STATISTICS APPLIED TO CLINICAL TRIALS: SELF-ASSESSMENT BOOK

Statistics Applied to Clinical Trials: Self-Assessment Book

by

TON J. CLEOPHAS, MD, PhD, Associate-Professor,
President American College of Angiology,
Co-Chair Module Statistics Applied to Clinical Trials,
European Interuniversity College of Pharmaceutical Medicine Lyon, France,
Internist-clinical pharmacologist,
Department Medicine, Albert Schweitzer Hospital, Dordrecht, Netherlands

AEILKO H. ZWINDERMAN, Math D, PhD, Professor,
Co-Chair Module Statistics Applied to Clinical Trials,
European Interuniversity College of Pharmaceutical Medicine Lyon, France,
Professor of Statistics,
Department Biostatistics and Epidemiology, Academic Medical Center Amsterdam, Netherlands

and

TOINE F. CLEOPHAS, D Techn,
Technical University, Delft, Netherlands

Springer Science+Business Media, LLC

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-1096-5 ISBN 978-94-010-0285-1 (eBook)
DOI 10.1007/978-94-010-0285-1

Printed on acid-free paper

TABLE OF CONTENTS

CHAPTER 3 / POWER, SAMPLE SIZE

CHAPTER 4 / PROPORTIONAL DATA ANALYSIS, PART I

CHAPTER 5 / PROPORTIONAL DATA ANALYSIS, PART II

CHAPTER 6 / META-ANALYSIS

CHAPTER 7 / INTERIM-ANALYSES

CHAPTER 8 / MULTIPLE TESTING

CHAPTER 9 / PRINCIPLES OF LINEAR REGRESSION

CHAPTER 10 / SUBGROUP ANALYSIS USING REGRESSION MODELING

CHAPTER 11 / RELATIONSHIP AMONG STATISTICAL DISTRIBUTIONS

CHAPTER 12 / STATISTICS IS NOT BLOODLESS ALGEBRA

PREFACE

The authors have taught statistics and given statistics workshops in France and the Netherlands for almost 4 years by now. Their material, mainly on power point, consists of 12 lectures that have been continuously changed and improved by interaction with various audiences. For the purpose of the current book simple English text has been added to the formulas and figures, and the power points sheets have been rewritten in the format given by Kluwer Academic Publishers. Cartoons have been removed, since this is not so relevant for the transmission of thought through a written text, and at the end of each lecture (chapter) a representative number of questions and exercises for self-assessment have been added. At the end of the book detailed answers to the questions and exercises per lecture are given. The book has been produced with the same size and frontpage as the textbook "Statistics Applied To Clinical Trials" by the same authors and edited by same publishers (2nd Edition, Dordrecht/Boston/London, 2002), and can be applied together with the current self-assessment book or separately.

The current self-assessment book is different from the texbook, because it focuses on the most important aspects rather than trying to be complete. So, it does not deal with all of the subjects assessed in the texbook. Instead, it repeats on and on the principle things that are needed for every analysis, and it gives many examples that are further explained by arrows in the figures.

The authors were very enthousiastic to prepare the manuscript since the workshops and Statistics Module for the European College of Pharmaceutical (Socrates Project of the European Community) are their passion, and it is very stimulating for them to find that students respond so well. The students and participants at the workshops are currently offered an exam at the end of the module, and many of them are doing extremely well.

The readership of the self-assesment book should be the students of the European Interuniversity College in Lyon (4th Academic Year starting September 2002, at the Claude Bernard University). The EC Socrates Program, Brussels, is now considering a larger financial support for this college after the first 3 successful years. The next year's module statistics in Lyon is planned for the first week of February 2003. Readership of the self-assessment book should also be the students participating at the Statistics Workshops in the Netherlands: we will have 1-3 day workshops June 13 Rotterdam, June 17 Rotterdam, June 24 Roosendaal, September 3-30 Tilburg, September 16-23 Amsterdam, October 2-16 Veldhoven, January 16-23, 2003, Amsterdam, and, so the authors have been said, there are many to come. Third, one of the authors, Professor Zwinderman who is head of the division Statistics of the Department of Biostatistics and Epidemiology at the Academic Medical Center of the University of Amsterdam has prepared the self-assesment book in such a manner that it can also be used for his students at the Amsterdam University. Finally, the readership of the self-assesment book should be everyone who is going to buy the current textbook. We often hear from the participants at the workshops that there are many staff physicians in Dutch hospitals who want to learn more about statistics but who are just to busy to make the time reservation for a full course. The two books can be used by busy staff physicians in their own time schedule. We will try to focus and remove the difficult mathematics, in order for readers to even enjoy reading the books while on holiday.

FOREWORD

Many doctors are notoriously bad at understanding statistics. Some even believe that either you are a good clinician or you are good at statistics. This, of course, is pure nonsense. It stems from the fact that, historically, statistics was poorly taught and books on the subject were as dry as a bone. The truth is, that without a good comprehension of statistics, clinicians cannot fully understand clinical trials. And without an understanding of clinical trials, evidence-based medicine becomes a farce.

This book is different. Rather than being comprehensive, it concentrates on the most relevant aspects. Rather than being theoretical and boring, it uses real life examples and is entertaining. Rather than being overloaded with formulas, it uses a language that physicians are used to.

The book is meant as a companion to a more in-depth textbook. As its subtitle points out, it is a 'self-assessment book'. For most purposes, however, it can be used as a stand alone reference text. It will be equally as helpful for the student as for the seasoned clinician or researcher.

Collectively, the authors of this book have considerable experience in teaching statistics and enough background in clinical medicine to know what is relevant and how to get it across to doctors. Much of statistics for clinical trials is essentially based on good old common sense. The authors have understood that a common sense approach is often the most constructive for problem-solving in clinical trials. It is also an approach that is close to the heart of most clinicians.
In my view, this book is a very readable, easy to understand text for doctors and scientists involved in clinical trials. It is as authoritative as it is to the point. I am sure it is an extremely valuable addition to the literature on medical statistics. It deserves to be a great success.

Univ. Prof. Dr. Dr. Edzard Ernst, FRCP,
Editor-in-Chief, Perfusion, the official journal of the Deutsche
Gesellschaft für Arterioskleroseforschung, University of Exeter, School
of Postgraduate Medicine and Health Sciences, Exeter, United Kingdom.

CHAPTER 1

INTRODUCTION TO THE STATISTICAL ANALYSIS OF CLINICAL TRIALS, CONTINUOUS DATA ANALYSIS

1. SCIENTIFIC RIGOR

Scientific rigor requires: strict and consistent scientific rules:

1. Prior hypothesis.

This hypothesis is tested with a probability of 5% (5% chance it is untrue). Why not posterior? Posterior hypotheses can easily generate hundreds of P-values: significances are then found by chance (compare: gamble 20x at 5% chance: you get about up to 40% chance of a significant results by chance).

2. Valid designs.

It reduces chance of biases (= systematic errors) and placebo effects. Blinded, randomized, controlled, objective measurements, adequate sample sizes make for a valid design.

3. Strict description of methods.

Describe validity criteria in detail, including methods of recruitment, randomization etc.

4. Data analysis uniform and thoroughly.

This should be done as described in the methods.

No clinical trial without proper statistics.

2. TWO TYPES OF DATA

Clinical trials: 2 types of data.

Efficacy data, e.g., blood pressures,
80, 81, 82, 80, 84, ...(continuous variables),
t-statistic, analysis of variance (ANOVA) .

Safety data, e.g., division sum: patients with side effect / all of patients,
(fractions between 0.0- 1.0) ,
Chi-square or McNemar statistic.

3. HISTORICAL CONTROLS

A special point: Historical controls.
A randomized controlled trial with major differences between old and new treatment is unethical, because half of patients received an inferior treatment.
So, why not use historical data as control?
Problem is, of course, the risk of asymmetries, e.g., different times, populations, equipments.

4. FACTORIAL DESIGNS

Another special point: factorial designs.
Mostly, randomized controlled trials answer a single question.
Research is costly, why not test 2 or more modalities.
E.g., patients with angina pectoris are treated with calcium channel blocker with or without beta-blocker.

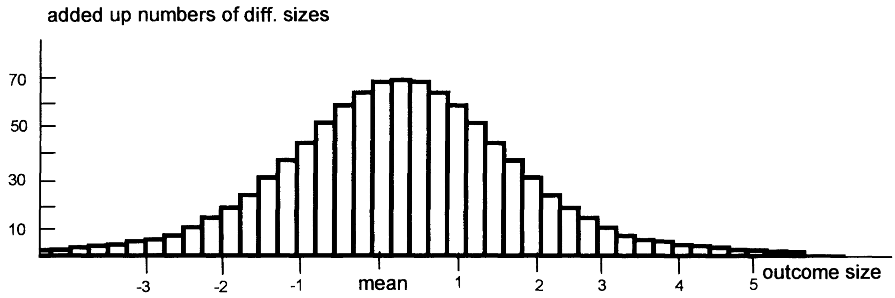
	Calcium channel blocker	no calcium channel blocker
Beta-blocker	regimen 1	regimen 2
No beta-blocker	regimen 3	regimen 4

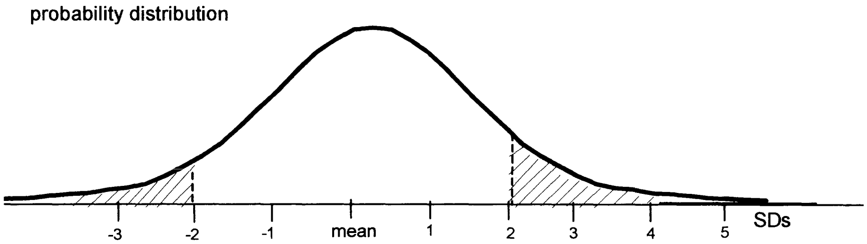
5. BIOLOGY IS FULL OF VARIATIONS

The remainder of this chapter involves continuous data only.
Because biological processes full of variations:
Statistics gives no certainties only chances.
What chances?: chances that hypotheses are true/untrue.
What hypotheses? E.g., (1) no difference from a 0 effect,
(2) real difference from a 0 effect,
(3) worse than a 0 effect.
Statistics is about estimating such chances / testing such hypotheses.

Note: trials often calculate difference between test treatment and control, and, subsequently, test whether this difference is larger than 0. So, a simple way to reduce a study of two means to one of a single mean and single distribution of data, is to take the difference and compare it with 0.

6. SUMMARIZE THE DATA





1.Histogram
- On the x-axis individual data/ on the y-axis "how often" (e.g., mean occurs most frequently, bars both sides gradually grow shorter).
- This method is not adequate for testing hypotheses.

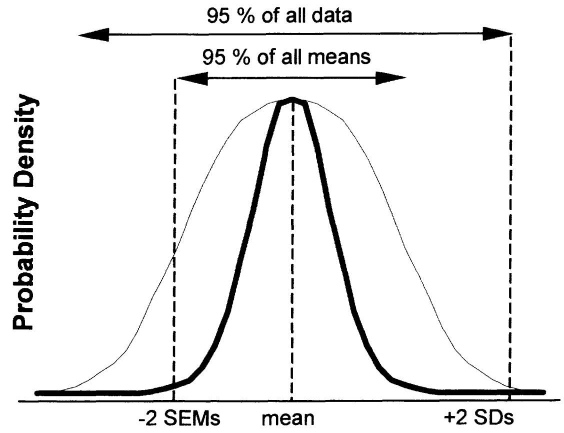
2.Gaussian curve
- On the x-axis individual data or SDs* -distant-from-mean.
- On the y-axis bars are replaced by continuous line.
- Now, it is impossible to determine from graph how many patients had a particular outcome!
- Instead, important inferences (conclusions) can be made, e.g.:

Total area under the curve (AUC)	= 100% of the data of our trial,
AUC left from mean	= 50% of data,
AUC left from –1 SDs	= 15% of data,
AUC left from –2 SDs	= 2.5% of data.

- This method is not yet adequate for testing hypotheses, but better.

* SD = standard deviation = is estimate of spread in data, is calculated according to $SD = \sqrt{[\Sigma\ (x- \bar{x})^2\ /\ (n-1)]}$
where x = individual data, $\bar{x}$ = mean, n number of data.

7. TWO GAUSSIAN CURVES



Two gaussian curves are given: a narrow and a wide one, both based on the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes means of many trials similar to ours. We won't try to make you understand why so. Still, it's easy to conceive that distribution of means of many trials is narrower, and has fewer outliers than the distribution of the actual data from our trial.

Believe it or not:
The narrow curve with SEMs* on the x-axis can be effectively used for testing important statistical hypotheses: 1. No difference between new and standard.
2. Real difference " " " " .
3. New treatment is worse than old.
4. Two treatments are equivalent.
SEM-curve is narrower than SD-curve because SEM=SD/√n.

* SEM= standard error of the mean.

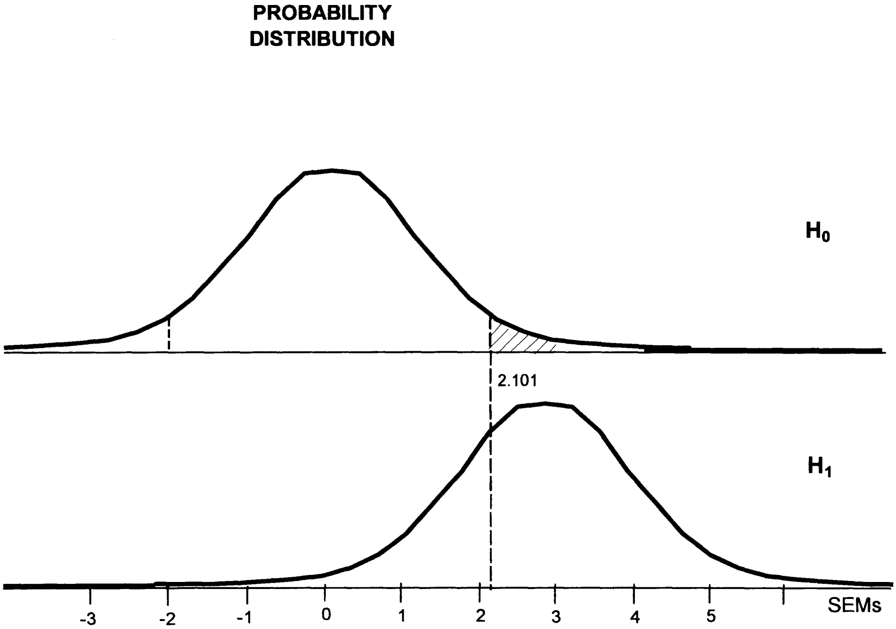
8. HUMAN BRAIN

The human brain excels in making hypotheses. We make hypotheses all the time, but they may be untrue. E.g., once you might have thought that only girls can become doctors. Later, this was, obviously, untrue. Hypotheses must be assessed with hard data.

9. NULL-HYPOTHESIS

Important hypothesis: Hypothesis 0.

No difference from a 0 effect.
We will now try and make a graph of hypothesis 0 = null hypothesis.



What does 0-hypothesis look like in graph?

H1 = graph based on the data of our trial with SEMs on the x-axis (in statistics often called z-axis).

H0 = the same graph with mean 0 (mean ± SEM = 0 ± 1).

Now we will make a giant leap from our data to the entire population (data are representative).

H1 = also summary of means of many trials similar to ours (this assumption is correct, because our study is representative for population, and because we have SEM-units on the x-axis).

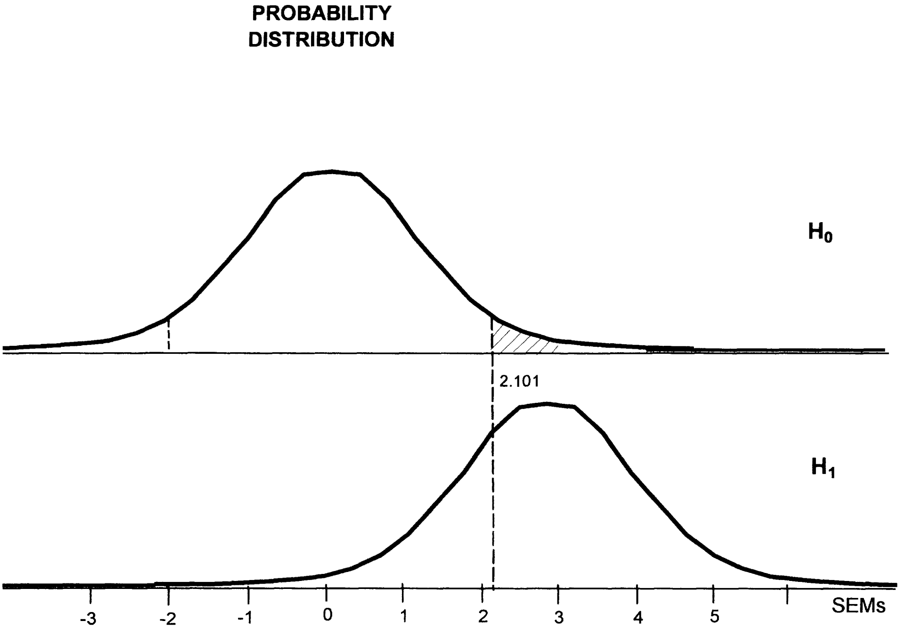
H0 = summary of means of many trials similar to ours but with overall effect 0 (our mean is not 0, but 2.9 SEMs. It, still, could be an outlier of many studies with an overall effect of 0).

So think from now on of H0 as the distribution of the means of many trials with an overall effect of 0. If hypothesis 0 is true, then the mean of our study is part of H0. We can't prove but can calculate chance / probability of this possibility.

A mean result of 2.9 SEMs is far distant from 0. Suppose it belongs to H0. Only 5% of the H0-trials are >2.1 SEMs distant from 0, because the AUC here = only 5% (Figure: striped area). The chance that our trial belongs to H0 is thus <5% (reject null hypothesis of no effect).

Conclude here:
<5% chance to find this result.
In usual terms: we reject the null hypothesis of no effect at Probability (P) < 0.05 or P < 5%.

10. ALPHA, THE TYPE I ERROR



Note:
Alpha: small AUC right from 2.101 SEMs.
Alpha: area of rejection of H0.
Alpha: type I error or the chance of finding a difference where there is none.
Note:
2.9 SEMs is far from 2.1 SEMs: so probability of finding 2.9 may be lot <5%.

11. T-TABLE

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	6.314	12.706	63.656	636.58
2	2.920	4.303	9.925	31.600
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	$\boxed{2.845}$	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.689
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.660
30	1.697	2.042	2.750	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
100	1.660	1.984	2.626	3.390
200	1.653	1.972	2.601	3.340
5000	1.645	1.960	2.577	3.293

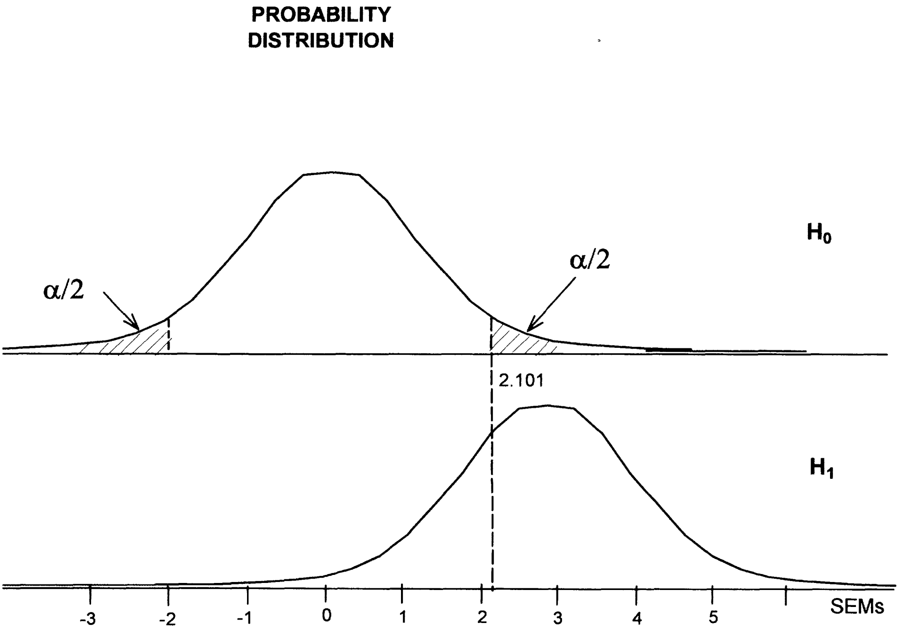
The above t-table tells us the exact % AUC right from 2.9 SEMs.

Four right columns	-Trial results: expressed in SEM units distant from 0 (=also T-values)
Upper row	-AUC values right from trial results.
Left column	-Adjustment for numbers of patients.

with a sample of n = 20 the AUC right from 2.9 SEMs is right from 2.845 → AUC<0.01 (Probability not <0.05, but even<0.01).

T-distribution = adjustment of normal distribution = just a bit wider for small samples (but with sample size of 120 or more identical to normal distribution).

12. REJECT THE NULL-HYPOTHESIS

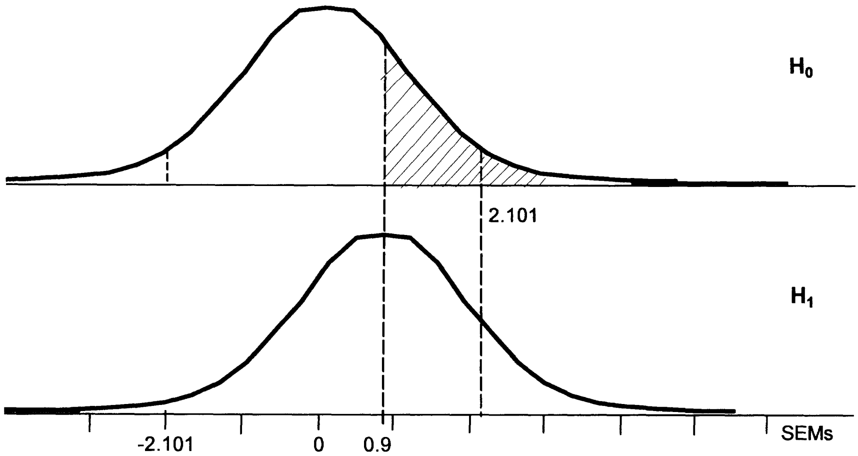


Alpha = outlier AUC of H0= area of rejection of H0= usually 5% = rather 2 x 2.5% = 2 x α / 2 (if a trial-mean is within this AUC: null hypothesis is rejected)

Alpha = chance of finding a difference where there is none.

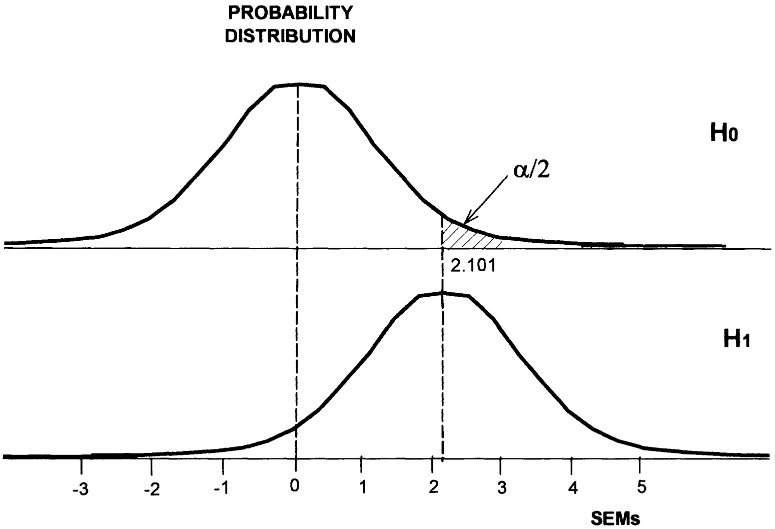
Alpha = therefore, so-called, type I error.

13. NEGATIVE TRIAL



Example of negative trial(= trial unable to reject null hypothesis).
Mean of our trial is 0.9 SEMs distant from 0.
Not on right side of 2.1 SEMs.
Null hypothesis not rejected.
AUC right from 0.9 = not 5 %, but no less than 35% of total AUC.
Do not reject 0 hypothesis, because P = 0.35 or 35 %.

14. BORDERLINE RESULT



Example of borderline result.

Mean = exactly 2.101 SEMs distant from 0.
Alpha level of rejection = 2.101 SEMs.
AUC right from 2.101 only 5%.
Reject null hypothesis at P = 0.05 or P = 5%.
P of 5% borderline result: 5% chance untrue.

15. TESTING TWO MEANS

So far, we have analyzed a single mean versus 0, now we will analyze 2 means versus each other. F.e., a parallel-group study of two groups tests the effects of two beta-blockers on cardiac output.

	Mean ± SD			$SEM^2 = SD^2/n$
group 1 (n=10)	5.9	± 2.4	liter/min	5.76/10
group 2 (n=10)	4.5	± 1.7	liter/min	2.89/10

Calculate: $mean_1 - mean_2 = 1.4$
Then calculate pooled SEM = $\sqrt{(SEM_1^2 + SEM_2^2)} = \sqrt{0.433} = 0.658$
Note: for SEM of difference: take square root of sums of squares of separate SEMs and, so, reduce the analysis of two means to one of a single mean.

$$T = \frac{mean_1 - mean_2}{\text{Pooled SEM}} = 1.4 \;/\; 0.658 = 2.127 \text{ with degrees of freedom } 20\text{-}2\text{=}18^*$$

2.127 is larger than 2.101. Thus, a t-value of 2.127 indicates an AUC < 5% as can be concluded from the t-table below.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	6.314	12.706	63.656	636.58
2	2.920	4.303	9.925	31.600
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	$\boxed{2.101}$	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.689
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.660
30	1.697	2.042	2.750	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
100	1.660	1.984	2.626	3.390
200	1.653	1.972	2.601	3.340
5000	1.645	1.960	2.577	3.293

Reject the 0 hypothesis of no difference at a probability (P) <0.05 or 5%. Conclude that there is a true difference between the samples (in clinical terms).

*Calculations have to be adjusted for degrees of freedom: with 2 groups of each 10 patients, we have 2x10 – 2 = 18 degrees of freedom.

16. TESTING PAIRED SAMPLES

Another example of two means is given.
Crossover trial to test efficacy of sleeping pill versus placebo.
Unlike previous sheet, there is only one group treated twice instead of two groups treated once.

	hours of sleep		
patient	drug	placebo	difference
1	6.1	5.2	0.9
2	7.0	7.9	–0.9
3	8.2	3.9	4.3
4	7.6	4.7	2.9
5	6.5	5.3	1.2
6	7.8	5.4	3.0
7	6.9	4.2	2.7
8	6.7	6.1	0.6
9	7.4	3.8	3.6
10	5.8	6.3	–0.5
Mean	7.06	5.28	1.78
SD			1.79

Simply calculate mean and SD of the various differences.
Next find SEM by taking SD / $\sqrt{n}$ = 0.56.
Mean difference ± SEM= 1.78 ± 0.56 .

$$T= \frac{\text{Mean difference}}{\text{SEM}} = \frac{1.78}{0.56} = 3.18$$

with a sample size of 10 (with 1 group of 10 patients we have 10-1= 9 degrees of freedom).

T-Table: v= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left of the total area under the curve

v	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

T-table shows that with T = 3.18 = between 2.821 and 3.250
→AUC right from 2.821 is 2 %.
→Conclude: $P < 0.02$ or < 2 %.
→Reject 0 hypothesis.

17. UNPAIRED TESTING OF PAIRED SAMPLES

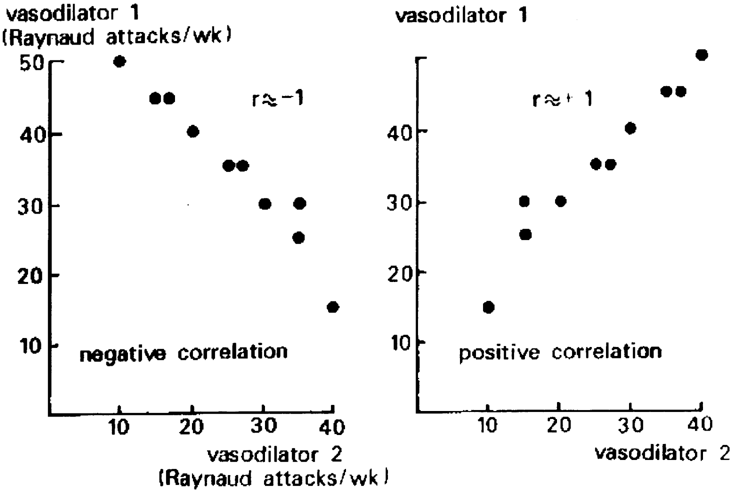
The same data as on the previous page are given.

	hours of sleep		
patient	drug	placebo	difference
1.	6.1	5.2	0.9
2.	7.0	7.9	–0.9
3.	8.2	3.9	4.3
4.	7.6	4.7	2.9
5.	6.5	5.3	1.2
6.	7.8	5.4	3.0
7.	6.9	4.2	2.7
8.	6.7	6.1	0.6
9.	7.4	3.8	3.6
10.	5.8	6.3	–0.5
Mean	7.06	5.28	1.78
SD	0.76	1.26	1.79
SEM	0.24	0.40	0.56

Difference: $\text{mean}_1 - \text{mean}_2 \pm \sqrt{(\text{SEM}_1^2 + \text{SEM}_2^2)} = 7.06 - 5.28 \pm \sqrt{[(0.24)^2 + (0.40)^2]} = 1.78 \pm 0.48$.
T = Mean / SEM = 1.78 / 0.48 = 3.71 (degrees of freedom = 20-2 = 18).
P = 0.005 .
This is approximately the same result as with paired testing. With strong positive or negative correlations this will not be so, however!!

18. POSITIVE AND NEGATIVE CORRELATIONS

The figure below shows two crossover studies of patients with Raynaud treated with two different vasodilators, the left study has a strong negative, the right study a strong positive correlation.



(Take the following few lines on faith; you can calculate it for yourself; chapter 9).
Strong positive correlation : Right graph.
If treatment 1 performs well, treatment 2 will do equally so.
Paired T-test will provide a T = 4.... and a P of <0.001.
Unpaired T-test will provide a T = 2.. and a P of only <0.05.

Strong negative correlation: Left graph.
If treatment 1 performs well, treatment 2 will not do so.
Paired T-test will provide a T = 1.7... and is thus not significant.
Unpaired T-test will provide a T = 2... and a P of <0.05.

19. UNPAIRED ANALYSIS OF VARIANCE (ANOVA)

So far everything relaxingly simple.
Now something really complicated is coming up.
If we want to analyze 3 groups rather than 2 we will need unpaired ANOVA.

Total variation
| |
Between group variation within group variation

In ANOVA:
Variations expressed as sums of squares (SS) and can be added up to obtain total variation.
Assess whether between-group variation is large compared to within-group variation.

Group	n patients	mean	SD
1	-	-	-
2	-	-	-
3	-	-	-

Grand mean = (mean 1 + 2 +3) / 3

SS between groups = n_1 (mean_1 – grand mean)^2 + n_2 (mean_2 – grand mean)^2 +....
SS within groups = (n_1 -1)($SD_1{}^2$) + (n_2-1) $SD_2{}^2$ +.....

$$F = \frac{\text{SS between groups / degrees of freedom}^*}{\text{SS within groups / degrees of freedom}^*}$$

F-table gives P-value (see section Tables).

*Degrees of freedom equals $n_1 + n_2 + n_3$ -3 for SS within, and 3 - 1 = 2 for SS between.

Example: Effect of 3 compounds on hemoglobin levels.

Group	n patients	mean	SD
1	16	8.7125	0.8445
2	10	10.6300	1.2841
3	15	12.3000	0.9419

Grand mean = (mean 1 + 2 +3) / 3 = 10.4926

SSbetween groups = 16 (8.7125-10.4926)^2 + 10(10.6300 – 10.4926)^2 .... (adjust for degrees of freedom)
SSwithin groups = 15 x 0.84452 + 10 x 1.28412 +........ (adjust for degrees of freedom)

F = 49.9 and so P <0.001.
In case of 2 groups: ANOVA= T-test (F = T^2) .

20. PAIRED ANALYSIS OF VARIANCE (ANOVA)

ANOVA can make lot of mess from only a few numbers.
Paired ANOVA is for 3 treatments in single group.

Total variation
| |
Between subj variation Within-subj variation
| |
Between treatment variation Residual variation (random)

Variations expressed as sums of squares (SS) and can be added up.
Assess whether between treatment variation is large compared to residual variation.

Subject	treatment 1	treatment 2	treatment 3	SD^2
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
Treatment mean	-	-	-	

Grand mean = (treatment mean 1 + 2 + 3) / 3 = ...

SS within subject = $SD_1^2 + SD_2^2 + SD_3^2$
SS treatment=(treatment mean_1 -grand mean)^2 + (treatment mean_2 -grand mean)^2+..
SS residual = SS within subject - SS treatment

$$F= \frac{\text{SS treatment / degrees of freedom}^*}{\text{SS residual / degrees of freedom}^*}$$

F table gives P-value.

*Degrees of freedom equals 3 -1 = 2 for SS treatment , and 4-1= 3 for SS residual.

Example: Effect of 3 treatments on vascular resistance (mm Hg.min / l).

Person	treatment 1	treatment 2	treatment 3	SD^2
1	22.2	5.4	10.6	147.95
2	17.0	6.3	6.2	77.0
3	14.1	8.5	9.3	18.35
4	17.0	10.7	12.3	21.45
Treatment mean	17.58	7.73	9.60	

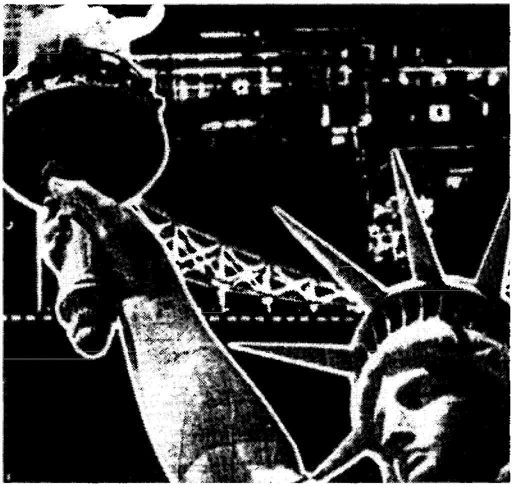
Grand mean = 11.63
SS within subject = 147.95 + 77.05 +....
SS treatment = (17.58 – 11.63)2 + (7.73 – 11.63)2 +....

SS residual = SS within subject - SS treatment.

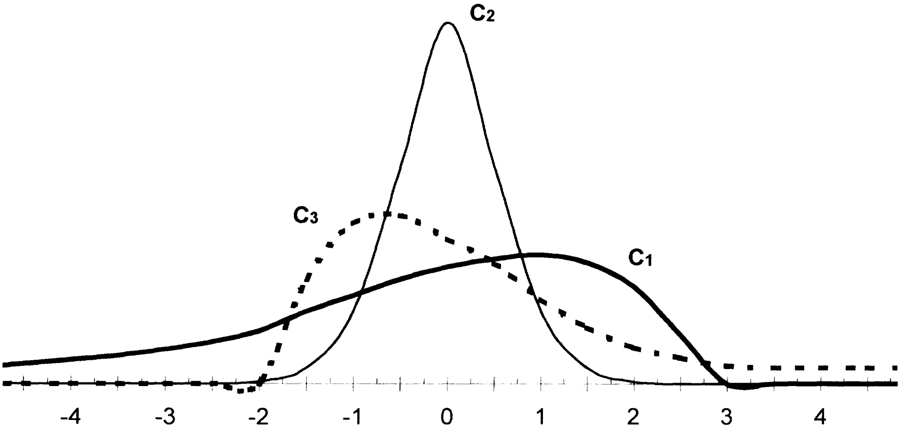
F= 14.31 and, so, P<0.01 (see section Tables).

In case of 2 treatments: ANOVA=T-test ($F = T^2$).

21. NON-PARAMETRIC TESTING FOR SKEWED DATA



Statue of liberty to tell you're free to use non-parametric test for any type of data.



When data are skewed you **must** use a non-parametric test.
A non-parametric test "normalizes" skewed data but can also appropriately used for non-skewed data.

For paired comparisons:
Mann-Whitney test(= Wilcoxon signed rank test= paired Wilcoxon test).

For unpaired comparisons:
Wilcoxon rank sum test(=unpaired Wilcoxon test).

22. PAIRED NON-PARAMETRIC TEST: MANN-WHITNEY TEST

Placebo-controlled crossover trial to test efficacy of sleeping drug from few pages ago.

	Hours of sleep			rank
Patient	drug	placebo	difference	(ignoring sign)
1	6.1	5.2	0.9	3.5x
2	7.0	7.9	-0.9	3.5
3.	8.2	3.9	4.3	10
4.	7.6	4.7	2.9	7
5.	6.5	5.3	1.2	5
6.	8.4	5.4	3.0	8
7.	6.9	4.2	2.7	6
8.	6.7	6.1	0.6	2
9.	7.4	3.8	3.6	9
10.	5.8	6.3	-0.5	1

patients are tested twice: with sleeping pill and with placebo.
Put differences in sleeping time in rank of ascending order.
Patient 1 and 2 are equal (0.9 hours different from 0) , so give them rank number 3.5 instead of rank numbers 3 and 4.

1. add up the rank numbers of positive and negative differences separately;
   + rank numbers = 3.5+10+7+5+8+6+2+9=50.5
   - rank numbers = 3.5+1= 4.5
2. For testing: table below uses smaller of two rank numbers = 4.5: with n_1 and n_2 both=10 → $P<0.02$.

Paired non-parametric test is called Mann-Whitney test,
the table uses smaller of the two rank numbers.

N pairs	P<0.05	P<0.01
7	2	0
8	2	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19

23. UNPAIRED NON-PARAMETRIC TEST: WILCOXON RANK SUM TEST

Table: two groups of patients (thin print group 1, fat print group 2) treated with different beta-blockers that reduce heart rate (beats/min).

Reduction heart rate (beats/min)	rank number
16	1
17	**2**
18	**3**
19	4
20	5
21	6
22	7
23	8
24	**9**
25	10
26	11
28	**12.5**
28	12.5
29	**14.5**
29	**14.5**
30	**16**
31	17
32	**18**
35	**19.5**
35	**19.5**

2 steps:

1. The data from both groups ($n_1 = 10$, $n_2 = 10$) ranked together in ascending order of magnitude. Average equal values. Instead of 12 and 13 we take 12.5 and 12.5.
2. Add up the rank numbers per group. We have 81.5 thin data and 128.5 fat data. Table for Wilcoxon rank sum test:
   difference >71 → P<0.05. Therefore: we cannot reject null hypothesis, groups not significantly different.

Unpaired non-parametric test: Wilcoxon rank sum test.
Table uses difference of added up rank numbers between group 1 and group 2.

n_1 → n_2 ↓	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

24. SUMMARY

What you should know:

1. Scientific rules (prior hypothesis, valid design, strict description, uniform analysis).
2. Efficacy and safety data are generally continuous and proportional data respectively.
3. Historical controls, factorial designs.
4. “Mean ± SEM" summarizes many trials similar to ours, and can be used for statistical testing.
5. Difference between paired and unpaired t-test.
6. Use of t-table, calculate examples of data for yourself.
7. Notion of negative and positive correlation in paired comparisons.
8. ANOVA appropriate for more than 2 groups or treatments (don't learn by heart).
9. Non-parametric tests; don't learn by heart but make sure that you can do it when looking it up.

25. EXERCISES TO CHAPTER 1

1. Give the four scientific rules for randomised controlled trials:
   A. prior hypothesis, strict description of methods, uniform data analysis, strict inclusion criteria,
   B. prior hypothesis, strict description of methods, uniform data analysis, valid design,
   C. strict description of methods, uniform data analysis, strict inclusion criteria, valid design,
   D. uniform data analysis, strict inclusion criteria, prior hypothesis, valid design.

   Which alternative is correct?

2.	Efficacy data	and	safety data in a trial.
A.	continuous data	"	proportional data,
B.	proportional data	"	continuous data,
C.	binary data	"	continuous data,
D.	binary data	"	ordinal data.

Which alternative is correct?

3. Factorial trial designs are for:
   A. multiple groups,
   B. historical data,
   C. continuous monitoring,
   D. multimodal therapies.

   Which alternative is correct?

4. Mean ± SEM summarizes
   A. many data similar to ours,
   B. many trials similar to ours,
   C. many means similar to ours,
   D. many standard deviations similar to ours.

   Which alternative is correct?

5. Using unpaired statistical test for paired data is wrong because:
   A. with + correlation the analysis is flawed,
   B. with – correlation the analysis is flawed,
   C. with 0 correlation the analysis is flawed,
   D. with – correlation power is lost.

   Which alternative is correct?

6. Calculate p-value of unpaired data using t-test
   data group 1: 6.0, 7.1, 8.1, 7.5, 6.4, 7.9, 6.8, 6.6, 7.3, 5.6
   group 2: 5.1, 8.0, 3.8, 4.4, 5.2, 5.4, 4.3, 6.0, 3.7, 6.2
   A. n.s.
   B. $0.05<P<0.10$
   C. $P<0.05$
   D. $P<0.01$

   Which alternative is correct?

7. Calculate p-value of paired data using t-test
   data sample 1: 6.2, 7.0, 8.1, 7.5, 6.5, 7.9, 6.8, 6.7, 7.3, 5.9
   data sample 2: 5.1, 7.8, 3.9, 4.5, 5.3, 5.4, 4.9, 6.1, 3.8, 6.3
   A. n.s.
   B. $0.05<P<0.10$
   C. $P<0.05$
   D. $P<0.01$

   Which alternative is correct?

8. Multiple groups ANOVA assesses whether:
   A. between groups sums of squares (SS) is large compared to within groups SS,
   B. within subject SS is large compared to within group SS,
   C. treatment SS is large compared to residual SS,
   D. within subject SS is larger compared to residual SS.

   Which alternative is correct?

9. Paired ANOVA assesses whether:
   A. between groups sums of squares (SS) is large compared to within groups SS,
   B. within subject SS is large compared to within group SS,
   C. treatment SS is large compared to residual SS,
   D. within subject SS is larger compared to residual SS.

   Which alternative is correct?

10. Calculate p-value of unpaired data using Wilcoxon rank sum test
    data group 1: 6.0, 7.1, 8.1, 7.5, 6.4, 7.9, 6.8, 6.6, 7.3, 5.6,
    group 2: 5.1, 8.0, 3.8, 4.4, 5.2, 5.4, 4.3, 6.0, 3.7, 6.2.
    A. n.s.
    B. 0.05<P<0.10
    C. P<0.05
    D. P<0.01

    Which alternative is correct?

