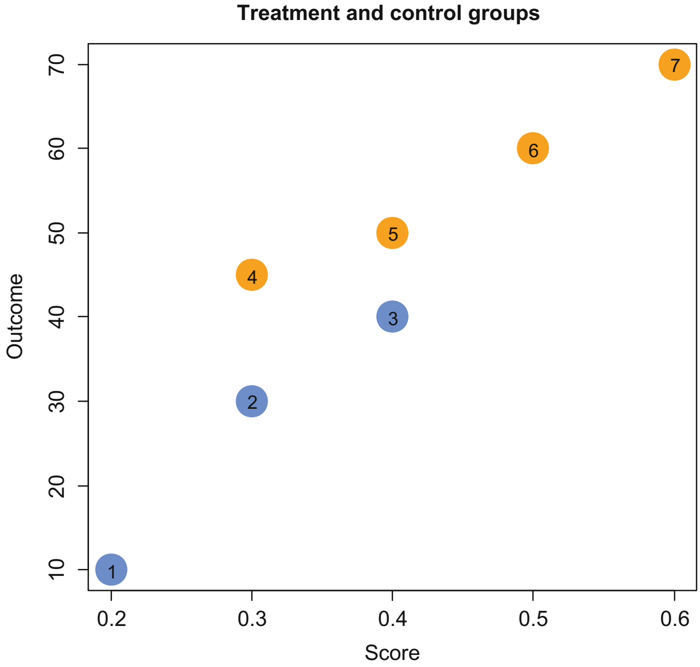


Fig. 14.4 Scatter plot for example 2

With the ATC method, the focus is reversed and the blue dots of the control group are matched to their nearest neighbor treated units, which leaves treated units 6 and 7 unmatched. Table 14.7 shows how the corresponding differences in means are calculated. More generally, one would prefer to combine both approaches and compute the average treatment effect (ATE) instead. Hand calculations from Table 14.6 give the following differences in outcome means between the two groups:

$$\widehat{E}(ATT) = \frac{(45+50+60+70)}{4} - \frac{(30+40+40+40)}{4} = 18.75$$

$$\widehat{E}(ATC) = \frac{(45+45+50)}{3} - \frac{(10+30+40)}{3} = 20$$

$$\begin{aligned}\widehat{E}(ATE) &= \frac{(45+50+60+70)+(45+45+50)}{7} \\ &\quad - \frac{(30+40+40+40)+(10+30+40)}{7} \\ &= 19.286\end{aligned}$$

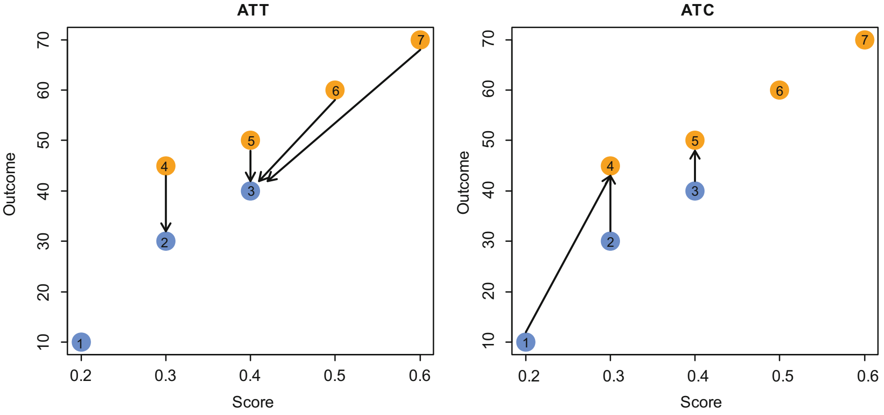


Fig. 14.5 ATT and ATC matching methods: example 2

Table 14.7 ATT, ATC and ATE average outcomes

ATT	Outcome	Matched outcome	Matched control unit
Unit	Treatment group	Control group	Unit
4	45	30	2
5	50	40	3
6	60	40	3
7	70	40	3
ATC	Outcome	Matched outcome	Matched treated unit
Unit	Control group	Treatment group	Unit
1	10	45	4
2	30	45	4
3	40	50	5

Mean outcome differences that take into account matching do differ from the unmatched result ($\widehat{E}_{\text{unmatched}} = 29.58$).

Figure 14.6 provides the coding to be used in R-CRAN to run ATT, ATC and ATE analyses. The package *Matching* is uploaded via the *library* command. The first step consists in creating the data to be matched. By order of appearance, *S* is a vector stating which units have been treated or not. In our example, the first three units are coded 0 which designates them as controls, code 1 implies that the corresponding unit is treated; *Score* denotes the variable we wish to match on (in this illustrative case, its values are exogenous); *Outcome* is our variable of interest; and *Units* indexes the individuals. The command *Match* is then used to compute the average treatment effect for the treated (*mymatch*1), for the controls (*mymatch*2), and the ATE measurement (*mymatch*3). In each case, we have to

```
> library(Matching)
> S=c(0,0,0,1,1,1,1)
> Score=c(0.2,0.3,0.4,0.3,0.4,0.5,0.6)
> Outcome=c(10,30,40,45,50,60,70)
> Units=c(1,2,3,4,5,6,7)
>
> mymatch1=Match(Y=Outcome,Tr=S,X=Score,estimand="ATT")
> summary(mymatch1)

Estimate...  18.75
AI SE......  4.8914
T-stat.....  3.8333
p.val......  0.00012646

Original number of observations..............  7
Original number of treated obs...............  4
Matched number of observations...............  4
Matched number of observations  (unweighted).  4

> mymatch1$index.treated
[1] 4 5 6 7
> mymatch1$index.control
[1] 2 3 3 3
>
> mymatch2=Match(Y=Outcome,Tr=S,X=Score,estimand="ATC")
> summary(mymatch2)

Estimate...  20
AI SE......  7.2008
T-stat.....  2.7775
p.val......  0.0054786

Original number of observations..............  7
Original number of control obs...............  3
Matched number of observations...............  3
Matched number of observations  (unweighted).  3

> mymatch2$index.treated
[1] 4 4 5
> mymatch2$index.control
[1] 1 2 3
>
> mymatch3=Match(Y=Outcome,Tr=S,X=Score,estimand="ATE")
> summary(mymatch3)

Estimate...  19.286
AI SE......  5.4761
T-stat.....  3.5218
p.val......  0.00042862

Original number of observations..............  7
Original number of treated obs...............  4
Matched number of observations...............  7
Matched number of observations  (unweighted).  7

> mymatch3$index.treated
[1] 4 4 5 4 5 6 7
> mymatch3$index.control
[1] 1 2 3 2 3 3 3
```

Fig. 14.6 ATT, ATC and ATE with R-CRAN: example 2

specify the outcome of interest (Y), the treatment variable (Tr), the score (X), and the type of measure (*estimand*). By running $\$index\,.\,treated$ and $\$index\,.\,control$ we display the seven observation numbers from the original dataset used for the matching, both for the treated and non-treated units, respectively. As can be seen, ATT focuses on those units that were exposed (4, 5, 6, 7), ATC on the non-treated units (1, 2, 3), while ATE takes into account all of them. The estimates correspond to the previous hand calculations. The three methods conclude to a smaller effect than without matching by giving more weights to central observations. R-CRAN also provides the standard error (*AI SE*) and the results of a test of difference between means (*T*-*stat* and *p*-*val*). When the *p*-value is lower than 5%, the observed difference is statistically significant, which is the case here for all three matching methods.

A difficulty arises sometimes when two or more units from the same group have the same score, which results in a tie problem. For instance, a treated unit may match with two or even more control units. Table 14.8 provides the data for example 3 as an illustration of the tie problem. A possible way to deal with this issue is to consider all the possible matched units and then use adequate weights to reflect the multiple combinations. For example, in Fig. 14.7 where ATT is estimated, treated units 6 and 7 each match both controls 2 and 3. There is a single match from 5 to unit 1. Unit 4 does not provide any match. The analysis thus considers a total of five matching couples, as shown in the R-CRAN program of Fig. 14.8. The ATT mean difference estimate is 43.33. Another but less preferable strategy is to break the ties randomly, and to consider only three couples. The code is *ties=FALSE*. As shown in Fig. 14.8, if the ties are randomly broken, treatment units 6 and 7 are both matched with one and only one of the closest control units (in terms of score). For instance, in Fig. 14.8, the algorithm randomly assigns unit 2 to units 6 and 7 (see *mymatch*2). In that case, the mean difference estimation rises to 53.33. Other iterations of the *Match* command with the argument *ties=FALSE* would generate the other possible combinations. For instance, unit 6 or unit 7 can be randomly assigned to unit 3.

Let us now consider a more general application of the method (example 4). Assume that one aims to compare two alternative treatments for lung cancer, namely strategy 0 and strategy 1. The data is cross-sectional (outcomes are observed at the same time after treatment) and presented in Table 14.9. It consists of 60 patients of different age (*Age*) who used to smoke or not (*Smoker*). The outcome of interest is survival at two years, represented by the variable *Death*. Survival is expressed by $Death = 0$. All patients underwent treatment, either with strategy 0 ($S = 0$) or with strategy 1 ($S = 1$). Table 14.10 provides the summary statistics for each group. At first sight, strategy 0 seems to yield the lower mortality rate, but patients who underwent strategy 1 are also on average both older and more likely to be smokers, thus pointing out a plausible selection bias: the treatment was not randomly assigned. A matching strategy has to be implemented to better assess the difference between the two strategies.

The codes used in R-CRAN are presented in Fig. 14.9. The first step consists in downloading the data via the $read\,.\,table$ command. Then, propensity scores are

Table 14.8 Data for example 3

Unit	Score	Outcome	Average outcome
	Control group		27.50
1	0.10	10	
2	0.35	20	
3	0.35	50	
4	0.60	30	
	Treatment group		70
5	0.10	60	
6	0.30	80	
7	0.40	70	

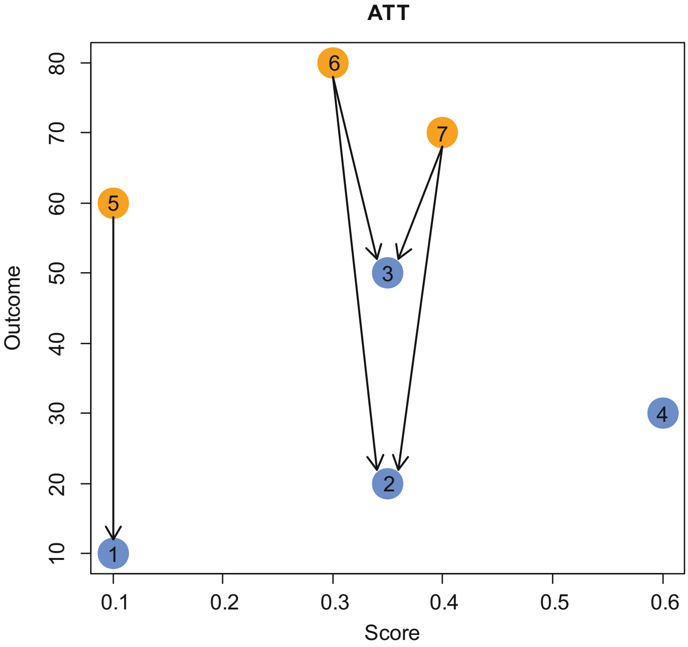


Fig. 14.7 ATT with ties: example 3

estimated via a logit regression using the *glm* function. The model specifies *S* as a function of *Age* and *Smoker*. It can be seen from the estimation results that both the age and being a smoker increase the probability of belonging to group $S = 1$. Option $x = TRUE$ in the *glm* command indicates whether the exogenous variables should be saved for subsequent analysis. This is important if we want to use the package *erer* and the command *maBina* to compute marginal effects. Since the output of a logit regression cannot be readily interpreted, the computation of marginal effects may be indeed helpful. We can deduct from Fig. 14.9 that, on average, being one year older increases the probability of undergoing strategy 1 by 6.7%. On average,

```
> library(Matching)
> S=c(0,0,0,0,1,1,1)
> Score=c(0.1,0.35,0.35,0.6,0.1,0.3,0.4)
> Outcome=c(10,20,50,30,60,80,70)
> Units=c(1,2,3,4,5,6,7)

> mymatch1=Match(Y=Outcome,Tr=S,X=Score,estimand="ATT")
> summary(mymatch1)

Estimate...  43.333
AI SE......  4.8432
T-stat.....  8.9472
p.val......  < 2.22e-16

Original number of observations..............  7
Original number of treated obs...............  3
Matched number of observations...............  3
Matched number of observations  (unweighted).  5

> mymatch1$index.treated
[1] 5 6 6 7 7
> mymatch1$index.control
[1] 1 2 3 2 3

# Random tie break
> mymatch2=Match(Y=Outcome,Tr=S,X=Score,estimand="ATT",ties=FALSE)
> summary(mymatch2)

Estimate...  53.333
SE.........  2.7217
T-stat.....  19.596
p.val......  < 2.22e-16

Original number of observations..............  7
Original number of treated obs...............  3
Matched number of observations...............  3
Matched number of observations  (unweighted).  3

> mymatch2$index.treated
[1] 5 6 7
> mymatch2$index.control
[1] 1 2 2
```

Fig. 14.8 ATT and ties with R-CRAN: example 3

being a smoker increases this probability by 59.1%. The fitted values of our regression are then saved in database *D* using *D*$*Score* = *mylogit*$*fitted*.*values*.

Rather than matching on all the exogenous characteristics, individual units are compared on the basis of their propensity scores. The matching is implemented using the package *Matching* and the function *Match*. We specify that the outcome variable (*Y*) is *Death*, the treatment variable (*Tr*) is *S*, the variable on which the observations are matched (*X*) is *Score*. A caliper is set to 0.2 in this analysis, although other values could be chosen. It means that all matches not equal to or within 0.2 standard deviations of the propensity score are dropped. The command *mymatch*$*ecaliper* shows that the tolerance distance is set at 6.4%.

By default, potential ties are taken into account in the *Match* command (we have *ties* = *TRUE*). If, for example, one treated observation matches more than one control observation, then all of them are taken into account for estimating the effect. From the results, we can see that 25 observations were excluded using the caliper command. We finally have 63 matched observations which include multiple matching. The estimated impact is significant and amounts to −0.41. This means that strategy 1 yields in fact a mortality rate 41% lower than strategy 0. According

Table 14.9 Data for example 4

Patient	S	Age	Smoker	Death
1	0	63	1	1
2	0	63	1	1
3	0	56	0	0
4	0	61	1	1
5	0	64	0	1
6	0	61	1	1
7	0	64	1	1
8	0	61	1	1
9	0	69	0	1
10	0	61	1	1
11	0	67	0	0
12	0	49	0	0
13	0	57	0	0
14	0	59	0	0
15	0	69	0	0
16	0	61	1	1
17	0	63	0	0
18	0	69	0	1
19	0	59	0	0
20	0	64	0	0
21	0	64	0	1
22	0	41	0	0
23	0	61	0	0
24	0	54	1	0
25	0	64	1	1
26	0	61	0	0
27	0	68	0	0
28	0	40	1	0
29	0	65	0	0
30	0	69	0	1
31	0	64	0	0
32	0	64	0	1
33	1	61	1	0
34	1	73	0	0
35	1	65	1	1
36	1	79	0	1
37	1	56	1	0
38	1	69	1	1
39	1	62	1	0
40	1	73	1	1
41	1	93	1	1
42	1	61	1	0
43	1	62	1	0

(continued)

Table 14.9 (continued)

Patient	S	Age	Smoker	Death
44	1	69	1	1
45	1	63	0	0
46	1	87	0	1
47	1	64	0	0
48	1	71	1	1
49	1	59	1	1
50	1	68	0	0
51	1	68	0	0
52	1	68	1	1
53	1	67	1	1
54	1	72	1	1
55	1	59	1	1
56	1	87	1	1
57	1	81	1	1
58	1	76	0	0
59	1	68	1	0
60	1	73	0	0

Table 14.10 Summary statistics for example 4

	Strategy	Death	Age	Smoker
Non-treated group	$S = 0$	46.87%	61.09	34.37%
Treated group	$S = 1$	53.57%	69.79	67.85%

to the *p*-value, this difference is statistically significant. Strategy 1 is therefore associated with a significant reduction in the mortality rate. Notice that the standard error provided by the *Match* command is the Abadie-Imbens standard error, which takes into account the uncertainty of the matching procedure.

At this stage, a question could be raised: why not use a regression model instead of the matching procedure? For instance, we could estimate *Death* as a function of *Age*, *Smoker*, and *S* to assess the impact of the strategies. In theory, this would yield the impact of *S ceteris paribus*, i.e. everything else being equal. Yet, this is true if and only if the regression model is not miss-specified. A key feature of any parametric regression analysis is that the shape of the functional relationships between the explained and the explanatory variables are predetermined. On the contrary, a matching strategy does not presume any functional form (except for estimating the propensity score). It has thus some non-parametric aspects in this respect, the main feature being that the effect is estimated only on a selected number of units with comparable characteristics. Moreover, when the selection bias is large, traditionnal econometrics will fail to estimate the true effect of the intervention due to multicollinearity problems (correlation among regressors).