11. Calculate p-value of paired data using Mann-Whitney test:
    data sample 1: 6.2, 7.0, 8.1, 7.5, 6.5, 7.9, 6.8, 6.7, 7.3, 5.9
    data sample 2: 5.1, 7.8, 3.9, 4.5, 5.3, 5.4, 4.9, 6.1, 3.8, 6.3
    A. n.s.
    B. 0.05<P<0.10
    C. P<0.05
    D. P<0.01

    Which alternative is correct?

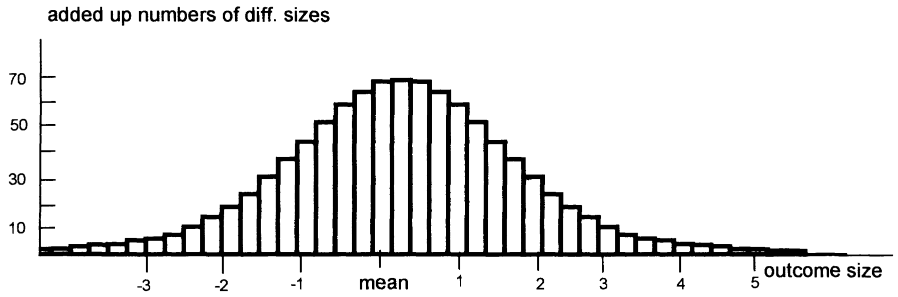
CHAPTER 2

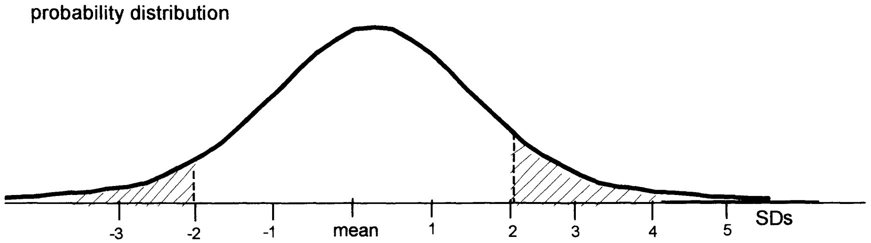
EQUIVALENCE TESTING

1. A NEGATIVE STUDY ≠ EQUIVALENT STUDY, WHY SO?

A study unable to find difference is not same as an equivalent study.
For example, a study of 3 patients finds no difference because the study is too small. Equivalence testing is important for diseases for which a placebo control group is unethical. In this case the new treatment has to be compared with standard treatment, and this comparison is at risk of finding little difference.

2. SUMMARIZE THE DATA





1.Histogram
- On the x-axis individual data/ on the y-axis "how often" (e.g., mean occurs most frequently, bars both sides gradually grow shorter).
- This method is not adequate for testing hypotheses.

2.Gaussian curve - On the x-axis individual data or SDs* -distant-from-mean.
- On the y-axis bars are replaced by continuous line.
- Now, it is impossible to determine from graph how many patients had a particular outcome!
- Instead, important inferences (conclusions) can be made, e.g.:

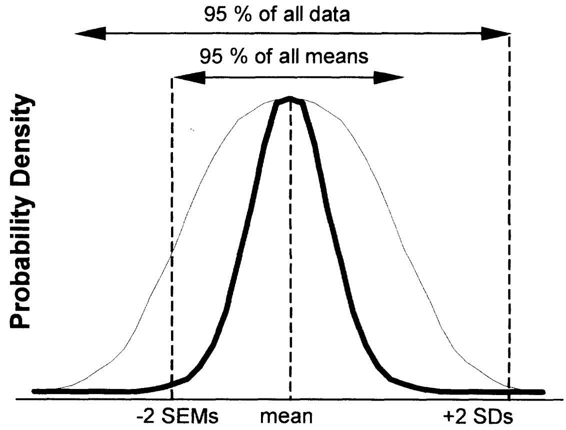
Total area under the curve (AUC) = 100% of the data of our trial

AUC left from mean	=	50%	of	data
AUC left from –1 SDs	=	15%	of	data
AUC left from –2 SDs	=	2.5%	of	data

- This method is not yet adequate for testing hypotheses, but better.

* SD = standard deviation = is estimate of spread in data, and is calculated according to SD = $\sqrt{}$ [Σ (x- $\bar{x}$)^2 / (n-1)]
where x = individual data, $\bar{x}$ = mean, n number of data).

3. TWO GAUSSIAN CURVES



Two gaussian curves are given: a narrow and a wide one, both based on the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes means of many trials similar to ours.

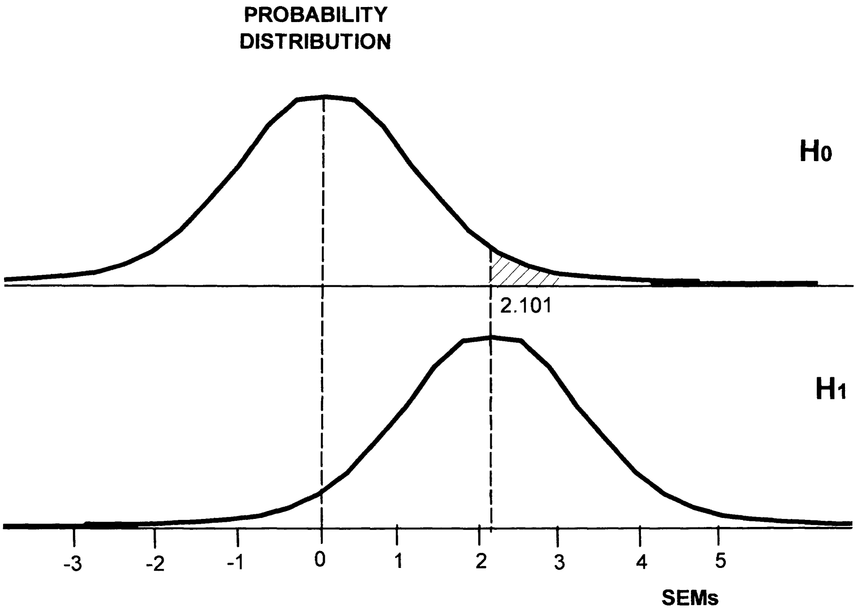
We won't try to make you understand why so. Still, it's easy to conceive that distribution of means of many trials is narrower, and has fewer outliers than the distribution of the actual data from our trial.

Believe it or not:
The narrow curve with SEMs* on the x-axis can be effectively used for testing important statistical hypotheses:

1. No difference between new and standard.
2. Real difference “ “ “ “ .
3. New treatment is worse than old.
4. Two treatments are equivalent.

SEM-curve is narrower than SD-curve because SEM=SD / √ n.
*SEM= standard error of the mean.

4. NULL-HYPOTHESIS



What does 0-hypothesis look like in graph?

H1 = graph based on the data of our trial with SEMs on the x-axis (in statistics often called z-axis).

H0 = the same graph with mean 0 (mean ± SEM = 0 ± 1).

Now we will make a giant leap from our data to the entire population (we can do so because data are representative for entire population).

H1 = also summary of means of many trials similar to ours (this assumption is correct, because our study is representative for population, and because we have SEM-units on the x-axis).

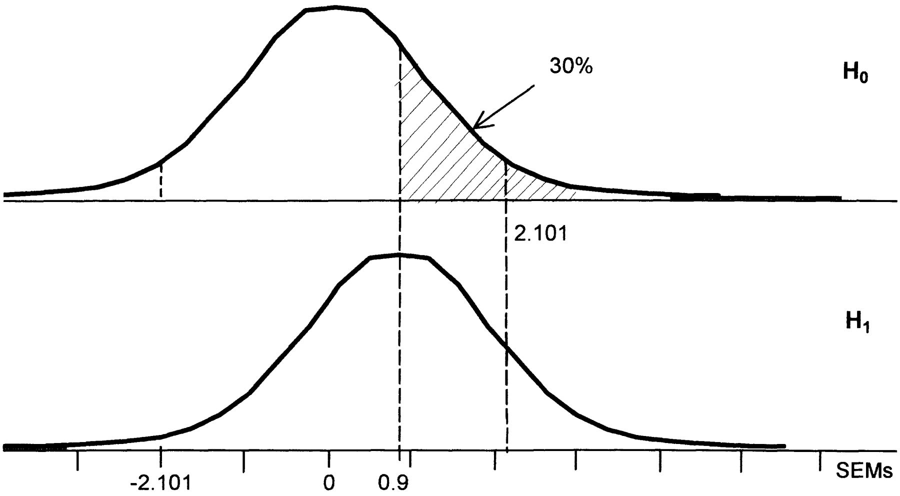
H0 = summary of means of many trials similar to ours but with overall effect 0 (our mean is not 0, but 2.101 SEMs. It, still, could be an outlier of many studies with an overall effect of 0).

So think from now on of H0 as the distribution of the means of many trials with an overall effect of 0. If hypothesis 0 is true, then the mean of our study is part of H0. We can't prove but can calculate chance / probability of this possibility.
A mean result of 2.9 SEMs is far distant from 0. Suppose it belongs to H0.

Only 5% of the H0-trials are ≥2.1 SEMs distant from 0,
because the AUC here = only 5% (Figure).
The chance that our trial belongs to H0 is thus ≤ 5%
(reject null hypothesis of no effect).

Conclude here:
≤ 5% chance to find this result.
In usual terms: we reject the null hypothesis of no effect at Probability (P) ≤ 0.05 or P ≤ 5%.

5. NEGATIVE STUDY

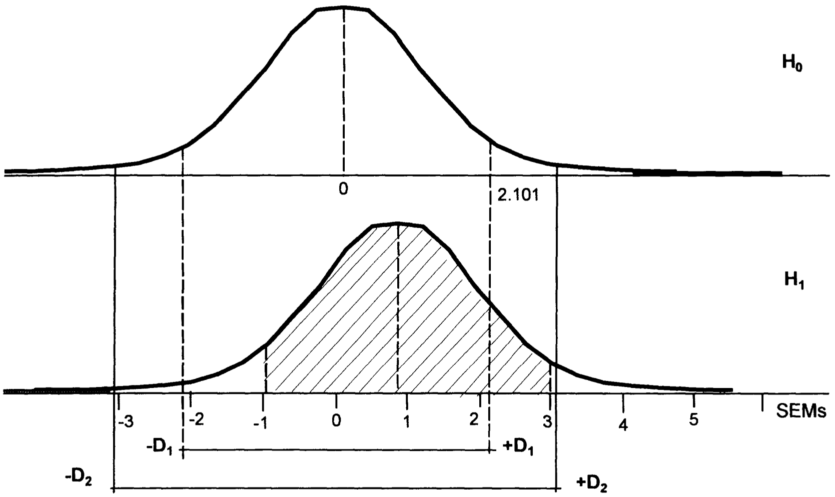


If our mean result is not 2.1 but only 0.9 SEMs distant from 0,
we say **not significantly different from 0. Do not interpret in terms of statistical equivalence.**
What it does mean: suppose our result belongs to H0 trials.

Up to 30% of H0 trials are 0.9 SEMs or more distant from 0,
because AUC right from 0.9 is 30%.
Chance that our result belongs to H0 is up to 30%.
This is not good enough to reject 0 hypothesis.
This is not good enough to accept equivalence either.

there is a subtle difference between "not different" and "similar".
So far we have talked about differences.
Now we are going to talk about similarities.

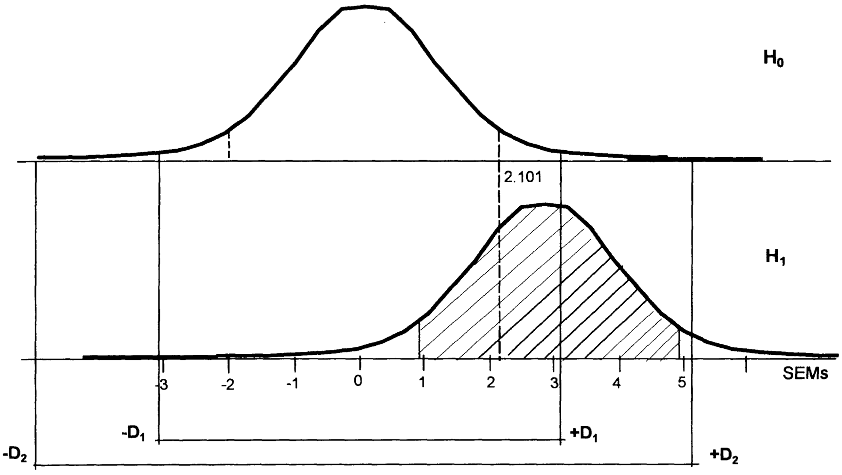
6. EQUIVALENCE TESTING



How similar versus (vs) control or vs 0 is our trial? The mean result of our trial 0.9 SEMs distant from 0, so not significantly different from 0. Is it then equivalent?? This depends on criterion of equivalence. The so-called D sets defined interval of equivalence. If our trial is completely within this interval: equivalence demonstrated. With D_1 boundaries: no equivalence demonstrated. With D_2 boundaries: yes, equivalence demonstrated.

Note: striped area under the curve = so-called 95% confidence intervals (CIs) = interval between approximately – 2 SEMs and +2 SEMs distant from the mean.

7. EQUIVALENT AND AT THE SAME TIME SIGNIFICANTLY DIFFERENT



Another example. The mean result of our trial is 2.9 SEMs distant from 0, and so significantly different from 0. Is it not equivalent then?? With D_1 the trial not completely within D-boundaries: we have no equivalence. With D_2 the trial is completely within D-boundaries: yes, we have equivalence.

Note: with D_2 there is a significantly different result and yet equivalence.
Conclude:
Equivalence testing is not complicated.
Just set your boundaries of equivalence,
check whether 95% CIs intervals are...

- completely within: equivalence is demonstrated;
- partly within: unsure;
- completely without: no equivalence demonstrated.

8. OVERVIEW OF ALL POSSIBILITIES

```
Study  Statistical                          equivalence
(1-8)  significance                         demonstrated
       demonstrated
________________________________________________________________________

1. Yes--------------------------------------------------------------< not equivalent >
2. Yes-------------------------------------------------------< uncertain >------------
3. Yes --------------------------------------------< equivalent >----------------------
4. No ---------------------------------< equivalent >-------------------------------
5. Yes--------------------------< equivalent >-----------------------------------------
6. Yes-----------< uncertain >--------------------------------------------------------
7. Yes-< not equivalent >------------------------------------------------------------
8. No--------<_______________ uncertain _______________>--------
______________________________!_____________________________!___________________
                             -D             O              +D
                                      true difference
```

Overview of all possibilities is given above: between brackets are 95%CIs of trials, on x-axis the place of 0- effect and D boundaries are given.

Study
1. Completely without D-boundaries: no equivalence,
2. Partly without.......... : uncertain,
3. Completely within.... : yes equivalence,
4. " " .... : yes equivalence,
5. " " .... : yes equivalence,
6. Partly without........... : uncertain,
7. Completely without... : no equivalence,
8. Partly without........... : uncertain.

Also an overview is given of whether or not a significant difference from 0 is demonstrated:

Particularly, possibilities 3 and 5 are remarkable: studies are equivalent because 95% CIs do not cross D-boundaries and significantly different because 95% CIs do not cross 0 value.
95% CIs interval = interval between approximately – 2 SEMs and +2 SEMs distant from the mean, so not crossing 0 indicates that mean result is > 2SEMs distant from 0 (statistically significant at $p<0.05$).

9. DEFINING D-BOUNDARIES

Hardest part: define D-boundaries.
For a rabbit -50 to + 50 miles/hour
turtle -2 to + 2 " "
snail -2 to + 2 meters/ " .
D-boundaries indicate area-of-undisputed-clinical-relevance.

10. ROBUSTNESS OF EQUIVALENCE TRIALS

Robustness of equivalence trials
-Equivalence testing is increasingly getting routine in clinical trials.
-E.g., for comparing test treatment with standard treatment:
efficacy parameters are then often equivalent,
while safety parameters (side effects) may be largely different.
In many trials patients are lost!
Intention-to-treat (ITT) population includes patients lost.
Completed-protocol-population includes only patients who completed study.
Usual hypothesis testing uses ITT population:

1. It makes differences look more similar.
2. It mirrors what will happen in practice (including non-compliants).
3. It shifts study towards negative result.

Equivalence testing using ITT population:

1. Idem.
2. Idem.
3. It shifts study towards positive result.

Note: Perform both ITT and completed protocol analysis. If difference is little, we have a robust study.

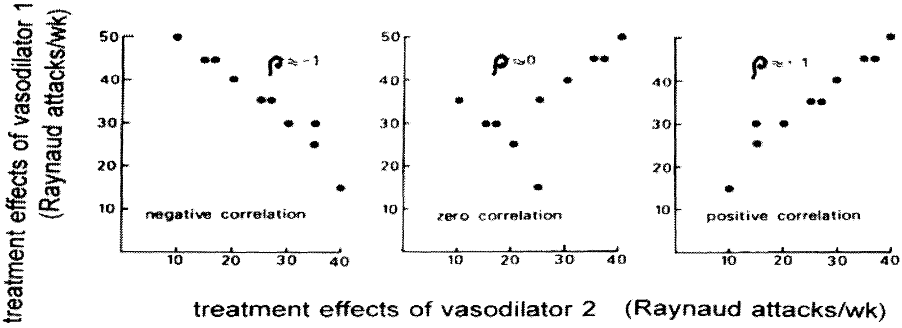
11. EXAMPLE

Example: Compound A and B are used for the treatment of asthma attacks, using peak expiratory flow (l/min) as primary outcome. The boundaries of equivalence were set at ± 15 l/min , which means that D= 15 l/min. The results were the following:

	Mean expiratory flow
Treatment A	420 l/min
Treatment B	417 l/min
Mean difference	3 l/min

The estimated standard error of this difference was calculated to be SEM = 4
The 95% CIs for this difference range between –1.96 and +1.96 SEM.
This means that the interval is approximately between -5 and +11 SEMs and is thus entirely within the range of equivalence of –15 l/min and + 15 l/min, and, so, equivalence is confirmed.

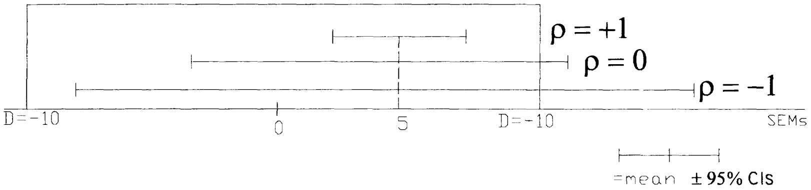
12. CROSSOVER EQUIVALENCE STUDIES WITH DIFFERENT LEVELS OF CORRELATION



Example:
Three equivalence trials for patients with Raynaud's phenomenon treated with vasodilator 1 for one week and vasodilator 2 for another week. Data are the numbers of Raynaud attacks per week.
Left trial shows a strong negative correlation between treatments: every time one treatment performs well, the second does not so.
Middle trial: the correlation is approximately zero.
Right trial: strong positive correlation between treatments exists: every time one treatment performs well, the other performs well too.

Observe that mean difference is every time 5, 5, and 5, but SEMs are very different 6.46, 2.78, and 0.76.
The graph below shows 95% CIs of 3 studies and the given D-boundaries.

	$\rho = -1$			$\rho = 0$			$\rho = +1$	
	vasodilator			vasodilator			vasodilator	
one	two	paired differences	one	two	paired differences	one	two	paired differences
45	10	35	45	40	5	10	10	0
40	15	25	40	35	5	20	15	5
40	15	25	40	35	5	25	15	10
35	20	15	35	30	5	25	20	5
30	25	5	30	25	5	30	25	5
30	25	5	30	10	20	30	25	5
25	30	-5	25	15	10	35	30	5
25	35	-10	25	15	10	40	35	5
20	35	-15	20	20	0	40	35	5
10	40	-30	10	25	-15	40	40	5
means		5			5			5
SEMs		6.46			2.78			0.76



Only the positive-correlation-study demonstrates equivalence!
Note: crossovers mostly have a positive correlation and are therefore very suitable for equivalence testing.

13. CONCLUSIONS

1. The use of placebos is unethical when an effective active comparator is available.
2. With active comparator new treatment may simply match standard treatment.
3. Equivalence trials have to be at least twice as large as comparative trials, you will understand after reviewing power analysis (chapter 3).
4. Predefined area of equivalence between -D and +D is based on clinical arguments.
5. Equivalence testing is indispensable in drug development(for comparison of new treatment versus an active comparator).

14. EXERCISES TO CHAPTER 2

1. A. study unable to find difference demonstrates therapeutic equivalence,
   B. ¨ ¨ ¨ ¨ ¨ has little power,
   C. ¨ ¨ ¨ ¨ ¨ does not have little power,
   D. ¨ ¨ ¨ ¨ ¨ has adequate power.

   Which alternative is correct?

2. Therapeutic equivalence in a trial indicate:
   A. 95 % confidence intervals (CIs)completely within predefined boundaries of equivalence,
   B. .... partly within ....,
   C. .... completely without.................................,
   D. .... partly without.......................................

   Which alternative is correct?

3. A. presence of equivalence includes the possibility of significant difference,
   B. presence of equivalence and significant difference cannot be found simultaneously,
   C. significant difference means that 95 % CIs cross 0 value,
   D. equivalence means that 95 % CIs cross D boundary of equivalence.

   Which alternative is correct?

4. A robust or sensitive equivalence trial indicates that:
   A. intention to treat (ITT) analysis yields better sensitivity than completed trial (CT) analysis,
   B. ITT analysis yields less sensitivity than CT analysis,
   C. ITT analysis yields similar sensitivity as CT analysis,
   D. ITT analysis shifts trial towards negative result.

   Which alternative is correct?

5. Example of therapeutic equivalence trial yields mean result in group 1 of 415 l/min (n=100), in group 2 of 421 l/min (n=100); SEM of the mean differences between the two groups 4 l/min; D boundaries are set at + and – 10 l/min.
   A. equivalence demonstrated,
   B. equivalence unsure,
   C. no equivalence,
   D. significant difference demonstrated.

   Which alternative is correct?

6. Paired data are very sensitive to equivalence testing:
   A. with + correlation,
   B. with – correlation,
   C. with 0 correlation,
   D. always.

   Which alternative is correct?

7. Equivalence studies are adequate for:
   A. comparison of new treatment vs standard treatment,
   B. ....vs baseline,
   C. ....vs placebo,
   D. testing small differences.

   Which alternative is correct?

CHAPTER 3

POWER, SAMPLE SIZE

1. DEFINITION OF STATISTICAL POWER

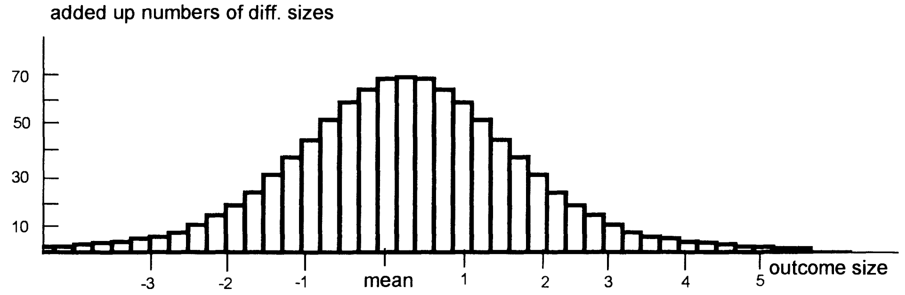
Clinical trials are for testing possible differences between new and standard treatment (or placebo).Statistical power is the chance of finding difference where there ís one, and is thus very relevant, because it assesses thé underlying hypothesis of most research. Biostatistics is at the interface of mats and biology. Biostatistics is based on approximations rather than exactnesses(e.g. normal distributions, linear relation etc).
Big power means a big chance of finding a difference where there ís one. Large trials have big power. Less relevant possibilities: chance of finding nó difference where there is one (type II error), and the chance of finding a difference where there is none (type I error). How to calculate size of power?

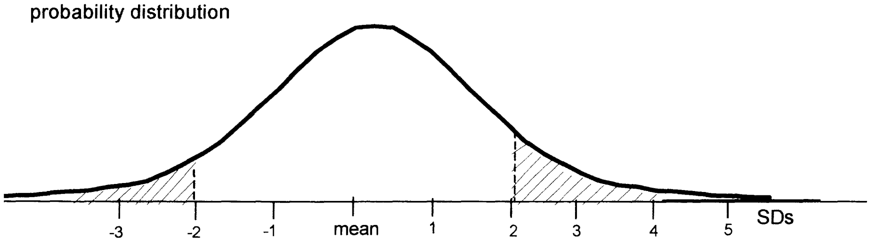
2. STATISTICS GIVES NO CERTAINTIES

what does statistics do for you? Statistics gives no certainties, only chances
What chances ? Chances that hypotheses are true/untrue (we accept 95% truths).
What hypotheses ?
E.g. 1. no difference from a 0 effect,
2. real difference from a 0 effect,
3. worse than a 0 effect.
Statistics is about estimating such chances / say testing such hypotheses.

Note: Trials often calculate differences between test treatment and control (for example, standard treatment, placebo, baseline), and, subsequently, test whether difference-between-two is different from 0.

3. SUMMARIZE THE DATA





1.Histogram
- On the x-axis individual data/ on the y-axis "how often" (e.g., mean occurs most frequently, bars both sides gradually grow shorter).
- This method is not adequate for testing hypotheses.

2.Gaussian curve
- On the x-axis individual data or SDs* -distant-from-mean.
- On the y-axis bars are replaced by continuous line.
- Now, it is impossible to determine from graph how many patients had a particular outcome!
- Instead, important inferences (conclusions) can be made, e.g.:

Total area under the curve (AUC) = 100% of the data of our trial,
AUC left from mean = 50% of data,
AUC left from –1 SDs = 15% of data,
AUC left from –2 SDs = 2.5% of data.

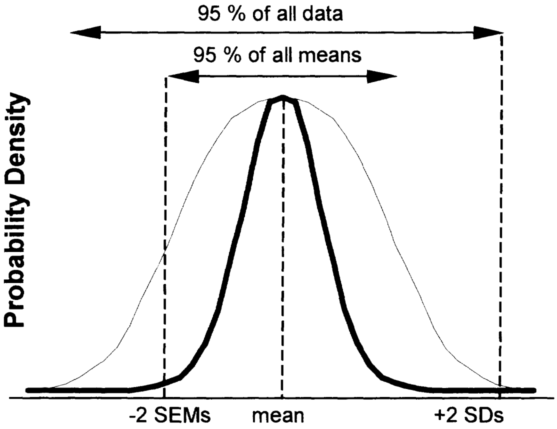
- This method is not yet adequate for testing hypotheses, but better.

* SD = standard deviation = is estimate of spread in data, is calculated according to

$$SD = \sqrt{\ [\ \Sigma(x- \bar{x})^2 \ / \ (n-1)\ ]}$$

where x = individual data, $\bar{x}$ = mean, n number of data).

4. TWO GAUSSIAN CURVES



Two gaussian curves are given: a narrow and a wide one, both based on the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes means of many trials similar to ours. We won't try to make you understand why so. Still, it's easy to conceive that distribution of means of many trials is narrower, and has fewer outliers than the distribution of the actual data from our trial.

Believe it or not:
The narrow curve with SEMs* on the x-axis can be effectively used for testing important statistical hypotheses:

1. No difference between new and standard.
2. Real difference " " " " .
3. New treatment is worse than old.
4. Two treatments are equivalent.

SEM-curve is narrower than SD-curve because SEM=SD/ √n.

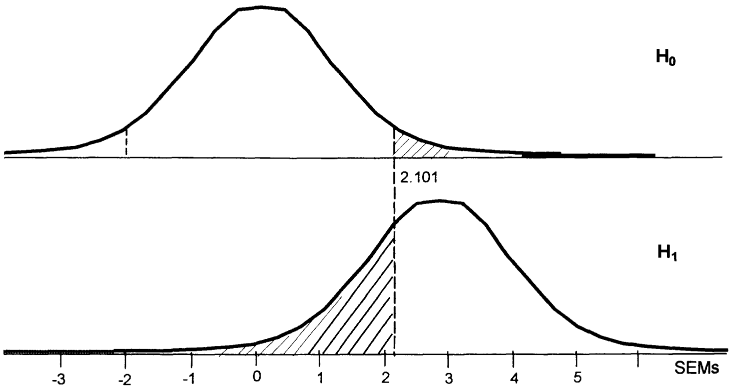
* SEM= standard error of the mean.

5. IMPORTANT HYPOTHESES

-Important Hypotheses are
Hypothesis 0,
No difference from a 0 effect,
Hypothesis 1,
Real difference from a 0 effect.

-We will nów, particularly, emphasize hypothesis 1.

6. HYPOTHESIS 1



-What do 2 hypotheses look like in graph?

-H1= graph based on the data of our trial (mean ± SEM= 2.9± 1).

-H0 = same graph with mean 0(mean ± SEM= 0± 1).

-Now we make a giant leap from our data to the entire population (our data were representative).

-H1 = also summary of means of many trials similar to ours (if we repeated trial, difference would be small, and distribution of means of many such trials would look like H1.

-H0 = summary of means of many trials similar to ours, but with overall effect 0 (our mean not 0 but 2.9. Still, it could be outlier of many studies overall effect 0.

-So, think from now on of H0 and H1 as summary of means of many trials.

-If hypothesis 0 is true, then mean of our study is part of H0.

-If hypothesis 1 is true, then mean of our study is part of H1.

-So, mean of our study may be part of H0, or may be part of H1.

-We can't prove anything, but we can calculate the chance of either of these possibilities.

-A mean result of 2.9 is far distant from 0:

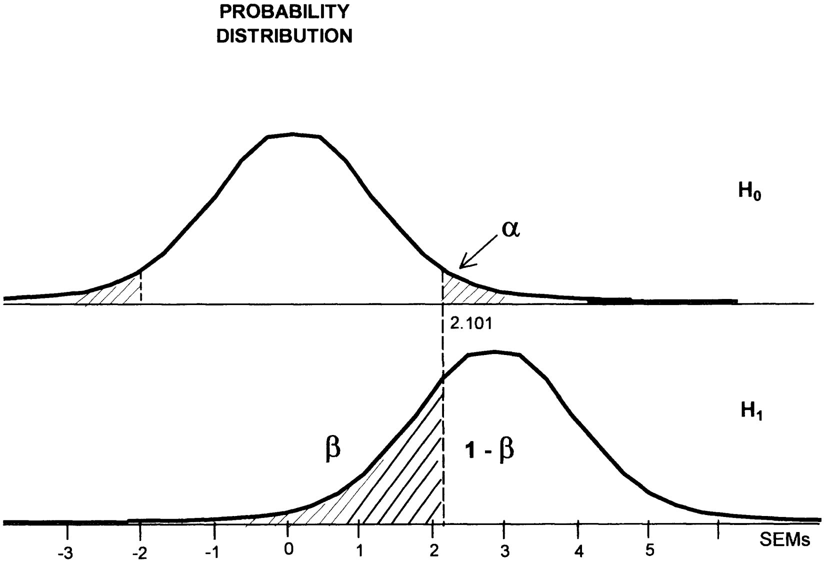
Suppose it belongs to H0.
Only 5% of the H0 trials >2.1 SEM distant from 0.
Chance that it belongs to H0 is thus < 5%.
We reject this possibility if probability < 5%.
(We say reject null hypothesis of no effect).
Suppose it belongs to H1.
Only 30% of the H1 trials <2.1 SEM distant from 0. These 30% cannot reject null hypothesis, only 70% can.
Conclude here if H0 true, <5% chance to find it,
if H1 true, 70% chance to find it.
Or in usual statistical terms: we reject null hypothesis of no effect at P<0.05 and with a statistical power of 70%.

7. ALPHA, BETA, 1-BETA

Alpha: small area under the curve (AUC) right from 2.1.
Alpha: level of rejection of H0.
Beta: AUC left from 2.1.
Beta: chance of finding nó difference where there is one.
Beta= type II error.
1-Beta = chance of finding a difference where there is one.
1-Beta= statistical power of a trial.

Please remember the little words.
Alpha = chance to find a difference where there is none.
Beta = chance to find no difference where there is one.
1-Beta = chance to find a difference where there really is one.
1-Beta = STATISTICAL POWER.

8. POWER GETS LARGER WHEN THE MEAN GETS LARGER



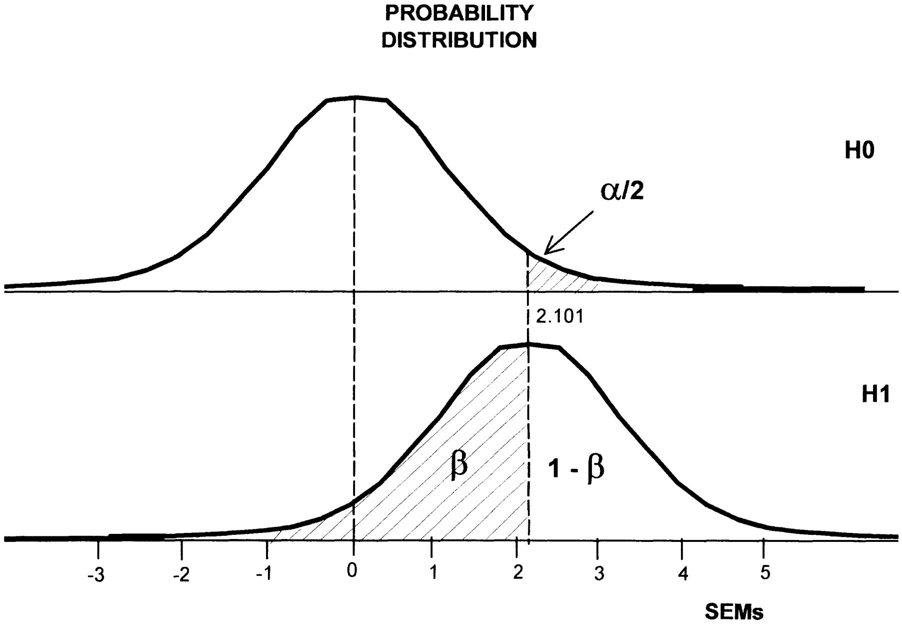
again:Alpha = outlier AUC of H0=area of rejection of H0 =usually 5% (trials within this AUC: null hypothesis is rejected),
= chance to find a difference where there is none,
= type I error.

Beta = left AUC of H1 (trials within this AUC: H1 is true but H0 cannot be rejected),
= chance to find no difference where there is one,

= type II error.
1-Beta = AUC right from 2.101.
= STATISTICAL POWER.

As the mean result of our trial gets larger, the alpha level remains 5%, but the AUC of 1-beta gets larger, and so the trial gets more power. As the mean result of our trial gets smaller, alpha again remaining 5%, the AUC of 1-beta gets smaller, and so the trial loses power.

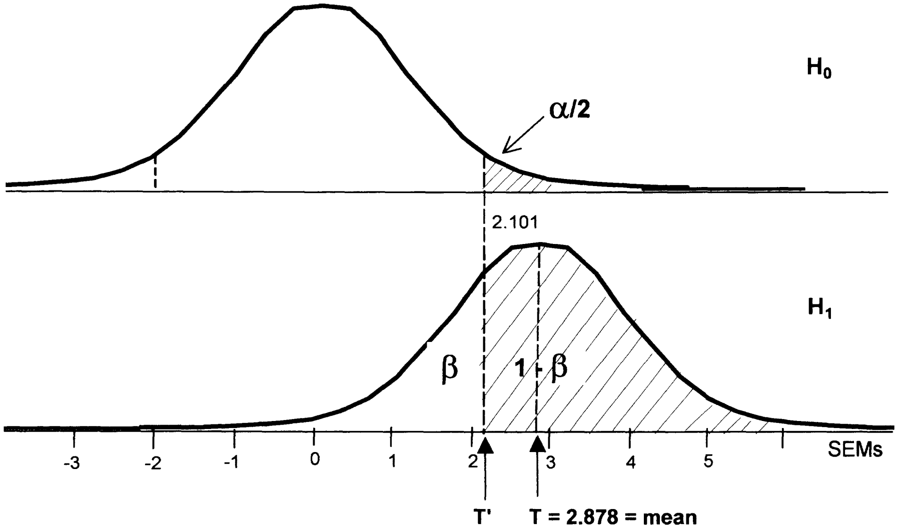
9. EXAMPLE OF POOR POWER



Example of poor power.
Mean result 2.101 SEMs distant from 0.
AUC right from 2.101 5% of total AUC.
Reject the null hypothesis of no effect .
However, 1-beta: covers only 50% of the AUC of alternative hypothesis H1.
Power is, thus, only 50%.
Note: a power of 50% is unacceptable for reliable testing (type II error 50%).

10. HOW TO CALCULATE POWER

PROBABILITY DISTRIBUTION



How to calculate power?

1. Estimate from graph (AUC in the above example: 1-beta= approximately 70%),
2. extrapolate from statistical table,
3. use computer.

It is useful to master method 2. Mean result of our trial 2.878 SEMs distant from 0 = T-value of our trial; 2.878 is, thus, the T-value of our trial.
Find beta by subtracting T-T^1 where T^1 is the T yielding AUC of 5% =2.101.
T-T^1= 2.878-2.101=0.777.
Now use T-table to find 1-beta .

11. USE OF T-TABLE TO FIND POWER

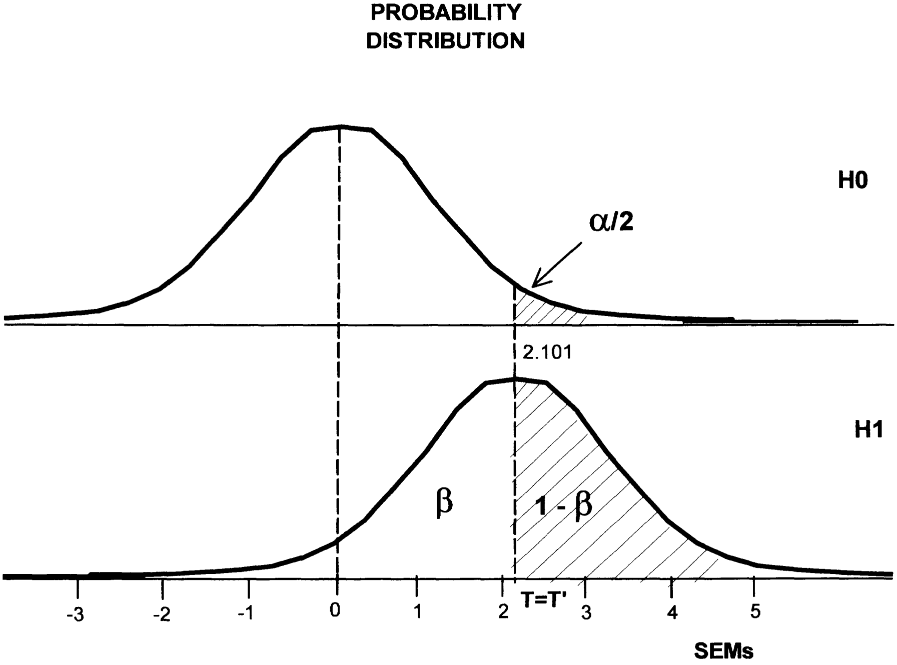
-8 Columns of T-values =mean results of trials in numbers of SEMs distant from 0.
-Left-hand column gives degrees of freedom (adjustments for number of patients and number of groups, for example, with 20 patients consisting of 2 groups we have 20-2= 18 degrees of freedom (dfs)).
-Upper two rows: AUCs right from T-values.
-E.g., a T of 2.11 and 20 subjects and 2 groups (18 degrees of freedom) means that the AUC right from 2.101 is < 0.05 (tested 2 sided, testing 2-sided means testing both right and left end of total AUC simultaneously).

Now for power analysis. Our T= approximately 2.9. T1 = approximately 2.1. T-T1= 0.777, AUC right from 0.777 means right from 0.688, is thus close to 0.25 = 25%. Beta (always tested 1-sided)= 25%, 1-Beta = statistical power is close to 1-25%=75%.

T-Table: v= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

v	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

12. USE OF T-TABLE TO FIND POWER, ONE MORE EXAMPLE



Different example.
Mean result is 2.1 SEMs distant from zero.
2.1 the T from the T-table for our trial.
Find Beta by subtracting T-T^1 where T^1 is the T yielding AUC of 5% =2.101.
T-T^1= 0.0.
Now use T-table to find 1-beta.

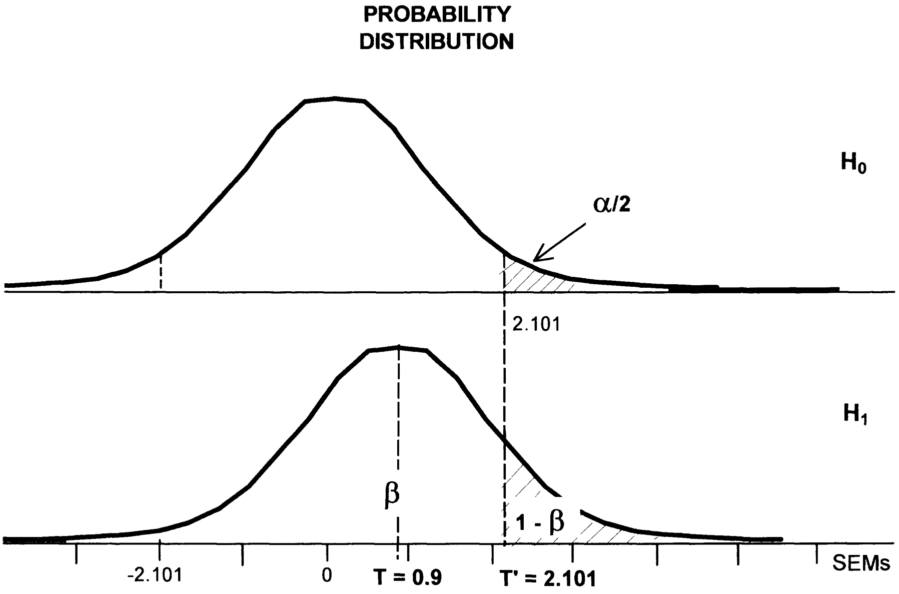
T= 2.101, T^1 = 2.101, T-T^1= 0.0 , which is a bit less than 0.257, AUC right from 0.0 = bit more than 0.40, For example, 0.50= 50%, beta (always tested 1-sided) = 50%, 1-Beta = statistical power = 1-0.50= 0.50= 50%, Power of only 50% is unacceptable for testing.

T-Table: v= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

Q = 0. 5	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16 0.0	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

13. USE OF T-TABLE TO FIND POWER, ONE MORE EXAMPLE

Things may get worse.
Mean result is 0.9 SEMs distant from zero.
T = 0.9. Find Beta by subtracting $T-T^1$ where T^1 is the T yielding AUC of 0.05 = 2.101.
$T-T^1$= -1.20.



Our T is 0.9. T^1 is 2.101. $T-T^1$= -1.2. 1.2 is again between 0.68 and 1.3, and it is close to 1.3 which corresponds with a little bit more than an AUC of 10%: 15% or so, -1.2 corresponds with AUC of 100% -15% = 85%, Beta = 85% , 1-Beta= 15%= STATISTICAL POWER. You already noticed that the procedure is rather imprecise with extreme values.