Another asset of matching is that one can formally assess the quality of the match by using two-sample *t*-tests of exogenous variables or by calculating the average standardized bias *SB*:

```
> D=read.table("C://mydataPSM.csv",head=TRUE,sep=";")

> mylogit=glm(S~Age+Smoker,D,family=binomial,x=TRUE)
> summary(mylogit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.88718    5.77310  -3.272  0.00107 **
Age           0.26964    0.08532   3.160  0.00158 **
Smoker        2.37366    0.79753   2.976  0.00292 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(erer)
> maBina(mylogit)$out
            effect error t.value p.value
(Intercept) -4.701 1.440  -3.264   0.002
Age          0.067 0.021   3.140   0.003
Smoker       0.531 0.143   3.159   0.000

> D$Score=mylogit$fitted.values

> library(Matching)
> mymatch=Match(Y=D$Death,Tr=D$S,X=D$Score,
+ estimand="ATE",caliper=0.2)

> mymatch$ecaliper
[1] 0.06365222

> summary(mymatch)

Estimate...  -0.41714
AI SE......  0.11858
T-stat.....  -3.5177
p.val......  0.00043533

Original number of observations..............  60
Original number of treated obs...............  28
Matched number of observations...............  35
Matched number of observations  (unweighted).  63

Caliper (SDs)........................................   0.2
Number of obs dropped by 'exact' or 'caliper'  25
```

Fig. 14.9 Matching strategy with R-CRAN: example 4

$$SB = \frac{\bar{x}^{S=1}_{\text{matched}} - \bar{x}^{S=0}_{\text{matched}}}{\sigma^{S=1}_{\text{matched}}}$$

For each covariate x, the standardized bias is defined as the difference of sample means in the matched database, divided by the standard deviation of the matched treatment group. While there are no formal rules, a standardized bias after matching between 3 and 5% is usually considered as sufficient, while a standardized bias of 20% is considered large. Notice that the literature uses also other measures for the denominator, such as the square root of the average of sample variances, or the pooled standard deviation.

Figure 14.10 explains how to judge whether the matched units are comparable. First, the package *Matching* is uploaded, and the previous results are reproduced by estimating the propensity score (*glm* command), then saving the fitted-values of that regression in *D*$*Score*, and finally implementing the *Match* function with a caliper

```
> library(Matching)
> D=read.table("C://mydataPSM.csv",head=TRUE,sep=";")
> mylogit=glm(S~Age+Smoker,D,family=binomial,x=TRUE)
> D$Score=mylogit$fitted.values
> mymatch=Match(Y=D$Death,Tr=D$S,X=D$Score,
+ estimand="ATE",caliper=0.2)

> MatchBalance(S~Age+Smoker,D, match.out=mymatch,ks=FALSE)

***** (V1) Age *****
                       Before Matching          After Matching
mean treatment........     69.786                   63.429
mean control..........     61.094                   64.171
std mean diff.........     95.413                  -16.752

mean raw eQQ diff.....     9.0714                   1.5397
med  raw eQQ diff.....          7                        1
max  raw eQQ diff.....         24                        5

mean eCDF diff........    0.21603                  0.07215
med  eCDF diff........    0.17857                 0.063492
max  eCDF diff........    0.45982                  0.22222

var ratio (Tr/Co).....     1.6984                   1.7926
T-test p-value........ 0.00014981                  0.42543

***** (V2) Smoker *****
                       Before Matching          After Matching
mean treatment........    0.67857                  0.48571
mean control..........    0.34375                  0.37143
std mean diff.........     70.401                   22.537

mean raw eQQ diff.....    0.35714                 0.079365
med  raw eQQ diff.....          0                        0
max  raw eQQ diff.....          1                        1

mean eCDF diff........    0.16741                 0.039683
med  eCDF diff........    0.16741                 0.039683
max  eCDF diff........    0.33482                 0.079365

var ratio (Tr/Co).....    0.97135                   1.0699
T-test p-value........   0.009057                  0.28471

Before Matching Minimum p.value: 0.00014981
Variable Name(s): Age  Number(s): 1

After Matching Minimum p.value: 0.28471
Variable Name(s): Smoker  Number(s): 2

> head(mymatch$index.treated)
[1] 33 42 47 33 42 34
> head(mymatch$index.control)
[1] 4 4 5 6 6 7
> head(D$Score[mymatch$index.treated])
[1] 0.4836383 0.4836383 0.1638015 0.4836383 0.4836383 0.6892361
> head(D$Score[mymatch$index.control])
[1] 0.4836383 0.4836383 0.1638015 0.4836383 0.4836383 0.6777499
```

Fig. 14.10 Quality of the match in R-CRAN: example 4

equal to 0.2. The output is saved under the name *mymatch*. The *MatchBalance* command is then used to determine if the matching was successful in achieving balance on the observed covariates. We have to specify first the list of the variables we wish to obtain univariate balance statistics for (namely, $S \sim Age + Smoker$), then the database we used (D), and finally the output object from the *Match* function ($match\,.\,out = mymatch$). If no *Match* output is included, balance statistics will only be reported for the raw unmatched data. The option $ks = FALSE$ excludes a set of statistics that are not necessary for our analysis.

The full output for the *MatchBalance* call is presented in Fig. 14.10. Two different sets of statistics are provided. First, in interpreting these results, we would like the after-matching means to be as close as possible for each covariate. The most important statistic in this respect is the T-test p-value which provides the result of a test of difference between means, both before and after matching. As can be seen, there was a significant difference in *Age* and *Smoker* before the matching. This is not the case anymore after matching, which supports our analysis. It is also possible to examine the standardized bias, *std mean diff*, which indicates the percentage difference in means between the treated and control groups ($\times 100$). Before matching, those differences are high. After matching, they decrease and amount to -16.75 for *Age*, and 22.54 for *Smoker*. These values are still large (close to or higher than 20%), which means that we could refine our analysis using a smaller caliper.

Second, we may also examine the difference in distribution between the two matched groups. In this matter, the *MatchBalance* output contains summary statistics based on empirical QQ-plots. A QQ-plot is a graphical method for comparing two distributions by plotting their quantiles one against each other. To illustrate, assume that our dataset consists of three matched couples: three non-treated units aged 65, 63 and 74 are matched with three treated units aged 67, 65, and 64, respectively. A QQ-plot first orders these vectors and then plots the non-treated group $(63, 65, 74)$ against the treated group $(64, 65, 67)$. Should the size of the groups be different, R-CRAN would linearly interpolate data points so that the vector sizes match. The differences obtained in absolute value $(1, 0, 7)$ are then used to provide additional statistics. For instance, the mean (*mean raw eQQ diff*), median (*med raw eQQ diff*) and maximum (max *raw eQQ diff*) differences are provided. In our 3-couple example, this would for instance yield a mean value equal to (1+0+7)/3, a median value equal to 1, and a maximum value equal to 7. An additional set of statistics consists of summary statistics based on the standardized empirical-QQ plots: *mean eCDF diff*, *med eCDF diff* and max *eCDF diff*. What matters most is that those QQ-plot statistics decrease after matching, approaching 0. Last, the *MatchBalance* call provides the variance ratio of treatment over control *var ratio* (*Tr/Co*), which should be equal to 1 were there perfect balance.

It is also possible to assess the quality of the matching by comparing manually the observations that have been matched, using $*index . treated* and $*index . control*. In Fig. 14.10, the command *head* provides the first six observation numbers in each group. Unit 4 ($S = 0$) has been matched with units 33 and 42 ($S = 1$), which seems rather relevant given their characteristics (see Table 14.9). They all are 61 years old smokers. Unit 5 has been matched with unit 47, and they are both 64 years old and non-smokers. The matching is however not always this precise. For instance unit 7, a 64 years old smoker, is matched with patient 34, a 73 years old non-smoker. Their scores however are similar (67% versus 68%) as their characteristics compensate each other.

14.4 Regression Discontinuity Design

Regression discontinuity design elicits the effect of an intervention by comparing a treatment group and a comparison group around a threshold above or below which the intervention is dispensed. Assume for instance that a central government makes funds available for municipalities with less than five thousands inhabitants. To estimate the effect of such a policy one would have to examine municipalities with comparable characteristics. To do so, regression discontinuity design exploits the discontinuity in treatment by comparing only the municipalities in the vicinity of the cutoff point, i.e. those with a population slightly lower than 5000 (the treatment group) and those with a population slightly higher (the comparison group). The underlying assumption of the method is the following: by examining observations lying close to either side of the threshold, one should eliminate selection biases.