T-Table: v= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

v	*Q* = **0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
	2Q = **0.8**	**0.5**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.002**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

14. POWER FORMULAS

The above calculations made use of the formula:
Power= 1- prob ($z< T-T^1$), (learn by heart)
T=T of data,
T^1=T yielding AUC of 5%,
z= an interval on the Z-axis,
Prob = AUC between T and T^1.

For proportions different formula:
Z power= 2(arcsine $\sqrt{p_1}$-arcsine $\sqrt{p_2}$)$\sqrt{n/2}$-z^1. (don't learn by heart)

For equivalence testing:
Power=1-prob $z<D/SEM-z_{(1\text{-alpha})}$. (don't learn by heart)

Take note of z-values:
z^1 and $z_{(1\text{-alpha})}$
are intervals on the z-axis (x-axis) which are 1.96 SEMs wide with normal distributions and bit wider with T-distributions.

15. SAMPLE SIZE REQUIREMENTS

How many data are required in a sample ?
Just pulling the sample sizes out of a head gives rise to
ethical problems (too many patients given a potent inferior treatment unethical),
scientific " (negative studies require repetition),
financial " (costs involved in too small or too large studies).
Essential part of planning a clinical trial is to decide: how many people need to be studied in order to answer the study objectives.

16. A SIMPLE METHOD TO CALCULATE REQUIRED SAMPLE SIZE

Mean should be at least 1.96 or approximately 2 SEMs distant from 0 to obtain statistical significance.
Assume mean = 2 SEM.
Then mean/ SEM=2.
Then mean/ SD/ $\sqrt{n}$ = 2.
Then $\sqrt{n}$= 2.SD/mean.
Then n= 4. (SD/mean)^2 .
For example, with mean 10 and SD 20.
We will need a sample size of at least n= 4 (20/10)^2= 4.4=16.
P-value is then 0.05 but power is only 50%.

17. A MORE ACCURATE METHOD TO CALCULATE REQUIRED SAMPLE SIZE, POWER INDEX METHOD

The statistical power (1) of a trial assessing a new treatment vs control is determined by 3 major variables:

(2) D (mean difference or mean result).

(3) Variance in the data estimated as SD or SEM.

(4) Sample size.

It follows that we can calculate (4) if we know the other 3 variables.

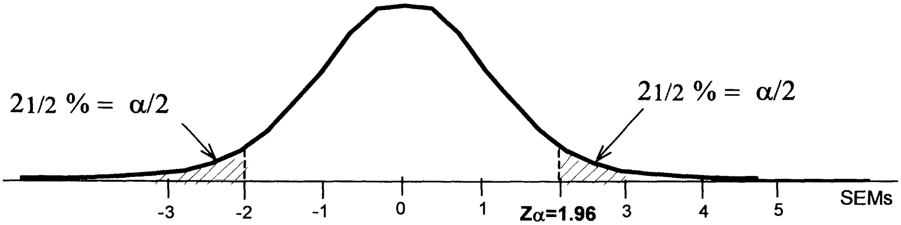
The relationship between (4) and the 3 other variables can be expressed in fancy formulas with $(z_\alpha + z_\beta)^2$ = power index as an important element in all of them.

18. SAMPLE SIZE COMPUTATIONS FOR CONTINUOUS VARIABLES, EXAMPLE

Formula for continuous variables:

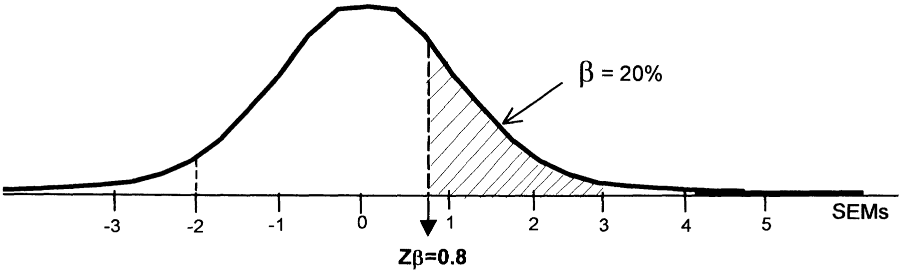
$n = 2. (SD/mean)2\ (z_\alpha + z_\beta)^2$

What is the size of this $(z_\alpha + z_\beta)^2$?



Z_α means " a place" on the Z-line. If alpha is defined 5%, or rather 2 x 2 1/2 % , then right from this place on the Z-line AUC=5%, or rather 2x2 1/2 %. So this place must be 1.96 SEMs distant from 0, or a bit more with T-distribution.

So, z_α = 1.96 = approximately 2.0.



If beta is defined 20%, what is place on Z-line of z_β ? Right from this place AUC= 20% of total AUC. This means that this place must be approximately 0.8 SEMs distant from 0. So z_β = approximately 0.8. Thus, $z_{(alpha)}$ = approximately 2.0,

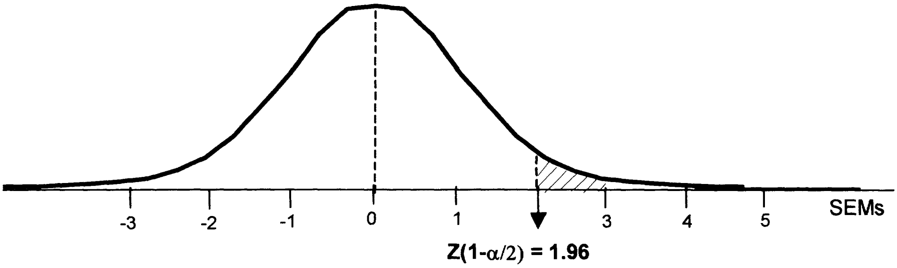
and z $_{\text{(beta)}}$ = approximately 0.8.
Power index = $(z_\alpha + z_\beta)^2$ = $(2.8)^2$ = 7.8 → 7.8.
Required sample size n= 2. $(\text{SD/mean})^2 (z_\alpha + z_\beta)^2$. (learn by heart)
Conclude that with α = 5% and power = 1- β = 80%, the required sample size is n = 15.6 $(\text{SD/mean})^2$.
E.g., with SD 20 and mean 10, we will need sample size of n= 15.6 $(20/\ 10)^2$ = 62.

19. OTHER FORMULAS TO CALCULATE REQUIRED SAMPLE SIZE

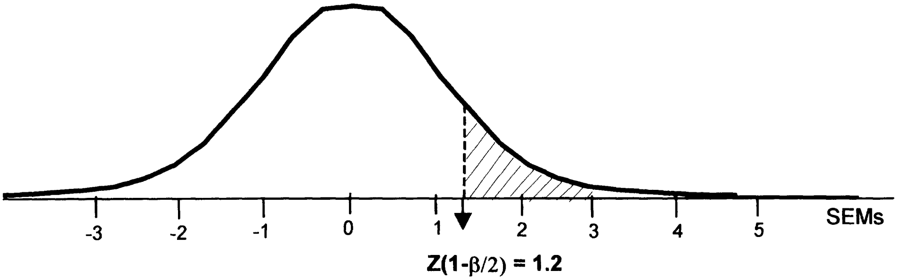
Required sample size formula for proportions:
$N= 2\ p_{\text{average}}\ (1- p_{\text{average}}\ )\ (z_\alpha + z_\beta)^2\ / D^2$ (don't learn by heart)

Required sample size formula for equivalence testing (" " " ")
$N = 2(\text{between subject variance})\ (z_{1-\frac{1}{2}\alpha} + z_{1-\frac{1}{2}\beta})^2 / D^2$
(where D is minimal difference we wish to detect).

What is size of power index of equivalence testing $(z_{1-\frac{1}{2}\alpha} + z_{1-\frac{1}{2}\beta})^2$?



If alpha is defined 5%, then ½ alpha = 2 ½ %. What is the place on the Z-line of $z_{1-\frac{1}{2}\alpha}$? Left from this place: AUC = 1- ½ alpha = 100- 2 ½ %= 97 ½ % of total AUC. So place is, just like z_α , 1.96 SEMs distant from 0, or bit more with T-distribution. So, $z_{1-\frac{1}{2}\alpha}$ = 1.96 or app 2.0.



Now, if beta is defined 20%, then ½ beta = 10% ,what is the place on the Z-line of $z_{1-\frac{1}{2}\beta}$? Left from the place the AUC = 100% -10% = 90% of total AUC.

This means that this place must be approximately 1.2 SEMs distant from 0, or bit more.

$z_{1-\frac{1}{2}\beta}$ = approximately 1.2.

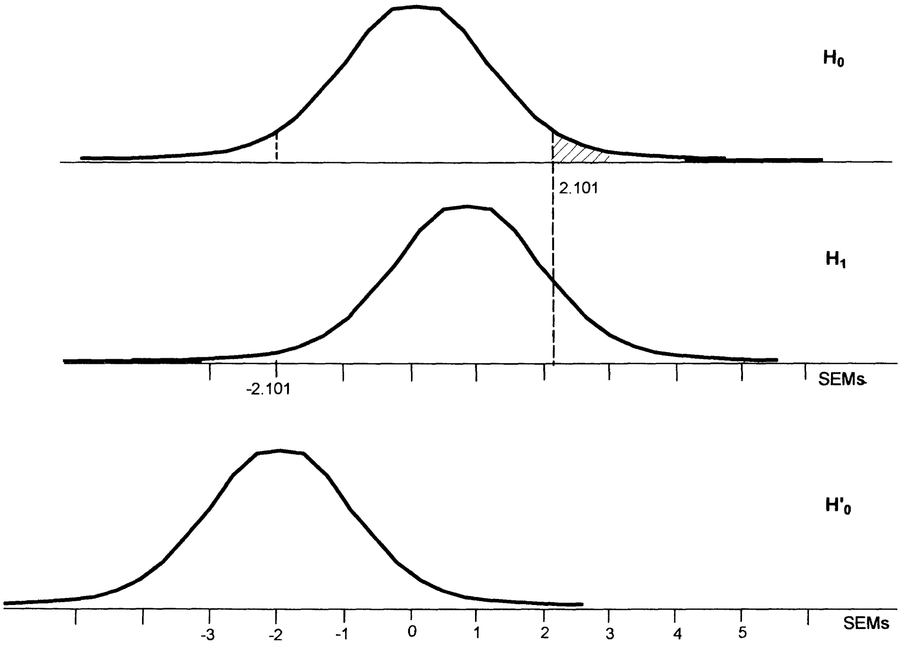
$z_{1-\frac{1}{2}\alpha}$ = approximately 2.0.

Power index for equivalence testing = $(2.0 + 1.2)^2$ = app 10.9.

NOTE: Power index null hypothesis testing = 7.8.
" " equivalence testing = 10.9.
Obviously, for equivalence testing much larger sample sizes are required !

20. TYPE I, TYPE II AND TYPE III ERRORS

We now address something really useful and very simple too.
If you can't demonstrate superiority of a new product,
maybe you 're interested to test whether it is inferior.
Testing inferiority = testing the chance of type III error.
Note: type I error = alpha = chance of finding a difference where there is none,
type II error =beta =chance of finding no difference where there really is one.



How it works?
Suppose: mean result approximately 1 SEM distant from zero,
which is not enough to reject the null hypothesis.

So, we have a negative trial (not able to reject its null hypothesis).Is this trial able to reject the chance of a type III error?
It is simple: we just need a new null hypothesis at approximately -2SEMs distant from 0 ($H0^1$). Our mean result is now approximately 3 SEMS distant from the new null hypothesis. 3 SEMS means a P value of 0.001.
And so, we have strong evidence to reject the null hypothesis of worse than zero. Our new treatment is, thus, nót significantly worse than control.

21. CONCLUSIONS

1. If underlying hypothesis in research is that one treatment is really different from control, power analysis is a more reliable approach to statistically evaluate the data than null hypothesis testing; a power level of at least 80% is recommended. Power = chance of finding a difference where there actually is one.
2. Despite the sometimes speculative character of prior estimates of presumable results of a trial, it is currently appropriate to calculate required sample size based on expected results.
3. The type III error can demonstrate in a negative trial whether the new treatment is significantly worse than the control treatment.
4. Important formulas:
   Power = 1- prob ($z < t\text{-}t^1$),
   Power index needed for calculating sample size($z_\alpha + z_\beta)^2$ is generally 7.8,
   Required sample size = 2. (SD/mean)2 ($z_\alpha + z_\beta)^2$.
5. Required knowledge for the exam of the module statistics of European College of Pharmaceutical Medicine Lyon france:
   be prepared to calculate power from simple example of (continuous) trial data using T-table. Be prepared to calculate required sample size of continuous data with alpha=0.05 and beta = 0.20 using power index (of proportional data see page 88).

22. EXERCISES TO CHAPTER 3

1. Statistical power
   A. chance of finding a difference where there is none,
   B...one,
   C.....................no difference ....none,
   D......................no difference..................one.

   Which alternative is correct?

2. Statistical power is
   A. α
   B. $1-\alpha$
   C. β
   D. $1-\beta$

   Which alternative is correct?

3. The mean result of a trial is 3.6 SEMs distant from 0. Find power using formula power= 1- prob($z<t-t^1$), where z is interval on the z-axis and t^1
   is the t of AUC of 5%, and $n_1=n_2=n=10$ (2 parallel groups).
   A. 90% < power < 95%,
   B. power> 80%,
   C. power <75%,
   D. power >75%.

   Which alternative is correct?

4. Calculate required sample size of a trial of continuous data that is expected to have normal distribution, a mean of 5 and SD of 15 and should produce a P-value of at least P=0.05 (mean result versus 0).
   A. 16,
   B. 36,
   C. 64,
   D. 100.

   Which alternative is correct?

5. The amount of statistical power is determined by
   A. mean effect-variance-sample size-power index,
   B. mean effect-variance-sample size,
   C. mean effect-variance,
   D. sample size- power index.

   Which alternative is correct?

6. The required sample size in unpaired trial is determined by:
   A. mean - SD - alpha level,
   B. mean - SD - beta level,
   C. mean - SD - power index,
   D. mean - SD - correlation level.

   Which alternative is correct?

7. Chance of finding a difference where there none
   A. type I error,
   B. type II error,
   C. type III error,
   D. null hypothesis.

   Which alternative is correct?

8. Chance of finding that new treatment is worse than control treatment
   A. type I error,
   B. type II error,
   C. type III error,
   D. null hypothesis.

   Which alternative is correct?

9. Chance of finding no difference where there is one
   A. type I error,
   B. type II error,
   C. type III error,
   D. null hypothesis.

   Which alternative is correct?

10 Calculate required sample size of a trial of continuous data that is expected to have normal distribution, a mean of 5 and SD of 15 and should produce a P-value of at least P=0.05 (mean result vs 0), and a power of at least 80%.

A. 115,
B. 51,
C. 205,
D. 320.

Which alternative is correct?

11. A parallel-group study of 40 subjects measures the effect of drug versus placebo on systolic blood pressure. We include the possibility of testing therapeutic equivalence and set the d-boundaries of equivalence between 0 to 6 mm Hg.

	N	mean(mm Hg)	SD(mm Hg)
Group 1 (drug)	20	9	4
Group 2 (placebo)	20	6	2

1. Calculate the level of statistical significance.
2. Calculate the level of statistical power.
3. Is statistical equivalence demonstrated?

12. In a study we expect a fall in cholesterol of 2 mmol/l with a SD of 4 mmol/l from baseline.

1. How many subjects have to be included to demonstrate a p-value of 0.05 and a power of 50%?
2. How many subjects have to be included to demonstrate a significant fall at α=0.05 and β=0.20 (power index= $(Z_\alpha + Z_\beta)^2 = 7.8$)?

CHAPTER 4

PROPORTIONAL DATA ANALYSIS: PART-1

1. SAFETY DATA ARE, GENERALLY, SUMMARIES OF PATIENTS WITH SIDE EFFECTS

-For efficacy data we, generally, use null-hypothesis testing.
-For safety data, more often, simply summaries of patients with side effects are given (±95% CIs= ± 2x p(1-p)/n x 100%, where p= proportion of patients with side effect).

2. EXAMPLE

-Sleepiness occured differently, 33% of the patients in left, 60% in right group. Is the difference true or due to chance?

	Alpha blocker n=16		alpha plus beta blocker n=15	
side effect	yes	no	yes	no
nasal congestion	10	6	10	5
alcohol intolerance	2	12	2	13
urine incontinence	5	11	5	10
disturbed ejaculation	4	2	2	2
disturbed potence	4	2	2	2
dry mouth	8	8	11	4
tiredness	9	7	11	4
palpitations	5	11	2	13
dizziness at rest	4	12	5	10
dizziness with exercise	8	8	12	3
orthostatic dizziness	8	8	10	5
sleepiness	5	10	9	6

In left group 5/10 sleepy, in right group 9/15. Is this difference in proportions real or due to chance? For that purpose an indication of certainty is required (e.g., the standard deviation).

3. STANDARD DEVIATION OF PROPORTION

To test whether there is a significant difference between two proportions standard deviations (SDs) of either of the proportions is required.

(1) SD continuous data $\surd\,[\,\Sigma\,(x-\bar{x})^2\,/\,(n-1)]$

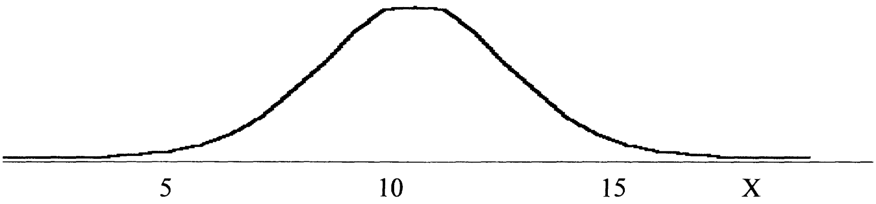
(2) SD proportional data $\surd\,[\,(\,p\,(1-p)\,]$

Important difference between formula (1) and (2) is that (2) is independent of n (=sample size).

4. WHY IS SD OF PROPORTION $\surd\,[(p(1-p)]$

This is not easy to prove. Yet, it is easy to conceive that the formula must be close to the truth.

For example, many samples of 15 patients are assessed for sleepiness. The proportion of sleepy people in the population is 10 out of every 15. Thus, in a representative sample from this population 10 sleepy patients will be the number most frequently encountered. It also is the mean proportion, and left and right from this mean proportion proportions grow gradually smaller, according to a binomial distribution (which becomes normal distribution with large samples). The figure below shows that the chance of 8 or fewer sleepy patients is 15% (AUC left from 8.3=15%). The chance of 6 or less sleepy patients is 2.5 % (AUC left from 6.6= 2.5%). The chance of 5 or less sleepy patients = 1%. This is a so-called binomial frequency distribution with mean 10 and a standard deviation of p (1-p)= 10/15 (1-5/15)= 1.7. -1SD means AUC of 15%, -2SDs means AUC of 2.5%. And, so, according to the curve below SD= p(1-p) is close to the truth.



Note: for null-hypothesis-testing SE rather than SD is required, and SE = SD/ $\surd$n.

5. METHOD-1 TO TEST DIFFERENCE BETWEEN TWO GROUPS OF PROPORTIONAL DATA

Normal test (= z-test for binomial or binary data) looks very much like T-test for continuous data. T= d/SE , z= d/SE, where d= mean difference between two groups or difference of proportions and SE is the pooled SE of this difference.

What we test is, whether this ratio is larger than approximately 2 (1.96 for proportions, a little bit more, e.g., 2.1 or so, for continuous data).

Example of continuous data (testing two means).

	Mean ± SD			$SEM^2 = SD^2/n$
group 1 (n=10)	5.9	± 2.4	liter/min	5.76/10
group 2 (n=10)	4.5	± 1.7	liter/min	2.89/10

Calculate: $mean_1 - mean_2 = 1.4$.
Then calculate pooled SEM = $\sqrt{}(SEM_1^2 + SEM_2^2) = \sqrt{}\,0.433 = 0.658$.
Note: for SEM of difference: take square root of sums of squares of separate SEMs and, so, reduce the analysis of two means to one of a single mean.

$$T = \frac{mean_1 - mean_2}{\text{Pooled SEM}} = 1.4 / 0.658 = 2.127 \text{ with degrees of freedom (dfs)} 18^{*} \; p<0.05.$$

*We have 2 groups of n=10 which means 2x10=20-2=18 dfs.

Example of proportional data (testing two proportions).

2x2 table	Sleepiness	No sleepiness
Left treatment (left group)	5	10
Right treatment (right group)	9	6

z = difference between proportions of sleepers per group / pooled standard error difference,
$z = (5/15 - 9/15) / \sqrt{}(SE_1{}^2 + SE_2{}^2)$,
SE_1 (or SEM_1) = $p_1 (1-p_1)/ n_1$ where $p_1 = 5/15$ etc........,
$d = |(5/15 - 9/15)| = 4/15$ pooled SE= $\sqrt{}(SE_1{}^2 + SE_2{}^2)$,
d/SE = 1.45, and difference is not statistically significant, because for a $p<0.05$ a d/SE of at least 1.96 is required.

6. NORMAL TABLE RATHER THAN T-TABLE MUST BE USED FOR PROPORTIONAL DATA

T-Table: v= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

v	*Q* = **0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
	2Q = **0.8**	**0.5**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.002**
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

The normal table is, actually, the bottom row of the T-table.

7. MORE EASY WAY TO TEST PROPORTIONS IS THE χ^2 TEST

More easy way to test proportional data is the χ^2 test:

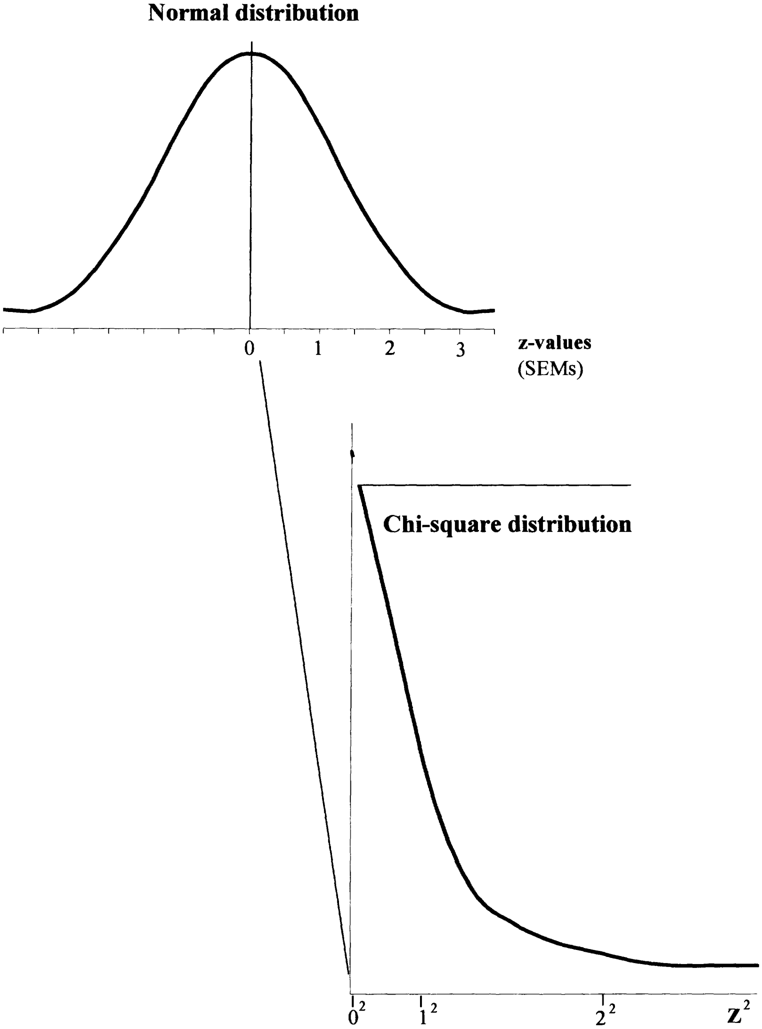
-Upper graph below presents a normal distribution.

-Lower graph: same distribution, but z-values have been squared, y-values are unchanged.

-Because z-values have been squared we have no negative values on z-axis anymore.

-Curve is skewed to right.

-Interpretation of chi-square (χ^2) curve: total AUC presents 100% of squared data.



8. HOW TO USE SQUARED CURVE

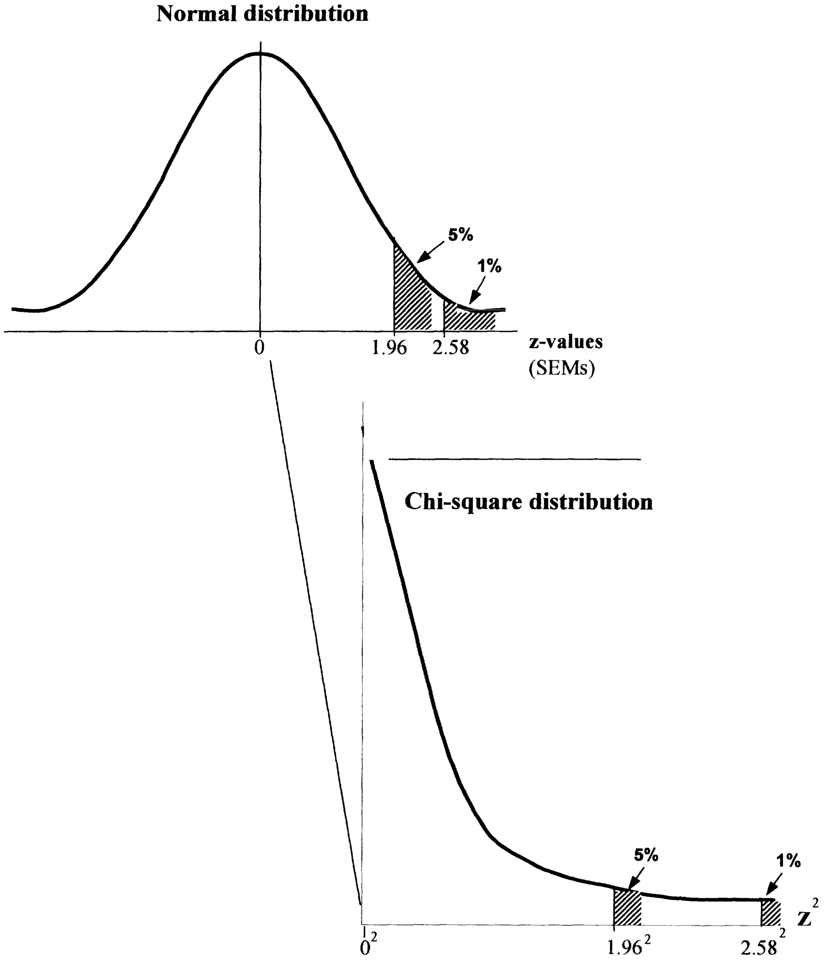
z-value = $\sum d/n$ = mean result of trial,

z-value^2 = $\sum d^2/n$ = variance of trial,

-Upper graph= frequency distribution of means (review chapter 1) of many trials similar to our trial: if mean trial result > 2 (1.96) distant from 0, and thus $p<5\%$.

-Lower graph= frequency distribution of variances of many trials similar to our trial: if variance > 1.96^2 distant from mean result, and thus $p<0.05$.

Interpretation: A chi-square value > 1.96^2 indicates that our variance is larger could happen by chance.



9. HOW THE χ^2 TEST FOR PROPORTIONS WORKS IN PRACTICE: 1X2 TABLE

Sleepiness	No-sleepiness	Sleepiness	No-sleepiness
observed		expected from population	
number	number	number	number
a (5)	b (10)	α (10)	β (5)

Is our observed sample significantly different from population?

a- α = 5-10 = -5
b- β =10-5 = $\underline{+5}$ +
0

So adding up differences from expected values does not tell us. Alternative: takes the square differences instead of differences.

$(a-\alpha)^2 = 25$ divide by α to standardize $= 2.5$
$(b-\beta)^2 = 25$ " " β " " $= 5$ +
$= 7.5$

Add-up sum of squared distances from supposed mean of population follows χ^2 distribution, $\chi^2 = 7.5$, 1 df $p<0.01$ (see section Tables).

10. HOW THE χ^2 TEST FOR PROPRTIONS WORKS IN PRACTICE: 2x2 TABLE

	Sleepiness	No-sleepiness	Sleepiness	No-sleepiness
	observed		expected	
Left treatment (left group)	5 (a)	10 (b)	....(α)	(β)
Right treatment (right group)	9(c)	6 (d)	... (γ)	....(δ)

cell 1: $(O-E)^2 / E = (a-\alpha)^2 / \alpha = (5 - 14/30 \times 15)^2 \ / \ 14/30 \times 15 = ..$
2: $= (b-\beta)^2 / \beta =$
3: $= (c-\gamma)^2 / \gamma =$
4: $= (d-\delta)^2 / \delta =$ ____________________ +
$= 2.106$

O= observed number;
E= expected number=(proportion sleepers /total number) x $\text{number}_{\text{group}}$.

Add-up sum of squared distances from expected number = best estimate of mean of population, and follows χ^2 distribution. χ^2 =2.106, and with 2-1=1degrees of freedom (dfs), and thus p>0.1.

11. ALTERNATIVE WAY TO FIND THE ADEQUATE χ^2 -VALUE: 2x2 TABLE

	Sleepiness	no sleepiness	total
Left treatment (left group)	5 (a)	10 (b)	a+b
Right treatment (right group)	9(c)	6 (d)	c+d
	a+c	b+d	

Calculating χ^2 value.

$$\frac{(ad-bc)(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

Value of chi-square again 2.106 at 2-1=1 degree of freedom, and thus p>0.1.

12. ONE MORE WAY TO FIND ADEQUATE THE ADEQUATE χ^2 -VALUE, FISHER-EXACT TEST: 2x2 TABLE

Fisher-exact test. It uses faculties. 5! indicates 5x4x3x2x1.

	Sleepiness	no sleepiness
Left treatment (left group)	5 (a)	10 (b)
Right treatment (right group)	9(c)	6 (d)

$$P = \frac{(a+b)!\ ((c+d)!\ (a+c)!\ (b+d)!}{(a+b+c+d)!\ a!b!c!d!} = 0.2 \quad \text{(much larger than 0.05)}$$

Computer can calculate faculties within 1-2 seconds.

Fisher-exact test is more adequate for testing small samples than the chi-square test (additional examples and discussion is given on page 87).

13. WITH χ^2 WELCOME TO THE REAL WORLD OF STATISTICS BECAUSE IT CAN BE USED FOR Kx2 TABLES

	Sleepiness	no sleepiness
Group 1	5 (a)	10 (b)
Group 2	9 (c)	6 (d)
Group 3	.. (e)	...(f)
Group 4	..	
Group 5		

cell a: $(O-E)^2 / E = (5 - 14/30 \times 15)^2 \ / \ 14/30 \times 15 =$ ..
b: $(O-E)^2 / E$
c: $(O-E)^2 / E$
d: $(O-E)^2 / E$
e: ..
f: ..
__________________ +
$\chi^2 =$ ..

The only difference is the degrees of freedom: they are 2-1 with 2x2 table and 5-1 with 5x2 table.

Note:- χ^2 looks weird at first.
-Main difference from normal-test or t-test: it uses squared values.
-Basis modern statistics.

14. MCNEMAR'S TEST FOR PAIRED YES/NO OBSERVATIONS

315 subjects tested for hypertension with:
automated device (test 1) and sphygmomanometer (test 2)
Finding discordant pairs.

		Test 1		
		+	-	total
Test 2	+	184	54	238
	-	14	63	77
total		198	117	315

$$\chi^2 = \frac{(54-14)^2}{54+14} = 23.5 \quad 1 \text{ df} \quad p< 0.001$$

184 subjects + twice, 63 - twice, no information. 54 and 14 subjects scored once + and once -. The latter two groups tell us which test is more likely to be positive. examples and discussion on this subject isgiven on page 90.

15. DIFFERENCES BETWEEN PROPORTIONS CAN ALSO BE ASSESSED BY CALCULATING THE ODDS RATIO (OR) AND ITS 95% CONFIDENCE INTERVALS (CIs) AND CHECKING WHETHER THE CONFIDENCE INTERVALS CROSS 1.0

Not crossing 1.0 means that the OR is significantly different from 1.0 at $p<0.05$. The 95% confidence intervals of an OR of unpaired data can be calculated as follows.

$\text{Ln OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$

For example

	Hypertension yes	hypertension no
Group 1	a n=5	c n=10
Group 2	b n=10	d n=5

OR=a/c / b/d = 0.25

95% confidence intervals of ln OR $= \ln \text{OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$

$= \ln 0.25 \pm 1.96 \sqrt{(1/5+1/10+1/10+1/5)}$

$= -1.3863 \pm 1.5182$

is between – 2.905 and 0.132,

95 % confidence interval of OR is between antiln –2.905 and antiln 0.132,

is between 0.055 and 1.14.

This confidence interval crosses 1.0, and is, thus, not significantly different from 1.0.

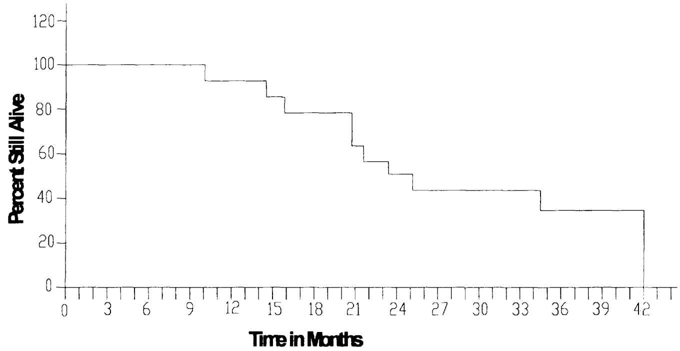
Additional examples and discussion on this subject is given on page 84.

16. HOW TO CALCULATE 95% CONFIDENCE INTERVALS OF AN ODDS RATIO WITH PAIRED OBSERVATIONS

$\text{Ln OR} \pm 1.96 \sqrt{(1/R + 1/S)}$ where R and S are the discordant pairs (see paragraph 14).

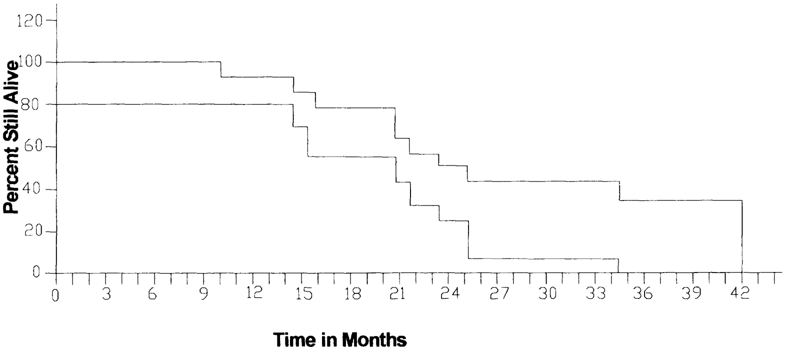
17. SURVIVAL ANALYSIS (KAPLAN-MEIER METHOD)

x-axis time, y-axis % survival.



-Fifteen patients followed for 36 months.
-At time 0 everybody is alive, At time 36 months 40% (6/15) alive.
-% does decrease whenever a patient dies.
-Problem: lost data (censored data), they live at the time they are lost, and so useful information.
-Solution:-recalculate fraction survivors at the end of every 2nd month,
-those who are lost are excluded,
-e.g., 1 lost and 1 death → 15/15 → 13/14 instead of 14/15.
-95% CIs of line can be calculated according to ± 2p(1-p)/n.100%.

18. TESTING SIGNIFICANCE OF DIFFERENCE BETWEEN 2 KAPLAN-MEIER CURVES



-In a certain 2-month period we have left the following numbers: a and b in curve 1, c and d in curve 2,

(i= 1,2,3,....)

	death	alive
curve 1	a_i	b_i
curve 2	c_i	d_i

Odds Ratio = a_i/b_i / c_i/d_i = $a_i\ d_i$ / $b_i\ c_i$

Significance of difference between curves calculated according to:
- Mantel-Haenszl (M-H) summary χ^2 test (=log rank test) :

$$\chi^2_{\text{M-H}} = \frac{(\Sigma\, a_i\ -\ \Sigma\,[(a_i+b_i)(a_i+c_i)/(a_i+b_i+c_i+d_i)]\ )^2}{\Sigma\,[\,(a_i+b_i)(c_i+d_i)(a_i+c_i)(b_i+d_i)\,/\,(\,a_i+b_i+c_i+d_i)^3]}$$

Note: alternative: Cox's proportional hazards model, analogous multiple linear regression for continuous data and logistic regression for multiple proportions, ± same result. Additional examples and disxussion on Kaplan-Meier methods is given on pages 94-99.

19. WHAT YOU SHOULD KNOW

1-For efficacy data null-hypothesis testing, for safety data summaries (95% CIs).
2-Test obvious differences in side effect scores between test and control using 2x2 tables.
3-Use chi-square or z-test for that purpose.
4-Paired data (each patient serves as his-her own control): Mc Nemar test is adequate.
5-Kaplan Meier curve : include lost patients.
6-Comparing Kaplan-Meier Curves: use Mantel-Haenszl chi-square = Log rank test.

20. QUESTIONS TO CHAPTER 4

1. A. Efficacy data analysis involves summaries and confidence intervals.
   B. Safety data analysis involves null-hypothesis testing.
   C. Efficacy data analysis mostly involves χ^2 - testing of proportional data.
   D. Safety data analysis mostly involves χ^2 - testing of proportional data.

   Which of the alternatives is the best answer?

2. SE of proportion = √ p (1-p) / n
   Is the difference between two unpaired proportions 4/16 and 12/16 different at
   A. 0.05<p<0.10
   B. p<0.05
   C. p<0.01
   D. 0<0.001
   using normal test: $z = d/SE = \text{proportion}_1 - \text{proportion}_2 \ / \ \sqrt{} \ (SE_1^2+SE_2^2)$.

   Which alternative is correct?

3. Is difference between the proportions from question 2 different at
   A. 0.05<p<0.10
   B. p<0.05
   C. p<0.01
   D. 0<0.001
   using the χ^2- test for 2x2 tables.

   Which alternative is correct?

4. Two groups of internists include 10 internists per group. 3 internists are burned out in group 1 while none is so in group 2.
   Is this difference significant at
   A. 0.05<p<0.10
   B. p<0.05
   C. p<0.01
   D. 0<0.001
   using the χ^2- test for 2x2 tables.

   Which alternative is correct?

5. A χ^2- curve is
   A. a squared Gaussian curve,
   B. a squared polynomial curve,
   C. a squared F-curve,
   D. a squared power curve.

   Which alternative is correct?

6. With χ^2- test and 1 df the null-hypothesis of no difference of our variance from a variance of 0 is rejected at $p<0.05$ if $\chi^2 >$ 1.960, (A)
   2.576, (B)
   3.484, (C)
   6.636, (D)

   Which alternative is correct?

7. Is the difference between two unpaired proportions 2/6 and 4/6 different at
   A. $0.05<p<0.10$
   B. $p<0.05$
   C. $p<0.01$
   D. not significantly different,
   using the Fisher-exact test.

   Which alternative is correct?

8. Our sample includes 4 blue eyed subjects out of 12, while the proportion in the general population is supposed to be 8 out of 12. At what probability is our sample different from the general population?
   A. $0.05<p<0.1$
   B. $p<0.05$
   C. $p<0.01$
   D. $p<0.001$

   Which alternative is correct?

9. Two antihypertensive drugs are tested for causing orthostatic hypotension in a single group of patients with hypertension. Is the difference between the numbers of patients suffering from orthostatic hypotension significant at
   A. $0.05<p<0.1$
   B. $p<0.05$
   C $p<0.01$
   D $p<0.001$
   using McNemar χ^2- test.

		Drug 1	
		yes	no
Drug 2	yes	65	28
	No	12	34

   Which alternative is correct?

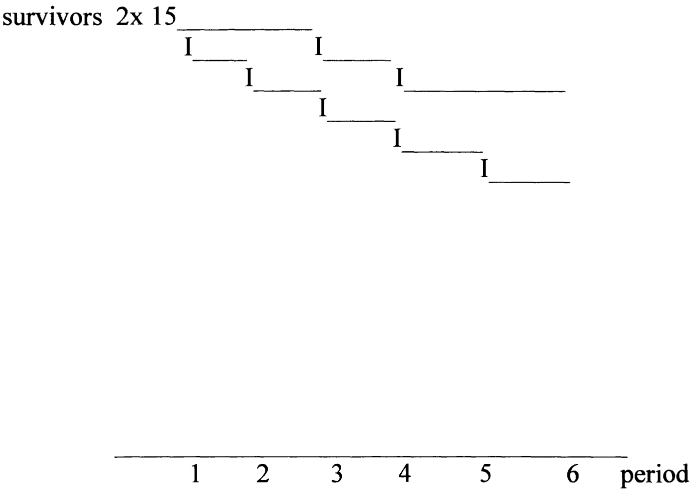
10. What is the odds ratio of discordant pairs and its 95% confidence intervals for orthostatic hypotension with drug-2 versus that of drug-1 from the above data?

11. In a parallel-group study the data are listed accordingly.

	orthostatic hypotension	
	yes	no
drug 1	77	62
drug 2	103	46

What odds ratio and what confidence intervals can be calculated and is this result significantly different from an odds ratio of 1.0?

12. Two groups of 15 subjects are followed for 6 periods, resp. 13 and 10 survive. Are the two Kaplan Meier curves significantly different from one another?



CHAPTER 5

PROPORTIONAL DATA ANALYSIS: PART 2

1. EXAMPLES

- How large is the response rate?
- How many patients do have side-effects?
- How many patients were alive (after 5 years)?

- Is the response rate under treatment A larger than under B?
- Are there more side-effects after than before treatment?
- What is the dose-effect curve?

Discrete data are encountered all over biomedical research. Basically a discrete variable is a characteristic that varies over patients but can occur in only a few different values. Gender is a typical example of a variable that can have two values "male" or "female", and "death" or "alive" is another example of a discrete variable. Variables that can attain only two values are called dichotomous or binary. Typical examples of discrete variables with more than two values are, e.g., blood type (A, B, AB, O), genotype in general, race.

In clinical (pharmacology) trials typical discrete data are sampled when investigating the response rate of drugs (response: yes or no), or the likelihood of side-effects (side-effect: yes or no), survival rate after a fixed time-period (alive: yes or no). In comparative trials the principal question is more complex like whether the response rate, or the survival rate is different between treatment A and B, or different before and after treatment, or different for varying doses.

A special type of discrete observation concerns the possibility that observations occur at different time points. This happens very often in trials with time-to-event as primary interest, e.g. time-to-death, or time-to-progression. Death or disease-progression varies between patients, and this variation requires special statistical analysis methods. A complication is caused by so-called censored data: when the event of interest is not observed in a patient during the trial, this does not entail that it will never occur in that patient, only that it did not occur in the time-period of the trial. Such an observation is called censored, and such data require special statistical handling. These methods and the methods required to analyze discrete data in general are subject of this chapter, but only the most basic methods will be discussed.