Formally, the approach consists in estimating the following model:

$$y_i = \beta_0 + \beta_1 S_i + \beta_2(x_i - x_c) + \beta_3 S_i(x_i - x_c) + \epsilon_i$$

where i stands for unit i; y_i denotes the outcome variable; S_i specifies whether unit i has been covered by the program; x_i is the assignment variable (so called running or forcing variable) upon which the treatment cutoff x_c is applied; the β's are the coefficients to be estimated; ϵ_i is an error term. Once estimated, the model yields on average:

$$\bar{y}^{S}_{\text{threshold}} = \widehat{\beta}_0 + \widehat{\beta}_1 S \text{ when } x_i \approx x_c$$

In other terms, around the threshold, the average outcome for the non-treated units (when $S = 0$) is $\widehat{\beta}_0$, while the average outcome for the treated ($S = 1$) is $\widehat{\beta}_0 + \widehat{\beta}_1$. The effect of the intervention is thus measured as:

$$\widehat{E} = \bar{y}^{S=1}_{\text{threshold}} - \bar{y}^{S=0}_{\text{threshold}} = \widehat{\beta}_1$$

From Fig. 14.11, we can see that the approach is equivalent to estimating two regression lines in a centered plan around the threshold. The estimated coefficient $\widehat{\beta}_0$ yields the intercept of the line for the non-treated group, and $\widehat{\beta}_2$ is the slope of that line. For the treated group, the intercept is given by $\widehat{\beta}_0 + \widehat{\beta}_1$ and the slope by $\widehat{\beta}_2 + \widehat{\beta}_3$.

Why go to the trouble of estimating an econometric model if we are interested only in a simple comparison of means around the threshold? In theory, comparing the treated units with the non-treated units in the neighborhood of x_c would be sufficient to establish impact. To do so, however, one would have to base the analysis on a very small set of observations, made only of those units that are in the close vicinity of x_c. This could yield highly inconsistent results. A regression analysis has the advantage of exploiting a higher number of observations by extrapolating the effect using estimated lines. However, it is not advised to use

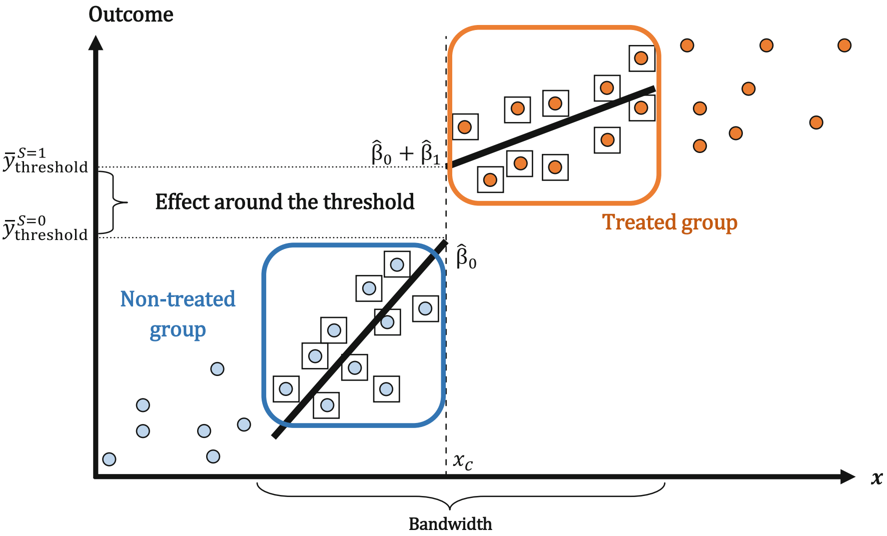

Outcome
$\bar{y}^{S=1}_{\text{threshold}}$
$\bar{y}^{S=0}_{\text{threshold}}$
Effect around the threshold
$\hat{\beta}_0 + \hat{\beta}_1$
$\hat{\beta}_0$
Treated group
Non-treated group
x_c
x
Bandwidth

Fig. 14.11 Principle of regression discontinuity design

the whole sample as the relationship between the outcome variable y and the running variable x could be unstable with respect to extreme observations. In practice, an optimal bandwidth around the threshold is selected, as we shall see later.

Regression discontinuity design requires that the units considered cannot manipulate their treatment status. Manipulation means that the running variable x for some units could be changed from their true values to influence treatment assignment. Graphically, this is analogous to expecting the probability density of x to be discontinuous at the cutoff point, with an unexpected high number of units on one side of the threshold, and a lower frequency on the other side, as illustrated in Fig. 14.12. Assume for instance that a hospital plans to evaluate the effect of a new treatment protocol for inpatients with a given diagnosis. Let us say that the running variable is defined as a quantitative indicator measuring the severity of disease, e.g., from 1 to 100, and that the cutoff is set arbitrarily to 50. If patients learn of the assignment mechanism, some of them might try to hide their true health status to increase the likelihood of participating in the new treatment. Similarly, municipalities may manipulate their demographic data in order to unduly benefit from central government funds. Ultimately, this might produce biased estimates (we would not have a situation that resembles randomization in the neighborhood of the threshold). Demonstrating the statistical integrity of the running variable x is thus of high importance. For this purpose, one may use statistical tests, such as the McCrary test, to establish the smoothness of the density of the running variable around the cutoff.

Let us now consider for instance the effect of mandatory attendance on exam performance in college (example 5). Imagine that students who scored below the

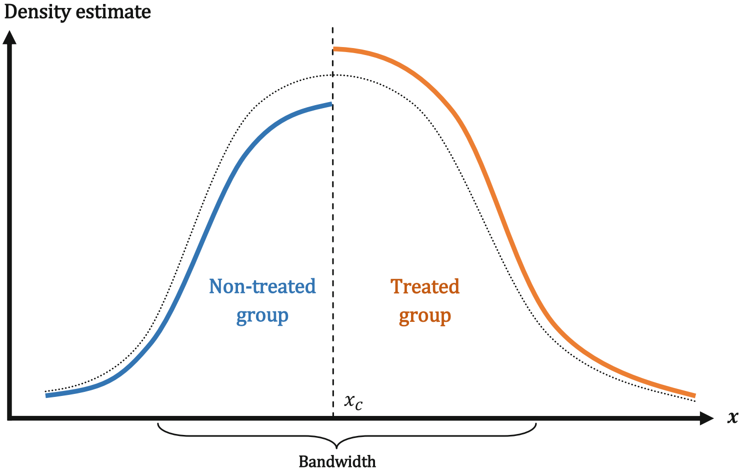


Fig. 14.12 Density of the running variable with manipulation

mean during semester 1 ($S = 1$) are assigned mandatory attendance during semester 2, while students above the mean ($S = 0$) are not. Another similar situation would be that of a disease treatment where self-administration of the drug ($S = 0$) would concern patients with a health status index above the mean, while patients with an index below the mean would receive the assistance of a nurse ($S = 1$) for better treatment compliance. Coming back to example 5, one would like to estimate the effect of mandatory course attendance on students' results. Table 14.11 presents the data. Variable *Ind* indexes the individuals; S is the treatment variable; $Grade1$ and $Grade2$ stand for the score obtained at the end of semester 1 and semester 2, respectively (it is a number between 0 and 100). $Grade1$ is thus the running variable and $Grade2$ is the outcome variable. Note that in this particular example, the groups are of the same size. The approach can however be applied indifferently to groups of different size.

Table 14.12 offers summary statistics. As can be seen, one cannot compare so easily the treated units with the non-treated units. Students belonging to group $S = 1$ have the lowest scores, respectively 41.20 and 45.46 on average for the two semesters, while students from group $S = 0$ have obtained 68.33 and 68.50 on average. Due to the selection mechanism itself, a naïve comparison of the grades obtained after the intervention would thus fail to estimate the true effect of the program. Using the difference-in-differences approach (see Sect. 14.2), we would obtain the following impact:

$$\widehat{E} = 45.46 + 68.33 - 68.50 - 41.20 = 4.09$$

However, the approach is not necessarily appropriate here as the selection bias is likely to vary through time: those with the highest scores, i.e. who already have

Table 14.11 Data for example 5

Ind	S	Grade1	Grade2
1	1	6	11
2	1	21	2
3	1	24	2
4	1	25	16
5	1	26	34
6	1	28	39
7	1	33	5
8	1	37	49
9	1	38	22
10	1	38	53
11	1	40	19
12	1	41	37
13	1	42	55
14	1	44	56
15	1	45	46
16	1	46	48
17	1	46	60
18	1	46	60
19	1	47	57
20	1	47	62
21	1	49	57
22	1	49	64
23	1	49	64
24	1	50	59
25	1	50	64
26	1	53	67
27	1	53	64
28	1	53	66
29	1	55	66
30	1	55	60
31	0	55	59
32	0	56	59
33	0	57	65
34	0	57	62
35	0	57	57
36	0	58	58
37	0	59	67
38	0	60	61
39	0	61	60
40	0	61	68
41	0	63	63
42	0	65	69
43	0	66	64

(continued)

Table 14.11 (continued)

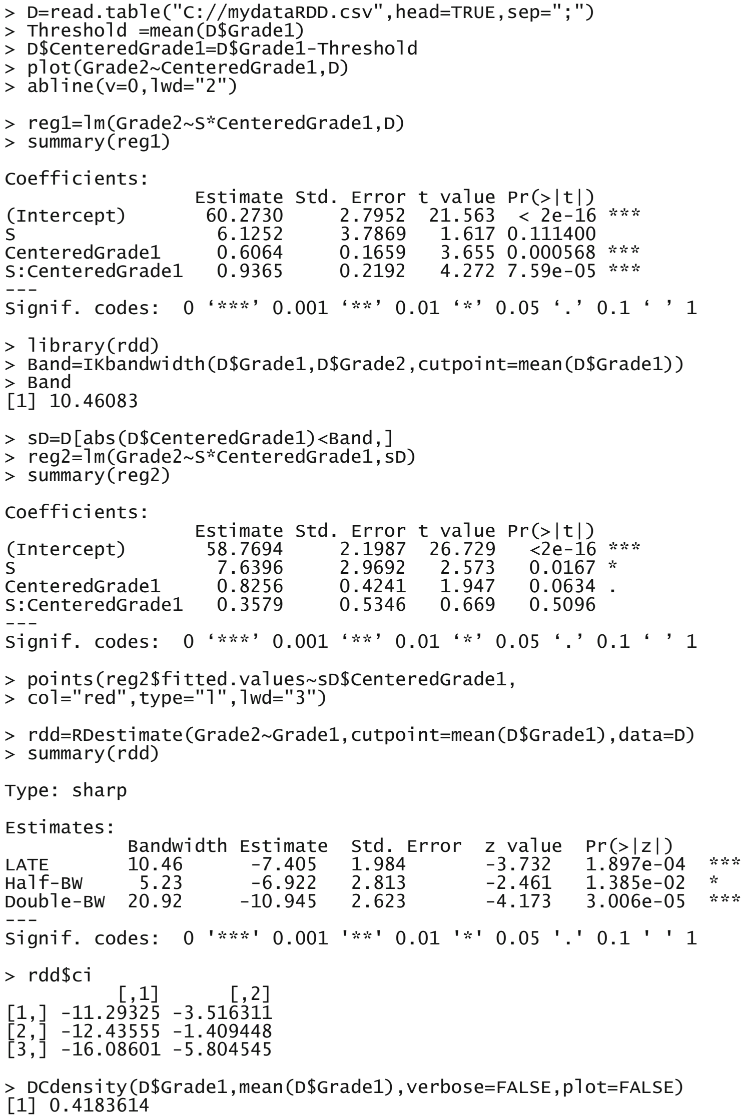
Ind	S	Grade1	Grade2
44	0	66	71
45	0	67	66
46	0	67	65
47	0	68	72
48	0	69	74
49	0	69	69
50	0	70	73
51	0	71	65
52	0	72	71
53	0	73	70
54	0	76	68
55	0	77	76
56	0	79	91
57	0	82	76
58	0	88	78
59	0	89	91
60	0	92	67

Table 14.12 Summary statistics for example 5

	Strategy	Grade1	Grade2
Non-treated group	$S=0$	68.33	68.50
Treated group	$S=1$	41.20	45.46
Grand mean		54.76 (threshold)	56.98

secured their degree, may relax their effort. On the other hand, those with the lowest scores may have incentives to study more. A matching strategy is not suitable either as the factors that affect exam performance, such as motivation and ability, can be unobservable. To overcome those issues, one can take advantage of the discontinuity in assignment. Around the threshold, students are likely to be similar in terms of individual characteristics, all the more so that students do not know ahead of time what the mean score will be, which reduces any potential manipulation of the running variable.

R-CRAN can be used to establish the impact of the intervention. In Fig. 14.13, the database is uploaded under the name D. The first step consists in plotting $Grade2$ as a function of $Grade1$ (see Fig. 14.14). The running variable $Grade1$ is centered to correspond to the usual framework of a regression discontinuity design ($CenteredGrade1 = D\$Grade1 - Threshold$). The threshold, which corresponds here to the mean grade obtained in semester 1 over the whole sample (0 in the centered plan), is added to the graph with the command *abline*. From Fig. 14.14, we can observe a slight break around the cutoff point between the grades of the students



```
> D=read.table("C://mydataRDD.csv",head=TRUE,sep=";")
> Threshold =mean(D$Grade1)
> D$CenteredGrade1=D$Grade1-Threshold
> plot(Grade2~CenteredGrade1,D)
> abline(v=0,lwd="2")

> reg1=lm(Grade2~S*CenteredGrade1,D)
> summary(reg1)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       60.2730     2.7952  21.563  < 2e-16 ***
S                  6.1252     3.7869   1.617 0.111400
CenteredGrade1     0.6064     0.1659   3.655 0.000568 ***
S:CenteredGrade1   0.9365     0.2192   4.272 7.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(rdd)
> Band=IKbandwidth(D$Grade1,D$Grade2,cutpoint=mean(D$Grade1))
> Band
[1] 10.46083

> sD=D[abs(D$CenteredGrade1)<Band,]
> reg2=lm(Grade2~S*CenteredGrade1,sD)
> summary(reg2)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       58.7694     2.1987  26.729   <2e-16 ***
S                  7.6396     2.9692   2.573   0.0167 *
CenteredGrade1     0.8256     0.4241   1.947   0.0634 .
S:CenteredGrade1   0.3579     0.5346   0.669   0.5096
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> points(reg2$fitted.values~sD$CenteredGrade1,
> col="red",type="l",lwd="3")

> rdd=RDestimate(Grade2~Grade1,cutpoint=mean(D$Grade1),data=D)
> summary(rdd)

Type: sharp

Estimates:
           Bandwidth Estimate  Std. Error  z value   Pr(>|z|)
LATE       10.46      -7.405   1.984       -3.732   1.897e-04  ***
Half-BW     5.23      -6.922   2.813       -2.461   1.385e-02  *
Double-BW  20.92     -10.945   2.623       -4.173   3.006e-05  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rdd$ci
          [,1]      [,2]
[1,] -11.29325 -3.516311
[2,] -12.43555 -1.409448
[3,] -16.08601 -5.804545

> DCdensity(D$Grade1,mean(D$Grade1),verbose=FALSE,plot=FALSE)
[1] 0.4183614
```

Fig. 14.13 Regression discontinuity design with R-CRAN: example 5

who were selected for mandatory attendance (on the left) and the grades of those who were not (on the right).

The impact is first estimated via a regression analysis on the whole sample (*reg*1). The estimated effect amounts to 6.12 (difference between the treated group and the non-treated group) but it is not significant. However, extrapolation from observations far from the cutoff may not be valid. An appropriate bandwidth should

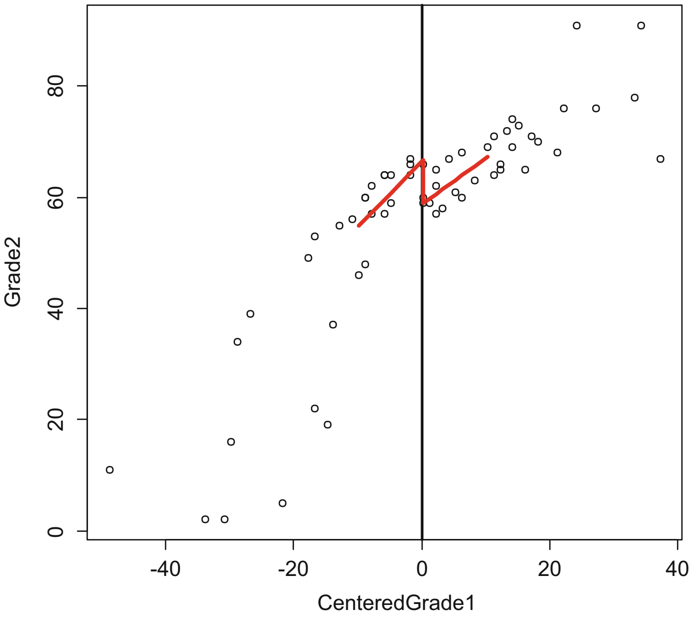


Fig. 14.14 Regression discontinuity design: example 5

be selected to better assess whether non-significance is due to the presence of extreme observations. The package *rdd* is very useful in this matter. The command *IKbandwidth* provides the Imbens-Kalyanaraman optimal bandwidth. The first entry of the command is the running variable, the second entry is the outcome variable, and the last entry is the cutoff point. The resulting bandwidth is *Band* = 10.46 and is used afterwards to estimate the impact of the intervention on a subsample *sD*, such that the distance (in absolute value) between the grade of semester 1 and the threshold is lower than that bandwidth: *sD*= *D*[*abs* (*D*$*GradeC*) < *Band*,]. The impact now amounts to 7.64 and is significant at a 5% level. The *points* function allows the regression lines to be drawn on Fig. 14.14 using the fitted values of the estimated regression model.