2. CHOICE OF STATISTICAL METHOD: A.

The type of "experiment",
and
the "type" of data sampled
determine
the required statistical method.

	1 sample **1 measurement**	**2 samples** **1 measurement**	**>2 samples** **1 measurement**
Discrete	Z- or chi-squared test	Z- or chi-squared test	chi-squared test
Censored	(kaplan-meier)	logrank test	logrank test
Quantitative	one sample t-test/ Wilcoxon test	unpaired t-test / Mann-Whitney test	ANOVA, Kruskal-Wallis test

There are many different statistical techniques, and most students and applied researchers are confused by the abundance, and the necessity to make choices among them. Luckily, there are many aids for making the choice somewhat easier. One is based on the rule that "the type of experiment, and the type of data together determine the required statistical analysis technique". This is a simplification naturally, the type of research question is important too, but the rule works quite well in most situations. The next two tables summarize the most important combinations of type of experiments, and type of data.

3. CHOICE OF STATISTICAL METHOD: B

	1 sample **2 measurements**	**1 sample** **>2 measurements**	**>1 samples** **>1 measurement**
Discrete	Mc Nemar test	Cochran's Q test	logistic regression
Censored	stratified logrank test	stratified logrank test	stratified logrank test
Quantitative	paired t-test / Wilcoxon test	ANOVA/ Friedman test	ANOVA

Quantitative data were discussed in previous chapters, in this chapter we look at discrete and censored data. There are other data types (ordinal data for instance, where observations denote a rank order only), but most data can be interpreted as quantitative or discrete (qualitative).

There are many different types of experiments, and we discuss only the simplest designs here: where we have only one group of individuals, two groups of individuals, or three of more groups of individuals, and each individual measured only once. When data are quantitative, the required statistical techniques are often analysis of variance techniques (of which the t-test is a special case), and these have been discussed in previous chapters. Equivalent techniques for discrete and censored data are the chi-squared test, and the logrank test.

4. CHOICE OF STATISTICAL METHOD: C

	1 sample 2 measurements	**1 sample >2 measurements**	**>1 samples >1 measurement**
Discrete	Mc Nemar test	Cochran's Q test	logistic regression
Censored	stratified logrank test	stratified logrank test	stratified logrank test
Quantitative	paired t-test / Wilcoxon test	ANOVA/ Friedman test	ANOVA

Other research designs concern one or more groups of individuals who are measured repeatedly. For quantitative data mixed-models (and the paired t-test) are used, for discrete and censored data the corresponding methods are stratified analysis procedures which are quite complex. If there is one group of individuals measured twice or more on a dichotomous variable, these stratified methods are known as McNemar′s or Cochran's test. The other methods involved are not discussed here.

5. ELEMENTS OF STATISTICAL ANALYSIS

- Operationalisation of research question in quantitative hypothesis,
- quantitative effect estimation: 'average' + variability,
- indication of the certainty (standard error (SE= SEM), confidence interval (CI)),
- hypothesis testing.

The basic steps taken in the statistical analysis of discrete and censored data are similar to those for quantitative data. Step 1 is always the translation of the research question into a quantitative hypothesis: statistics is about numbers. The research question is often formulated verbally, but for statistical analysis it must be translated into observable numbers. Usually the research question is translated into the so-called alternative hypothesis (H_a or H_1) , and its negation, otherwise called the null-hypothesis (H_0).

The specific type of numbers sampled in an experiment is determined by the research question, but in most cases many numbers, i.e., a lot of data is sampled. The second step, therefore, is to summarize the great amount of data into meaningful statistical entities. In general, these so-called statistics will be interpreted as a quantification of the effect of interest: examples are the mean, the median, and the proportion, or percentage. When data are quantitative, the variability of the data is important too, but this is less so with discrete data.

These statistics are based on finite samples, and it is easily imagined that in a new sample a somewhat different statistic will be found; this uncertainty must be quantified in the third step, and it entails calculation of the SE or a CI, usually a 95% CI.

Finally, we define the prior hypothesis which is going to be tested; a decision must be taken as to whether the hypothesis is true or false. This is often difficult, but in most cases we can calculate the likelihood of the data that will be found, if the (null-)hypothesis is correct: if this likelihood is small (say less than 0.05), then we will decide that the hypothesis is false.

6. EXAMPLE 1: ONE GROUP OF PATIENTS MEASURED ONCE

side-effects of ACE-inhibitors

- 135 diabetic patients with nephropathy,
- one year treatment with ACE-inhibitor,
- 10 patients experienced episodes of dry cough,

⇨ dry cough event rate = 10/135 = 0.074 = 7.4% = p,

- 95% confidence interval: 0.030 - 0.118.

The four elements of statistical analysis may be illustrated with the above data from a clinical trial on the side-effects of ACE-inhibitors. In this trial 135 diabetic patients with nephropathy were treated for one year with an ACE-inhibitor, and it was recorded in each patient whether or not specific side-effects occurred, and in this case we looked at episodes of dry cough. Hence, we observed a dichotomous variable 'dry cough' with outcomes 'yes' or 'no' for all of the 135 patients. Dry cough was observed in 10 patients, and the event rate was therefore 10/135 = 0.074 or 7.4%, and the 95% confidence interval was calculated to be between 0.030 and 0.118. The proportion or percentage is often denoted by "p".

7. EXAMPLE 1: QUANTIFICATION

p = proportion or percentage. When n is large, p is normally distributed
(binomial distribution becomes normal distribution).
with mean p (the true probability of side-effects) and a variance of
$SE^2(p) = p \times (1-p) / n$ we can calculate the $(1-\alpha)\%$ CI.

8. EXAMPLE 1: CONFIDENCE INTERVAL

- If α = significance level= 0.05, then (1-α)% CI= 95% CI,
- 95% CI $= p \pm z_\alpha \sqrt{\dfrac{p(1-p)}{n}}$,
- 95% CI $= 0.074 \pm 1.96 \sqrt{\dfrac{0.074(1-0.074)}{135}}$.

With the standard error of p, SE(p), it is easy to calculate confidence intervals of p. The formulas are not much different from those used for the means of quatitative data as given in chapter 1: add or subtract from p the standard error multiplied by z_α. This latter factor is the ordinate of the standard normal distribution associated with α, and determines the width of the confidence interval. Usually 95% confidence interval are used where z_α equals 1.96.

The normal approximation of the binomial distribution works fine when the number of observations is large, say 30 or more, but less so when n is small (say less than 10). The approximation is less good also when p is very small or very large, and it may easily happen that the confidence interval is larger than 1 or smaller than zero. In these cases, it is better to construct the CI on the basis of the exact binomial distribution (paragraph 11). The calculations required are more complex, but several computer programs are available, for instance CIA (Confidence Interval Analysis) provided by the book of Altman, Practical Statistics for Medical Research, Chapman & Hall, London, UK, 1991.

9. EXAMPLE 1: HYPOTHESIS TEST

Review chapter 4, paragraphs 1-6.

Greek letters are often used in statistics to represent population names, Roman lettter to represent sample names.

H_0: $\pi = \pi_0 = 0.10$ H_1: $\pi \neq \pi_0$

test:

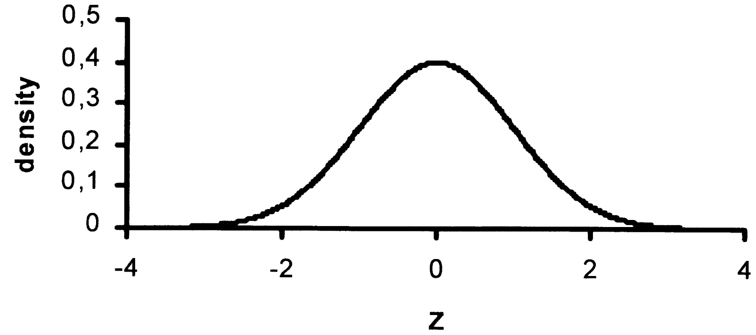
$$Z_o = \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.074-0.10}{\sqrt{0.10(1-0.10)/135}} = -1.01$$

p-value: $P(|Z|>|Z_o|) = 0.31$

In the present example hypothesis-testing is not obvious, but a natural concern would be whether the event-rate of this drug is better or worse than the event-rate of another drug, e.g., another ACE-inhibitor with an established event-rate of 10%. Then the research question might be whether or not the present ACE-inhibitor has lower event-rate. The alternative hypothesis (H_1) is therefore that the true event-rate of the present ACE-inhibitor (π) is unequal to the known event-rate of the other ACE-inhibitor (π_0=0.10). H_1 is defined as $\pi \neq \pi_0 = 0.10$, and the null-hypothesis is its negation, therefore H_0: $\pi = \pi_0 = 0.10$.

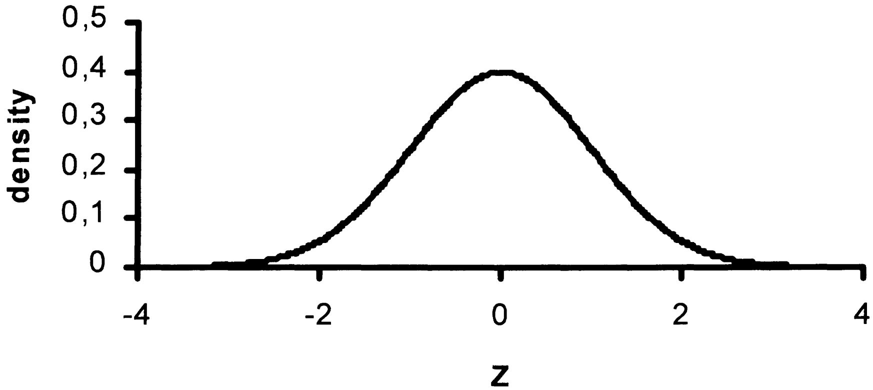
This null-hypothesis can be tested using the Z-test as given. This test-statistic is the ratio of the difference between the observed proportion p and its null-hypothesis value π_0 over the corresponding standard error (if the null-hypothesis is true): here we find the Z=-1.01. If H_0 is true, the test-statistic follows a standard normal distribution. This means that, in principle, every value between minus and plus infinity is possible, but the likelihood of very small or very large z-values is much less than of values close to zero. The basic decision strategy is similar to that discussed in earlier chapters: the null-hypothesis is rejected when the probability of the test-statistic value that was calculated or even more extreme, is small (usually smaller than α=0.05). This probability is denoted as "p-value", and in this case the p-value is 0.31.

10. STANDARD NORMAL DISTRIBUTION



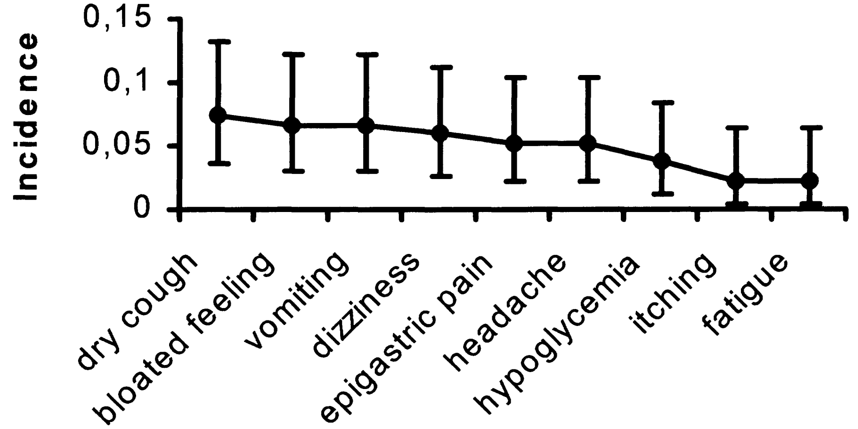
The standard normal distribution is illustrated here. On the x-axis (generally called z-axis in statistics) we have a normal distribution of proportions expressed as SEs (=SEMs) distant from the mean proportion (defined 0). On the y-axis we have"how often" the proportion will occur. The total area under the curve (AUC) represents 100% of all possible proportions.To the right of 1.96, and to the left of –1.96 lies 2.5% of the AUC. The AUC corresponds to the probability of observing that value of the z-statistic, hence the probability of Z>1.96 is 0.025, and similarly the probability of Z<-1.96 is also 0.025. When we add these two probabilities we find that the probability that the absolute value |Z| > 1.96 equals 0.050.

11. STANDARD NORMAL DISTRIBUTION: P-VALUE



The observed z-value was –1.01, and the probability of finding –1.01 or less equals 0.156, as can be found in the table of the standard normal distribution. Since our alternative hypothesis was formulated in a two-sided fashion (H_1: $\pi \neq \pi_0$), the p-value is defined as the probability of $|Z|>1.01$, and therefore the p-value must be doubled: p-value=2 x 0.156 = 0.312. When the alternative hypothesis was formulated –a priori- in a one-sided fashion (H_1: $\pi < \pi_0$), then the p-value would not have to be doubled.

12. EXAMPLE 1: GRAPHICAL ILLUSTRATION



A nice graphical display of dichotomous data is provided by drawing both the estimates of the proportions of side-effects, together with their 95% CIs. In the above graph both point-estimates (7.4% for dry cough) and the uncertainties of these estimates are given (it could as well be as high as 11.8% or as low as 3%). In the display the proportions of other side-effects are readily shown as well. Notice that the event-rates or incidences of these side-effects are very small, and, therefore, we used the exact approach to calculate confidence intervals. This approach* may give asymmetric intervals (never exceeding zero or one) in contrast to intervals based on the normal approximation which are always symmetric around p.

* The chance that x<k can be calculated from the binomial formula $n!k!(n\text{-}k)!(p)^k\ (1\text{-}p)^{n\text{-}k}$ where x=number of patients with side effects, n= total number of patients, p= (n-k)/n, and ! indicates faculty, e.g., 5!=5x4x3x2x1.

13. EXAMPLE 2: COMPARING TWO GROUPS OF PATIENTS

progression in the so-called REGRESS Study

- 884 patients with proven coronary artery disease (CAD),
- randomized between placebo (n=337) and pravastatin (n=322),
- after two year patients showed progression of CAD (angiographically or having had events (death, infarction, coronary intervention, stroke)).

Slightly more complex, but also more interesting, methods are needed to compare two groups of different patients with respect to a discrete variable. We introduce as an example a clinical trial on patients with proven CAD randomized to either placebo (n=337) or a statin (n=322), and of each patient we recorded after two years of treatment whether disease progressed. Progression was defined as the occurence of an event (death, infarction, stroke, coronary interventions) or as a decrease of the diameter of the coronary arteries which was measured by angiography. Thus, there are two groups of patients, measured once on a discrete variable: yes/no progression after two years. According to the overview given in the next paragraph we will discuss the chi-squared test.

14. EXAMPLE 2: DATA

Effect quantification:

Placebo:	progression 220/337 = p_1 = 0.653
Pravastatin:	progression 168/322 = p_2 = 0.522

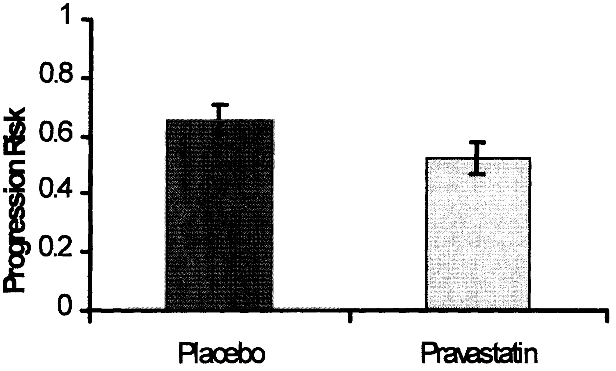
95% Confidence Intervals for the progression risks:

Placebo:	(0.602 - 0.704)
Pravastatin:	(0.467 - 0.576)

Progression was found in 220 out of 337 placebo-patient, hence the progression-rate was p_1=220/337=0.653, and in the statin-group it was p_2=168/332=0.522. Using the formulas given in paragraph 7 we can immediately calculated the corresponding 95% confidence intervals, and these vary between 0.602 and 0.704 in the placebo group, and between 0.467 and 0.576 in the statin-group.

The research question is whether statin-treatment is effective, or in other words whether the progression-rate is smaller under statin-treatment than under placebo-treatment. The two confidence intervals do not overlap, and this fact is strong evidence for a significant difference, but formally a somewhat different approach is needed to answer the research question.

15. EXAMPLE 2: GRAPHICAL ILLUSTRATION



The research question involves the difference in progression-rate between the two treatment groups, and it is very natural to display the two progression-rates to answer that question. Of course, it is necessary to indicate the certainty of the (estimated) progression-rates, and therefore the confidence intervals are indicated with the small black error-bars.

16. EXAMPLE 2: RISK DIFFERENCE

Alternative Quantification:
Risk Difference
d = p_1 - p_2 = 0.653 - 0.522 = 0.131

$$SE(d) = \sqrt{SE^2(p_1) + SE^2(p_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.653(1-0.653)}{337} + \frac{0.522(1-0.522)}{322}} = 0.0380$$

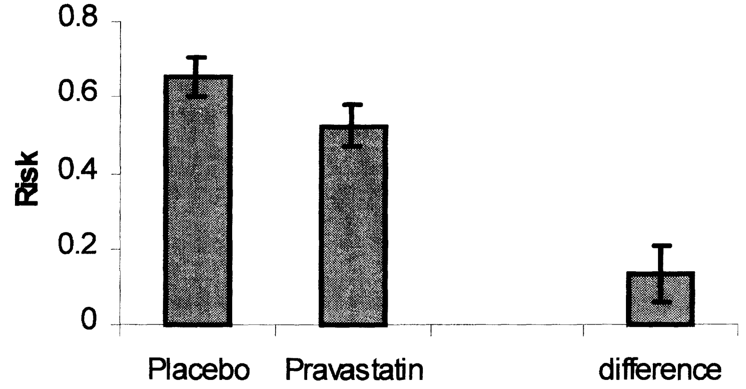
For the comparison of two rates several effect-quantifications are possible, but the difference of the two rates of risks is straightforward. The effect of statin-treatment is quantified then as a decrease of the progression-risk by d = p_1-p_2 = 0.653 – 0.522 = 0.131 .

The uncertainty of this estimate, expressed as SE of d, is easily calculated as the square root of the sum of the squared standard errors of p_1 and p_2 which was 0.038 in this case. Again this standard error is based on the normal approximation to the binomial distribution.

17. EXAMPLE 2: CONFIDENCE INTERVAL

95% CI in the above example is given by
$d \pm 1.96$ SE(d) $= 0.131 \pm 1.96 \times 0.038 =$
interval is between 0.0565 and 0.2055.
A confidence interval for d is, thus, calculated in similar fashion as before: subtract from or add to d the product of z_α x SE(d), and for a 95% CI z_α equals 1.96 (compare page 50). Thus $d \pm 1.96 \times 0.038$ yields an interval from 0.0565 to 0.2055. In other words, the risk-reduction of disease-progression by statin-treatment is estimated to be 0.131 with confidence interval (0.0565 – 0.2055).

18. EXAMPLE 2: GRAPHICAL ILLUSTRATION



The graphical illustration of the risk-difference can be done on the same scale as is used for the risks themselves, and the uncertainty of the estimate can be indicated with a confidence interval. Notice that the risk-difference varies between –1 and +1 (whereas the risks themselves vary between 0 and +1), and the confidence interval may well be less than zero. In fact, the observation that zero lies inside or outside the confidence interval is the basis of the statistical hypothesis test to be discussed next: when zero is outside of the confidence interval of d, d is significantly different from zero.

19. EXAMPLE 2: HYPOTHESIS TESTING

hypothesis test

H_0: $\pi_1 = \pi_2$ H_1: $\pi_1 \neq \pi_2$

test:

$$Z_o = \frac{d}{SE(d)} = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0.653-0.522}{\sqrt{\frac{0.653(1-0.653)}{337} + \frac{0.522(1-0.522)}{322}}} = 3.44$$

Z is standard normally distributed:

p-value: $P(|Z|>Z_o) = 0.001$.

The relevant research question is whether the progression rates differ in the two treatment groups. The alternative hypothesis is therefore that the true progression rates, denoted by π_1 and π_2, are different: H_1: $\pi_1 \neq \pi_2$. This alternative hypothesis is formulated in a two-sided fashion, but one could argue that only an improvement of the progression rate is of interest leading to a one-sided hypothesis: H_1: $\pi_1 > \pi_2$. Whichever method chosen, the null-hypothesis is the negation: H_0:$\pi_1 = \pi_2$.

This null-hypothesis can be tested with a generalization of the Z-test discussed in paragraph 9: the Z-statistic equals the ratio of d versus its standard error SE(d), and yields 3.44 here. Again, when the null-hypothesis is true, Z follows a standard normal distribution meaning that all values between minus and plus infinity are possible, but the likelihood of extremely low or high values is small. When the probability of finding Z (i.e. 3.44) or a more extreme value is as small as 5% (α = 0.05), we decide that the difference is statistically significant. In our example the two-sided p-value is 0.001, much less than 0.05, and we therefore may conclude that the progression risk under statin-treatment is much less than under placebo-treatment.

20. EXAMPLE 2: 2-BY-2 TABLE

Alternative Presentation (review chapter 4, paragraphs 15, 16):

Observed 2-by-2 table

	Placebo	**Pravastatin**	**total**
Progression	220 (65.3%)	168 (52.2%)	388 (58.9%)
no	117 (34.7%)	154 (47.8%)	271 (41.1%)
total	337	322	659

Relative Risk (RR) = 0.653/0.522 = 1.25 (95% CI: 1.10-1.43),

Odds Ratio (OR) =(0.653x0.478)/(0.522x0.347) = 1.72,

(95 CI: 1.26 - 2.36).

An alternative way of presenting the progression data is the two-by-two table. In such a table the different groups are represented by different columns and the different outcomes by different rows. Again 220 out of 337 patients showed progression in the placebo-group(65.3%), whereas 168 out of 322 (52.2%) showed progression in the statin-group. A nice feature of this presentation is that row-wise interpretation is also possible: 220 out of 388 patients with progression was treated with placebo in contrast to 117 of 271 without progression.

Often the effect of placebo treatment is quantified by the relative risk instead of the risk-difference. The relative risk (RR) is equal to the risk of progression under placebo treatment (65.3%) divided by the risk of treatment under statin treatment (52.2%): RR = 65.3/52.2 = 1.25. Thus, the risk of progression is 25% larger when treated with placebo compared to treated with statin.

The relative risk varies from zero to infinity, and is not defined when the denominator equals zero. The uncertainty is usually defined using the natural logarithm (ln) of RR, instead of RR itself: the standard error of ln(RR) equals

$SE(\ln RR) = \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$ and in this case we found SE(lnRR)=0.0665.

Subsequently, the 95% confidence interval is defined as lnRR±1.96xSE(lnRR), and we find ln(1.25) ± (1.96 x 0.0665): 0.0928 to 0.3535. By taking the antiln transformation we find the 95% CI of RR itself: $e^{0.0928}$ to $e^{0.3535}$ (1.10 – 1.43).

Another effect-quantification that is used in many studies is the odds ratio (OR). Instead of calculating the ratio of the progression-risks, the ratio of the odds is calculated:

	Risk	odds
placebo group	p_1=0.653 (a)	$p_1/(1-p_1)$=1.88 (b)
statin group	p_2=0.522 (c)	$p_2/(1-p_2)$=1.09 (d)
ratio	1.25	1.72

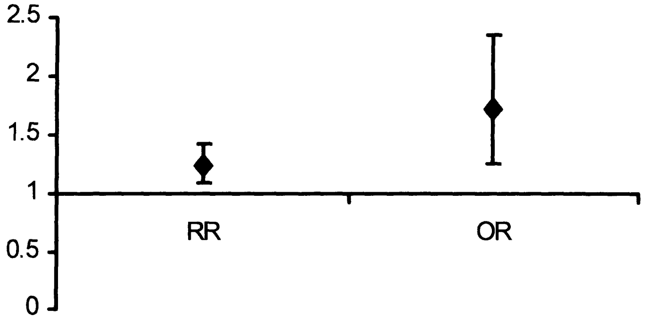
Similarly to the relative risk, the standard error of natural logarithm of the odds ratio is calculated. Define in the above 2-by-2 table the entries in the four cells as a, b, c, and d, then the standard error of ln(OR) is given by

$$SE(\ln OR) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{220}+\frac{1}{168}+\frac{1}{117}+\frac{1}{154}} = 0.1598^{\cdot}$$

Again the 95% CI of ln(OR) is calculated as: lnOR±1.96xSE(lnOR), and the 95% CI of OR is found by taking antiln of the interval of lnOR.

Both intervals are based on the normal approximation, which is adequate only when sample sizes are sufficient. Exact methods are available for small samples.

21. EXAMPLE 2: GRAPHICAL ILLUSTRATION OF RR AND OR



The graphical display of the estimated RR and OR is similar to the display of the risks or the risk difference, except for the scales that vary from zero to infinity. When there is no effect of statin treatment, the RR and the OR equal 1.0. Positive effect means that RR or OR is less than 1.0, but this rule depends on the specific definitions of RR and OR. When the CI does not contain 1.0, the risks or odds are significantly different from 1.0.

OR is generally larger than RR, but the CI too. When the risks (p_1 and p_2) are low, the difference between RR and OR become smaller and that is the reason why OR is sometimes called an approximation of the RR. There are no definite arguments to prefer the RR above the OR, but in general the RR seems clinically more useful than the OR. The RR can be calculated in cohort studies only, not in case-controls studies, whereas the OR can be calculated in both. Since most regression models for discrete data use odds instead of risks, the odds ratio is encountered very often in biomedical and epidemiological literature.

22. EXAMPLE 2: HYPOTHESIS TESTING

Expected 2-by-2 table
if H_0 is correct

	Placebo	Pravastatin	total
Progression	0.589x337	0.589x322	388 (58.9%)
no Progression	0.411x337	0.411x322	271 (41.1%)
total	337	322	659

	Placebo	Pravastatin	total
Progression	198.5	189.5	388 (58.9%)
no Progression	138.5	132.5	271 (41.1%)
total	337	322	659

The null-hypothesis of no difference can be tested with an alternative hypothesis test, the so-called chi-squared test. This test is statistically equivalent to the Z-test discussed above, but the chi-squared test can be generalized to discrete data with more than two classes, and also to research-designs involving three or more patient-groups.

The idea of the chi-squared test is based on a comparison of the observed and expected 2-by-2 table. When the null-hypothesis is true (H_0: $\pi_1 = \pi_2$), the total number of patients with progression (388, 58.9%) will be distributed over the two groups proportionally to the two sample sizes. Thus the expected number of patients with progression in the placebo group will be 0.589x337=198.5, and the expected number of patients with progression in the statin-group will be 0.589x322=189.5. Similarly, the expected number of patients without progression in both groups can be calculated.

23. EXAMPLE 2: CHI-SQUARED TEST

Review chapter 4, paragraphs 7-11.

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 11.83$$

X^2 is chi-squared (χ^2)-distributed with one degree of freedom.

For 2-by-2 tables X^2 equals Z^2 (compare page 177).

The chi-squared test is also applicable in RxC tables when 3 or more (=R) samples are compared for 3 or more (=C) variables.

When the null-hypothesis of no difference between the treatment groups is true, the observed and expected tables should be identical. Obviously, there will be small differences and the larger the difference the more indication for a real existing risk-difference.

The extent of difference between the observed and expected tables is quantified by the test statistic, which is a sum over the four cells of the 2-by-2 table of the squared difference of the observed and expected numbers divided by the expected cell number. When the null-hypothesis is true, the test-statistic is chi-squared - distributed with 1 degree of freedom. The chi-squared distribution is more complex than the normal distribution, but when there is one degree of freedom Z^2 is equal to X^2. The two-sided p-value of the chi-squared distribution is therefore the same as the p-value of the Z-test.

The advantage of the chi-squared test is its applicability to comparing R groups on a discrete variable with C levels. When there is no difference between the R groups, the corresponding chi-squared test X^2 is chi-squared distributed with (R-1)x(C-1) degrees of freedom. This test is often called the Pearson Chi-squared test.

24. EXAMPLE 2: FISHER EXACT TEST

Review chapter 4, paragraph 12.

	Placebo	**Pravastatin**	**total**
Progression	a	b	388 (58.9%)
no Progression	c	d	271 (41.1%)
total	337	322	659

The Fisher-exact test follows a socalled hypergeometric distribution:

$$P(a=220)=\frac{\binom{337}{220}\binom{322}{388-220}}{\binom{659}{388}}$$

The chi-squared test is valid only with large samples (the expected numbers must be 5 or larger). In small samples an exact test is needed, and this is Fisher's exact test. It is based on the hypergeometric distribution of the cell counts given the marginal counts. The one-sided p-value is calculated as the probability that cell count a equals the observed value or less: in our case Pr(a≤220).

$$\Pr(a \le 220) = \Pr(a = 0) + \Pr(a = 1) + \ldots + \Pr(a = 220)$$

The hypergeometric distribution is a complex function of binomial coefficients $\binom{n}{k}$ and is laborious to calculate when the cell counts are large, but it provides an exact p-value.

25. SAMPLE SIZE CALCULATION

Sample size for a two-group clinical trial.
Suppose standard treatment has efficacy p_1, say $p_1 = 0.5$,
and the new treatment increases this to p_2, say $p_2 = 0.6$.

How many patients must be included in the two samples to achieve 80% / 90% power for this expectation?

Until now we have discussed the statistical analysis of data, but in the design of trials an important question, for many reasons, is how many patients should be included in the trial. Suppose that a trial is to be designed comparing two treatments with progression rates p_1 and p_2, say p_1=0.5 and p_2=0.6. How many patients must be sampled to have 80% or 90% power to find this assumed difference to be significant? There is a straightforward formula for this question ....

$$N = (z_\alpha + z_\beta)^2 \frac{p_1(1-p_1)+p_2(1-p_2)}{(p_1-p_2)^2}$$

α=0.05: z_α = 1.96

β = 0.10: z_β = 1.28

β = 0.20: z_β = 0.84

$$N = (1.96+0.84)^2 \frac{0.5(1-0.5)+0.6(1-0.6)}{(0.5-0.6)^2} = 384 / group.!$$

In this equation "N" is the required number of patients in each group. z_α is the standard normal ordinate associated with significance level α; the significance level is almost always 0.05, in which case z_α is 1.96 (compare page 50). z_β is the standard normal ordinate associated with (1-β)% power. When power is 80%, z_β is 0.8, and when power is 90% z_β equals 1.3 (compare page 50).

In our case 384 patients per group (768 in total) are needed.

In many cases an old therapy with known event-rate (p_1) will be compared to a new therapy of which the event-rate is unknown. Instead of specifying p_2 a clinically relevant difference is considered (p_1-p_2).

26. DISCUSSED SO FAR....

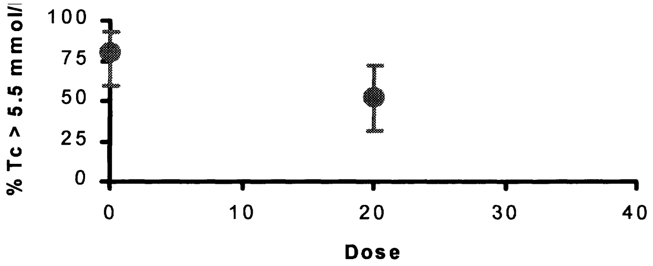
	1 sample **1 measurement**	**2 samples** **1 measurement**	**>2 samples** **1 measurement**
Discrete	Z- or chi-squared test	Z- or chi-squared test	chi-squared test
Censored	(kaplan-meier curve)	logrank test	logrank test
Quantitative	one sample t-test/ Wilcoxon test	unpaired t-test / Mann-Whitney test	ANOVA, Kruskal-Wallis test

Discussed so far were the statistical methods to analyze discrete data from experiments involving one, two, or more groups of patients measured once. Next we will discuss methods for discrete data from one group of patients measured twice.

27. EXAMPLE 3: 1 SAMPLE, TWO MEASUREMENTS

1 sample, two measurements,
dose-finding.

- 25 patients with familial hyperlipidemia,
- response to 2 months treatment with 0 or 20 mg simvastatin: < 5.5 mmol/L?



In clinical pharmacology a typical example of a study in which the same patients are measured twice, is a dose-finding study. This phase-2 type of research is done to find the optimal dose of a new drug, and involves almost always finding the balance between efficacy and side-effects. Here we consider a study in 25 patients with familial hyperlipidemia. Their total cholesterol (Tc) was measured before and after 2 months of treatment with 0 or 20 mg of a specific statin. Tc was dichotomized into high or low (<5.5 mmol/L), and the latter observation was defined as a response.

The graphical display show that before treatment almost 80% of the patients had high Tc, whereas after treatment about 50% of the patients had high Tc. The small bars represent the 95% CIs. The quantification of the effect (i.e., the proportion of patient with Tc<5.5 mmol/L) is largely similar to that of the above example where we compared two groups of different patients. But there is a difference. In our case Tc of all patients was measured twice. Data were dichotomized. The fact that we have repeated measurements complicates the statistical analysis because of the likely correlation of the data before and after treatment. This means that the standard chi-squared test approach cannot be used because for that purpose it is required that all data are independent of each other. Here we have paired data which need not be independent, and the correlation must be taken into consideration. This is done by using McNemar's test.

28. McNEMAR′S TEST

Hypothesis test according to McNemar (review chapter 4, paragrpah 14).
H_0: $\pi_{0\ \mathrm{mg}} = \pi_{20\ \mathrm{mg}}$

		Tc after dose 20 mg	
		< 5.5 mmol/L	**> 5.5 mmol/L**
Tc after dose 0	**< 5.5 mmol/L**	**a**	**b**
	> 5.5 mmol/L	**c**	**d**

$$X^2 = \frac{(b - c)^2}{b + c}$$

X^2 is chi-squared distributed with one degree of freedom.
McNemar′s test can be readily performed using SPSS statistical software (command: “Exact” test).

Again we can present the data in a 2-by-2 table as is shown here, but the four cells have different entries. The left upper cell (indicated by “a”) contains the patients whose Tc was low on both occasions, and the cell right-below contains the patients whose Tc was high on both occasions (“d”). In these patients the discrete variable (yes/no Tc<5.5 mmol/L) did not change, and one can argue that therapy was ineffective or unnecessary in these patients. As a consequence information on treatment-efficacy can only be obtained from patients whose Tc-status changed: the upper-right cell contains the patients whose Tc deteriorated, and the left-below cell the patients who improved. If the null-hypothesis of no treatment-effect is true, then the proportion of patients with low Tc will be equal before and after therapy: H_0:$\pi_{0\ \mathrm{mg}}$= $\pi_{20\ \mathrm{mg}}$. And also if H_0 is true, then every change of Tc-status is affected by chance alone, and one may expect as much patients who deteriorated as those who improved.

This expectation is tested by McNemar’s test. This test-statistic is the ratio of the squared difference $(b–c)^2$ divided by their sum b+c. If H_0 is true, b and c are expected to be the same, and hence the test-statistic will be close to zero. The test-statistic follows a chi-squared distribution with one degree of freedom if H_0 is true and if the sample is sufficiently large, and on this basis the p-value is calculated.
If the sample size is small, and exact p-value can also be calculated.

29. EXAMPLE 3: McNEMAR'S TEST

		Tc after dose 20 mg		
		< 5.5 mmol/L	> 5.5 mmol/L	total
Tc after dose 0	**< 5.5 mmol/L**	**4**	**1**	**5**
	> 5.5 mmol/L	**8**	**12**	**20**
	total	**12**	**13**	**25**

$$X^2 = \frac{(b-c)^2}{b+c} = \frac{(8-1)^2}{8+1} = 5.44$$

$\alpha = 0.05$; $\chi^2_{\alpha} = 3.841$ => P-value < 0.05

thus reject H_0.

In our case we found that in 8 patients Tc decreased to below 5.5 mmol/L, and that Tc increased to above 5.5 mmol/L in 1 patient. McNemar' test gives value $(8-1)^2/(8+1)=5.44$, with p-value 0.020 (exact p-value is 0.039).
The effect of statin-treatment must be quantified with the difference of the two response-proportions: $d=p_2-p_1 = (12/25) - (5/25) = 0.48 - 0.20=0.28$. The standard error of the difference is given by

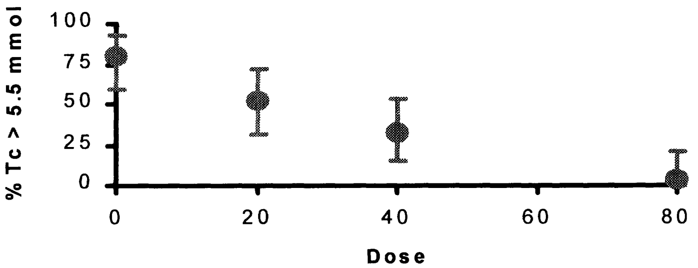
$$SE(d) = \sqrt{\frac{b+c-n\times(p_2-p_1)}{n(n-1)}} = \sqrt{\frac{8+1-25\times 0.28}{25\times 24}} = 0.058$$

and the 95% CI by d ± (1.96 x SE(d)) : 0.166 – 0.394.

30. EXAMPLE 4: >2 REPEATED MEASUREMENTS

1 sample, four measurements,
dose-finding:

- 25 patients with familial hyperlipidemia,
- response to 2 month treatment with 0 or 20 mg simvastatine: < 5.5 mmol/L?



These statistical techniques can be easily generalized to research design with 3 or more repeated measurements. In dose-finding studies often many doses are tested, and in our example the response proportions of two additional doses were found to be about 30% for 40 mg, and about 1% for 80 mg.

Although data are pretty clear here, one might want to test the overall null-hypothesis of no effect by testing each dose-pair individually using the McNemar's test: this involves 6 tests here. A better approach is to perform one single overall statistical test which compares all four proportions all at once with the null-hypothesis H_0:$\pi_{0\ mg}$ =$\pi_{20\ mg}$ =$\pi_{40\ mg}$ =$\pi_{80\ mg}$.

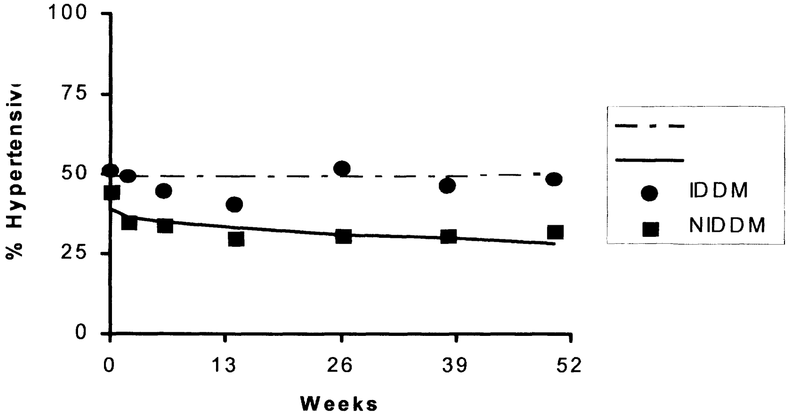
31. EXAMPLE 4: COCHRAN'S TEST

Hypothesis test:
H_0: $\pi_{0\ mg} = \pi_{20\ mg} = \pi_{40\ mg} = \pi_{80\ mg}$,
test: Cochran's Q,
Q = 14.53 is chi-squared distributed with k degrees of freedom: $p<0.001$.

Cochran's test may be used to compare k (2 or more) repeated measurements of a dichotomous variable. When k is 2 Cochran's test is equal to McNemar's test. Cochran's test-statistic is somewhat complex to calculate but all major statistical computer programs have it available. When the null-hypothesis is true, it follows a chi-squared distribution with k-1 degrees of freedom.

32. OTHER REPEATED MEASUREMENTS DESIGNS

>1 sample,
>1 measurements.



IDDM=insulin dependent diabetics; NIDDM= non-insulin dependent diabetics

We discussed repeated measurements only in case of one group of patients measured repeatedly. This is the most simple case, and very often one has several different patient subgroups measured repeatedly. Especially when patients are followed in time repeated (discrete) data will occur. Below an example is given of NIDDM and IDDM patients who were followed weekly for one year, and their hypertension status was recorded. The research question is whether the proportion of patients with hypertension changes, and whether the changes are different for NIDDM and IDDM patients. Such research designs require specialized statistical techniques ....

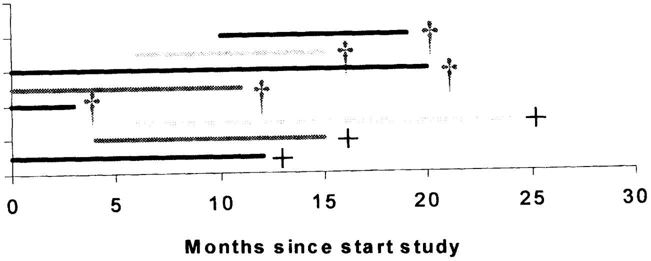
33. OTHER REPEATED MEASUREMENT DESIGNS: SPECIAL TECHNIQUES

- Marginal logistic regression models,
- random effect logistic regression models,
- complex mathematics but now widely available in dedicated and general purpose software: e.g. S-plus or SAS.

These techniques are based on the logistic regression model which will be discussed later, and use two different methods to account for the repeated measurements: either by using a population-average model, or by using a random-effect model. The mathematics, and statistical reasoning is complex, but software is widely available.

34. CENSORED DATA

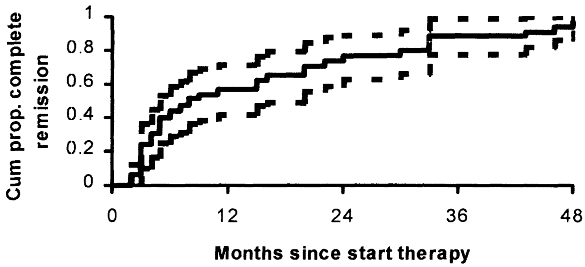
- Incompletely observed 'failure' times,
- occur everywhere in biomedicine, most notably in cancer research.



A special case of discrete data occurs when patients are followed until some event occurs. This kind of data arises often in follow-up trials. Complications happen because the follow-up periods of patients differ, for instance due to the fact that patients enter into the trial at different times, and also because events occur at different time-points in different patients. In some patients the event of interest did not occur, but this should be interpreted in terms of "did not yet occur". Such observations are called censored observations, and should be handled carefully, and the statistical techniques to do this are known as survival analysis methods.

35. KAPLAN-MEIER CURVE

Quantification (review chapter 4, paragraphs 17, 18):
Kaplan-Meier Curves



In a follow-up study of one group of patients the occurence of events is usually summarized with the Kaplan-Meier curve. For each individual one needs the time until the event of interest occured or the time until the date of the last follow-up visit. It requires also the discrete variable "yes/no, event occured", and a common zero time-point. In clinical trials this zero-time point may be the date when treatment commenced. Important assumption is that censoring occurs at random. This is difficult to ascertain, but censoring is usually random when it occurs because the study is randomized and blinded. When patients in whom events are imminent, are likely to end their participation to the study (causing censored observations), then these censored data are not random. In the latter case statistical analysis is much more complicated, and here we assume that censoring is random.

The Kaplan-Meier curve is a nonparametric estimate of the cumulative event-free survival distribution with characteristically rectangular shape. At each time-point it gives the cumulative proportion of patients with an event. Therefore the curve starts at zero at time-point zero, and will increase at each time-point where one or more patients had an event, thus each jump in the curve represents one or more patients with events.

Often the Kaplan-Meier is given in inverse fashion, starting at 1 or 100% at time-point zero, representing the cumulative proportion patients without events, hence the name survival function. Like all other statistics, the Kaplan-Meier curve is an estimate and its uncertainty is illustrated with confidence intervals.

A nice feature of the curve is that the median (and any other percentile point) time-to-event can be read immediately from the graph: at the y-axis choose the value 0.5, take a straight horizontal line, and where it intersects the curve, go straight down to the x-axis. In our case the median time to remission (of depression) is about 8 months after start of treatment.

36. KAPLAN-MEIER CURVE: DEFINITION

At time t:

$$S(t) = S(t-1)(1 - \frac{\# events}{\# atrisk})$$

The value of the Kaplan-Meier curve at time-point t, denoted S(t), is given as $S(t) = S(t-1) \times \left(1 - \frac{\# events}{\# at\ risk}\right)$, where “#events” denotes the number of patients with events at time-point t, and “#at risk” denotes the number of patients that are still under observation at time-point t. This definition leads to the possibility that the curve may go down to zero even if not all patients had events. A small example may illustrate the calculations.

Suppose one has 8 patients with the following time-to-event times: 1, 2, 3^+, 4, 6, 6, 6^+, 8. Six patients had events at time-points 1, 2, 4, 6, 6, and 8, whereas two patients were followed for 3 and 6 months but did not have events (denoted as 3^+, and 6^+).

Time	at risk	events	S(t)
0	8	-	1
1	8	1	1x(1-1/8)=0.875
2	7	1	0.875x(1-1/7)=0.75
3^+	6	0	0.75
4	5	1	0.75x(1-1/5)=0.6
6	4	2	0.6x(1-2/4)=0.3
8	1	1	0.3x(1-1/1)=0

At t=0 the survival is 100% naturally, or the proportion alive is 1. At t=1 there are 8 patients at risk, and 1 patient died, thus survival is 1-1/8=0.875=S(1). At t=2 there are only seven patients left (one died at t=1), and 1 of these seven died, thus survival equals (1-1/7) times the preceding survival (S(1)=0.875) giving S(2)=0.75. At t=3 there were 6 patients left, one patient was censored, but none died, thus S(3) is the same as S(2). At t=6 there were still 4 patients left, two died, and one was censored, thus S(6)=0.3. Finally, at t=8 there was only one patient at risk, and he/she died, thus S(8) equalled zero.

This little example illustrates that the survival curve may become zero, even in a sample where some patients did not have the event of interest. This happens always when the patient with the longest follow-up did have an event. It illustrates also that interpretation of survival curves must be done carefully: S(t) is the proportion patients *who were followed at least until time-point t*, that were still event free.

37. HAZARD, CUMULATIVE HAZARD, SURVIVAL CURVE

Alternative quantification:

- Hazard at time t = risk of event at time t given that no event occurred up to t = λ (t),
- cumulative hazard up to t =

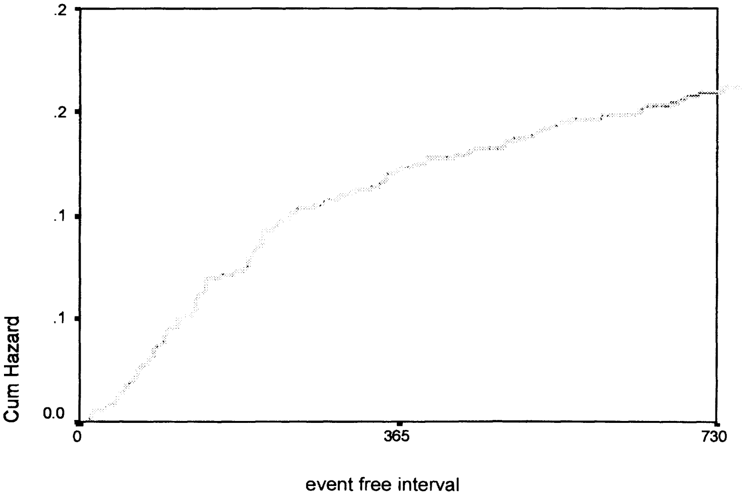
$$\lambda\ (t) = \lambda\ (1) + \lambda\ (2) + \ldots + \lambda\ (t)$$

$$\lambda\ (t) = {}_0\int^t \lambda\ (u)\ du$$

- S(t) = exp(- λ (t)) or λ (t) =-ln(S(t)).

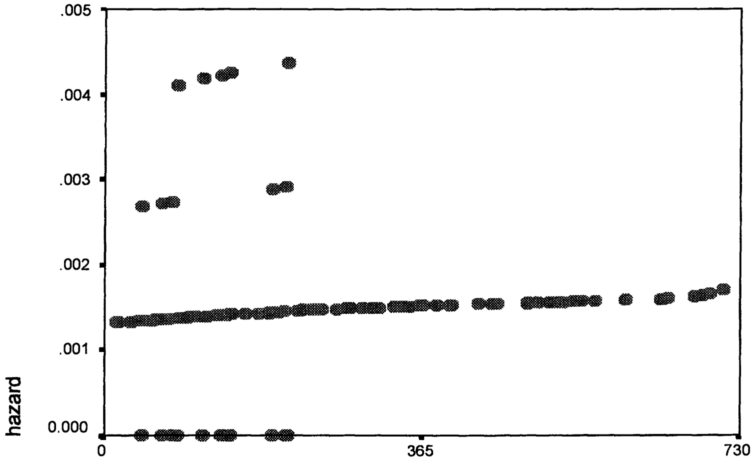
The Kaplan-Meier survival curve is found in numerous epidemiological papers. Related statistical quantities that are also often used, are the hazard at time t, and the cumulative hazard at time t. The hazard at time t, denoted as λ(t), is the proportion patients who were event-free until time t, and had the event at time t: basically it is the ratio (#events) divided (# at risk) as was calculated in the preceding slide in the computation of the Kaplan-Meier curve. The cumulative hazard up to t, denoted as Λ(t) is the summation of all hazards until t: Λ(t)= λ(1)+ λ(2)+ ...+ λ(t), and when there are many time-points this specializes to the integral $\Lambda(t) = \int_0^t \lambda(u)\, du$. The survival can also be estimated as the exponent of minus Λ(t). This quantification is used in regression models (chapter 10).

38. CUMULATIVE HAZARD FUNCTION: EXAMPLE



The cumulative hazard function looks very much like a survival function: it starts at zero, and is rectangularly shaped like the Kaplan-Meier curve.

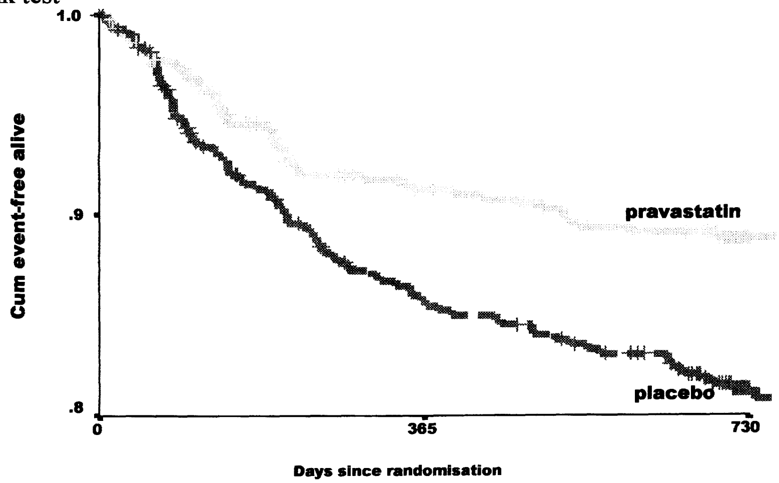
39. HAZARD FUNCTION: EXAMPLE



The hazard function is very jumpy with zero values at all time-points where patients were censored. This means that the hazard function cannot be used directly, which is disappointing because the concept itself is of direct clinical value as it represents the probability of getting the event at time-point t. When analyzing death, it is called the *instantaneous death rate*.

40. COMPARING SURVIVAL CURVES: LOGRANK TEST

Comparison of Curves:
Logrank test



logrank=9.69, df=1, p=0.0019.

The null-hypothesis of no difference between two or more survival curves of different groups of patients can be tested by the logrank test. This is a generalization of nonparametric tests, discussed in earlier chapters, and the same reasoning applies therefore here: when the null-hypothesis of no difference among k groups is true, the logrank test follows a chi-squared distribution with k-1 degrees of freedom.

41. COMPARING SURVIVAL CURVES: COMMENTS

- When no censored data are present, (a special case of) the logrank test is equal to the Mann-Whitney test(chapter 1, paragraph 22).
- Quantification can best be described using the difference of the median survival time, and not by using the mean survival time.
- Alternatively, the difference can be described using the relative risk, calculated from a censored-regression model, e.g. the Cox proportional regression model (chapter 10, paragraphs 6-8).
- Life table method.

There are several different tests for comparing survival curves to each other, other tests are known as the Breslow statistic, the Tarone-Ware statistic. These tests differ slightly in weighing of early and late events, but in general the logrank test is used. When there are no censored observations, the Breslow statistics equals the Mann-Whitney test discussed earlier.

Due to the existence of censored observations, the quantification of treatment effect is best done with the difference of median survival times, and with means. When the survival curve remains above the 50^{ste} percentile, however, the median does not exist either, and another percentile must be chosen.

An alternative effect-quantification is the relative risk, defined as the ratio of the hazard at time t under treatment B divided by the hazard at time t under treatment A. This effect-quantification is also used very often, especially with regression models for survival data (i.e. the Cox regression model), but it is sensible only when the hazard ratio is more or less the same for all time-points.

42. FINALLY: SOFTWARE

- SAS
- SPSS
- S-plus
- Stata
- Statgraphics
- Excel
- ....

* help function
* statistics coach

Nowadays all these computations are carried out with computer-programs for statistical analyses. Much used programs are SAS (www.sas.com), often in pharmacology, and SPSS (www.spss.com) often used at university research institutes. Both programs have been used for decades, and are improved and expanded continuously. But there are many other packages, such as S-plus (www.splus.com), and Stata (www.stata.com). These programs are quite expensive, but common procedures are also available in microsoft's excel, and in excel-add'ins. There are also many websites offering online statistical analyses for free.

43. QUESTIONS TO CHAPTER 5

In a clinical trial diabetes patients with microalbuminuria were randomized to either placebo (n=100) or treatment with ACE-inhibitors (n=100) during one year. In the placebo-group 15 patients (15%) suffered from severe headaches, and 40 patients (40%) in the ACE-inhibitor group.

1. The best statistical test to compare these two percentages is:
   A. the two-sample Student's t-test,
   B. the chi-squared test for a two-by-two table,
   C. the McNemar test.

2. The 95% confidence-interval for the proportion patients with severe headaches in the ACE-inhibitor group is:
   A. $0.4 \pm 1.645 \sqrt{(0.4\,(1\text{-}0.4) / 40)}$
   B. $0.4 \pm 1.960 \sqrt{(0.4\,(1\text{-}0.4)/40)}$
   C. $0.4 \pm 2.021 \sqrt{(0.4\,(1\text{-}0.4)/40)}$

   A new trial was designed to investigate the occurrence of severe headaches during ACE-inhibition. It was decided to design a randomized controlled parallel-group trial comparing patients on placebo with patients on ACE-inhibitor therapy. It was assumed that severe headaches would occur in 10% of patients on placebo and in 20% of the patients on ACE-inhibition. The significance level was set to 0.05.

3. How many patients must be included in each treatment group to ensure 80% power for this expectation?
   A. 200 per group,
   B. 33 per group,
   C. 16 per group.

   In an retrospective case-control investigation of chronic leukemia in first complete remission, 30 adult patients who had received bone marrow transplantation (BMT: cases) were compared to 30 adult patients who were treated with maintenace chemotherapy (CT: controls). Each BMT-patient was matched to a CT-patient such that their age, gender, disease duration, and prior treatment were comparable. The treatments (BMT and CT) were evaluated by comparing the percentage patients alive after five years. The following data were found

		CT		
		Died	**Alive**	**Total**
BMT	**Died**	12	9	21
	Alive	2	7	9
	Total	14	16	30

4. For comparing between treatments the percentages of the patients alive after five years the best statistical test is
   A. the chi-squared test for a two-by-two table,
   B. the logrank test,
   C. the McNemar test.

All of the above tests are test-statistics that are chi-squared distributed when the null hypothesis (H0) is true. The following value of the test-statistic was found: χ^2 =4.45 with one degree of freedom.

5. The following conclusion may be drawn.
   A. 0.10 < p-value < 0.05; H0 may be rejected,
   B. 0.05 < p-value < 0.025; H0 may be rejected,
   C. 0.05 < p-value < 0.025; H0 may not be rejected.

6. The odds ratio for death of BMT versus CT is
   A. (12 x 7) / (9 x 2),
   B. (21 x 9) / (14 x 16),
   C. this study does not permit the estimation of the odds ratio.

In a clinical trial comparing the efficacy of lipid-lowering treatment using statins in patients with age>70 years, 6000 patients were randomized between placebo or pravastatin treatment during five years. One of the efficacy criteria was the cognitive function after five years of treatment as measured using the minnesota mental state examination (MMSE). This score was dichotomized into a score of 25 or higher or below 25. Scores below 25 are indication for dementia. The following results were found:

	MMSE		
	<25	**>25**	**Total**
Placebo	352	2648	3000
Pravastatin	248	2752	3000
Total	600	5400	6000

7. The 95% confidence interval for the proportion patients with MMSE<25 in the placebo group is
   A. 0.106 - 0.129
   B. 0.108 - 0.127
   C. 0.113 - 0.121

8. The relative risk for MMSE<25 of placebo versus pravastatin equals:
   A. 1.03
   B. 1.42
   C. 1.48

9. The best statistical test for comparing the percentages of patients with MMSE<25 in both treatment groups is
   A. the Mantel-Haenszel test,
   B. the Chi-square test for a two-by-two table,
   C. the McNemar´s test.

The two treatment groups appeared to be inbalanced with respect to age; the mean age of the pravastatin patients was 75.2, and of the placebo patients 78.4. Since age is the most important risk factor for dementia (MMSE<25), it was important to evaluate the effect of statin-treatment adjusted for the age-differences.

10. This is best achieved by
    A. a cox regression analysis,
    B. a linear regression analysis,
    C. a logistic regression analysis.

CHAPTER 6

META-ANALYSIS

1. REVIEW OF THE PAST

Meta-analyses are post-hoc analyses.
What they test, is, however, close to the primary hypotheses.
With the established guidelines, probability statements, may therefore, be as valid as with randomized controlled trials.
JAMA, BMJ, and Lancet published, in 1998, 14, 24, and 7 meta-analyses. All of them met standard Cochrane- guidelines. N Engl J Med had no single meta-analysis in its index (instead, it published 61 review monographies, which is another, maybe, less objective way to communicate state of the art knowledge). Specialists'journals (JACC, Diabet Care, Oncol J, Gastroenterology, Ang, J Neur Neurosurg Psychiatr, J Am Geriat Soc, J Clin Endocrin Metab) published in the same year 8 meta-analyses, 1-3 each, none of them accounting all of the standard guidelines: publication bias was accounted never, heterogeneity only twice, and robustness only twice.

- The method of meta-analysis is increasingly appreciated.
- It is made use of by regulatory bodies.
- Meta-analysis can reduce our boundaries of uncertainty.
- The development of new drugs can benefit from it.

2. A LOT OF MISUNDERSTANDING

Mathematics of meta-analyses is not complicated.
Master basic principles:
Publication bias means that negative trials are not published.
Heterogeneity means that trials are often different because of different methods, and different populations.
In the field of meta-analyses, there is a lot of misunderstanding.

-anecdote 1

Groningen pharmacologists performed a meta-analysis of efficacy of ACE-inhibitors.
They reported that they excluded publication bias by thoroughly searching Medline.
They excluded heterogeneity by strict inclusion criteria.

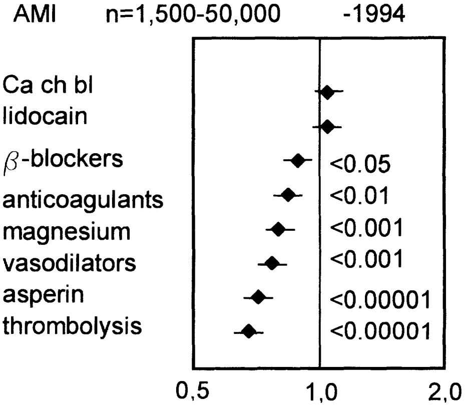
-anecdote 2

Prof vd Broucke from Leyden (Neth) had a major controversy with important international meta-analists after publishing that he believed negative trials were irrelevant. Negative trials are necessary to complement meta-analyses.

3. WHAT IS A META-ANALYSIS?

What is?
A systematic review of trials with pooled data.
How long?
1970, psychologists were the inventors.
First, there were systematic reviews only(no pooling).
Since 1995 more homogeneous trials were published: pooling became a possibility.
How?

4. AN EXAMPLE OF A SUMMARY OF META-ANALYSES HELPFUL TO CARDIOLOGISTS FOR EVERYDAY DECISION-MAKING



The above figure shows the pooled data of many large meta-analyses of optimal treatment of acute myocardial infarction (AMI). On the x-axis are the pooled odd ratios = chances of mortality in users of the compound / chances of mortality in the non-users. <1.0 indicates the compound is efficaceous; >1 not efficaceous. P-values are also given: P=0.05 means 5% chance to find this odds ratio if a odds ratio of 1.0 is true.

5. PROPORTIONS, STANDARD ERRORS OF PROPORTIONS, ODDS, ODDS RATIOS

Odds = likelihood = chance = probability = risk that an event will occur divided by the chance that it won't.

Contingency table	numbers of subjects who died	numbers of subjects who did not die
Test treatment (group1)	a	b
Control treatment (group2)	c	d

The proportion of subjects who died in group1 (or the risk (R) or probability of having an effect)

= p = a / (a+b) , in group 2 p= c / (c+d),

the quotient of a / (a+b) and c / (c+d) is called risk ratio (RR) .

Another approach is the odds approach, where a/b and c/d are odds, and their quotient is odds ratio(OR).

In clinical trials we use ORs as surrogate RRs, because, here, a/a+b is simply nuts.

For example,

	treatment group		control group		whole population
Sleepiness(n)	32	a	4	b	4000
No sleepiness(n)	24	c	52	d	52000

We assume that the control group is just a sample from the population, but its ratio of b/d is that of population.
So suppose 4 = 4000, and 52 = 52000, then $\frac{a/a+b}{c/c+d}$ is close to $\frac{a/b}{c/d}$ = RR of the population.

Like with continuous data we can calculate SDs and SEMs and 95% confidence intervals of rates (or numbers, or scores) and of proportions or percentages.

SD of sample of n = $\sqrt{}$n,
SD of difference between two samples of n_1 and n_2 = $(n_1-n_2)/\sqrt{}\,(n_1+n_2)$,
SD proportion = $\sqrt{}$ [p(1-p)],
SEM proportion = $\sqrt{}$ [p/(1-p)] / $\sqrt{}$n.

We assume, that the distribution of proportions of many samples follows a normal distribution (in this case called the z-distribution) with 95% confidence intervals between

$$p \pm 2 \sqrt{[p/(1-p)]} / \sqrt{n},$$

a formula looking very similar to the 95% CI intervals formula for continuous data,

$$\text{mean} \pm 2 \sqrt{(SD^2/ n)}.$$

Differences and sums of the SDs and SEMs of proportions can be calculated similarly to those of continuous data

$$\text{SEM of differences} = \sqrt{([p1/(1-p_1)] / n_1 + [p_2/(1-p_2)]/ n_2)}$$

$$\text{with 95\% CI intervals : } p_1 - p_2 \pm 2. \text{ SEMs.}$$

The odds approach is not much different from the RR approach, particularly, not with rare diseases. Odds ratios are used in mortality studies, meta-analyses of them, retrospective epidemiological case control studies. RRs in epidemiological cohort studies (common diseases).

6. HOW TO CALCULATE 95%CONFIDENCE INTERVALS OF AN ODDS RATIO

$\text{Ln OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$.

For example

	Hypertension yes	hypertension no
Group 1	a n=5	c n=10
Group 2	b n=10	d n=5

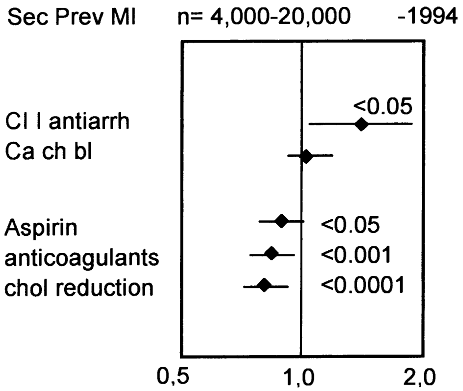
OR=a/c / b/d = 0.25.

95% confidence intervals of ln OR $= \ln \text{OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$
$= \ln 0.25 \pm 1.96 \sqrt{(1/5+1/10+1/10+1/5)}$
$= -1.3863 \pm 1.5182$
is between – 2.905 and 0.132.

95 % confidence interval of OR is between antiln –2.905 and antiln 0.132
is between 0.055 and 1.14.

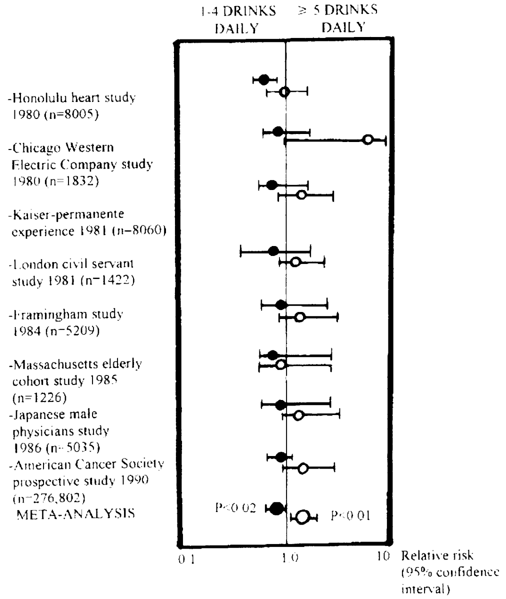
This confidence interval crosses 1.0 and is thus not significantly different from 1.0.

7. ANOTHER EXAMPLE OF A SUMMARY OF META-ANALYSES HELPFUL TO CARDIOLOGISTS FOR EVERY-DAY DECISION MAKING



The above figure shows the pooled data of many large meta-analyses of secundary prevention of myocardial infarction. On the x-axis are the pooled odd ratios = chances of a second myocardial infarction in users of the compound / chances of second infarction the non-users. <1.0 indicates the compound is efficaceous; >1 not efficaceous. P-values are also given: P=0.05 means 5% chance to find this odds ratio if a odds ratio of 1.0 is true.

8. EXAMPLE OF AN EPIDEMIOLOGICAL META-ANALYSIS



Epidemiologists borrowed the technique from clinical pharmacologists, and started to meta-analyze themselves. The above figure shows an epidemiological meta-analysis of cohort studies comparing carriers of a risk factor (alcoholic beverages) versus non-carriers of risk factor. On the x-axis are the RRs instead of ORs, which means here chances of myocardial infarction in drinkers / chances of myocardial infarction in non-drinkers. RR< 1 means that the risk factor protects, >1 risk factor does not protect.

9. IMPORTANT MATTERS NEED FEW WORDS

Important matters need few words. The ten commandments has only 279 words, the American Declaration of Independence has exactly 300. Recent European Community Directives on the import of caramel and caramel products needed 26,921 words. Books have been written on guidelines for meta-analyses. However, meta-analyses are not so complex as selling caramel to the European Community. The only thing to account is the scientific method. Scientific rigor requires that we stick to

(1) a clearly defined prior hypothesis,
(2) to a thorough search of trials,
(3) to strict inclusion criteria for trials, and
(4) to uniform guidelines for data analysis.

10. CLEARLY DEFINED PRIOR HYPOTHESIS

Item (1) of scientific method is a clearly defined prior hypothesis.
Why prior and not posterior hypothesis?
Good research starts with prior hypothesis.
This hypothesis is tested at P=0.05.
The problem posterior hypotheses is that
often they are tested 20x or more times.
Significancies are found by chance.
This procedure is called data dredging, and can be compared with
gambling (e.g., gambling 20 times at 5% chance gives up to 40 % chances of success).

11. THOROUGH SEARCH OF TRIALS

Checklist for Medline Search

Step	search term
Indication	
1	tendinitis.sh.
2	elbow.sh.
3	elbow joint.sh.
4	2 or 3
5	1 and 4
6	tennis elbow.sh.
7	5 or 6
8	epicondylitis.tw.
9	elbow.tw.
10	7 or 8 or 9
Intervention	
11	injections.sh.
12	inject$.tw.
13	infiltr$.tw.
14	exp glucocorticosteroids.sh.
15	triamcinolon$.tw.
16	hydrocortison$.tw
17	methylprednisolon$.tw.
18	betamethason$.tw.
19	lidocain$.tw.
20	bupivacain$.tw.
21	11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20
Publication type	
22	randomized controlled trial.pt.
23	controlled clinical trial.pt.
24	randomized controlled trials.sh.
25	random allocation.sh.
26	double blind method.sh.
27	single blind method.sh.
28	22 or 23 or 24 or 25 oe 26 or 27
29	(animal not (human and animal)).sh.
30	28 or 29
31	clinical trial.pt.
32	exp clinical trials.sh.
33	(clin$ adj25 trial$).tw.
34	((singl$ or doubl$ or tripl$) adj25 (blind$ or mask$)).tw.
35	placebos.sh.
36	placebo$.tw.
37	random$.tw.
38	research design.sh.
39	volunteer$.tw.
40	31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39
41	40 not 29
42	41 not 30
43	30 or 42
Indication, intervention and publication type	
44	10 and 21 and 43

Item (2) of scientific method is a thorough search of trials.
A systematic procedure is required.
It is helpful just like an aircraft-checklist is.
Without it, things go wrong,
e.g., an airplane-door not appropriately locked.
With a checklist we are on the right track in no time, but, MEDLINE still requires 45 steps to be taken.
Start with logging in indication (could be diagnosis group).
Then log in intervention (could be a medicine).
Then take steps to arrive at randomized controlled trials.
Trick have to be learnt: sh means subject headings (main words, e.g., diagnosis groups, medicines etc),
Tw means free text words (connective words, make you communicate with the software),
$ means no money, but word-endings (e.g., injections, able, ing etc),
pt means not patient, but type of publication (parallel, crossover, selfcontrolled etc).
Without checks or tricks, search doesn't make any sense.

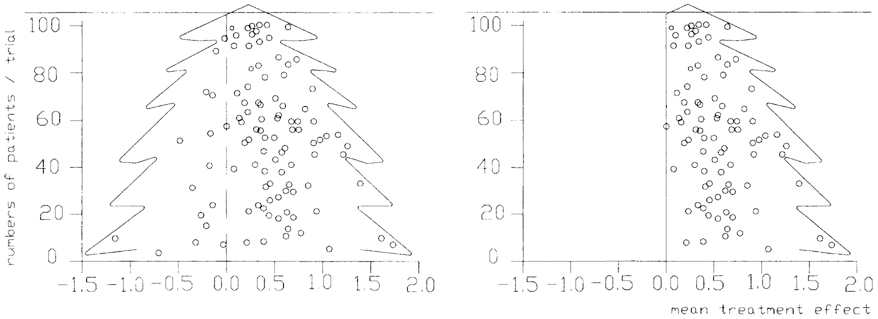
12. STRICT INCLUSION CRITERIA

Item (3) of scientific method is strict inclusion criteria.
It reduces the chance of bias. Bias = systematic error that no one is aware of.
A trial includes less bias if blinded, statistics is proper, ethics is proper, description of methods is proper.

13. UNIFORM GUIDELINES FOR DATA ANALYSIS

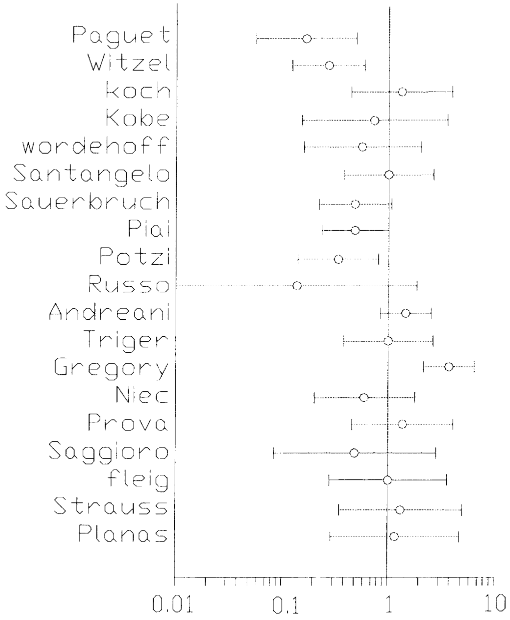
Item (4) of scientific method is uniform guidelines for data analysis.
Here statistics comes in. Professor Hills, the famous statistician from London, UK, once said: investigators use statistics as a drunk uses a lamppost, for support rather than illumination. This is of course inapporopriate use.
Adequate statistics is a powerful aid to prevent erroneous conclusions.
It should not be too complicated, and data dredging for significances should be rejected. Prior hypotheses should be tested rather than posterior hypotheses.

14. DATA ANALYSIS: FIRST PITFALL, PUBLICATION BIAS



First we plot the data with on the x-axis results of studies, and on the y-axis size of the studies. We, thus, get a funnel plot or christmas tree plot. A statistical necessity is that small studies have a large variance, and large studies have a small variance. It is also a statistical necessity that the pattern to be obtained is symmetrical.
A cut-off pattern indicates that a number of studies have not been published. Publication bias means that small studies with a negative result are les likely to be published. The pattern may be obvious from the graph, but in case of doubt can be statistically tested, e.g., by comparing the means of small studies versus the means of large ones).

15. DATA ANALYSIS: SECOND PITFALL, HETEROGENEITY



The second pitfall of meta-analyses is the heterogeneity.
The above example shows on the x-axis odds ratios = chances of fatal esofageal varices bleeding with sclerotherapy/ chances of the same without sclerotherapy. An odds ratio <1 indicates that sclerotherapy is helpful, >1 that it is not so.
We can test heterogeneity by testing whether differences between trials are larger than could happen by chance, and use for that purpose multiple-groups-ANOVA for continuous data, and Chi-square for odds ratios or risk ratios.

16. TESTING HETEROGENEITY

For continuous data multiple-groups-ANOVA can be applied.
Assess whether between-study variance is large compared
to within-study variance.

$SS_{\text{between studies}} = n_1 (\text{ mean}_1 - \text{grand mean})^2 + n_2 (\text{ mean}_2 - \text{grand mean})^2 + \ldots.$
$SS_{\text{within studies}} = (n_1\text{-}1)(SD_1^2) + (n_2\text{-}1)\, SD_2^2 + \ldots..$

$$\text{F value} = \text{ test statistic}= \frac{SS_{\text{ between studies}} / \text{dfs}}{SS_{\text{ within studies}} / \text{dfs}}$$

F-table gives P-value.

For odds ratios multiple groups chi-square test can be applied.
Assess whether variance in ORs is large compared to variance in 95% CIs.
Normalize ORs and 95% CIs of the ORs because CIs around ORs are skewed.
E.g.,OR (95% CI s) = 0.17 (0.02 to 1.55),

ln OR (ln 95% CI s) = -1.8 (-3.9 to 0.4) (nicely symmetric around –1.8)
chi-square value = test statistic=
$\sum (\ln OR_i - \ln \text{pooled } OR)^2 \, (1/1.96 \, . \ln 95\% \, CI_i)^2$ k-1 dfs
chi-square table gives P-value.

CI= confidence interval,
k= number of studies in the meta-analysis.

The above approach is called the fixed model for testing heterogeneity. Random effect model for heterogeneity of Dersimonian and Laird assumes heterogeneity and introduces separate variable. If both are not significant, no heterogeneity can be assumed with confidence, if either of them is positive, pooling will be a problem.

17. HOW TO TEST HETEROGENEITY, CALCULATE AND POOL ODDS RATIOS OF VARIOUS STUDIES AND TO TEST WHETHER POOLED ODDS RATIOS IS DIFFERENT FROM 1.0 , EXAMPLE

For example, 4 studies assessed odds ratios of all cause deaths in patients with heart failure treated with beta-blockers. In order to meta-analyze these studies, first calculate from 95% CIs s=standard error.
s= (ln upper value minus ln lower value)/1.96.
For example:
with a 95% CI of 0.97-1.43
s= (0.3576 minus –0.0305)/1.96 = 0.1980,
then s^2 = 0.0393,
then $1/s^2$ = 25.510.

	OR	95% CI	lnOR	$1/s^2$	$\ln OR/s^2$	$(\ln OR)^2/s^2$
Waagstein	1.18	0.97-1.43	0.16	25.510	4.08	0.653
Packer	0.41	0.39-0.80	-0.89	13.33	-11.86	10.56
CIBIS	0.66	0.54-0.81	-0.42	100	-42	17.64
MERIT	0.66	0.53-0.81	-0.42	100	-42	17.64
pooled data				238.84	-91.78	46.493

Test if pooled OR is significantly different from 1.0

$$= \frac{(\ln OR_1/s_1{}^2 + \ln OR_2/s_2{}^2 + ...)^2}{1/s_1{}^2 + 1/s_2{}^2 ....} =$$

$$= \chi^2_{\text{pooling}} \quad \text{for 1 df}$$

$$= (-91.78)^2 / 238.84 = 39.65 \quad p<0.0001.$$

Test if heterogeneity is between the studies $=(\ln OR_1)^2/s_1^2+(\ln OR_2)^2/s_2^2 +..-\chi^2_{\text{pooling}}$

$$= 46.493 - 39.65 \qquad \text{for } 4\text{-}1=3 \text{ dfs}$$

$$= 6.843 \qquad 0.05 <p< 0.10.$$

Calculate pooled 95% CIs $= e^{OR \pm 1.96 / \sqrt{(1/s_1^2 + 1/s_2^2 + ...)}}$

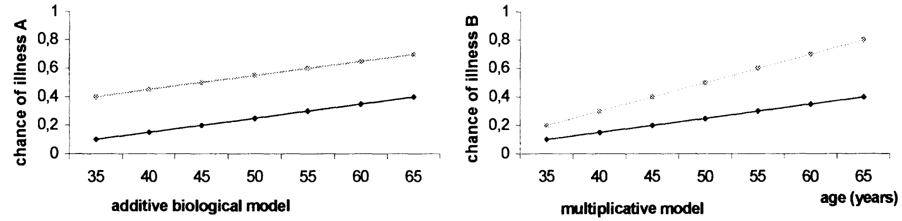
$$= e^{-91.78 / 238.84 \ \pm 1.96/ \sqrt{238.84}}$$

$$= e^{-0.3842 \pm 0.127}$$

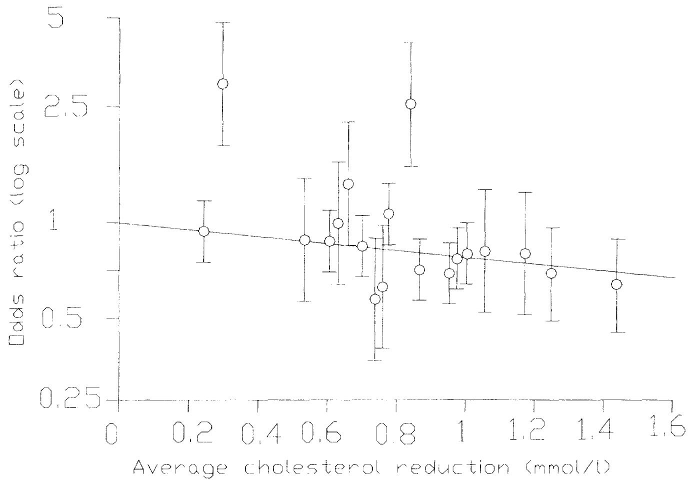
$$= 0.68 \ (0.59\text{-}0.77)$$

significantly different from 1.0.

18. WHAT TO DO IN CASE OF HETEROGENEITY?



Significant heterogeneity in a meta-analysis is not a disaster.
What you should do, is find the cause. E.g.: chance of disease increases when age increases as demonstrated in the above figure. Right hand graph shows a pattern raising heterogeneity in a meta-analysis because there is no not linear pattern anymore.



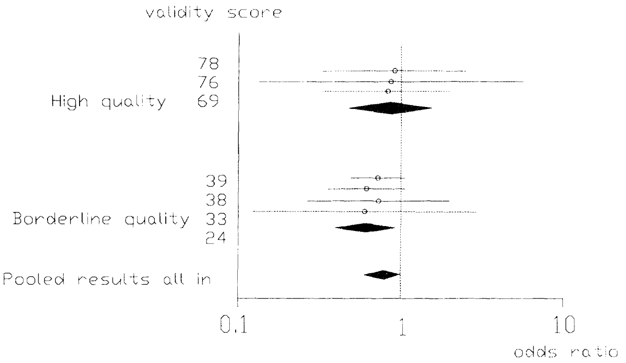
Another cause of heterogeneity is outlier studies. The above figure shows that the chance of myocardial infarction decreases when plasma-cholesterol decreases. However, two outlier studies disturb the otherwise strong linear relationship.

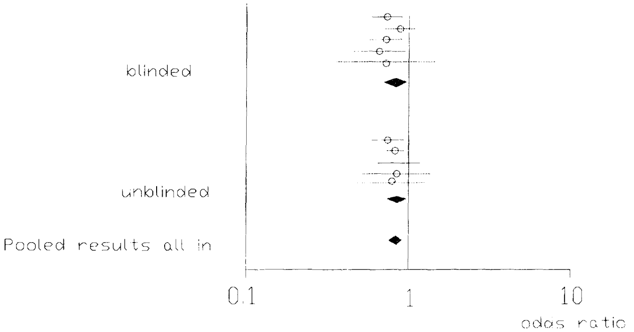
	No of cancer cases	No of studies	Mean (SE) difference in serum cholesterol (mmol/l)*	Heterogeneity
All studies:				
Overall	12 516	33	−0·041 (0·009)	$\chi^2=53$, df=32, P=0·01
Socioeconomic status:				
High	619	4	+0·032 (0·048)	$\chi^2=37$, df=30, P=0·18
Mixed	10 378	20	−0·030 (0·010)	
Low	1 519	9	−0·130 (0·025)	
Studies with lung cancer data:				
All cancers	8 062	19	−0·043 (0·012)	$\chi^2=40$, df=18, P=0·002
Lung cancers	2 239	19	−0·101 (0·022)	$\chi^2=36$, df=18, P=0·007
Cancers other than lung	5 823	19	−0·023 (0·014)	$\chi^2=32$, df=18, P=0·02

* Mean cholesterol in those who subsequently developed cancers minus mean in those who did not.

A third cause of heterogeneity is shown in the above figure: outlier-populations. The chance of cancer increase when cholesterol increases. However, only in the lowest-social-group this effect was highly significant, and so the heterogeneity was probably due to a social factor rather than to differences in plasma-cholesterol-levels.

19. DATA ANALYSIS: THIRD PITFALL, LACK OF ROBUSTNESS





Lack of robustness or lack of sensitivity is the third pitfall of meta-analyses.
It is defined as the phenomenon that studies with borderline quality produce more spectacular results than do high quality studies. The main cause is placebo effects, which may also be doctor-mediated.
The above examples show in the left upper graph that the pooled result is mainly determined by the 4 lower quality studies. In the lower graph it is shown that lower and higher quality studies do not necessarily have different results.
What to do?: (1)remove low quality studies,
(2)look whether pooling changes results,
(3)Yes?, then don't pool.

Leaving out the studies at this stage is impossible (according to item (2) of the scientific method (strict inclusion criteria)).

20. CRITICIZMS OF META-ANALYSES

Criticizms of meta-analyses.
-They do not predict large trials.
-They do not predict adverse effects.
Cause: 3 pitfalls of meta-analyses:
(1) publication bias,
(2) heterogeneity,
(3) lack of robustness.

Initiatives against heterogeneity.
CONSORT=consolidated standards rcts (randomized controlled trials);
70 editors of international journals participate.

Initiatives against publication bias.
1-CONSORT.
2-Unpublished Paper Amnesty Movement (Lancet and BMJ participate).
3-World Association of Medical Editors (hundreds of journals participate).

Cochrane-Group and Evidence-based Movement has offices in every western country, and is in favor of meta-analyses as gold standard in any aspect of clinical research including.

1. Reporting randomized experimental research.
2. Development of new drugs.
3. Determination of individual therapies.
4. Leading the way for regulatory organs.
5. Epidemiological research.

Meta-analysis is simple and straight
Just remember the scientific method:

1. Clearly defined prior hypothesis.
2. Thorough search of trials.
3. Strict inclusion criteria.
4. Uniform guidelines for analysis.

21. EXAMPLE OF PUBLISHED META-ANALYSIS

ANGIOTENSIN II ANTAGONISTS FOR HYPERTENTION: ARE THERE DIFFERENCEW IN EFFICACY?

Paul R. Conlin, M.D.[1], J. David Spence, M.D.[2], Bryan Williams, M.D.[3], Arthur B. Ribeiro, M.D., Ph.D.[4] , Ikuo Saito, M.D.[5], Claude Benedict, M.D.[6], Antonius M.G. Bunt, M.D., Ph.D.[7]

[1]Endocrinology-Hypertension Division, Brigham and Women's Hospital and Harvard Medical School, Boston, M.A; [2]Siebens-Drake/Robarts Research Institute, University of Western Ontario, London, ON, Canada; [3]Cardiovascular Research Institute, University of Leicester, Leicester LE27LX, United Kingdom; [4]Nephrology Division, UNIFESP-EPM, Sao Paulo, Brasil; [5]Health Center, Keio University, Tokyo, Japan; [6]Universaty of Texas Medical School, Houston, TX; [7]Merck & Co. Inc., Whitehouse Station, NJ.

Running Head: Antihypertensive efficacy of angiotensin II antagonists

Correspondence and reprint requests to:

Paul R. Conlin, M.D.
Endocrinology-Hypertension Division
Brigham and Women's Hospital
221 Longwood Avenue
Boston, MA 02115
Phone: 1-617-732-5661
FAX: 1-617-732-5764
E-mail: pconlin@rics.bwh.harvard.edu

Meta-analysis of 43 AII-antagonist trials in patiens with hypertension.
Item 1 of scientific method: prior hypothesis?: yes, how large effect of losartan versus the rest.
Item 2 of scientific method: thorough search?: yes, Medline.
Item 3 of scientific method: strict inclusion criteria?: yes only randomized controlled trials.