More generally, with the command *RDestimate*, it is possible to estimate the impact of the intervention using a nonparametric procedure of local regressions that are implemented on both sides of the threshold. As previously, the bandwidth is calculated using the Imbens-Kalyanaraman method. Then the model is estimated with that bandwidth, half that bandwidth, and twice that bandwidth. The use of local regressions with small bandwidths mitigates the potential problem of incorrect functional form assumptions. For the selected bandwidth (in our example 10.46), the results are similar to the previous one. The local average treatment effect (LATE) amounts to −7.40 and is now significant at a 1% level. A word of warning: the software provides negative values as the treatment group is by default located at the right-hand side of the threshold; we observe the difference between the non-treated group and the treated group, and not the other way round. Significant results are also found with half and twice the bandwidth. Confidence intervals are provided using *rdd$ci*, where *rdd* is our previously estimated design. The first row yields the confidence interval for the first bandwidth: [−11.29; −3.52]. Mandatory attendance thus has a significant impact on exam performance.

DCdensity implements the McCrary sorting test. If there is a discontinuity in the density of the assignment variable at the cutoff point, then this may suggest that some students were able to manipulate their treatment status. Under the null hypothesis of the test, discontinuity is zero. As can be seen from the high p-value (41.8% is much higher than 5%), the test does not detect any manipulation of the design.

14.5 Instrumental Variable Estimation

The instrumental variable estimation addresses the problem of endogeneity in individual participation and can be applied to those situations where the exposure to a policy is determined to some extent by unmeasured factors (hidden bias). For instance, when individuals select themselves, treatment exposure can be related to unmeasured characteristics (e.g., personal, health or educational status) that also affect the outcome of interest, thereby creating a selection bias. Instrumental variables methods intend to overcome this problem by extracting variations in the treatment variable that would be purely exogenous. It consists in a two-step procedure that first examines the selection process itself (what are the factors influencing compliance), and then estimates the effect of the intervention.

Formally, a distinction has to be made between units that are eligible to receive treatment, $S \in \{0, 1\}$, and units that comply with it, $T \in \{0, 1\}$. We would like to assess the impact of an intervention using the following econometric model:

$$\mathrm{y_i} = \mathrm{\alpha}_0 + \mathrm{\alpha}_1 T_i + \beta_1 x_{1i} + \ldots + \beta_K\, x_{Ki} + \epsilon_i$$

where i stands for unit i, y is the variable upon which the influence of the intervention is explored, T_i denotes individual exposure to treatment, the x's are additional control variables, and ϵ_i is an error term. Standard regression models, such as ordinary least squares, make the assumption that the regressors T and x are uncorrelated with the errors in the model (which errors are meant to contain all the factors the model omits). This yields the following path analysis diagram:

$$\begin{tikzcd}
x \arrow[r] & y \\
\epsilon \arrow[ur] & T \arrow[u]
\end{tikzcd}$$

In other words, the regressors are assumed to be exogenous. If this condition is verified and if the functional form of the model is well specified, then the average treatment effect $\widehat{E}$ is directly obtained from the estimated coefficient $\widehat{\alpha}_1$.

In some situations, however, compliance may be explained by unobserved characteristics that also influence the outcome variable y. If this is the case, the treatment variable T cannot be considered as exogenous anymore. Consider for instance the impact on exam performance of non-compulsory evening classes for high-school students who would like to improve their skills. The error term ϵ embodies factors such as motivation, which may affect exam performance (y) but also the decision (T) to attend the courses. The diagram becomes:

$$\begin{tikzcd}
x \arrow[r] & y \\
\epsilon \arrow[r] \arrow[ur] & T \arrow[u]
\end{tikzcd}$$

In our example, if one were to use a conventional econometric model, like the one described previously, one would overestimate the average treatment effect. A spurious relationship would be observed between the outcome variable and the treatment variable.

Instrumental variable estimation aims to avoid inconsistent parameter estimation by generating only variations in T that are exogenous. To do so, one must find an instrument or instrumental variable z that is correlated with T but supposedly not correlated with ϵ:

$$\begin{tikzcd}
x \arrow[r] & y & \\
\epsilon \arrow[r] \arrow[ur] & T \arrow[u] & z \arrow[l]
\end{tikzcd}$$

The error term ϵ comprises omitted explanatory factors, some of which influence compliance T. By explicitly identifying a number of those factors, so called instruments, that are correlated with T, we diminish omitted or hidden influences that would otherwise go through the channel of ϵ (the Δ^- of the path analysis diagram below) and augment our understanding of factors leading to compliance (the Δ^+):

$$\begin{tikzcd}
x \arrow[r] & y & \\
\epsilon \arrow[r, "\Delta^-"'] \arrow[ur] & T \arrow[u] & z \arrow[l, "\Delta^+"]
\end{tikzcd}$$

In our example, we need for instance an instrument that is correlated with evening class attendance, but uncorrelated with motivation or any other factors from the vector of x's that may affect exam performance. One possible candidate is the distance from the school to family home. Those who live far away from school may find it difficult to attend the sessions. Concurrently, travel costs are less likely to be associated with factors such as motivation, difficult to measure and as such likely to be captured by ϵ.

A popular form of instrumental variables estimator is the two-stage least squares (2SLS) method. In a first stage, the treatment variable T is regressed on the instruments as well as on all the exogenous variables previously selected. The equation below specifies one instrument z of coefficient γ_1:

$$T_i = \gamma_0 + \gamma_1 z_i + \delta_1 x_{1i} + \ldots + \delta_K \, x_{Ki} + u_i$$

Note that the method does not require the endogenous variable T to be continuous (OLS can be used to estimate the first stage regression). In a second stage, the model of interest is estimated as usual, except that the T covariate is replaced with its fitted values $\widehat{T}$ obtained from the first stage estimation:

$$y_i = \alpha_0 + \alpha_1 \widehat{T}_{\mathrm{i}} + \beta_1 x_{1i} + \ldots + \beta_K \, x_{Ki} + \epsilon_i$$

The fitted values $\widehat{T}$ have been computed from supposedly exogenous variables only and should thus be cleaned of their association with ϵ.

Finding instruments that meet the conditions for their application is a key concern. The task is greatly eased if the exposure to a policy is determined by some external selection criteria (e.g., random assignment or exogenous eligibility threshold). If eligibility S to receive treatment is not correlated to the outcome in question, then S can serve as an instrument for treatment T. This is best exemplified with the Wald estimator:

$$\widehat{E} = \frac{\bar{y}^{S=1} - \bar{y}^{S=0}}{\textit{compliance rate}}$$

The average treatment effect is obtained by scaling up the difference between the eligible and ineligible groups. Econometrically, the approach amounts to using S as an instrument for T. The differences in S induce variations in the probability of participation T, which facilitates the identification of the causal effect of the program. Equivalently, the first-stage of a 2SLS regression yields:

$$\widehat{T} = \textit{compliance rate} \times S$$

Being ineligible ($S = 0$) means a zero chance of participating, while being eligible ($S = 1$) means a probability equal to the compliance rate. Without any additional covariates, the second-stage regression gives on average:

$$\bar{y}^S = \widehat{\alpha}_0 + \widehat{\alpha}_1 \widehat{T} = \widehat{\alpha}_0 + \widehat{\alpha}_1(\textit{compliance rate} \times S)$$

In other words, we have $\bar{y}^{S=0} = \widehat{\alpha}_0$ and $\bar{y}^{S=1} = \widehat{\alpha}_0 + \widehat{\alpha}_1(\textit{compliance rate})$. The average treatment effect $\widehat{\alpha}_1$ with 2SLS is thus equal to the Wald estimator. Furthermore, the model can be extended to a more complex functional form that better controls for group disparities, provided that additional covariates are available.

Table 14.13 The Wald estimator: example 6

Student	S	Grade	T
1	0	10	0
2	0	20	0
3	1	10	0
4	1	30	1

The Wald estimator applies only when the quasi-experimental design approaches a randomized experiment, i.e. when the units share similar characteristics on average. To illustrate (example 6), Table 14.13 provides information about two classes of students that were randomly assigned to a treatment group ($S = 1$) and a comparison group ($S = 0$). The number of students is reduced drastically for simplicity of exposition. Only individuals 3 and 4 were allowed to attend the evening courses ($S = 1$) but only individual 4 participated ($T = 1$). The between-subjects estimate of the treatment effect is:

$$\widehat{E} = \bar{y}^{T=1} - \bar{y}^{T=0} = 30 - \frac{10 + 20 + 10}{3} = 16.66$$

By doing so, we compare student 4 (a potentially good student) with all the other students and may overestimate the effect of the intervention. On the other hand, as only half of the eligible units participated in the program, a comparison of the class averages would underestimate the effect:

$$\widehat{E} = \bar{y}^{S=1} - \bar{y}^{S=0} = 20 - 15 = 5$$

With the Wald estimator, we obtain:

$$\widehat{E} = \frac{\bar{y}^{S=1} - \bar{y}^{S=0}}{compliance\ rate} = \frac{20 - 15}{0.5} = 10$$

The Wald estimator thus accounts for non-compliance by weighting the difference observed between the treatment group and the comparison group. Yet, the method holds only if the groups are similar on average with respect to the exogenous factors that influence the outcome in question. In other words, the re-scaling is correct if the assignment S is random or close enough to random. If not, then a 2SLS estimation should be implemented with additional regressors to control for any disparity in the exogenous characteristics.

Let us now consider the dataset for example 7, as presented in Table 14.14. In this quasi-experiment, all the students were eligible to participate in the evening sessions (S=1 for all students). The Wald approach cannot be used. As can be seen from Fig. 14.15, using R-CRAN, the difference between the treated and

Table 14.14 Dataset for example 7

Student	S	T	Grade	Distance	Gender
1	1	1	53	1	0
2	1	1	48	9	1
3	1	0	14	9	0
4	1	1	31	2	0
5	1	0	27	17	1
6	1	1	34	6	1
7	1	1	30	7	1
8	1	0	19	15	1
9	1	1	59	2	0
10	1	0	45	16	0
11	1	1	56	1	1
12	1	1	53	1	0
13	1	0	60	18	0
14	1	0	46	12	1
15	1	0	41	20	0
16	1	1	63	1	0
17	1	1	89	6	1
18	1	1	64	7	1
19	1	1	77	1	0
20	1	1	56	5	1
21	1	1	75	6	1
22	1	1	56	5	1
23	1	1	38	2	1
24	1	1	46	1	0
25	1	1	38	2	0
26	1	1	89	10	1
27	1	1	77	1	0
28	1	0	15	8	1
29	1	1	73	5	0
30	1	1	71	7	1
31	1	0	7	6	0
32	1	1	13	3	1
33	1	1	52	9	0
34	1	0	7	20	0
35	1	1	21	5	1
36	1	0	22	8	0
37	1	0	43	18	0
38	1	1	24	2	0
39	1	0	43	15	1
40	1	1	40	4	0
41	1	0	47	14	1
42	1	0	3	6	0
43	1	1	57	4	0

(continued)

Table 14.14 (continued)

Student	S	T	Grade	Distance	Gender
44	1	1	49	9	0
45	1	1	50	11	0
46	1	1	52	12	1
47	1	1	67	1	0
48	1	1	69	2	0
49	1	1	67	8	0
50	1	1	68	10	0
51	1	0	4	16	1
52	1	1	84	1	1
53	1	0	72	17	0
54	1	0	41	20	1
55	1	1	88	7	1
56	1	0	13	7	0
57	1	1	52	9	0
58	1	1	89	10	0
59	1	1	71	7	1
60	1	1	40	4	1

non-treaded units amounts to 26.86. The same result is obtained when one regresses *Grade* on T. However, such an estimation strategy is not appropriate, as those who have decided to comply with the evening program may also be those with the highest levels of motivation. Instrumental variables estimation thus appears as a more suitable method to establish the impact of the intervention.

Distance to home is used as an instrument for treatment. The function *ivreg* from the package *AER* estimates the relationship between y and T by two-stage least squares. Be aware that all statistics computed "on their own" (two consecutive OLS regressions) would be biased as the second stage needs also to encompass the uncertainty of the first-stage. With the *ivreg* command on the other hand, the standard errors of the second-stage regression include the fact that we are using an estimated regressor. The variable *Gender* (0 for female and 1 for male) has been included in the model as an additional covariate. As can be seen from the results of *reg*2, the average treatment effect amounts to $\widehat{E} = 14.49$ and is significant at the 10% significance level.

Several validity tests exist to verify the legitimacy of the instruments employed and they are displayed through the *diagnostics* = *TRUE* code. First, an instrument must be a good predictor of the endogenous variable. This means that the instrument z must be sufficiently correlated with the treatment variable T (i.e. the covariance must be nonzero: $\sigma_{z,T} \neq 0$). In practice, this assumption is checked by reporting the F-test on all instruments to see if instruments are jointly significant in the first-stage regression.

```
> D=read.table("C://mydataIV1.csv",head=TRUE,sep=";")

> #T
> mean(D$Grade[D$T==1])-mean(D$Grade[D$T==0])
[1] 26.85751

> #OLS
> reg1=lm(Grade~T,D)
> summary(reg1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.947      4.519   6.627 1.24e-08 ***
T             26.858      5.467   4.913 7.70e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #2SLS
> library(AER)
> reg2=ivreg(Grade~T+Gender,~Distance+Gender,D)
> summary(reg2,diagnostics=TRUE)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.2439     6.3660   6.008  1.4e-07 ***
T            14.4977     7.8670   1.843   0.0706 .
Gender        0.3317     5.3867   0.062   0.9511

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   1  57    65.719 4.57e-11 ***
Wu-Hausman         1  56     6.328   0.0148 *
Sargan             0  NA        NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #First-stage regression
> reg3=lm(T~Distance+Gender,D)
> summary(reg3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.109799   0.078670  14.107  < 2e-16 ***
Distance    -0.060678   0.007485  -8.107 4.57e-11 ***
Gender       0.104048   0.084744   1.228    0.225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 14.15 Instrumental variable with R-CRAN: example 7

$$H_0 : \text{all the instrument coefficients are jointly zero}$$

$$H_1 : \text{at least one instrument coefficient is nonzero}$$

The usual rule of thumb is that the F-statistic should not be lower than 10. For instance, the diagnostic in Fig. 14.15 yields a statistic equal to 65.719, which is found to be significant (p-value lower than 5%). The F-test thus rejects the null hypothesis of "weak" instruments. Note that when only one instrument is in use, the approach simply amounts to implementing a t-test of significance in the first-stage regression. For instance, the third regression in Fig. 14.15 provides the results of the first-stage regression (see *reg*3) for example 7. We can see that participation in treatment is significantly associated with distance from home. The p-value is the same as that previously found in the diagnostic output, i.e. 4.57e-11. By construction, the t-statistic (–8.107) is also the square root of the F-statistic previously found (i.e. $\sqrt{65.719}$).

Second, it is possible to assess the extent of the endogeneity problem using the Wu-Hausman test. The test does not focus solely on the results of the 2SLS estimations, but examines instead the overall differences between the 2SLS and OLS coefficients. Under the null hypothesis that there is no endogeneity, the estimators will not be systematically different:

$$H_0 : \text{Treatment is exogenous (the coefficients are not different)}$$

$$H_1 : \text{Treatment is endogenous (2SLS is more appropriate)}$$

If one does not reject the null hypothesis, one should decide not to use an instrumental variable. Roughly speaking, the problem of endogeneity is not serious enough to justify the use of 2SLS. Rejecting the null hypothesis on the other hand indicates the presence of endogeneity. For instance, the Wu-Hausman test in Fig. 14.15 yields a p-value equal to 0.0148, which is lower than 5%. The test does reveal an endogeneity problem, which overall gives support to the 2SLS estimator.

Finally, the exogeneity of the instruments with respect to the error term ϵ of the second-stage regression can be assessed with the Sargan over-identification test. The latter is used when there are more instruments (the z's) than endogenous regressors (the T's). Under the null hypothesis of the test, all instruments are uncorrelated with ϵ:

$$H_0 : \text{Instruments are exogenous (i.e.not correlated with the residuals)}$$

$$H_1 : \text{Not all instruments are exogenous}$$

If the p-value falls below the 5% significance level, we reject H_0 and conclude that at least some of the instruments are not exogenous. In that case, one must find other instruments. Note that the Sargan test returns NA if there is one instrument per endogenous regressor (the system is said to be exactly identified). The reason for this impossibility lies in the fact that the error term is unobserved and must be estimated. The test procedure is as follows. First, estimate the second-stage regression and obtain the 2SLS residuals, denoted $\widehat{\epsilon}$ hereafter. Second, regress $\widehat{\epsilon}$ on all exogenous variables including the instruments. The test statistic is then defined as the number of observations n times the coefficient of determination R^2. Under the null hypothesis, this statistic (nR^2) is distributed according to a chi-square distribution with $L - K$ degrees of freedom, where L is the number of instruments and K is the number of endogenous variables. When the system is exactly identified, by construction, it is not possible to compute this statistic because there is not enough information available to implement the test (the residuals $\widehat{\epsilon}$ are computed based on $\widehat{T}$, which itself is based on z). For instance, in Fig. 14.15, the tests returns NA because only one variable (*Distance*) is used to instrument the treatment variable T,

Table 14.15 Dataset for example 8

Student	S	T	Grade	Distance	Gender
1	0	0	36	1	0
2	0	0	39	9	1
3	0	0	9	9	0
4	0	0	18	2	0
5	0	0	27	17	1
6	0	0	24	6	1
7	0	0	21	7	1
8	0	0	18	15	1
9	0	0	42	2	0
10	0	0	42	16	0
11	0	0	39	1	1
12	0	0	36	1	0
13	0	0	57	18	0
14	0	0	39	12	1
15	0	0	42	20	0
16	0	0	45	1	0
17	0	0	72	6	1
18	0	0	51	7	1
19	0	0	57	1	0
20	0	0	42	5	1
21	1	1	75	6	1
22	1	1	56	5	1
23	1	1	38	2	1
24	1	1	46	1	0
25	1	1	38	2	0
26	1	1	89	10	1
27	1	1	77	1	0
28	1	0	15	8	1
29	1	1	73	5	0
30	1	1	71	7	1
31	1	0	7	6	0
32	1	1	13	3	1
33	1	1	52	9	0
34	1	0	7	20	0
35	1	1	21	5	1
36	1	0	22	8	0
37	1	0	43	18	0
38	1	1	24	2	0
39	1	0	43	15	1
40	1	1	40	4	0
41	1	0	47	14	1
42	1	0	3	6	0
43	1	1	57	4	0

(continued)

Table 14.15 (continued)

Student	S	T	Grade	Distance	Gender
44	1	1	49	9	0
45	1	1	50	11	0
46	1	1	52	12	1
47	1	1	67	1	0
48	1	1	69	2	0
49	1	1	67	8	0
50	1	1	68	10	0
51	1	0	4	16	1
52	1	1	84	1	1
53	1	0	72	17	0
54	1	0	41	20	1
55	1	1	88	7	1
56	1	0	13	7	0
57	1	1	52	9	0
58	1	1	89	10	0
59	1	1	71	7	1
60	1	1	40	4	1

variable *Gender* being considered as exogenous as it appears in both regressions (first and second stages).

Let us finally examine the database of example 8 (Table 14.15) which differs from that of example 7 by the fact that students were allowed to attend the evening sessions ($S = 1$), while others were not ($S = 0$). The Wald approach can now be used, which simplifies greatly the search for an instrument. Results from the R-CRAN program of Fig. 14.16 show that the difference between the treated units ($T = 1$) and non-treated units ($T = 0$) amounts to 24.18. As the students who attended the lessons are likely to be the most motivated, this may yield an overestimation of the effect of the intervention. Furthermore, because only 70% of those assigned to treatment complied with it, a comparison of the eligible with the ineligible will yield an underestimation of the effect (the difference amounts to 10.52). To overcome the selection bias, we rely on S to serve as an instrument for T. According to the Wald estimator, the average treatment effect is equal to 15.03. As expected, similar results are obtained with the *ivreg* function (see *reg*1). The effect is significant at the 10% level.

It is also possible to include additional covariates (e.g., *Gender*) to try to reduce the bias due to the selection process itself. Last, to improve the quality of the analysis, it is possible to include additional instruments, like the one we have used in example 7, provided that this data is available. This is done for instance in Fig. 14.16 with *reg*2. Both *Distance* and S are used as instruments for T. From the 2SLS results of *reg*2, the average treatment effect amounts to $\widehat{E} = 14.22$ and is now significant at the 5% significance level (instead of 10% with the Wald estimator).