Minor flaw: Oddoustock's data size.

		DBP (mmHg)			SBP (mmHg)			responders
First Author	n	Mean	SD	95% CI	Mean	SD		(%)
Trimarco[45]	72	-11.9	6.8	[-13.5,-10.3]	-14.6	11.8	[-17.4,-11.8]	
Gradman[22]	79	-10.1	7.0	[-11.7,-8.5]	-13.0	12.7	[-15.8,-10.2]	
Dalof[30]	132	-9.0	7.7	[-10.3,-7.7]	-11.8	12.2	[-13.9,-9.7]	49
Weir[48]	110	-8.9	7.5	[-10.3,-7.5]	-9.6	13.0	[-12.9,-7.1]	
Mackay[47]	138	-8.8	7.6	[-10.1,-7.5]	-10.7	14.3	[-13.1,-8.3]	52
Chan[28]	89	-8.8	7.7	[-10.4,-7.2]	-12.6	13.8	[-15.5,-9.7]	56
Townsend[27]	132	-8.8	7.7	[-10.1,-7.5]	-8.6	13.8	[-11.0,-6.2]	
Roca-	192	-8.7	7.7	[-9.8,-7,6]	-12.0	13.8	[-14.0,-10.0]	56
Cusachs[48]								
Tikkanen[25]	200	-8.4	7.1	[-9.4,-7.4]	-10.6	13.0	[-12.4,-8.8]	51
Wilson[29]	36	-8.4	5.9	[-10.4,-6.4]	-10.0	9.2	[-13.1,-6.9]	
Oddou-	534	-8.0	7.7	[-8.7,-7.3]	-10.5	13.8	[-11.7,-9.3]	44
Stock[8]								
Ikeda[48]	125	-7.7	9.0	[-9.3,-6.1]	-9.2	13.8	[-11.6,-6.8]	46
Oparil[50]	97	-7.3	9.0	[-9.1,-5.5]	-6.1	14.4	[-9.0,-3.2]	
Mallion[25]	109	-7.0	8.6	[-8.3,-5.7]	-9.3	11.9	[-11.6,-7.0]	46
Byyny[51]	29	-6.7	7.8	[-9.7,-3.7]	-11.7	17.6	[-18.4,-5.0]	
Andersson[9]	83	-6.6	8.7	[-8.5,-4.7]	-11.1	21.2	[-15.7,-6.5]	
Oparil[11]	192	-6.2	7.7	[-7.3,-5.1]	-8.3	13.8	[-10.3,-6.3]	44
Martina[52]	10	-4.0		9-9.2,1.2]	-7.0	8.2	[-12.9,-1.1]	
Hegner[38]	82	-13.4	7.7	[-15.1,-11,7]	-16.1	15.8	[-19.6,-12.6]	74
Mallion[34]	94	-13.2	7.6	[-14.8,-11.8]	-17.2	11.9	[-19.6,-14.8]	61
Corea[35]	84	-11.5	6.8	[-13.0,-10.0]	-13.1	14.8	[-16.3,-9.9]	67
Holwerda[32]	136	-9.5	6.2	[-10.6,-8.4]	-12.4	12.7	[-14.5,-10.3]	55
Oddou-	545	-8.3	12.0	[-9.3,-7.3]	-11.0	17.5	[-12.5,-9.5]	46
Stock[8]								
Oparil[23]	150	-7.2	14.6	[-9.6,-4.9]	-8.6	25.1	[-12.6,-4.6]	43
Black[33]	384	-7.1	14.7	[-8.6,-5.6]	-8.0	17.5	[-9.8,-6.2]	42

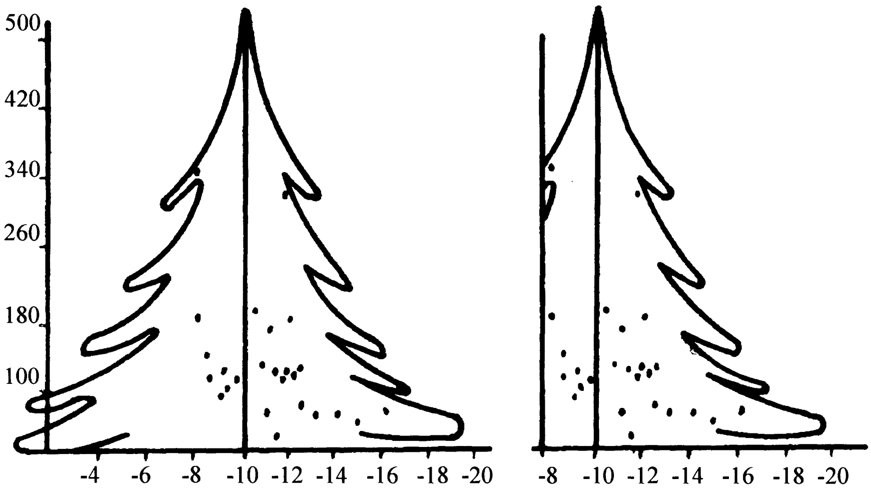
Data analysis.

1st Pitfall. publication bias.

- -Trials are divided into 2 groups: large trials involving > 100 patients, small ones < 100 patients.
- -Small trials gave best results.
- -Have small trials with negative results been excluded from publication?

Table Publication bias

	trials n> 100	p	trials n< 100
Fall systolic blood pressure	10.2	<0.05	10.8
Fall diastolic blood pressure	8.1	<0.05	8.5



-Christmas tree plot: x-axis results studies, y-axis size of studies.
-Small trials large range of results, large studies small range of results.
-Cut the Christmas tree, the lower left-hand side is empty indeed.
-Some publication bias (difference 3 small studies).

Data analysis.
2nd Pitfall: heterogeneity.
-Wide range between effects of various studies.
-Test for heterogeneity significant for systolic pressures.
-95% CIs pooled data (also estimate of heterogeneity) not wider 5% of treatment effect(eg 10.5 mm Hg (9.8-10.9).

	Tabel heterogeniteit
monotherapy	
Fall in systolic pressure (range)	6.1 to 17.2 mm Hg[1]
Fall in diastolic pressure (range)	4.0 to 13.4 mm Hg
duplicate therapy	
Fall in systolic pressure (range)	11 to 21.5 mm Hg[1]
Fall in diastolic pressure (range)	9.0 to 15.5 mm Hg

Data analysis.
3rd pitfall: Lack of robustness.
-Definition: low quality studies have more spectacular results.
-Table high and low quality studies.
-Low quality studies have, indeed, a better mean result (ns, but pooled it would have been significant).
-So not 100% robustness, yet differences clinically irrelevant.

Tabel lack of robustness

	Overall mean results	mean results no SDstudies (lower quality studies)
Low-dose data		
Fall in systolic pressure (mm Hg)	10.8	11.2
Fall in diastolic pressure (mm Hg)	8.5	8.9
High-dose data		
Fall in systolic pressure (mm Hg)	13.3	13.7
Fall in diastolic pressure (mm Hg)	8.9	9.9
Duplicate-therapy data		
Fall in systolic pressure (mm Hg)	13.3	13.9
Fall in diastolic pressure (mm Hg)	9.9	10.8

Conclusions of this example.

1. Some publication bias, heterogeneity, and lack of robustness.
2. Cause heterogeneity: lack of dose titration in some studies.
3. Effects of 3 pitfalls small (5-6% of total effect).
4. Maybe, testing publication bias, heterogeneity, robustness not always needed with large meta-analyses!!
5. Efficacy of different AII-antagonists is not different.
6. Large doses almost as effective as low doses.

22. EXERCISES TO CHAPTER 6

Indicate which alternative is **not** correct.

1. Scientific rules for meta-analyses include:
   - A. prior hypothesis, thorough search of trials, strict inclusion criteria, uniform data analysis,
   - B. prior hypothesis, thorough search of trials, testing pitfalls, uniform data analysis,
   - C. prior hypothesis, thorough search of trials, valid design, uniform data analysis,
   - D. prior hypothesis, thorough search of trials, pooling the data, uniform data analysis.

2. Frequent cause of heterogeneity:
   - A. heterogeneous ages of study participants,
   - B. outlier data,
   - C. social factors,
   - D. low quality trials.

3. Frequent cause of lack of robustness:
   - A. heterogeneous ages of study participants,
   - B. low quality trials,
   - C. placebo effects patient-mediated,
   - D. placebo effects doctor-mediated.

4. Initiatives against pitfalls of meta-analyses:
   1. CONSORT (Consolidated Standards Randomized Trials),
   2. Unpublished Paper Amnest Movement,
   3. World Association of Medical Editors,
   4. Evidence-based Movement.

5. The most important pitfalls of meta-analyses are:
   - A. publication bias, heterogeneity, lack of robustness, post hoc analysis,
   - B. publication bias, heterogeneity, post hoc analysis,
   - C. publication bias, lack of robustness, post hoc analysis,
   - D. publication bias, heterogeneity, lack of robustness.

6. Publication bias:
   A. cannot be excluded,
   B. can be excluded by thoroughly searching for trials,
   C. cannot be adjusted,
   D. is due to negative trials being published.

7. Heterogeneity:
   A. cannot be excluded,
   B. can be excluded by strict inclusion criteria,
   C. can be adjusted,
   D. can be excluded by thoroughly searching for trials.

8. Odds is a surrogate for risk in clinical trials because:
   A. calculated risk would overestimate true risk,
   B. calculated risk would underestimate true risk,
   C. calculated risk ratio would overestimate true risk ratio,
   D. calculated risk ratio would underestimate true risk ratio.

9. A. Describe the four most important scientific rules for meta-analysis.
   B. Name and explain the three most important pitfalls of meta-analyses.

CHAPTER 7

INTERIM ANALYSIS

1. INTERIM ANALYSIS: LOOKING AT DATA BEFORE CLOSURE

In clinical trials, especially in trials involving many patients or with very long duration, it is tempting to look at the data to see whether expectations come true, whether differences are already significant. In general this should be discouraged because the validity of trial and its results is endangered: if participants know what happened to patients treated so far, they might change the protocol knowingly or unknowingly. This should not be allowed and, therefore, data should not be looked at before formal end of the trial.
Sometimes there are however valid reasons to look at data before closing the trial. When the new treatment is much more efficacious than expected it is unethical to randomize patients to a placebo treatment (or the other way around), and if a new treatment has much more side-effects than expected this may be reason to stop the trial as well. In order to check this, it may be worthwhile to look at the data, and this is called an interim analysis. Interim analyses must be specified in the research protocol; it must be specified what will be looked at and what decisions will be taken dependent on results found. Often this looking at data requires unblinding of treatment given to the patients so far.
Also it is worthwhile to look at data to ensure that protocol is adequately followed by the participants. This type of looking at data does not require unblinding the treatment, and is called monitoring. Monitoring is of utmost importance to increase the likelihood of a successful trial.

2. MONITORING

Purposes:

- In order to maintain quality,
- to ensure that the protocol is followed,
- to ensure that in-/exclusion are appropriate,
- to check accrual rate,
- to check the availability and consistency of the data sampled.

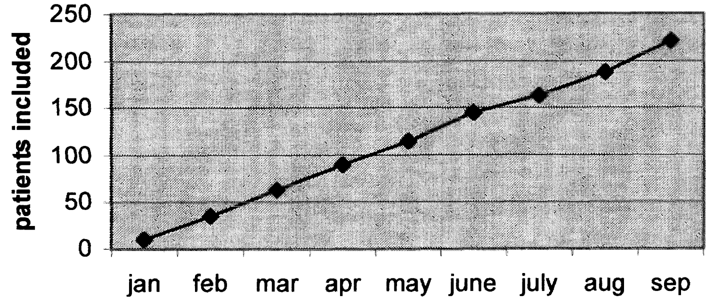
Monitoring clinical trials is used to ensure that quality standard are maintained. When monitoring it is checked whether (1)patients and physicians adequately follow the trial protocol, (2)informed consent is obtained, (3)the inclusion and exclusion criteria are met, (4)the included patients are truly randomized, (5)the required data are adequately sampled, and (6) the data input in the database is correct. Important also is the monitoring of the accrual rate. The number of patients included in the trial directly influences the power of the trial, and often the required number of patients included in the trial is computed sharply. Thus, it is of great importance that the target number of patients is obtained.

3. DATA CONSISTENCY AND AVAILABILITY

Success of a trial depends entirely on sampling the correct information, and that means that it is of paramount importance to check whether the data entered in CRFs (case report forms) are correct. All CRFs must be checked therefore against hospital files. This will minimize the chance of fraud, but more important it will minimize the number of false entries into the database, and, thus, will minimize residual variance in the statistical analysis, increasing the power of the trial. By checking CRFs missing data are identified which are always difficult to handle, but less so when identified early instead of late in the trial. Correspondence of data in the CRF and the database is more efficiently checked by double data-entry, but any other solution may be used for the monitoring process instead.

4. PATIENT ACCRUAL

Monitors should report regularly to all participating centers.



Number of patients included directly determines the power and, therefore, the success-probability of a trial, and it important that projected numbers are obtained. It is empirically shown that, initially, physicians are participating with much enthusiasm, but this decreases with time. One way of stimulating centers to continue to include patients in the trial is regular reporting of the progress of the trial, for instance by reporting the numbers of patients included in the trial, and the final goal. Monitoring these numbers can also be the basis for changing the protocol along the way, for instance by enlarging inclusion criteria.

5. CHANGING INCLUSION CRITERIA

When changing in-/exclusion criteria or when adjusting sample size,

- always make protocol amendments,
- consider the statistical consequences such as type-I error rate, sample size, power,
- monitoring is essential for good quality.

Any change in the study protocol, for instance change of inclusion or exclusion criteria, or any other adjustment of sample size must be accompanied by a formal protocol amendment. In such amendment one must consider the statistical consequences of the change for the type-I error rate, sample size and power.
It cannot be stressed too much that monitoring is essential for good quality of a clinical trial.

6. INTERIM ANALYSIS

Interim analyses should be done

- for analyzing efficacy and/or side-effects
  - for ethical concerns,
  - for efficiency reasons,
  - and to check assumptions made while designing,
- only when decisions can be taken.

In contrast to monitoring where it is not necessary to know which treatment was given to whom, interim analysis requires unblinding the results. Interim analysis is an analysis of effect or side-effect before formal end of the trial. It is performed mainly for ethical reasons, efficiency reasons, or to check whether the prior assumptions made are met. Ethical reasons play an important role when there is valid suspicion that the new tested treatment is really much more effective than standard treatment. If this is the case, it is not ethical to treat patients with an inferior therapy, and it is important to know as early as possible. This can be observed at an interim analysis.
Time, energy, and financial resources are similarly reasons to perform an interim analysis: when a new treatment is truly more effective than the standard treatment, to continue such a trial means unnecessary spending of scarce money, time, energy, and patients. At an interim analysis it may be decided to end the trial accordingly. Similarly, it may be observed that the new treatment is no more effective than standard treatment. When interim calculations indicate that the trial is likely to be successful, the trial may be stopped, and the standard treatment

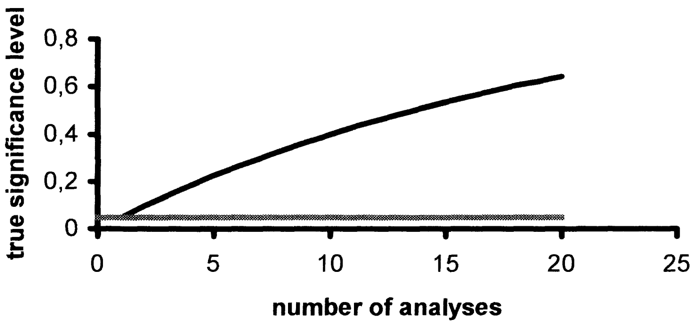
favored. It is important to realize that interim analyses are only sensible when decisions can be taken. If only a few patients are entered in the trial, an interim analysis is useless, and when all patients have been entered and follow-up is almost completed, an interim analysis is equally so.

7. DANGERS: RANDOM HIGH

- Every look at the data increases the type-I error rate and may introduce bias.
- Suppose null-hypothesis is correct, and k analyses are performed, with type I error=α =0.05, then the true significance level = 1-(0.95)k.

Interim analyses cannot be performed at libitum, there is an important price, because every statistical analysis runs the risk of a type-I error of accepting there is a significant difference where there is actually none. When analyzing the data more than once, which is done in studies with interim analyses, the type-I error will increase, and, obviously, a risk of bias is introduced. This is easy to conceive when realizing that the standard type-I error rate is usually specified to be α=0.05. When k analyses are performed (that is 1 final analysis, and k-1 interim analyses), and a significance level of 0.05 is used at each analysis, then the actual type-I error rate will be must larger than 0.05, and will be $1-(0.95)^k$.

8. SIGNIFICANCE LEVEL



In the above graph it is illustrated that the type-I error rate may increase dramatically with increasing number of interim analyses. In general it will not be as dramatical as shown here, but it may increase until the likelihood of a type-I error is almost certain. This is not acceptable for adequate clinical trials.

9. CORRECTION FOR INCREASING TYPE-I ERROR RATE

Methods for correcting the increassed risk of type I error:

- Bonferroni corrected significance level $\alpha^*=0.05/k$,
- Group-sequential methods (for comparing means):
  - $k=2$; $\alpha = 0.0294$
  - $k=3$; $\alpha = 0.0221$
  - $k=4$; $\alpha = 0.0182$
  - $k=5$; $\alpha = 0.0158$
- Pocock's method (Biometrika 1977; 64: 191-9).

There are several ways of correcting for the increase of the type-I-error rate. They are all based on lowering the nominal significance level α.

The best known method is the Bonferroni method. This method entails lowering the nominal significance level to α divided by the number of interim analyses: $\alpha^*=\alpha/k$. This method works because the type-I error rate $(1-(1-\alpha^*)^k)$ remains close to $\alpha=0.05$ for any k:

k=1	type-I error=	0.05
k=2		0.0494
k=3		0.0492
k=4		0.0491
...		...
k=10		0.0489

But the formula given $(1-(1-\alpha^*)^k)$ applies only when the interim analyses are independent and that is clearly not the case as data in the first analysis are also considered in the second analysis, and so on. This means that the Bonferroni correction method is very conservative: the actual significance level will be less, and sometimes much less, than the required or desired significance level.

Several investigators have looked at actual significance levels when analyzing data on an interim basis, among who is Pocock. He suggested a group-sequential method for interim analysis. This means that a specified significance level is to be used depending on the number of interim analyses planned: .02994, .0221, .0182, .0158 for 2, 3, 4, or 5 analyses respectively.

10. SOME RULES

- Make always a formal report of the interim analysis,
- look only at the most important evaluation criterion (only 1) and use the others to confirm the result,
- do the analysis on up-to-date material,
- do the analysis only when there is a substantial number of patients,
- keep results secret if the trial continues,
- use a triple blind procedure and an independent committee,
- predefine (in the protocol) "when to decide what".

Some common-sense rules about interim analyses include that the analysis be done by an independent committee, and that this committee be blinded to the actual treatment (triple blind procedure). This means that if the trial is continued, the trialists will not be aware of the results so far. In general, results should be kept secret if the trial continues, but a formal report of the analysis must be made in any case.

It is obvious that interim analyses should only be done when there are sufficient patients already included in the trial, and that they should be done on up-to-date material: this will guarantee optimal power for the analysis. In order not to increase type-I error rate, it is wise to do the analysis on a single efficacy criterion only. Others may be used to confirm results when a certain decision is reached. Decisions must be specified in detail in the protocol: the protocol must contain the rules when to decide what.

11. DECISION RULES

- Stop and reject "the H0 of no effect" when the difference in efficacy is $\geq \theta$ and statistically significant,

  and

- stop and accept "the H0 of no effect" when the difference in efficacy is $\leq \xi$, and the p-value ≥ 0.5 (for example),

  or

- stop the trial when the adverse event rate $> \lambda$.

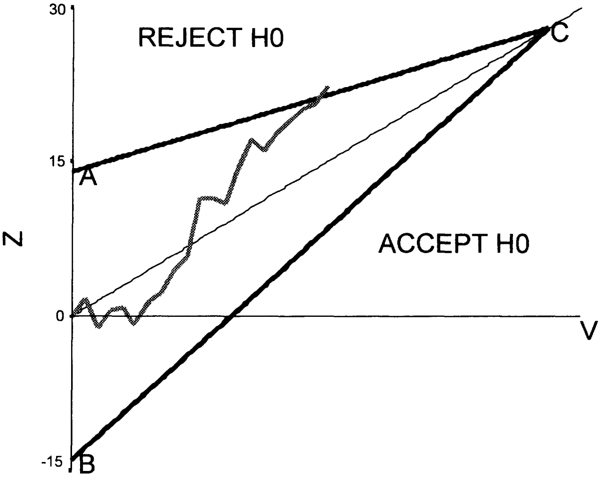
Two decisions can be taken on the basis of interim analyses: to continue or to stop the trial. The latter should be done only when the null-hypothesis of no effect is rejected and when, at the same time, efficacy is greater than some pre-specified level. Stopping is also considered when efficacy level is less than an expected level, and, of course, not significant. Independently of efficacy the trial may be stopped when the side-effects are too numerous, but the protocol must define the acceptability of the side-effect rate.

Apart from stopping or continuing the trial, the assumptions made in designing the trial may be checked in the interim analysis. When these assumptions are observed to be invalid, the trial-design may be adapted, usually leading to an adaptation of the sample size.

12. SEQUENTIAL TRIALS

Another approach to stopping the trial.

- Calculate after every patient the treatment difference Z and its information content V.
- Stop the trial when the "stopping boundary"is crossed.



When many interim analyses are planned, it may be more efficient to perform a so-called continuous sequential trial. This means that a statistical analysis is planned after each patient that finishes the trial. Each time, the effect of treatment is quantified (denoted by Z), as well as the amount of information sampled so-far (denoted by V). This is plotted against each other, shown by the dotted line in the figure. Decisions rules must be defined in order to make decisions about significance of effect, or lack of it. Shown here is one way, called a triangular design. As long as the dotted trial-line remains within the boundary the trial continues, but as soon as it crosses the boundaries the null-hypothesis of "no effect" will be rejected (upper left area), or accepted (lower right area).

The definition of the boundaries is according to established formulas, for instance given by Whitehead (Evaluating sequential trials (PEST version 3) www.reading.ac.uk/mps/pest/pest.html).

13. EFFICACY (Z)

- Z is the quantification of efficacy of the above sequential trial:
  risk difference: $Z = p_1 - p_2$,
  relative risk: $Z = p_1 / p_2$,
  hazard ratio: $Z = \lambda_1(t) / \lambda_2(t)$,
  odds ratio: $Z= p_1(1-p_1) / p_2 (1-p_2)$,
  mean difference = $X_{.1} - X_{.2}$.

Efficacy (Z) is quantified differently based on the type of data sampled. If data is dichotomous (death/alive; ill/recovered; yes/no relapse) Z may be the risk-difference, the odds ratio, or the relative risk (p=proportion). When the variable ($X_{i,j}$, ith group, jth subject) is numerical (quantitative), Z is typically a difference of means. When data is of the survival-data type, Z is usually a hazard-ratio (λ = αδϖερσε εϖεντ ρατε).

14. DIFFERENCE OF TWO MEANS (V)

- V is a quantification of the certainty of Z.
- In general $V=1/SE^2(Z)$.
- If $Z = p_1 - p_2$: $V=n_1/(p_1(1-p_1)) + n_2/(p_2(1-p_2))$.
- If $Z = X_{.1} - X_{.2}$: $V = n_1/S_1^2 + n_2/S_2^2$.

The amount of information sampled so far, is usually the inverse of the squared standard error associated with Z: so if Z is the difference of two means V equals $\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}$ where n_1 and n_2 are the sample sizes of the two means, and S_1 and S_2 are the associated standard deviations. When Z is the difference of two proportions, V equals $\frac{n_1}{p_1(1-p_1)} + \frac{n_2}{p_2(1-p_2)}$. When Z is an odds ratio or relative risk, the formulas for V are given in chapter 6, paragraphs 21, 22.

15. CONCLUSION

- Monitoring is essential for a high quality trial.
- Interim analysis of efficacy is rarely required, but if it is planned:
  - plan only a few (Pocock (paragraph 9): never more than 5),
  - do it when a substantial number of observations are available,
  - specify (as much as possible) the comparisons to be made and their associated decisions in the protocol,
  - use an independent committee,
  - use a triple-blind procedure.

Again, monitoring is essential in clinical trials, interim analyses should be done only in specific circumstances. It endangers the type-I error rate and this must be taken care of in the statistical analysis. Best are the group-sequential trial procedures developed by Pocock (paragraph 9). If an interim analysis is performed, its rules must be specified as much as possible in the study protocol.

16. EXERCISES TO CHAPTER 7

Indicate which alternative is correct.

1. The primary goal of interim analyses is
   A. controlling the risk of a type I error of a clinical trial,
   B. controlling the risk of a type II error of a clinical trial,
   C. controlling design assumptions of a clinical trial.

2. The consequence of an interim analysis is
   A. increased risk of a type I error,
   B. decreased risk of a type I error,
   C. increased risk of a type II error.

3. An interim analysis should ideally be done by
   A. the trial's data committee,
   B. the trial's executive committee,
   C. an independent committee.

4. The best way of controlling the statistical consequences of interim analyses is
   A. a bonferroni adjustment of p-values,
   B. a group-sequential trial approach,
   C. using a larger power level.

5. A sensible stopping-rule in an interim analysis of a clinical trial could be
   A. "stop the trial when the p-value is less than 0.05",
   B. "stop the trial when the side-effects occur in more than 25% of the patients",
   C. "continue when the p-value >0.05".

6. The results of an interim analysis must be kept confidential because
   A. this is required by the governmental authorities,
   B. in that way the results of the analysis cannot influence treatment of new patients included in the trial,
   C. in that way statistical analysis is unbiased.

CHAPTER 8

MULTIPLE TESTING

1. TWO SITUATIONS

Two situations are given.

- Comparing many groups of different patients: multiple comparison.
- Using many evaluation criteria: multiple testing.

There are, thus, two different situations where the problem of multiple comparisons arise, (a.) where more than two groups of patients are compared with each other, and (b.) where two or more criteria are used to compare two (or more) groups. In both cases the problem is that the type-I error rate is endangered.

2. TYPE-I ERROR RATE

(1) Assume that the null-hypothesis of no-difference is true.
(2) Suppose 2 comparisons/tests are performed.
The chance of at least one significant test at $p < 0.05$
is the twice the chance with 1 test.
(3) With k tests and $\alpha = 0.05$ the chance of at least 1 significant test
increases to $1 - 0.95^k$.

The type-I error is the conclusion that a difference exists on the basis of the trial-data while it does not exist in reality. This risk of this error is the statistical significance, and, worldwide, the upper limit is accepted as α=0.05. Multiple comparisons, and multiple testing increase the actual significance level.
To further explain this, suppose that there is no difference in reality. Also, suppose that k comparisons or tests are performed each with significance level α=0.05. The actual chance (=probability=Pr) of a false conclusions equals

Pr(1 or more significant tests) = 1 – Pr(no significant tests).

Assuming that the k comparisons are independent, then the type-I error-risk can be written as

$$\text{Pr(type-I error)} = 1 - \binom{k}{0}\alpha^0(1-\alpha)k = 1-(1-\alpha)^k .$$

If the significance level α is set to 0.05, the risk will be larger for any value of k>1, and for large k the risk will be close to 1. Multiple comparisons and multiple tests will not be independent in most cases, therefore the error-rate will not be as dramatical as implied by the above formula, but even with small dependencies, the error rate will soon get large, and need correction. Both for multiple comparisons, and for multiple tests there are different ways to adjust the significance level, and some of them will be discussed here.

3. MULTIPLE COMPARISONS

- Suppose k treatment groups,
- H_0: $\theta_1 = \theta_2 = \theta_3 = ... = \theta_k$ where θ = treatment effect,
- k(k-1)/2 different comparisons are possible,
- example:
  compare 4 different selective serotonin reuptake inhibitors (SSRIs) with each other and with placebo with respect to the intravaginal ejaculation latency time (IELT) after 6 weeks of treatment in patients with ejaculation praecox. Baseline IELT<60 sec, on average 20 sec.

Multiple comparisons are encountered in trials where k treatment groups are compared with respect to some efficacy criterion θ. The standard null-hypothesis is that all theta's are equal: H_0: $\theta_1=\theta_2=...=\theta_k$. With k groups, there are k(k-1)/2 different comparisons possible: group 1 with 2, group 1 with 3, and so forth, until group k-1 with group k.

Patients with ejaculatio praecox have a very quick ejaculation upon intromission (within 60 seconds in over 90% of intromissions), and in our example the average IELT was 20 seconds. Latency times were skewed-distributed, and we therefore analyzed log(IELT) values.

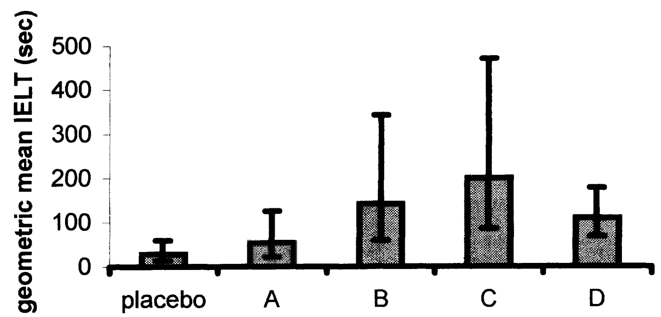
4. EXAMPLE 1: DATA SUMMARY

After 6 weeks of treatment:

Treatment	sample size n	mean x	standard deviation S
Placebo	9	3.34	1.14
SSRI A	6	3.96	1.09
SSRI B	7	4.96	1.18
SSRI C	12	5.30	1.51
SSRI D	10	4.70	0.78

The average log(IELT) values after six weeks varied between 3.34 for the placebo group and 5.30 for the third SSRI. There are several questions that a researcher may ask here. One may focus on differences between each active drug (SSRI A, B, C, and D) versus placebo, thereby inferring efficacy of each SSRI against no treatment. On the other hand one may also inquire whether the active drugs differ amongst each other in efficacy. In this example both questions are of interest, and this leads to 5*(5-1)/2 = 10 different paired comparisons.

5. EXAMPLE 1: GRAPHICAL DISPLAY



Significant differences ?

Although we summarized the IELT data on the log scale, it is very informative to display this summary on the original time scale. Back-transforming the average of log-transformed individual observations yields the geometric mean of the original observations, and is often very close to the median. The standard deviation of the log-transformed observations is often close to the variation coefficient (CV) of the original data points, and therefore the limits of the 95% confidence of the log-transformed data will often coincide with the 95% confidence interval of the median of the original scale. Naturally, this procedure can be applied only with positive data only. If zero values or negative data are observed, other transformation must be looked at in order to standardize and normalize distributions.

The confidence intervals of the placebo group and of the SSRI B, C, and D groups do not overlap and this suggests significant differences, but here there is a need for a multiple-comparison-correction. The intervals of the 4 SSRI's do overlap, and this suggests no-significant-differences, but this must be interpreted with care, because the relevant confidence is the confidence of the difference of the means, and not of the two means separately.

6. TWO STRATEGIES

Two strategies are possible:

(1) ANOVA,
if p-value >0.05 then accept HO.
<0.05 then perform a least significant difference (LSD) procedure.

(2) Direct multicomparisons:
Bonferroni t-test,
Tukey's highest significant difference(HSD) test,
Student-Newman-Keuls test,
Dunnett′s test.

For the comparison of k groups of different patients we can take two different scenarios. The first scenario is to start with analysis of variance (ANOVA), and inspect the global test of the null-hypothesis of no difference among the k groups. This test is a single test, and only one p-value is interpreted, hence there is no multiple comparisons problem: the type-I error risk is at the nominal level $\alpha=0.05$. If the p-value is larger than α, then the analysis stops, and the null-hypothesis of no difference is not rejected. When the p-value is less than α, then the null-hypothesis of no difference is rejected. Paired comparisons are, subsequently, done using the LSD procedure *without* further correction for multiple comparisons. The basic philosophy of this strategy is that the type-I error rate is controlled through the ANOVA procedure.

The second strategy ignores the p-value of the ANOVA procedure, and starts directly with paired comparisons. Since the type-I error rate is not controlled through a global test, a correction for the multiple testing is needed. There are many procedures, and some of the best known procedures are called: Bonferroni, Tukey's HSD, Student-Newman-Keuls, and Dunnett's procedures.

We will also discuss each of these multiple comparisons procedures.

7. LSD TEST

Tukey's least significant difference (LSD) test

- is to be used in connection with ANOVA,
- is (sort of) pairwise t-tests,

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{S_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

$S^2{}_W$ = residual variance of ANOVA

t_{ij} is distributed as a Student's t with N-k degrees of freedom

The LSD procedure consists of calculating a t-type statistic for each pair of groups: it is the ratio of the difference of the two means divided by the standard error of this difference. It looks very much like the Student's t-test, and, in fact, the only difference is the estimate of the within-groups-variance (or pooled variance: s_w^2). Here it is estimated on the basis of all k groups (via the ANOVA procedure), whereas in Student's t-test it is estimated on basis of data in the two groups only. The LSD-statistic follows a t-distribution (like Student's t-test) but with N-k degrees of freedom where N is the total number of patients in the k groups.

8. ALTERNATIVES

The alternative strategy to test multiple samples/groups is a pairwise test without ANOVA,

- with a Bonferroni corrected significance level: 0.05/(k(k-1)/2),
- considering the distribution of the various t_{ij}-values (HSD test),
- multiple ranging (SNK test),
- specialized treatment comparisons (e.g. against placebo) (Dunnett′s test).

The Bonferroni procedure consists of calculating Student's t-test for each pair of treatment groups, but instead of using α(=0.05) as significance level, α/(k(k-1)/2) is used as significance level. When k is 5, ten different pairs of groups can be made, and α*(=α/10=0.05/10)=0.005 is used as significance level. Bonferroni's correction is very simple, but, unfortunately, it is too strict, otherwise called conservative.

Less conservative is Tukey's highest-significant-difference (HSD) test. Again a t-type statistic is calculated as given in the previous sheet (LSD-statistic), but here the p-value is not calculated using the t-distribution with N-k degrees of freedom, but according to a much more complex distribution which entails the multiple comparison correction.

The Student-Newman-Keuls procedure is a multiple-ranging-procedure of which there are also much more variants. Dunnett's procedure entails focusing on special contrasts only, for instance the comparison of each active treatment with placebo.

9. MULTIPLE RANGING

How do we proceed with multiple testing:

1. test homogeneity of all k means,
when rejected,
2. test homogeneity in all possible sets of (k-1) means,
when rejected ,
3 ......

The Student-Newman-Keuls procedure is a multiple-ranging-procedure. This means a procedure to find homogeneous subsets of treatment groups. A homogeneous subset is a set of groups for which the global null-hypothesis of no difference cannot be rejected.

The strategy is basically a step-procedure with the global ANOVA test of no difference among the k groups as first step. When this global test is rejected (denoting heterogeneity), all possible subsets of k-1 groups are considered. If the overall test for a subset is not significant, the procedure stops for that subset, and homogeneity is not rejected. When there is at least one subset with a significant overall test, subsets of k-2 groups are considered, and this procedure continues until homogeneity for all subsets cannot be rejected.

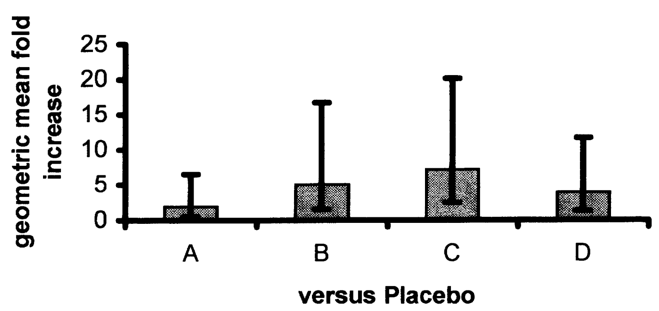
The advantage of multiple ranging is that less statistical tests need to be performed, and therefore a less stringent correction factor is needed.

10. EXAMPLE 1: RESULTS

	Difference	**P value**			
	Mean (SE)	**LSD**	**HSD**	**Bonferroni**	**Dunnett**
Placebo vs A	-0.62 (0.63)	0.33	0.86	0.99	0.73
B	-1.62 (0.60)	0.01	0.07	0.10	0.035
C	-1.96 (0.52)	0.001	0.005	0.006	0.002
D	-1.36 (0.55)	0.017	0.12	0.17	0.058
A vs B	-1.00 (0.66)	0.14	0.56	0.99	
C	-1.34 (0.60)	0.03	0.18	0.30	
D	-0.74 (0.61)	0.24	0.75	0.99	
B vs C	-0.34 (0.57)	0.56	0.98	0.99	
D	0.26 (0.59)	0.66	0.99	0.99	
C vs D	0.60 (0.51)	0.25	0.76	0.99	

A highly significant global ANOVA statistic was provided in our example, indicating that the null-hypothesis of no difference must be rejected. Paired comparisons clearly indicated that the p-values of the LSD-procedure were the smallest illustrating their liberality. P-values of the Bonferroni correction were highest, and Tukey's HSD values were in between. Dunnett's procedure only considered special contrasts, which entailed less inflation of p-values than seen with the HSD procedure.

11. EXAMPLE 1: ANOTHER GRAPHICAL DISPLAY



Results of our analysis may be displayed in many ways, but when differences of groups are of main interest, these differences must be displayed. A big advantage of showing the results in this way is that when the value zero is not contained in the interval, the difference must be significant. This requires the calculation of confidence intervals corrected for multiple comparisons, and indeed these can be calculated in association with the method chosen.

12. CORRECTED CONFIDENCE INTERVALS

- Confidence intervals (CIs) may be constructed using similar methods

Most computer programs allowing multiple comparisons procedures also provide associated confidence intervals.

13. NO METHOD IS BEST

- Which method is best? No preference, specify arguments for any method has to be provided in protocol.
- There are no multiple comparison tests available for discrete, or censored data or non-parametric methods. Best is to use an overall test, and perform pairwise comparisons only when the overall test is significant.

There are few compelling arguments for prefering one method for multiple comparisons over another. When the consequences of incorrectly deciding that a difference is for real, are large, a conservative method may be prefered (like the Bonferroni method), but the consequence of such choice is decrease of power.

We have discussed only methods for quantitative data, and unluckily these methods have not been developed as much for discrete or censored data. Also these methods require more-or-less normally distributed data, and multiple comparisons procedures have neither been developed for ranked data.

When discrete or censored data need to be corrected for multiple comparisons, or when a non-parametric statistical analysis tool is used, the best way to proceed is to use a global test first (as in the above first strategy), and continue with pairwise comparisons only when the global test is significant. The conservative Bonferroni method is, of course, also a possibility.

14. MULTIPLE TESTING

- In a decent trial there are often
  several primary evaluation criteria,
  " secondary " " ,
  " tertiary " " .
- Generally, a multitude of tests is performed and this, dramatically, increases the type-I error risk.

The second circumstance where the problem with repeated statistical analysis arises, is when efficacy in a clinical trial is characterized with two or more different variables. This happens everyday. Almost all clinical trials use so-called primary, secondary, and tertiary efficacy criteria, and each criterion is tested for significance between the two (or more) treatments that are compared in a clinical trial. The multitude of criteria increases the type-I error rate when the significance level is not adjusted.

15. WHAT TO DO?

How to handle multiple endpoints in a trial:

- Use as few as possible, say only one, efficacy criterion

There is a simple remedy to the problem of multiple testing; use as few efficacy criteria as possible, preferably only one. But this is hardly possible in many areas of biomedical research: cardiovascular disease, for instance, is characterized by myocardial, cerebral, or peripheral vascular conditions, and often many more relevant event-types. Rheumatoid arthritis is similarly characterized by a multitude of different symptoms and signs: swollen joints, pain, movement limitations, inflammation markers et cetera. In fact, few diseases can be typified with a single disease marker, and therefore many markers need to be included in most trials in order to adequately assess treatment efficacies..

16. CORRECTION

Corrections for multiple endpoints:

- Use the Bonferroni correction: $\alpha^* = \alpha/k$.
- Weigh each p-value (Hochberg): multiply
  - the largest with weight 1,
  - the second largest with weight 2,
  - the third largest with weight 3,
  - the smallest with weight k,
  - preserve the original ordering.

In case of multiple criteria the Bonferroni method can be used: basically when there are k variables, the Bonferroni correction means that $\alpha^*= \alpha/k$ is used as significance level. It is very conservative. Somewhat less conservative is Hochberg's method, which uses a Bonferroni type correction factor, but ranked by the level of the p-values, provided that the original order is maintained.

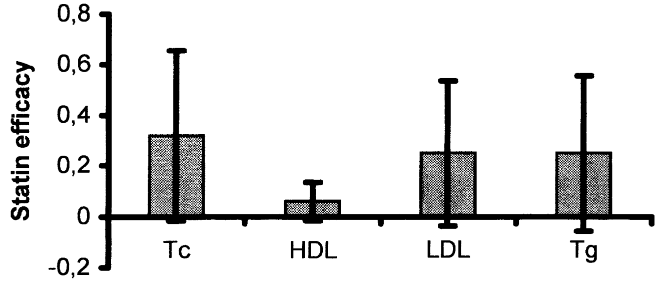
17. EXAMPLE 2: DATA

Change of:	Placebo (n=31)	Statin (n=48)	P^*	$P^\#$	$P^@$
Total cholesterol decrease	-0.07 (0.72)	0.25 (0.73)	0.06	0.24	0.11
HDL cholesterol increase	-0.02 (0.18)	0.04 (0.12)	0.07	0.28	0.11
LDL cholesterol decrease	0.34 (0.60)	0.59 (0.65)	0.09	0.36	0.11
Triglycerides increase	0.03 (0.65)	0.28 (0.68)	0.11	0.44	0.11

Take a small trial as an example, comparing the efficacy of statin and placebo of lowering total cholesterol (Tc), LDL cholesterol (LDL), and triglycerides (Tg), and increasing HDL cholesterol (HDL). Since two groups are compared with respect to mean values, the appropriate test statistic is the Student's t-test. The uncorrected p-value is given (p^*), as well as the Bonferroni corrected value ($p^\#$), which is given here as four times Student's p-value. Naturally, the p-value should remain less than 1. In Hochberg's procedure ($p^@$) the smallest p-value is multiplied by k = 4, the second smallest is multiplied by (k-1)=3, and so on, and the largest p-value is multiplied by unity. But since the original rank order is to be maintained, all Hochberg's corrected p-values are 0.11 here.

An important advantage of both Bonferroni's and Hochberg's method is that they can be applied with any type of statistical test.

18. EXAMPLE 2: GRAPHICAL DISPLAY



The differences between the mean values of the statin- and placebo-groups are displayed above, as well as the associated CIs. The CIs can be corrected for multiple testing as well.

19. ALTERNATIVES

How to correct CIs for multiple testing: two possibilities.

- Two steps:
  - 1. overall test: Hotelling's T-square (or another form), stop if not significant,
  - 2. t-tests without correction.
- Make a composite of variables on the same scale (when they are not too highly interrelated).

When two treatment groups are compared on a number of quantitative (normally distributed) variables, an overall test is available for testing globally the null hypothesis of no difference on all variables: this test statistic is called Hotelling's T-square. Using this statistic, a strategy similar to that of multiple comparisons can be followed. When the global test statistic is non-significant, the analysis is stopped, but when it is significant no further correction is needed for the individual test statistics.

Another possibility is to make first a composite variable which combines all efficacy criteria, and perform the statistical analysis on the composite only. In our example it is reasonable to believe that statin-treatment has similar effect on all four lipid variables, and, thus, a composite of these four variables might be adequate.

20. COMPOSITE

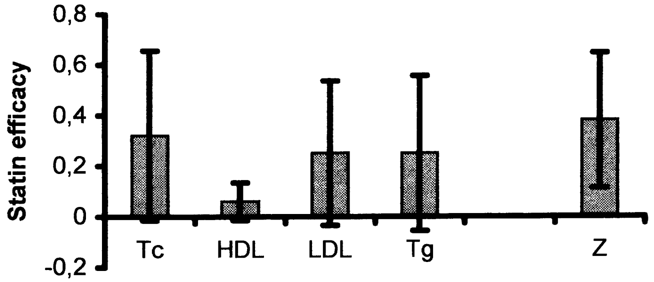
How to perform a composite analysis of the four variables.

Z = (Tc* + HDL* + LDL* + Tg*)/4
Tc* = standardized Tc: Tc* = (Tc-mean(Tc))/SD_{Tc}......

Placebo: mean Z = -0.23 (SD 0.59)
Statin: mean Z = 0.15 (SD 0.56)
p = 0.006

One possible composite is the mean value of the four variables. But in calculating the mean it is sensible to take care that the variables are measured on the same scale, and have the same direction. Standardization is easily obtained by subtracting data from the mean and dividing this by the standard deviation.
In our example the composite mean is lowered significantly more in the statin-group than in the placebo-group: hence the power to detect statin-efficacy is not sufficient for each variable on itself, but for the average the power is sufficient.

21. EXAMPLE 2: GRAPHICAL DISPLAY



For each of the four criteria the 95% CI for the difference of the two treatment groups contains the zero value, but not for the composite variable ....

21. CONCLUSION

- Beware of the multiple testing/comparison problem.

- Whatever you choose, may be acceptable, provided decisions are taken
  - a priori and
  - as specified in the protocol.

The important message of this chapter is that multiple comparisons and multiple testing pose a serious statistical problem. There are few arguments to prefer a specific method of correction; whichever method, the decision must be taken a priori and must be specified as much as possible in the study protocol. We should add that another, more philosophical, approach to the problem of multiple endpoints has been described (Cleophas et al, Statistics applied to clinical trials, Kluwer Academic Publishers, Boston, MA, 2002, pp 1-3).

22. EXERCISES TO CHAPTER 8

Indicate which alternative is correct.
In a study into genetic determinants of the effect of anti-TNF (tumor necrosis factor) treatment in patients with severe rheumatoid arthritis the role of the TNF-α marker was assessed. This marker had 10 polymorphisms in this sample. In this study all patients were treated with anti-TNF for one year. Before and after the disease-activity-score (DAS) was assessed in all patients. Below are the results of the change in this DAS-score.

TNF-α	sample size	average change	SD
99	13	1.65	0.47
101	11	1.05	0.32
103	7	1.12	0.55
105	17	0.75	0.41
107	6	0.88	0.39
109	14	1.01	0.44
111	10	0.84	0.35
113	5	1.44	0.42
115	13	0.95	0.27
121	4	0.66	0.31

1. Comparing statistically each genotype with all others in a pairwise fashion using a fixed significance value
   A. increases the power,
   B. decreases the risk of a type-I error,
   C. increases the risk of a type-II error.

2. The best statistical test for the hypothesis of no difference between these 10 genotypic groups is
   A. the Student t-test,
   B. the chi-square test for two-by-two tables,
   C. oneway analysis of variance.

3. Suppose the above statistical test yields a p-value of less than 0.05. The investigator is interested in genotypes in which the DAS-change is extremely large or small. He decides to compare statistically all genotype groups with each other. This is best done with
   A. the Student's t-test ,
   B. the Dunnett t-test,
   C. the LSD test.

4. Another investigator, using the same data, decides that when two mean DAS-changes of two genotype-groups are not significantly different, then other genotype groups with DAS-changes lying in between, cannot be significantly either. This investigator should use
   A. the modified t-test, such as the HSD test,
   B. a multiple range test, such as of Student-Newman-Keuls,
   C. oneway analysis on selected genotype groups.

5. In a clinical trial using 5 different efficacy criteria, the bonferroni correction means that the significance level must be set to
   A. 0.05
   B. 0.025
   C. 0.01.

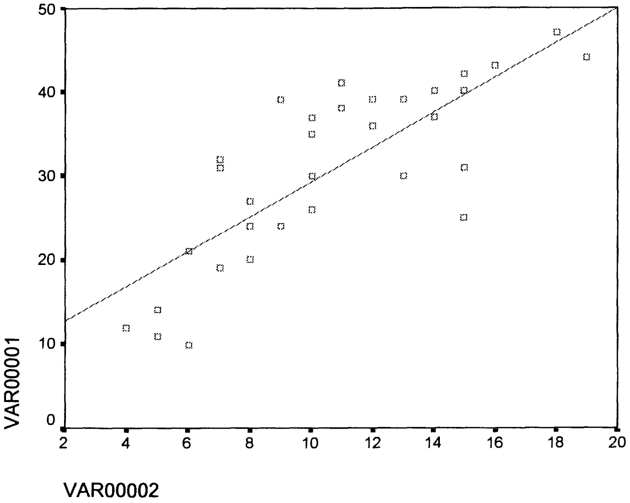
CHAPTER 9

PRINCIPLES OF LINEAR REGRESSION ANALYSIS

1. PAIRED OBSERVATIONS: REGRESSION ANALYSIS CAN BE USED FOR PREDICTING ONE OBSERVATIONS FROM ANOTHER

patient no.	new treatment	bisacodyl
	(y-variables)	(x-variables)
	(days with stool)	(days of stool)
	VAR 00001	VAR 00002
1	24	8
2	30	13
3	25	15
4	35	10
5	39	9
6	30	10
7	27	8
8	14	5
9	39	13
10	42	15
11	41	11
12	38	11
13	39	12
14	37	10
15	47	18
16	30	13
17	36	12
18	12	4
19	26	10
20	20	8
21	43	16
22	31	15
23	40	14
24	31	7
25	36	12
26	21	6
27	44	19
28	11	5
29	27	8
30	24	9
31	40	15
32	32	7
33	10	6
34	37	14
35	19	7

2. PAIRED DATA SHOULD BE FIRST PLOTTED



A linear correlation is obvious(x- variable gets larger, the y-variable gets larger). A regression line can be calculated from the data according to equation.

y=a+bx

The line drawn provides the best fit for the data given,
where y= socalled dependent, and x=independent variable, b = regression coefficient, a= intercept.

3. REGRESSION LINE, THE EQUATION

A regression line can be calculated from the data according to the equation

y=a+bx

The line drawn from this linear function provides the best fit for the data given, where y = socalled dependent, x = independent variable, b = regression coefficient.

a and b from the equation y=a+bx can be calculated.

$$b = \text{regression coefficient} = \sqrt{\frac{[\Sigma(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2}}$$

$$a = \text{intercept} = \bar{y} - b\bar{x}$$

r = correlation coefficient = is another important determinant and looks a lot like b.

$$r = \sqrt{\frac{[\Sigma(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}$$

r = measure for the strength of association between y- and x-data. The stronger the association, the better y predicts x.

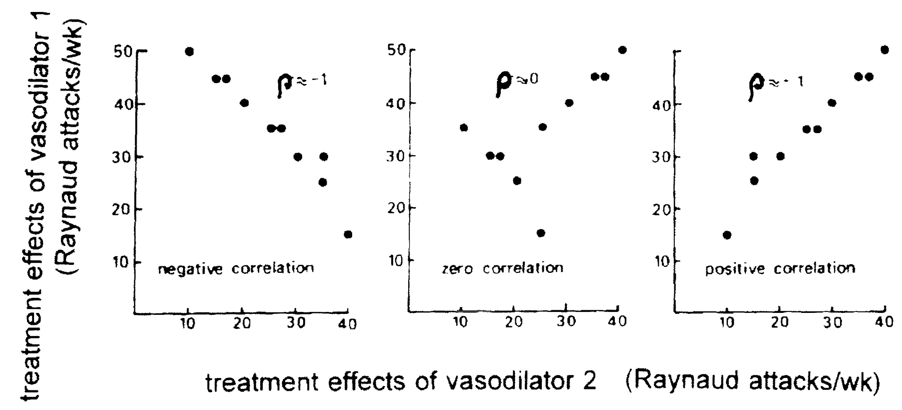
4. CORRELATION COEFFICIENT

r varies between -1 and + 1.

-Strongest association is either -1 or +1 (all data exactly on the line),

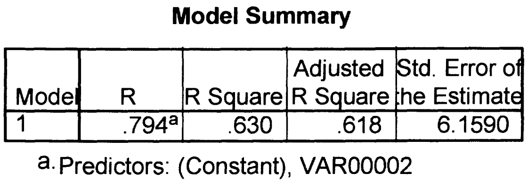
-weakest association 0 all data are parallel either to x-axis or to y-axis, or half one direction, half the other,

-for convenience standardize r by dividing it by standard deviation of y-values, then b=r.



5. USE SPSS 8 FOR WINDOWS 99 STATISTICAL SOFTWARE TO ANALYZE DATA FROM PARAGRAPH 1 (THE STOOL DATA)

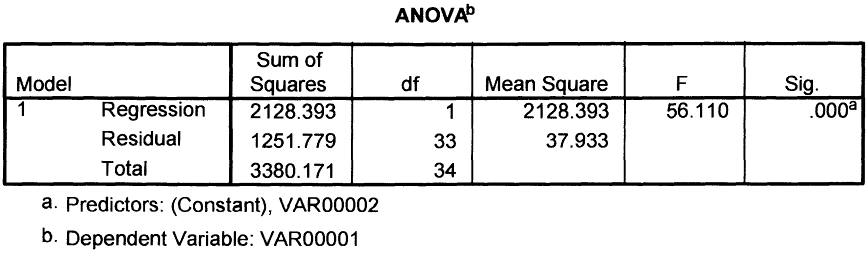
-Command: Statistics; Regression; Linear



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.794[a]	.630	.618	6.1590

[a] Predictors: (Constant), VAR00002

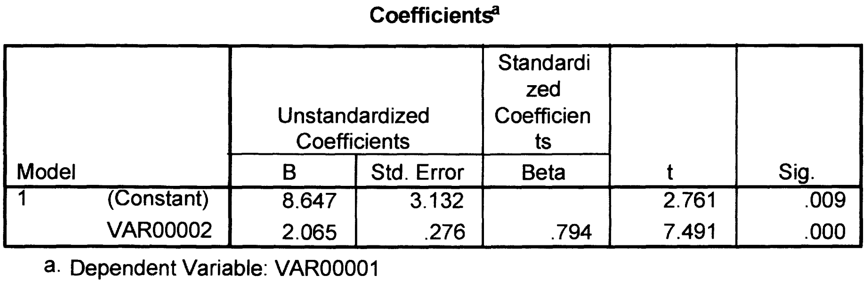


ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2128.393	1	2128.393	56.110	.000[a]
	Residual	1251.779	33	37.933		
	Total	3380.171	34			

[a] Predictors: (Constant), VAR00002

[b] Dependent Variable: VAR00001



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.647	3.132		2.761	.009
	VAR00002	2.065	.276	.794	7.491	.000

[a] Dependent Variable: VAR00001

Interpretation:

Model Summary:
0.630 ⇒63.0% of variation in y-data is explained by variation in x-data, the adjusted r-square for small samples.

ANOVA:
SS regression = $(\mathrm{SPxy})^2$) / SSx = 2128.393. SS total= SS y.
SS regression / SS total = 2128.393 / SStotal = 2128.393 / 3380.171 = 0.630 = r^2

Coefficients:
Regression equation→ new laxant = 8.647 +2.065.bisacodyl,
If SS y is defined to be 1, then r = b , t = 7.494 = $\sqrt{}$ F = $\sqrt{}$ 56.110

6. THREE COLUMNS OF PAIRED DATA INSTEAD OF TWO

patient no.	new tr y-variable	bisacodyl x_1-variable	age x_2 -variable	patient no.	new tr y-variable	bisacodyl x_1-variable	age $x_2$variable
1	24	8	25	19	26	10	27
2	30	3	30	20	20	8	20
3	25	15	25	21	43	16	35
4	35	10	31	22	31	15	29
5	39	9	36	23	40	14	32
6	30	10	33	24	31	7	30
7	27	8	22	25	36	12	40
8	14	5	18	26	21	6	31
9	39	13	14	27	44	19	41
10	42	15	30	28	11	5	26
11	41	11	36	29	27	8	24
12	38	11	30	30	24	9	30
13	39	12	27	31	40	15	20
14	37	10	38	32	32	7	31
15	47	18	40	33	10	6	29
16	30	13	31	34	37	14	43
17	36	12	25	35	19	7	30
18	12	4	24				

7. USE SPSS 8 FOR WINDOWS 99 STATISTICAL SOFTWARE TO ANALYZE THE ABOVE DATA

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.719	.701	5.4498

[a] Predictors: (Constant), VAR00003, VAR00002

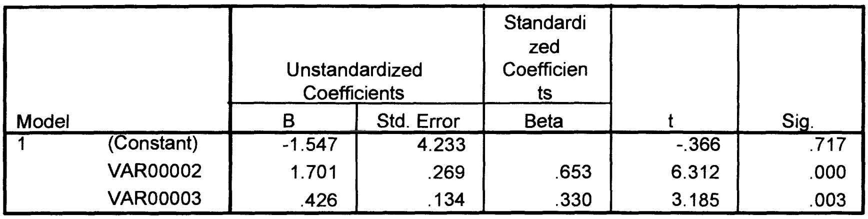
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2429.764	2	1214.882	40.905	.000[a]
	Residual	950.407	32	29.700		
	Total	3380.171	34			

[a] Predictors: (Constant), VAR00003, VAR00002

[b] Dependent Variable: VAR00001

Coefficients [a]



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.547	4.233		-.366	.717
	VAR00002	1.701	.269	.653	6.312	.000
	VAR00003	.426	.134	.330	3.185	.003

[a] Dependent Variable: VAR00001

Interpretation:
Model Summary
0.719⇒ 71.9 %variation in the y-data is explained by variation in the x-data. The adjusted r-square is for small samples. Addition of age produces 71.9 - 63.0 = 8.9% extra explanation of the variance in the y-data.

ANOVA
SS regression = (SPxy)^2 /SS x = 2429.764. SS total = Ssy. SS regression / SS total = 2429.764 / 3380.171 = 0.719 = r ^{2};

Coefficients
Regression equation→ new laxant = -1.547 + 1.701.bisacodyl, + 0.426.age. If SSy is defined 1, then r=b, both the efficacy of bisacodyl and age are significantly correlated with the efficacy of the new laxant.

8. ANOTHER EXAMPLE OF A MULTIPLE LINEAR REGRESSION MODEL

Quality of life in patients with coronary artery disease and angina pectoris has various determinants.
y-variable= quality of life
x-variables=1.Age
2.Gender
3.Rhythm disturbances
4.Peripheral vascular disease
5.Concomitant calc channel bl
6.Concomitant beta blockers
7.NYHA-classification
8.Smoking
9.body mass index
10.hypercholesterolemia
11.hypertension
12.diabetes mellitus
Index of quality of life = a + b$_1$ (age) +b$_2$ (gender) + ... b$_{12}$ (diabetes).

9. MULTICOLLINEARITY TESTING IN THE ABOVE EXAMPLE

Correlation between independent variables may be correlated but not too closely: e.g. bmi, body weight, body length should not be included all three (single linear regression is used for this purpose, r -values).

	age	gender	rhythm	vasc dis	ccb	bb	NYHA	smoking	bmi	chol	hypt
gender	0.19										
rhythm	0.12	ns									
vasc dis	0.14	ns	ns								
ccb	0.24	ns	0.07	ns							
bb	0.33	ns	ns	ns	0.07						
NYHA	0.22	ns	ns	0.07	0.07	ns					
smoking	-0.12	ns	0.09	0.07	0.08	ns					
bmi	0.13	ns	ns	ns	ns	0.10	-0.07				
chol	0.15	ns	ns	0.12	0.09	ns	0.08	0.09			
hypt	0.09	ns	0.08	ns	0.10	0.09	0.09	0.09	0.07		
diabetes	0.12	ns	0.09	0.10	ns	0.08	ns	0.11	0.12	0.10	

vasc dis= peripheral vascular disease; ccb= calcium channel blocker therapy; bb= beta-blocker therapy; bmi= body mass index; hypt= hypertension; ns= not statistically significantly correlated (Pearson correlation P-value>0.05).

10. RESULTS FROM THE ABOVE EXAMPLE

Regression Coëfficients and Standard Errors for a Multiple Linear Regression of index of quality-of-life on various concomitant variables (=covariates).

Covariate	estimated regression coëfficient	estimated standard error	test statistic (T)	Significance level (P-value)
Age	-0.03	0.04	0.8	0.39
Gender	001	0.05	0.5	0.72
Rhythm disturbances	-0.04	0.04	1.0	0.28
Peripheral vascular disease	-0.00	0.01	0.1	0.97
Calcium channel blockers	0.00	0.01	0.1	0.99
beta blockers	0.03	0.04	0.7	0.43
NYHA-classification	-0.08	0.03	2.3	0.02
Smoking	-0.06	0.04	1.6	0.08
body mass index	-0.07	0.03	2.1	0.04
hypercholesterolemia	0.07	0.03	2.2	0.03
hypertension	-0.08	0.03	2.3	0.02
diabetes mellitus	0.06	0.03	2.0	0.05

NYHA= New York Heart Association

Results nicely fit in prior hypotheses

1.Patients with angina pectoris and concomitant diabetes mellitus benefit better.

2. With cholesterolemia equally so.

3. Hypertension causes diastolic dysfunction but reduces endothelial function less so.
4. Nicotine causes vasoconstriction of resistance arteries but leaves endothelial function.
5. Mobility reducing obesitas and low NYHA give little anginal symptoms and thus little benefit.

11. CONCLUSIONS

If you are confused now by the complexity of this chapter, don't be: multiple linear regression analysis and its extensions like logistic regression and Cox's proportional hazard model is not as important for clinical trials as it is for observational research:

1. Regression analysis assesses associations not causalities.
2. Clinical trials assess causal relationships.
3. We believe in causality if factor is introduced and gives rise to a part outcome.
4. Always air of uncertainty with regression analysis

Multiple linear regression is interesting, but, in the context of clinical trials only exploratory.

12. QUESTIONS TO CHAPTER 9

1. Suppose in a multiple regression equation
   $y = 24.4 + 5.6\ x_1 + 6.8\ x_2$,
   y stands for weight (pounds)
   and x_2 for age (years. For each additional year of age, then, it can be expected that weight will increase by 24.4 pounds.

   1. Right 2. Wrong

2. A study of independent determinants of longevity provides the following results
   s = standard error = 13.4 R-square = 89.1%
   Analysis of variance

	Sums of squares(SS)	df	mean square(MS)	f	sig.
Regression	7325.33	4	1831.33	10.19	0.013
Residual	898.28	5	179.66		
Total	8223.60	9			

	Coeff	St-error	t-ratio	sig.
Constant	82.237	81.738	1.01	0.361
School	-1.553	4.362	-0.36	0.736
Age	-1.685	1.253	-1.35	0.236
Psychological score	0.110	0.291	0.38	0.720
Social score	6.876	7.658	0.89	0.410

The regression equation for the given data is

a. $y = 82.2 - 1.55\ x_1 - 1.69\ x_2 + 0.11\ x_3 + 6.88\ x_{4,}$
b. $y = 13.4 - 1.55\ x_1 - 1.69\ x_2 + 0.11\ x_3 + 6.88\ x_{4,}$
c. $y = 81.74 - 4.36\ x_1 + 1.25\ x_2 + 0.29\ x_3 + 7.66\ x_{4,}$
d. $y = 82.24 - 0.36\ x_1 - 1.35\ x_2 + 0.38\ x_3 + 0.90\ x_{4.}$

3. From question 2 how much of the variation in the longevity is explained by the regression?
   a. 94 %
   b. 82%
   c. 89 %
   d. 13 %

4. From question 2
   a. Is school an independent determinant of longevity?
   b. Is age an independent determinant of longevity?
   c. Is social score an independent determinant of longevity?
   d. Is longevity dependent on all of the x-variables?

5. From question 2, the proportion of variation in longevity explained by variation in the x-variables can be calculated from
   1. dividing SS reg by SS residual,
   2. dividing SS reg by SS total,
   3. dividing SS residual by SS total,
   4. dividing MS reg by MS residual.

6. The … test is a statistic used to test the significance of a regression as a whole.

7. In a multiple linear regression of longevity a negative regression coefficient of determinant x indicates that
   a. Longevity increases when the determinant increases,
   b. Longevity decreases when the determinant increases,
   c. None of these.

8. Signs of possible presence of multicollinearity in a multiple regression are
   a. Significant t values for the coefficients,
   b. Low standard errors for the coefficients,
   c. A sharp increase in a t value for the coefficient of an x-variable when another x-variable is removed from the model,
   d. All of the above.

9. Nine patients are tested subsequently with two different driugs for the urate-clearance-potential. Mean results are virtually the same, but also there seems to be a strong positive correlation, every time 1st drug performs well, the 2nd drug does so equally. The drugs are obviously from one and the same class. How strong does 1st drug predict the effect of 2nd drug?

patient	1st drug	2nd drug
1.	0.645	0.1117
2.	0.750	0.6296
3.	1.000	1.1475
4.	1.300	1.6654
5.	1.750	2.1833
6.	2.205	2.7013
7.	3.500	3.2192
8.	4.000	3.7371
9.	4.500	4.2550

regression analysis
s = standard error = 0.3957 $R^2 = 93.6\%$
Analysis of variance

	Sums of squares(SS)	df	mean square(MS)	f	sig.
Regression	16.094	1	16.094	102.77	0.000
Residual	1.096	7	0.157		
Total	17.191	8			

	Coeff	St-error	t-ratio	sig.
Constant	-0.4063	0.2875	-1.41	0.201
Drug 1	0.51792	0.0511	10.14	0.000

Provide the regression equation from the above data.

10. The analysis from 9 shows a strong positive , strong negative correlationship.
    1. positive 2. negative

11. From the data from question 10
    R^2 = 93.6% indicates that
    a. 93.6% of the variation in drug 2 is explained by the variation in drug 1,
    b. 93.6% of the variation in drug 1 is explained by the variation in drug 2,
    c. None of these.

12. Is the f-value from the ANOVA table identical to the squared t-ratio (drug 1) from the coefficient table?
    a. yes b. no

13. The r-square can be calculated from SS regression / Sstotal.
    a. yes b. no

CHAPTER 10

SUBGROUP ANALYSIS USING REGRESSION MODELING

1. SUBGROUP QUESTIONS

Subgroup questions include:

- Who has unusual large response? Is such occurrence associated with subgroups of patients?
- they can be used for refining patient- or dose-selection for future trials.
- Subgroup-analyses are -by nature- almost surely underpowered to definitely answer the above questions and therefore hypothesis generating rather than hypothesis confirming.

In addition to the primary questions of clinical trials, there are often many other scientific issues that may be addressed using the data sampled. Some of these issues may be completely different from the primary question of the trial, but others may be associated with it. When the trial indicates that the new treatment is significantly more effective than the standard treatment, a natural question is to wonder if specific patients may be identified for whom effectiveness is unusually larger or smaller than for others. The same applies with negative trials where no significant difference between two treatments is demonstrated. In such a case, it is often questioned whether this lack of efficacy-difference regards all of the patients or whether it results from averaging subgroups of patients that do have a real efficacy-difference from control. Questions like these are addressed in subgroup analyses. It must be stressed that subgroup analyses can only be interpreted as hypothesis-generating: they may point to specific aspects of the research-design that should be altered in a subsequent clinical trial to maximize power. Inclusion- and exclusion-criteria may have to be different from the current ones in order to better select those patients likely to have large efficacy, or to better select the most effective dosage.

Subgroups are usually defined on the basis of observable characteristics of the patients enrolled in the trial, for instance gender, age, body mass index (bmi). The basic question is then whether efficacy is different for women and men, for older and younger patients, or for patients with high and low bmi. The trial has almost surely little power to answer such questions, especially when subgroups are further refined, like "young men with high bmi". A statistical model providing better power than that of subgroup analysis is in this situation given by regression modeling. The use of the latter models, although they will not completely solve the power-problem, they suffer less from it, and, in addition, may be used to address questions on confounding, and interaction. Results of regression modeling can be used immediately to predict individual treatment-efficacy, and thus to tailor treatment to individual needs and possibilities.

2. DIFFERENT REGRESSION MODELS

Regression models: many possibilities:

- Quantitative data: linear/nonlinear regression models.
- Discrete data: (probit) logistic regression.
- Censored data: Cox regression.

The first step in subgroup analysis is the definition of treatment-efficacy. The data-nature determines the specific regression model to be used. When treatment-efficacy is defined in terms of events to occur, then the variable describing efficacy is typically dichotomous: the value 1 indicates efficacy, and the value 0 inefficacy. For dichotomous data the logistic regression model is generally used. When treatment-efficacy is scored in discrete (ordinal) categories such as "good-moderate-bad", ordinal or multinomial regression analysis is required.
When events vary in time, and the follow-up duration of the patients is variable as well, then the data are socalled censored event-times, and the usual regression model is the Cox proportional hazards regression model, or Cox model in short.
When efficacy is measured on a continuous scale, like amount of LDL-cholesterol or systolic blood pressure lowering, applicable models are the well known linear regression model, or nonlinear regression model depending on the assumed relation between efficacy and patient characteristics. In most cases, the linear model is used.

3. GENERAL FORM OF REGRESSION MODELS

General form:

$$E[Y_i|\,X_i] = g^{-1}\,(\,\beta_0 + \beta_1\,X_{1i} + \beta_2\,X\beta + \ldots + \beta_k\,X_{ki})$$

$$\mathrm{Var}[Y_i|\,X_i] = \sigma_e^{\;2}$$

Y_i is the dependent variable in the ith subject (primary efficacy variable),
X_i is a covariate, predictor or independent variable in the ith subject,
$E[Y_i|\,X_i]$ is the expected Y-value in the ith subject given the X-values in the same subject,
g^{-1} is the link-function,
β is a regression parameter, which must be estimated.

Regression models are used to estimate the expected value of the efficacy-variable E[Y_i| X_i]. So Y_i is the efficacy in patient i, and its expectation is supposed to depend on the characteristics of patient i: X_{1i} , X_{2i}, ..., X_{ki}. The dependence varies among different regression models (linear, logistic, or Cox), and is indicated by the link-function "g^{-1}". The expectation must be interpreted as follows: "take patients with characteristics x_1, x_2, ..., x_k. In such patients efficacy can be calculated to be E[Y|X]. In models for quantitative efficacy-data (linear, or nonlinear models) it is necessary to describe, in addition, how efficacy varies among patients with the same characteristics, and this is mostly assumed to be constant (Var(Y|X)=σ^2).

The efficacy-variable Y is called the dependent variable, and the patient characteristics are called independent variables, or predictors, or covariates. The parameters that will be estimated are indicated as "β" and are called regression weights, or regression parameters. Notice that when β-values are known, we can predict the expected treatment-efficacy for any patient of which we observed X_1, X_2,...,X_k -values: the linear predictor is then $\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}$.

4. LINK-FUNCTIONS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

$$P(Y_i = 1 \mid X_i) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}$$

P (Y=1/X) is the expected proportion Y given the I/X-variables.

$$S(Y_i \mid X_i) = S_0(y)^{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}$$

S(Y|X) is the expected survival Y given the X-variables.

β is a direct effect with linear regression, a log-odds-ratio with logistic regression, or a log-relative risk with Cox regression.

The link-function g^{-1} determines the nature of the regression model. For the linear model the link-function is 1, thus the mean of Y given X is modeled by: E[Y_i|X_i]= $\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}$.

For the logistic model the link-function is the logit-function= logistic function, which means that Y is transformed into ln (Y/ (1-Y)). and consequently the probability of effective treatment (Y=1) is modeled as a function of the covariates: exp(....). Exp. indicates "e to the power ...", where "e" is defined as 2.71828.... (ln(e)=1). In this logistic model the regression parameter β can be interpreted as the natural logarithm of the odds ratio.

For the Cox-model the link-function is the "log minus log"-function, and consequently the cumulative percentage of patients with effective treatment as time proceeds is modeled. In this model the regression parameter β can be interpreted as the natural logarithm of the relative risk.

5. ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

Linear regression model - assumptions:

- The relation between Y and $z= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ +e is linear , where z is the way our transformed Y-data can be written.
- Distribution of the residual e is normal.
- The residual variance is the same for all Z-variables (homoscedasticity).
- The residual e is independent.

Important assumptions when using the linear regression model are

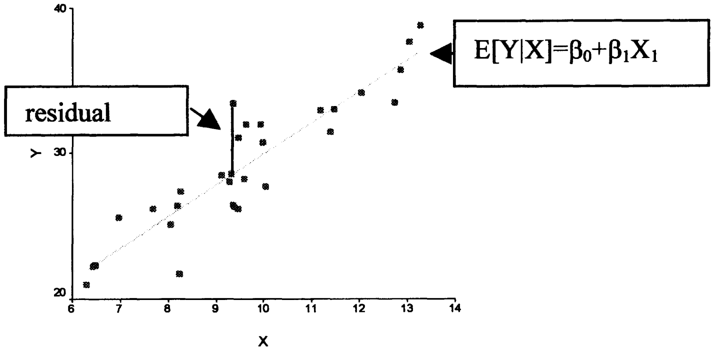
(i) linearity,
(ii) normal distribution,
(iii) homoscedasticity,
(iv) independence.

The name linear regression model already indicates that it is assumed that the relation between efficacy and patient characteristics is linear. This is obvious when a patient characteristic has only two values (e.g. gender), but need not be the case when a characteristic varies widely (e.g. age, or lipid level). The assumed linearity must be checked, and the easiest way to do so is to inspect scatter-plots of Y-values versus X-values.

A second assumption is that Y follows a normal distribution for each different combination of $X_1,X_2,...,X_k$, and it is also supposed that the variation for each different combination of $X_1,X_2,...,X_k$ is the same. This latter thing is called homoscedasticity, a pretty strong assumption in practice. Finally, the residual between the observation y_i and the expectation E[Y|X] must be independent of each other and of all covariates. This latter assumption is met in most cases. Normality can be inspected by inspecting the histogram of the residuals, and homoscedasticity by the scatter-plot of Y versus $\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_kX_{ki}$ and superimposing the regression line.

When these assumptions are violated (which to decide is often rather subjective), alternative regression models must be used. Many exist but their are statistically increasingly complex.

In the graph below a linear regression line is illustrated, as well as the key-concepts.

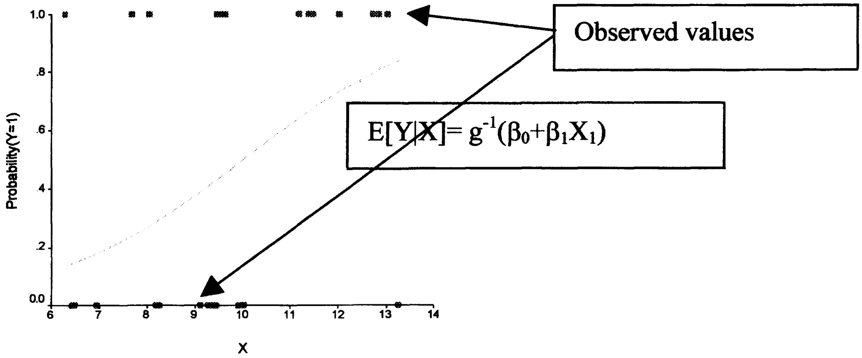


6. LOGISTIC AND COX REGRESSION MODEL

- Logistic regression model:
  relation between $P(Y=1|X)$ and X variables follow the logistic transformation.

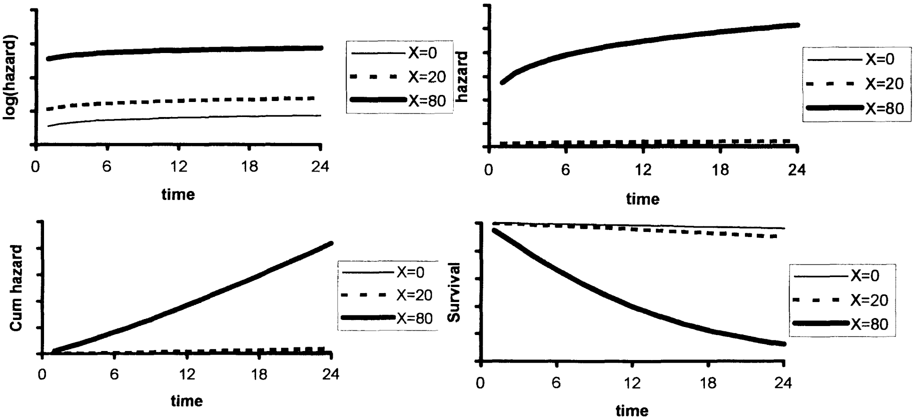
- Cox regression model:
  relation between $S(Y/X)$ and X-variables follow a log-relative risk transformation.

The basic assumption of the logistic curve is that the relation between the probability of observing an event ($Y=1$) and the combination of covariates follows the logistic transformation. In the graph below an example is given of a logistic curve. This assumption is not very strong. Indeed, the applicability of the appropriateness of the logistic model is rarely checked. One can show that the curve cannot be logistic if important covariates are not in the model, or when the functional form of covariates is wrong. Deviations from the logistic curve are the basis of socalled goodness of fit assessment (Hosmer-Lemeshow), but a good start is to make a scatter-plot.



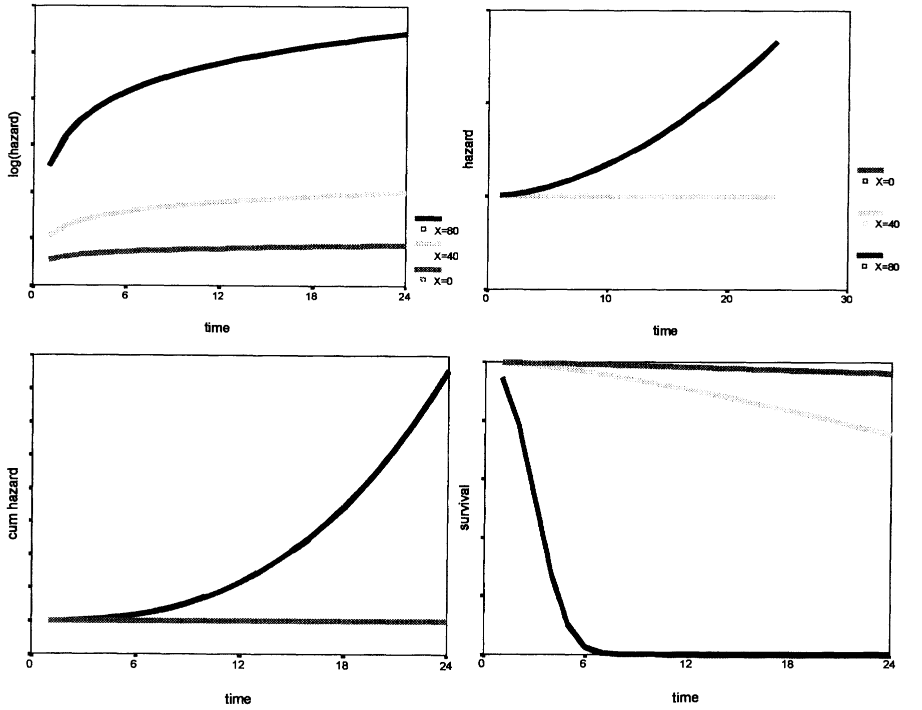
The basic assumption of the Cox regression model is that the risks are proportional over follow-up time: this means that if the risk of an event for males is R times larger than for females at month t, then it can be assumed that the same proportionality applies at month t+x. This assumption is, however, hard to check, at least not by inspecting the plots of the hazards, or survival curves.

7. AN EXAMPLE WHERE THE COX MODEL FITS



To illustrate that it is difficult to ensure the fit of the Cox model, look at the log-hazard curves (upperleft) associated with a covariate with three categories (X=0, X=20, or X=30). It is evident that the distances between the curves is constant over time, because the log-risks are increasing equally in all of the three groups. In the upperright graph the risks themselves are given, and it is clear that proportionality is hard to establish. The same applies for the cumulative hazards (lower left) and to the survival curves (lower right).

8. AN EXAMPLE WHERE THE COX MODEL DOES NOT FIT



In the upper left graph it can be clearly seen that the difference of the log-risks are not constant, thus proportionality does not apply. This cannot be concluded from the hazards themselves (upper right), or from the cumulative hazard (lower left), or from the survival curves (lower right).

Concluding: proportionality can only been seen directly in the log-hazard plot but this plot is unfortunately not readily available.

Instead, proportionality can be inspected visually by looking at so-called martingale residuals, or functions thereof., which is an increasingly important subject in current clinical research, however, far beyond the scope of this book.

In the sequel we will only discuss the linear regression model, but all of the aspects to be discussed almost without qualifications equally apply to the logistic and the Cox model.

9. INCREASING PRECISION: EXAMPLE

A real data example is given:

a parallel-group study of placebo (n=434) versus pravastatin (n=438),

- two years treatment,
- main endpoint average LDL-decrease:
  pravastatin: 1.23 (SD 0.68)
  placebo: -0.04 (SD 0.59),
- efficacy-result: 1.23 - -0.04 = 1.27
  standard error (SE)= 0.043.

The first application of regression models is the possibility to increase precision of the estimated treatment effect. Take as an example a randomized trial comparing placebo and statin treatment in 434 and 438 patients, respectively. LDL is measured before and after two years of treatment and the average decreases are

statin:	1.23 (SD 0.68)
placebo:	-0.04 (SD 0.59)

Obviously, LDL-decrease varies at lot in both treatment groups but –on average-treatment efficacy can be quantified as 1.23 – (-0.04) = 1.27. Since the patients in the two groups are independent of each other the standard error of this estimate equals $\sqrt{((0.68^2/438)+(0.59^2/434))} = 0.043$.

10. INCREASING PRECISION: USING A REGRESSION MODEL

$Y_i = \beta_0 + \beta_1 X_{1i} + e_i$
$X_1 = 1$ if a patient receives pravastatin and zero if placebo
=> β_1 is efficacy: 1.27 (SE = 0.043 is a function of σ_e^2).
Suppose there is a covariate X_2 which is related to Y, but not to X_1:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$
β_1 remains the same, but the variance σ_e^2 will become (much) smaller
=> SE(β_1) will be smaller => increased precision.

The same analysis (and results) can be obtained by using a regression model $Y_i=\beta_0+\beta_1X_{1i}+e_i$, where Y_i is the LDL-decrease of patient i (i=1,..,434+438), and X_{1i} equals 1 if patient i receives statin, or zero if patient i receives placebo. The term e_i represents the residual variation and has standard deviation σ_e.
This linear regression analysis yields an estimate of β_1 of 1.27 with standard error 0.043; hence, completely equal to the above analysis. It is important to realize that the standard error of β is a monotonic increasing function of σ_e.

Suppose further that there exists a second covariate X_2 which is related to Y, but not to X_1:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+e_i$$

In this case (X_2 not related to X_1, but related to Y), inclusion of X_2 in the regression model will not change the estimated value of β_1 but it will result in (much) smaller estimate of the residual standard deviation σ_e:

$$\begin{tikzcd} Y_i=\beta_0+\beta_1X_{1i}+\boxed{e_i} \arrow[d] \\ Y_i=\beta_0+\beta_1X_{1i}+\boxed{\beta_2X_{2i}+e_i} \end{tikzcd}$$

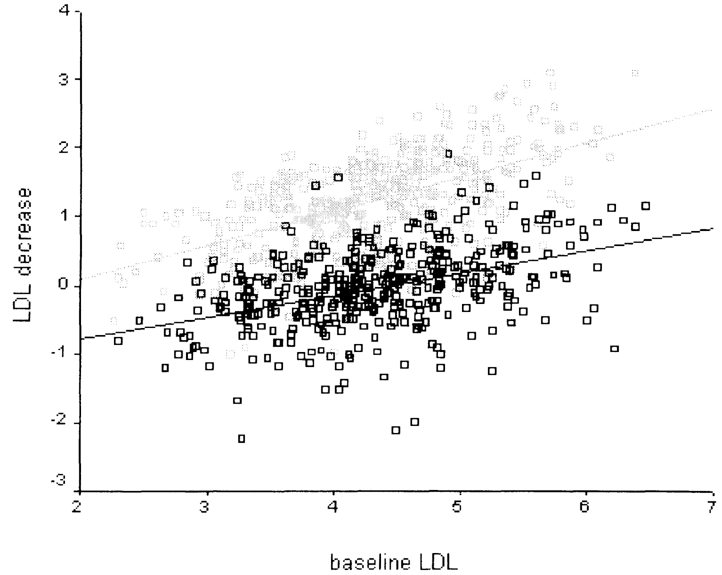
The decrease of σ_e results in a smaller standard error of β.

11. INCREASING PRECISION: EXAMPLE OF A REGRESSION MODEL

An example is baseline LDL.
Baseline LDL is not related to treatment (randomized trial).
placebo: 4.32 (SD 0.78)
pravast: 4.29 (SD 0.78) $p=0.60$
Baseline LDL is (almost surely) related to LDL-decrease.
$\beta_2 = 0.41$ (SE 0.024, $p<0.0001$)
=> efficacy: $\beta_1 = 1.27$ (SE 0.037, was 0.043).

The baseline LDL values are almost surely unrelated to treatment (in a randomized trial), but baseline levels are usually correlated to change-values (almost surely), thus baseline levels are a good example of a second covariate. In our case we find no difference between the two treatment groups on baseline levels, but a highly significant relation between baseline levels and change-values ($\beta_2=0.41$, $p<0.0001$). The estimated treatment-effect (β_1) was 1.27, the same as above, but the standard error was lowered to 0.037.

12. INCREASING PRECISION: GRAPHICAL ILLUSTRATION



The difference between the two regression line (grey and black) represents the treatment efficacy: for each level of the x-axis (baseline LDL) the average LDL-decrease is a constant amount larger in the statin (grey) group than in the placebo (black) group, even though the decrease is much less in patients with low baseline LDL than in patients with high baseline LDL.