```
> D=read.table("C://mydataIV2.csv",head=TRUE,sep=";")

> mean(D$Grade[D$T==1])-mean(D$Grade[D$T==0])
[1] 24.18304

> mean(D$Grade[D$S==1])-mean(D$Grade[D$S==0])
[1] 10.525

> compliance.rate=sum(D$T)/sum(D$S)
> compliance.rate
[1] 0.7

> #Wald
> (mean(D$Grade[D$S==1])-mean(D$Grade[D$S==0]))/compliance.rate
[1] 15.03571
> library(AER)
> reg1=ivreg(Grade~T,~S,D)
> summary(reg1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   37.800      4.504   8.393 1.35e-11 ***
T             15.036      7.880   1.908   0.0613 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #2SLS
> reg2=ivreg(Grade~T+Gender,~S*Distance+Gender,D)
> summary(reg2,diagnostics=TRUE)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   37.618      4.742   7.933 8.87e-11 ***
T             14.225      6.478   2.196   0.0322 *  
Gender         1.245      5.302   0.235   0.8152    

Diagnostic tests:
                 df1 df2 statistic  p-value    
Weak instruments   3  55    36.500 4.06e-13 ***
Wu-Hausman         1  56     8.674  0.00469 ** 
Sargan             2  NA     0.086  0.95786    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #First-stage regression
> reg3=lm(T~S*Distance+Gender,D)
> summary(reg3)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0182262  0.1124783  -0.162 0.871867    
S            1.1335693  0.1369514   8.277 3.09e-11 ***
Distance    -0.0003741  0.0107255  -0.035 0.972299    
Gender       0.0422890  0.0786985   0.537 0.593190    
S:Distance  -0.0551792  0.0140010  -3.941 0.000231 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 14.16 Instrumental variable with R-CRAN: example 8

The entry *diagnostics = TRUE* examines the legitimacy of the instruments employed. The F-test supports the analysis with a statistic equal to 36.5 and p-value lower than 5%. The Wu-Hausman test concludes that there was indeed an endogeneity problem (the test rejects the null hypothesis). Moreover, since we now have two instruments and one single endogenous regressor, the Sargan test can be implemented. We find a p-value equal to 0.958, which is higher than 5%. We thereby do not reject the hypothesis of exogeneity. Our instruments thus play the role they have to play. Note that one NA remains in the output, but only because this test is defined by only one value of degrees of freedom (*df*2 will be always NA). Last, the first-stage regression (*reg*3) points out that compliance with treatment depends both on eligibility (coefficient=1.13 and significant) and the distance from home (interaction term=−0.05 and significant). A logit or probit model could be estimated to further investigate these results.

Bibliographical Guideline

One of the first studies employing difference-in-differences was that of Ashenfelter and Card (1985) who wanted to analyze the impact of a training program for unemployed and low-income workers using longitudinal information on earnings for a treatment group and a comparison group. Since then, and given the few number of observations required for its application, difference-in-differences methods have become very popular in many fields. Propensity score matching has been originally developed by Rosenbaum and Rubin (1983). Most of its applications pertain to the case of a binary treatment, although recent developments have extended the method to other cases (see Hirano and Imbens 2004). The first application of regression discontinuity design can be traced back to Thistlewaite and Campbell (1960), who analyzed the impact of merit awards on students' later success, using the fact that the allocation of these awards was based on an observed test score. The McCrary test has been developed more recently, in 2008. A detailed presentation of RDD methods is available in Lee and Lemieux (2010). The idea that instrumental variables can be used to solve an identification problem was first introduced in Wright (1915) (see Stock and Trebbi 2003). Two-stage least squares were developed more or less independently by Theil (1953), Basmann (1957) and Sargan (1958). To go further, one may read Imbens (2014) who reviews recent work in the literature on instrumental variables methods.

For further references, the reader may also rely on two additional sources. The first is the "Handbook on Impact Evaluation" published by the World Bank and available online: it reviews most of the quantitative methods and models related to impact evaluation. Second, the European Commission provides several guides on the topic. One of them, "Evalsed Sourcebook: method and techniques" provides a very pedagogical description of the methods and techniques that are applied in the evaluation of socio-economic development.

References

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics, 67*, 648–660.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica, 25*, 77–83.

European Commission. (2013). *Evalsed sourcebook: Method and techniques*. Brussels: EC.

Hirano, K., & Imbens, G. (2004). The propensity score with continuous treatments. In *Missing data and Bayesian methods*. Hoboken, NJ: Wiley.

Imbens, G. (2014). *Instrumental variables: An econometrician's perspective* (NBER Working Papers 19983). National Bureau of Economic Research.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5–86.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: Quantitative methods and practices*. Washington, DC: World Bank.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*, 281–355.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*, 698–714.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Sargan, J. (1958). The estimation of economic relationships using instrumental variables. *Econometrica, 26*, 393–415.

Stock, J. H., & Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives, 17*, 177–194.

Theil, H. (1953). *Estimation and simultaneous correlation in complete equation systems*. Central Planning Bureau: The Hague.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology, 51*, 309–317.

Wright, P. G. (1915). Moore's economic cycles. *Quarterly Journal of Economics., 29*, 631–641.