The residual variance is a measure of the spread around the two regression lines, which is much less when conparing the spread of the black points around the black line with the spread of grey points around the grey line. This residual variance is much less than it is when considering the spread of all grey and black points together.

This decrease of the residual variance leads to smaller standard errors, by which the efficiency of the estimated treatment effect is increased. In the logistic and in the Cox models the same arguments apply, although somewhat more complicated arguments are needed due to the non-linearity of these models.

13. INCREASING PRECISION

- Usually there are many many many candidates to consider: specify which ones will be used in the protocol.
- In non-linear regression models β_1 always changes by including covariates, thus its interpretation changes (can be greatly inflated).

In most trials many baseline patient characteristics are sampled, and are candidates for inclusion in the regression model. In order to protect oneself against chance findings, one should only consider covariates that are specified in the protocol.

14. CONFOUNDING

Confounding

- can be defined a covariate Z that is associated with both Y and X_1 ,
- and troubles interpretation of the efficacy estimate β_1 .

- What is thought to be efficacy may just reflect the unbalance of Z between treatment groups.

Regression models can also be used to estimate treatment effect when there are characteristics asymmetrically distributed between treatment groups. Suppose the covariate X_1 is the covariate representing treatment (i.e. $X_{1i}=1$ if patient i receives active treatment, and $X_{1i}=o$ when patient i receives placebo). Suppose further that a covariate exists denoted by Z (e.g., gender or age) which is associated both with X_1 and with treatment-efficacy Y. Such covariate is called a confounding variable. Z will trouble unbiased estimation of the treatment-effect β_1, because part of the treatment effect may be due to the unbalance with respect to Z. For instance, when there are far more women in the active treatment group than in the placebo group, then the difference in treatment-efficacy may be due to the fact that (whatever) treatment is simply more effective in women than in men.

15. CONFOUNDING: SOME RULES

Confounding

- will not happen often in randomized trials,
- will happen always in non-randomized research.
- When it happens, adjustment of β_1 is required using the linear regression model:

  $Y_i = \beta_0 + \beta^*_1 X_{1i} + \beta_2 Z_i + e_i$

if $r_{XZ}>0$ and $r_{YZ}>0$	then	$\beta^*_1 < \beta_1$
if $r_{XZ}>0$ and $r_{YZ}<0$	then	$\beta^*_1 > \beta_1$
if $r_{XZ}<0$ and $r_{YZ}>0$	then	$\beta^*_1 > \beta_1$
if $r_{XZ}<0$ and $r_{YZ}<0$	then	$\beta^*_1 < \beta_1$.

In well-controlled trials confounding will rarely exist because all of the (baseline) patient characteristics will be equally distributed across treatment groups due to the randomization process, and this means that Z and X_1 will not be associated. Even when there is a patient characteristic that shows a significant difference between treatment groups one may consider this as a chance finding (because so many characteristics are evaluated). If an important covariate is unbalanced it may or will bias the estimated treatment-effect β_1. In non-randomized studies confounding will almost always occur.

Correction of the confounding effects of Z on the estimation of the treatment-effect β_1 may be done with regression models such as the linear regression model: $Y_i = \beta_0 + \beta_1^* X_{1i} + \beta_2 Z_i + e_i$. The estimate of β_1^* will in general be different from the estimate β_1 of the regression model without Z. The relation between β_1^* and β_1 depends on the correlations of X_1 and Z with Y (r_{xz} and r_{yz}, r=correlation coefficient): the treatment-effect can be too large, or too small when the confounding effect of Z is ignored.

16. CONFOUNDING: WARNING

- Check only the necessary (known) confounders.
- Beware of multiple testing.

In most trials there are huge numbers of covariates, all candidates for inspection of their confounding effects. This is not a sensible strategy because of chance findings due to the multiple testing problem (chapter 8). It is, therefore, sensible to specify in the protocol which covariates will be assessed.

When two or more confounders are found, the (linear) regression model can be extended with both or more confounders to remove their confounding effect. There are limitations to the number of covariates that should be incorporated in regression models; a rule of thumb suggests that at least 10 observations are required for each term in a regression model, thus if k confounders are considered and one treatment-variables one should aim for 10x(k+1) patients in the study.

17. INTERACTION/SYNERGISM

Looking for subgroups with different efficacy.

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}.X_{2i} + e_i$ Suppose X_2=0 or 1:

$X_2=1$: $Y_i = \beta_0 + (\beta_1 + \beta_3) X_{1i} + e_i$

$X_2=0$: $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$

The final use of regression models is to investigate whether treatment is equally effective in different subgroups of patients. This phenomenon is called interaction (mainly in statistical literature), or synergism in pharmacological literature.

Treatment (X_1) and a covariate X_2 are said to interact with respect to the treatment-efficacy when the regression parameter β_3 associated with the product X_1*X_2 in the regression model below is unequal to zero:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}xX_{2i} + e_i,$$

Suppose X_2 is a dichotomous covariate with values 0 or 1, the treatment-effect will be different for patients characterized by X_2=1 or X_2=0:

$$X_2=0 \quad \Rightarrow \quad Y_i = \beta_0 + \beta_1 X_{1i} + e_i,$$
$$X_2=1 \quad \Rightarrow \quad Y_i = (\beta_0+\beta_2) + (\beta_1+\beta_3) X_{1i} + e_i,$$

18. INTERACTION/SYNERGISM: EXAMPLE

Primary question: H_0: $\beta_3 = 0$

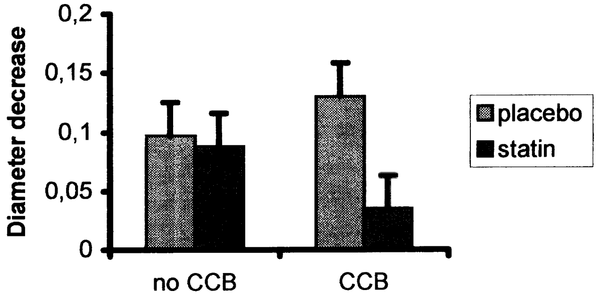
Example: is there interaction between statins and calcium channel blockers (CCBs)?

Efficacy criterion = change of diameter of coronary vessels.

placebo	no CCB	0.097 (0.20)
	CCB	0.130 (0.22)
statin	no CCB	0.088 (0.19)
	CCB	0.035 (0.19)

Take the example of a trial evaluating the effect of statins versus placebo to reduce progression of coronary artery disease as measured by the decrease of the minimum obstruction diameter of the coronary arteries. About half of the patients received a calcium channel blocker (CCB) in addition to the statin or placebo, and the question was whether effect of statins was mediated by concomitant CCB medication. The average MOD (maximal observed diameter)-decreases were 0.097, 0.13, 0.088, and 0.035 in the four groups of patients.

19. INTERACTION/SYNERGISM: GRAPHICAL ILLUSTRATION



The difference between the average MOD-decrease of the placebo- and statin-goups represent statin-efficacy, and this difference is much larger in patients who received a CCB than in patients who did not receive a CCB. This difference in efficacy can be estimated with a regression model as given above.

20. INTERACTION/SYNERGISM: GRAPHICAL ILLUSTRATION

Efficacy:
no CCB: $\beta_1 = 0.097 - 0.088 = 0.011$
CCB: $\beta_1 + \beta_3 = 0.130 - 0.035 = 0.095$
$\beta_3 = 0.095 - 0.011 = 0.084$, p=0.011

The estimates of β_1 and β_3 can be derived immediately from the means as observed:

no CCB: $\beta_1 = 0.097 - 0.088 = 0.011$
CCB: $\beta_1 + \beta_3 = 0.130 - 0.035 = 0.095 \Rightarrow \beta_3 = 0.084$.

The advantage of using the regression model instead of subgroup analysis, is the possibility to test statistically whether the estimate of β_3 differs significantly from zero. Another important advantage is that hypotheses on more complex interactions (and or confounding) can be estimated and tested simultaneously.

21. INTERACTION/SYNERGISM: WARNINGS

- Be careful investigating interactions: multiple testing problem.
- Do not enter too many covariates in a regression model: $(k<n/10)$.

Like with confounding, interactions of many many covariates with treatment may be evaluated. Since the likelihood of chance-findings is huge, it is again sensible to evaluate only those interactions that were specified a priori. Similarly, caution is needed regarding the total number of covariates in the regression model.

22. CONCLUSION

Good models:

- Check assumptions.
- Use selection algorithms sparsely.
- Instead use penalized methods, shrink regression weights.
- Caution against optimistic results: cross-validation may be helpful.

We conclude with some cautions. Application of regression models is very easy, since many computer programs are available. Successful application should

- always check fit of the regression models,
- use covariate selection sparsely,

23. QUESTIONS TO CHAPTER 10

Indicate which alternative is correct.

1. In randomized clinical trials subgroup analysis for identification of patients with unusually high or low response is most often
   A. unwanted because of bias in the data sampled,
   B. a hypothesis confirming analysis,
   C. a hypothesis generating analysis.

2. When using the standard linear regression model (Y=a+bX+e) it is assumed that the residual term of the regression model (e) is
   A. independent of the dependent variable Y,
   B. normally distributed given the independent variable X,
   C. the same for all values of the dependent variable Y.

3. When in the above linear regression model the standard deviation of e given X, S_e, is proportional to the mean of Y given X it is sensible to transform Y using
   A. the square-root transformation,
   B. the logarithmic transformation,
   C. the identity transformation.

4. When in a randomized clinical trial comparing two treatments, a linear regression model is used to evaluate the randomized treatments (X1) next to a covariate X2, and X2 is independent of X1, this analysis
   A. will correct for the confounding effect of X2,
   B. will increase precision of the effect-estimates of the treatments (X1),
   C. is pointless because X2 and X1 are independent.

5. When in the above clinical trial, the linear regression model $Y=b_0+b_1X1+b_2X2+b_3(X1.X2)+e$ is used and the regression weight b_3 is significantly larger than zero, this points to
   A. confounding of X2 on the efficacy estimate of X1,
   B. interaction between X1 and X2,
   C. the inappropriateness of the regression model for the data sampled.

CHAPTER 11

RELATIONSHIP AMONG STATISTICAL DISTRIBUTIONS

1. VARIABLES TO ASSESS CLINICAL DATA

Sample of clinical data mainly assessed through 3 variables:

1. The mean result.
2. The spread or variability of the data.
3. The sample size.

-F.e., in pharmacokinetics we want little-variability(1) in drug-levels(2) because too low not efficaceous, too high not safe, and so (1) is here more relevant than (2).
-We test variability using Chi-square-distribution.
-Chi-square distribution closely related to normal-distribution.
-It was invented by Pearson [1] 1900, 300 years after normal-distribution (De Moivre 1667-1754).
-Chi-square-distribution: heart of modern statistics (modern stats is more interested in variances of data and samples of data than in means of them).
-Chi-square provides simple device to analyze complex data: multiple groups and multivariate analyses.

2. CENTRAL TENDENCY AND SPREAD OF DATA

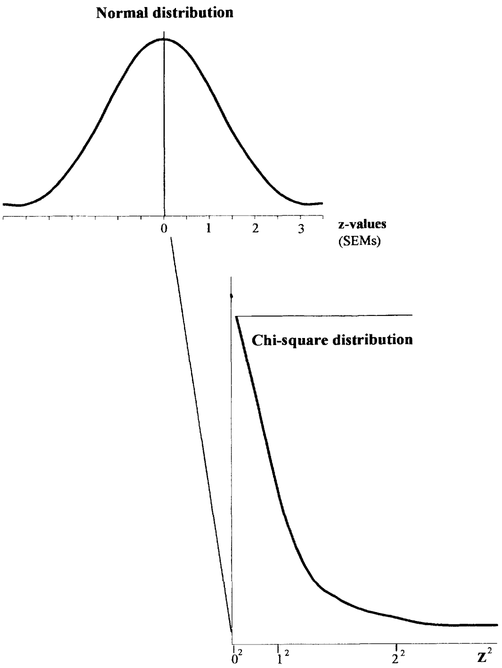
Repeated observations: Central-tendency and spread (departure from central tendency).
(Compare right and wrong bets).
The more wrong bets, the more spread in data.
We need index to estimate size of spread.

Why not: $\sum d / n$ = mean distances of our data from mean (doesn't work).
Better: $\sum d^2 / n$ = add-up sum of squared distances (works very well and is called variance of n observations.
Note: $\sum d / n$ also used to describe something else:
mean of a sample of data if d is defined from 0 rather than from mean.

-Considering the two formulas, means and variances of sample look a lot the same.
-Frequency distribution of variances nothing else than distribution of squared values of normal distribution.
-Chi-square-distribution is squared normal-distribution.

3. CREATING A CHI-SQUARE DISTRIBUTION

-Upper graph below shows a normal distribution.
-Lower graph shows the same distribution, but z-values have been squared, y-values are unchanged.
-Because z-values have been squared, we have no negative values on z-axis anymore.
-Curve is skewed to right (a socalled chi-square (χ^2) curve).
-Interpretation of skewed curve: total area under the curve (AUC) =100% of the squared data.

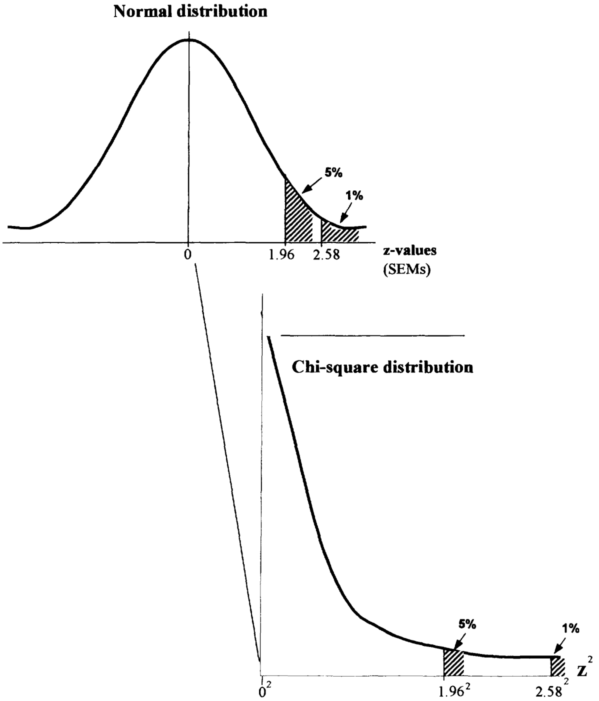


4. HOW TO USE THE SQUARED CURVE

$\Sigma d/n$ = mean = z-value upper graph.
$\Sigma d^2/n$ = variance (=sd^2) = z^2-value lower graph.

-Upper graph= frequency distribution of means, and presents mean results of many trials similar to our trial: if mean trial-result > 2 (1.96) distant from 0 → $p<0.05$, if >2.58 → $p<0.01$.

-Lower graph= frequency distribution of variances of many trials similar to our trial: if variance > 1.96^2 distant from mean →$p<0.05$, if > 2.58→$p<0.01$.



5. HOW χ^2 WORKS IN PRACTICE: 1x2 TABLE

Sleepines	No-sleepiness	Sleepines	No-sleepines
observed		expected from population	
number	number	number	number
a (5)	b (10)	α(10)	β(5)

Is our observed sample significantly different from population?

a- α = 5-10 = -5

b- β =10-5 = $\underline{+5}$ +

0

So adding up differences from expected values does not tell us.
Alternative: takes the square differences instead of differences

$(a-\alpha)^2 = 25$ divide by α to standardize = 2.5

$(b-\beta)^2 = 25$ " " β " " = 5 +

χ^2 = 7.5

Add-up sum of squared distances from supposed mean of population is larger than compatible with a χ^2 distribution, and we reject that our sample is not different from the population with $p<0.01$ (chi-square-table on page 205, 1 degree of freedom (df)).

6. HOW χ^2 WORKS IN PRACTICE: 2x2 TABLE

	Sleepiness	no-sleepiness	sleepiness	no-sleepiness
	observed		expected	
Left treatment (left group)	5 (a)	10 (b)	....(α)	(β)
Right treatment (right group)	9(c)	6 (d)	... (γ)	....(δ)

cell 1: $(O-E)^2 / E = (a-\alpha)^2 / \alpha = (5 - 14/30 \times 15)^2 \ / \ 14/30 \times 15 = ..$

2: $= (b-\beta)^2 / \beta =$

3: $= (c-\gamma)^2 / \gamma =$

4: $= (d-\delta)^2 / \delta =$ ______________________________ +

= 2.106

O= observed number;

E= expected number=(proportion sleepers /total number) x $\text{number}_{\text{group}}$

Add-up sum of squared distances from expected number = best estimate of variance of the data, and follows χ^2 distribution. With χ^2 =2.106 and 2-1=1 df → p>0.1 (chi-sqaure table on page 205).

7. WITH χ^2 WELCOME TO THE REAL WORLD OF STATISTICS BECAUSE IT CAN BE USED FOR k x 2 TABLES

	Sleepiness	no sleepiness
Group 1	5 (a)	10 (b)
Group 2	9 (c)	6 (d)
Group 3	... (e)	...(f)
Group 4	..	
Group 5		

cell a: $(O-E)^2 / E = (5 - 14/30 \times 15)^2 \ / \ 14/30 \times 15 = ..$

b: $(O-E)^2 / E$

c: $(O-E)^2 / E$

d: $(O-E)^2 / E$

e: ..

f : ..

______________________ +

$\chi^2 = ..$

The main difference from the 2x2 table test is the degrees of freedom (dfs): they are 2-1=1 with 2x2 table, and 5-1=4 with 5x2 table.

Note:- χ^2 looks weird at first.

-Main difference from normal or t-test: uses squared values.

-Basis modern statistics.

8. WHY NOT χ^2 FOR CONTINUOUS DATA

Any sample of data, either continuous or proportional, can be characterized by mean ± standard deviation (SD).

(1) SD continuous data = $\sqrt{[\frac{\sum (x-\bar{x})^2}{(n-1)}]}$ dependent on sample size,

(2) SD proportional data = $\sqrt{[p(1-p)]}$ independent of sample size.

Note: z-test = t-test but (1) replaced with (2).

Proportional data follow binomial distribution (≅ normal distribution).

Sleepines	No-sleepiness	Sleepines	No-sleepines
observed		expected from population	
number	number	number	number .
a (5)	b (10)	10 (α)	5 (β)

What is chance=probability=P of x = 5 sleepy patients with random sample of n = 15 patients?

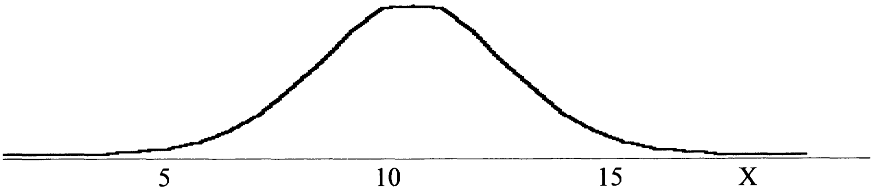
P (x=k) = n ! / x! (n-x)! (p)^x (1-p) $^{n-x}$ where <!> is faculty, e.g., 5!= 5x4x3x2x1).

P (x=5) = 15 ! / 5! (10 !) (10/15)^5 (5/15)^{10}= 0.000273

P (x ≤5) = < 0.000273= <0.001.

Mean proportion = n x p = 15 x 10/15 = 10.

SD = √[p(1-p)] independent of sample size.



Note: Why SD = √ [p(1-p)] . assume n=1 and x_1 =0 and x_2 =1.

SD^2 = $\sum[(x-\bar{x})^2 Px]$

Then SD^2 = $(x_1-\bar{x})^2 Px_1 + (x_2-\bar{x})^2 Px_2 = (0-p)^2(1-p)+(1-p)^2 p$

= p^2 (1-p) + $(1-p)^2$ p= (p+1-p) p(1-p)= p(1-p).

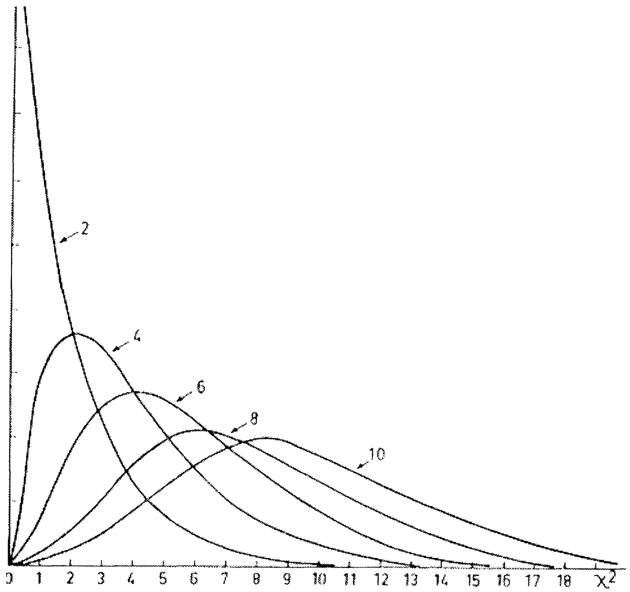
9. WHAT ARE THE ADVANTAGES OF A χ^2 -TEST COMPARED TO Z-TEST OR T-TEST

1. SD not dependent on sample size.
2. Like with variances squares are used: no-negative values.
3. In statistics variances encountered all the time (between subject, within subject, within group, between-........).
4. Mean variances of multiple samples can be simply added up.

With χ^2 we can play magic; analyses are bit messy; amazing that it works

10. WITH CHI-SQUARE WELCOME TO THE REAL WORLD OF STATISTICS

-Variances of multiple samples can be added up.
-(Variance, actually, is defined as add-up sum of squared distances from the mean).
-Add-up sum can be analyzed simultaneously.
-Possible both for continuous, proportional data, percentages, odds ratios, risk ratios et cetera.
-Only difference: breadth of the chi-square curve gets wider and wider the more samples of variances are added.



11. EXAMPLES

Example 1: (note σ = sd of population, s = sd of random sample).
Population with $\sigma^2 = 12$. We take an at random sample of 10. What is probability of variance = $s^2 > 20$? s^2 / σ^2 is standardized variation per datum. We have a sample of 10 and thus 10 data. Add-up sum provides χ^2 and equals $10 \times s^2 / \sigma^2$ or slightly better $(10-1) \times s^2 / \sigma^2$. $\chi^2 = (n-1)\ 20 / 12 = 15$ for 9 dfs → $0.05<p<0.1$.

Example 2: pill producing machine: sd of ≤ 8 mg requested,
random sample of n=25 produces s= 10 mg.
$\chi^2 = (n-1)\ 10 / 8 = 24\ .\ 100 / 64 = 37.5$ for 24 dfs $p<0.05$ Stop machine.

More examples:

1. Anxious people have high variability in performance, mean performance like non-anxious.
2. For hospitals more variability in stay-days requires more demanding care.
3. Diagnostic tests should have small test-retest variability.

12. MORE EXAMPLES, HOW TO CALCULATE 95%CONFIDENCE INTERVALS OF AN ODDS RATIO WITH UNPAIRED OBSERVATIONS

$\text{Ln OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$
where ln indicates natural or socalled Napierian logarithm.
For example

	Hypertension yes	hypertension no
Group 1	a n=5	c n=10
Group 2	b n=10	d n=5

OR=a/c / b/d = 0.25

95% confidence intervals of ln OR $= \ln \text{OR} \pm 1.96 \sqrt{(1/a+1/b+1/c+1/d)}$
$= \ln 0.25 \pm 1.96 \sqrt{(1/5+1/10+1/10+1/5)}$
$= -1.3863 \pm 1.5182$
is between – 2.905 and 0.132

95 % confidence interval of OR is between antiln –2.905 and antiln 0.132
is between 0.055 and 1.14

This confidence interval crosses 1.0 and is thus not significantly different from 1.0.

13. MORE EXAMPLES, HOW TO CALCULATE 95% CONFIDENCE INTERVALS OF AN ODDS RATIO WITH PAIRED OBSERVATIONS

Ln OR ± 1.96 $\sqrt{}$ (1/R + 1/S) where R and S are discordant pairs.

For example:
59 patients are treated with beta-blocker for angina pectoris.

	clinical benefit	no benefit
major side effects	9	2
no side effects	41	18

χ^2 -McNemar = $(18-9)^2$ / (18+9) = 81/27= 3 1 df p=0.08

OR benefit/risk = 20/50 / 11/59= 2.145
Calculation 95% confidence intervals:
ln 2.145 - 1.96 x 0.408 to ln 2.145 + 1.96 x 0.404 means from -0.123 to 1.509.
Find 95% confidence interval (CI) by taking the anti-ln 0.9-4.5 (95% CI crosses 1.0, thus p=0.08).

14. MORE EXAMPLES, HOW TO CALCULATE AND TO POOL ODDS RATIOS OF VARIOUS STUDIES (UNPAIRED DATA)

For example 4 studies assessed odds ratios of all cause deaths in patients with heart failure treated with beta-blockers

First, calculate from 95% CIs the standard error (s).
s= (ln upper value minus ln lower value)/1.96 ,
for example:
with 95% CI 0.97-1.43
s= (0.3576 minus –0.0305)/1.96= 0.1980
then s^2 = 0.0393
then $1/s^2$ = 25.510.

	OR	95% CI	lnOR	$1/s^2$	$lnOR/s^2$	$(lnOR)^2/s^2$
Waagstein	1.18	0.97-1.43	0.16	25.510	4.08	0.653
Packer	0.41	0.39-0.80	-0.89	13.33	-11.86	10.56
CIBIS	0.66	0.54-0.81	-0.42	100	-42	17.64
MERIT	0.66	0.53-0.81	-0.42	100	-42	17.64
pooled data				238.84	-91.78	46.493

Test if pooled OR is significantly different from 1.0

$$= \frac{(\ln OR_1 / s_1{}^2 + \ln OR_2 / s_2{}^2 + ..)^2}{1/s_1{}^2 + 1/s_2{}^2 ....} =$$

$$= \chi^2_{\text{pooling}} \quad \text{for 1 df}$$

$$= (-91.78)^2 / 238.84 = 39.65 \qquad p<0.0001$$

Test if heterogeneity between the studies is significantly different

$$= (\ln OR_1)^2 / s_1^2 + (\ln OR_2)^2 / s_2^2 + ... - \chi^2_{\text{pooling}}$$

$$= 46.493 - 39.65 \qquad \text{for } 4\text{-}1\text{=}3 \text{ dfs}$$

$$= 6.843 \qquad 0.05 < p < 0.10$$

Calculate pooled 95% CIs $= e^{OR \pm 1.96 / \sqrt{(1/s_1^2 + 1/s_2^2 + ...)}}$

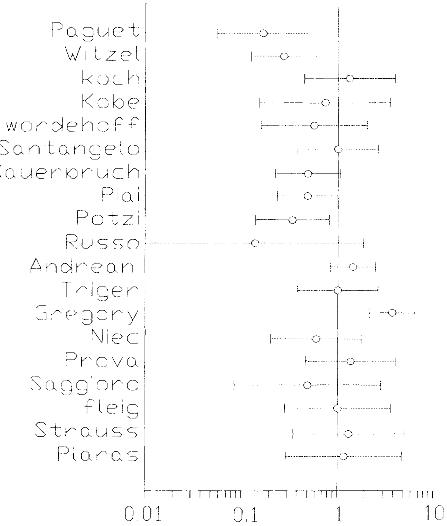
$$= e^{-91.78 / 238.84 \; \pm 1.96 / \sqrt{238.84}}$$

$$= e^{-0.3842 \pm 0.127}$$

$$= 0.68 \; (0.59\text{-}0.77)$$

significantly different from 1.0.

15. MORE EXAMPLES, HETEROGENEITY OF TRIALS IN A META-ANALYSIS



Meta-analysis of 19 trials of endoscopic sclerotherapy for esophageal bleeding is shown. On x-axis results (RR= bleeders on sclerotherapy/bleeders on sham sclerotherapy). Chi-square is calculated according to: ln RR_1/s_1^2 + ln RR_2/s_2^2 + ln RR_3/s_3^2 +..., where ln= natural logarithm is used to normalize the data and s^2 = variance of the means.

Result chi-square= 43 (18 dfs) →P<0.001.

Significant heterogeneity demonstrated, overall pooling of these data is not warranted.

16. MORE EXAMPLES, EXTENSION OF CHI-SQUARE IS F-DISTRIBUTION= DIVISION-SUM OF TWO CHI-SQUARE DISTRIBUTIONS(USED IN ANALYSIS OF VARIANCE (ANOVA))

Total variation

| |

Between-group-variation within-group-variation

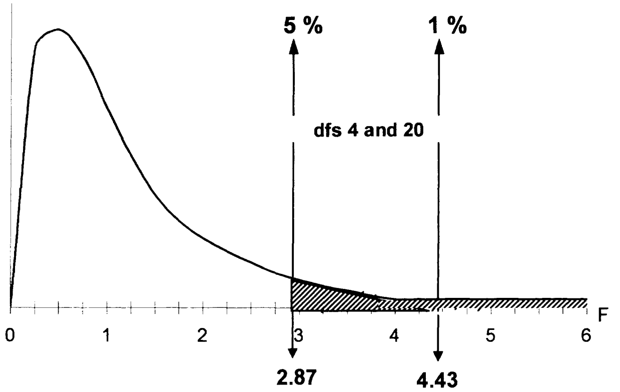
Variations are expressed as sums of squares (SS) and can be added up to obtain the total variation. We assess whether between-group-variation is large compared to within-group-variation.

Group	n patients	mean	sd
1	-	-	-
2	-	-	-
3	-	-	-

Grand mean = (mean 1 + 2 +3)/3
$SS_{\text{between groups}} = n_1$ (mean_1 – grand mean)^2 + n_2 (mean_2 – grand mean)^2 +....
$SS_{\text{within groups}}$ = $(n_1\text{-}1)(sd_1^{\,2}) + (n_2\text{-}1)\ sd_2^{\,2}$ +.....

$$F = \text{test-statistic} = \frac{SS_{\text{between groups}} / \text{dfs}^*}{SS_{\text{within groups}} / \text{dfs}}$$

The F-table gives P-value.



Degrees of freedom equals $n_1 + n_2 + n_3 + .. + n_k$ -k for SSwithin and k-1 for SSbetween.

17. LIMITATIONS OF STATISTICAL TESTS AS DISCUSSED AND CONCLUSIONS

T-test, normal-test, chi-square test, F-test, all assume that repeated observations follow normal distribution.
Mathematical formula $F(x) = 1/\sqrt{2\pi s^2}\ /\ e^{-(x-m)\ /2S}$
where s = standard deviation and m = mean value.

However:
-repeated observations in nature do not precisely follow this single formula,
-may even follow largely different patterns.
-formula approximation, amazing that it works.

-p-values for making predictions is tricky. A p-value of e.g. 0.001 means:
chance 0.001 if H0 true, chance ± 80% if H1 true.
only true if data follow normal distribution, and representative for population at large.

We wish that more often these limitations be accounted by the advocates of evidence-based medicine. If we accept the above limitations, normal distribution can be used to try and make predictions, on the understanding that statistical testing cannot give certainties, only chances.

18. QUESTIONS AND EXERCISES TO CHAPTER 11

1. A sample of clinical data is determined by the following variables
   a. mean,
   b. sample size,
   c. variance,
   d. statistical power.

   Which answer is wrong.

2. Why can variability in a sample of clinical data not be assessed as add-up sum of differences from the mean value of the sample?

3. χ^2-distribution is
   a. squared t-distribution,
   b. " normal distribution,
   c. " f-distribution,
   d. " binomial distribution.

   Which answer is correct?

4. χ^2-distribution is used to assess
   a. whether mean of data is significantly different from 0,
   b. whether variability of data is significantly different from 0,
   c. whether variability of data is significantly different from mean,
   d. whether mean of data is significantly different from baseline.

   Which answer is correct?

5. The χ^2-distribution is used to test the null-hypothesis that
   a. $\chi^2 > 1.96$ distant from 0
   b. $\chi^2 > 2.56$ distant from zero
   c. $\chi^2 > 3.84$ distant from zero
   d. $\chi^2 > 6.55$ distant from zero.

   Which answer is correct?

6. The χ^2 - test for 2x1 table tests the following data. In a sample from given population 60 patients have a borderline hypertension, 100 patients have not. From epidemiological surveys we know that the population distribution is 40 versus 120. Is this sample significantly different from the given population?

7. The χ^2 - test for 2x2 table is used to test whether two groups are significantly different from one another.

	hypertension	
	yes	no
Group 1	a n=60	c n=40
Group 2	b n=100	d n=120

8. The χ^2 - test for 3x2 table is used to test whether three groups are significantly different from one each other.

	hypertension	
	yes	no
Group 1	a n= 60	d n= 40
Group 2	b n=100	e n= 120
Group 3	c n= 80	f n= 60

9. The χ^2 - test for 2x3 table is used to test whether two groups are significantly different from one each other.

	hypertension		
	yes	no	don´t know
Group 1	a n=60	c n= 40	e n=60
Group 2	b n=50	d n= 60	f n=50

10. A pill should have a diameter of 6 mm with a standard error (SE) of 0.5 mm. We test and find an SE of 0.9. Is our result significantly different from the required SE?

11. Four studies assessed odds ratios of sudden deaths in patients with heart failure treated with beta-blockers. Odds ratios (ORs) of sudden death are given. Pool the data and test heterogeneity. Also calculate the pooled OR and 95% CIs (s is standard error).

	OR	95% CI	lnOR	$1/s^2$	$lnOR/s^2$	$(lnOR)^2/s^2$
Waagstein	1.46	0.72-2.68	0.378	9.26	3.037	0.996
Packer	0.45	0.21-0.99	-0.796	6.67	-5.31	4.226
CIBIS	0.58	0.40-0.83	-0.545	31.25	-17.03	9.280
MERIT	0.59	0.45-0.77	-0.528	35.71	-18.85	5.255
pooled data				82.29	-38.153	19.757

12. Four studies assessed odds ratios (ORs) of deaths due to progressive heart failure in patients treated with beta-blockers. Odds ratios of sudden death are given. Pool the data and test heterogeneity. Also calculate the pooled OR and 95% CIs (s is standard error).

	OR	95% CI	lnOR	$1/s^2$	$lnOR/s^2$	$(lnOR)^2/s^2$
Waagstein	1.03	0.42-2.53	0.0296	5.00	0.148	0.004
Packer	0.20	0.076-0.64	-1.514	3.57	-5.40	8.183
CIBIS	0.92	0.724-1.20	-0.083	71.4	-5.93	0.492
MERIT	0.50	0.33-0.79	-0.0693	3.33	-0.231	0.016
pooled data				83.3	-11.413	8.695

CHAPTER 12

STATISTICS IS NOT BLOODLESS ALGEBRA

1. BIOLOGICAL PROCESSES ARE FULL OF VARIATIONS

Statistics can not give certainties, only chances.
Chances that prior hypotheses true/untrue.
The human brain hypotheses all the time, may be untrue.
In clinical medicine: hypotheses may be untrue, must be assessed with hard data.
When statistics comes in, many a clinician becomes nervous.

Clinicians leave data to statistician.
Statistician runs data through SAS [2001, Chicago, IL] or SPSS [2001, New York, NY] who sees if any significancies: scenario bad practice, kills data.
Biostatistics can do more than provide irrelevant P-values.

2. STATISTICS IS FUN FOR CLINICAL INVESTIGATORS

- Is not maths.
- Proofs prior hypothesis.
- Discipline at interface of biology and maths.
- Maths used to answer biological questions.
- Above scenario does not answer reasonable biological questions, is data dredging.
- Source of lot of misinterpretations in clinical medicine.
- Statistical analysis: confine to prior hypotheses.
- Problem with multiple tests like gambling:

  20 times with chance 5%. After game $(1\text{-}0.05)^{20} = (0.95)^{20} = 0.36 = 36\%$ chance prize. Result not based on significant effect but play of chance.

3. USE SIMPLE TESTS

- Statistical result not confirming prior belief, don't trust.
- Simplest univariate test adequate for data.
- Fancy multivariate procedures not in place.
- Statistics confirms prior hypotheses.
- Appropriate because based on sound arguments.
- If not, find out why: imperfections in design or execution?

Another fun thing with statistics, although not as important, method of secondary analyses: it proves nothing, kind of sports, new ideas.

4. STATISTICAL PRINCIPLES IMPROVE QUALITY OF TRIAL

(1) Take care of symmetries in data,
(2) emphasis on statistical power,
(3) assess why drug works,
(4) accounting Type I, II, III errors,
(5) weighing benefits drug against risks.

5. STATISTICS PROVIDES EXTRAS

Parallel designs cannot:
(1) manage multimodal therapies,
(2) manage historical data,
(3) manage ethics and efficacy during long-term trials,
(4) study drugs, before toxicity information is available,
(5) account therapeutic equivalence,
(6) study multiple treatments/groups,
(7) adjust baseline levels.

6. STATISTICS PROVIDES EXTRAS, SPECIAL DESIGNS CAN MANAGE WHAT PARALLEL DESIGNS CANNOT

Special designs for such purposes:
(1) factorial design,
(2) historical controls design,
(3) group-sequential interim analysis design,
(4) sequential design for continuous monitoring,
(5) therapeutic equivalence design,
(6) multiple crossover-periods / multiple parallel-groups design,
(7) multivariate adjustment for age, gender, baseline differences.

7. FOR EXAMPLE, INTERIM ANALYSES

Goals:
(1) ethical concern,
(2) financial concern,
(3) check prior assumptions (sample size, expected efficacy, expected safety).

Problems:
(1) risk of finding differences by chance if you test many times (type I error),
(2) bias due to unblinding the interim result.

Rules:
(1) 1 variable,
(2) 1 or 2 interim analyses,
(3) predefined stopping rules,
(4) only if enough patients are included,
(5) performed by independent investigators,
(6) results must be as confidential as possible,
(7) adjust p-values ($p<0.01$ is safe advise).

Special form of interim analysis: continuous monitoring.
(1) Recalculate result-so-far after each new patient.
(2) Provide a-priori-stopping-boundaries.
(3) For the benefit of early studies, before toxicity information.
(4) Stop study any time.

8. STATISTICS IS NOT LIKE ALGEBRA AND REQUIRES BIOLOGICAL THINKING AND JUST A BIT OF MATHS

(1) Mathematically: representative samples required, biologically the first datum in complete ignorance greatest information (first case of disease great deal of information).

(2) Flexible alpha and beta required. When false+ is worse for patient than false-, respectively 5% and 20% levels are okey. With life-threatening diseases better reduce beta-level to 10% or less..

(3) Include " safety factor" with sample size. Sample size is based on pilot data or expectations. To reduce risk of type I/II make sample size larger, e.g., 10 % larger than required.

9. STATISTICS TURNS ART INTO SCIENCE

- Science of medicine consists of experiments.
- Art of medicine trust, sympathy, the threatened patient.
- Science of medicine estimated by statistical methods.
- Psychosocial and personal factors difficult to measure.
- Last 5 years quality-of-life assessments produce reproducible results and turns art into science

10. STATISTICS FOR SUPPORT RATHER THAN ILLUMINATION?

-1948 first randomized controlled trial [Medical Research Council 1948].
-Until then, observations uncontrolled.
-Initially (1) trials frequently negative,
(2) little sensitivity due to small samples,
(3) inappropriate hypotheses based on biased prior data.
-Subsequently flaws recognized and accounted
(1) interaction,
(2) time effects,
(3) negative correlations,
(4) asymmetries.

Now clinical trials rarely-negative/ rather confirmational.

Clinicians used to apply statistics as a drunk uses a lantern standard, for support rather than illumination. Not anymore. Statistics is now an important help to reliable conclusion.

11. STATISTICS HELPS CLINICIANS TO BETTER UNDERSTAND THE LIMITATIONS OF RESEARCH

- Medical literature snowed under with mortality trials.
- Invariably 10-30 % relative rise survival.
- Mortality important endpoint, may be so.
- Yet, a relative rise survival of 30 % means that your risk of death goes from 3 to 2% or less.
- Mortality is insensitive variable of preventive medicine begun at middle-age.
- Background noise associated senescence then high. More sensitive endpoint then morbidity.

Notes:
(1) Patients prefer better quality of life and reduced morbidity instead of 1-2 % increased survival in return for long-term-drug-treatment with side effects.
(2) Relative risk reductions are often overinterpreted in publications as though they were absolute risk reductions.
(3) So are underpowered P-values: 0.05 means chance of type I error 5%, type II error of 50%.

12. LIMITATIONS OF STATISTICS

(1) Type I / II errors,
(2) little clinical significance of statistically significant data,
(3) statistics gives no certainty but predicts a chance under the understanding that:
- -H0 is true
- -H1 is true
- -data follow a particular normal distribution
- -data are representative

Statistics leaves a lot of uncertainty, and, correspondingly, evidence-based medicine does so.

13. CONCLUSIONS

1. Statistics fun for clinical investigator, confirms hypotheses.
2. Accounting statistical principles helps reduce imperfections.
3. Getting command of non-classical study designs provides extras.
4. Statistics not like algebra, requires biological thinking and bit of maths.
5. Statistical analyses can be performed on quality of life.
6. Clinical investigator must know what statistics cannot answer.
7. Statistics helps to interpret limitations clinical research.
8. Statistics has limitations of its own: gives only chances, does not automatically indicate clinical relevance, can not test every possible bias.

Not being familiar with statistics raises a two-way risk: you're not only missing the benefit of it but also fail to adequately recognize its limitations. We hope that this book will be an incentive for participants to improve statistical skills in order to better understand the statistical data of others and themselves.

14. QUESTIONS TO CHAPTER 12

1. What alternative is correct?
   A . Statistics provides certainties
   B. Statistics provides hypotheses.
   C. Statistics provides chances.
   D. Statistics provides hard data.
   E. Statistics tests hard data.

2. What do we mean by data dredging?
   A. Ask a statistician to explore the data for any significances.
   B. Multiple posterior hypothesis testing for explorative purposes.
   C. Look into the data for any corrrelations and trends for confirmational purposes.

3. What is the main statistical problem with multiple testng?
   A. The chance of type I errors is increased.
   B. You have to test without a null-hypothesis
   C. You have to test with an alternative hypothesis.

4. Secondary analyses are:
   A. for exploring new hypotheses,
   B. for confirmation of primary endpoints,
   C. more reliable than primary analyses because instead of univariate they use multivariate assessments,
   D. require a study with larger sample sizes than do primary analyses.

5.How can we weigh efficacy versus safety?
   A. Division sum of scores.
   B. Product of scores.
   C. Subtraction sum of scores.
   D. Add-up sum of scores.

6. For a multimodal therapy, what study design is adequate?
   A. Factorial design.
   B. Historical control design.
   C. Group-sequential interim analysis design.
   D. Equivalent study design.
   E. None of these.

7. For a efficacy study prior to availibility of toxic data, what design is required?
   A. Factorial design.
   B. Historical control design.
   C. Group-sequential interim analysis design.
   D. Equivalent study design.
   E. None of these.

8. For a crossover study with multiple periods what study design is required?
   A. Factorial design.
   B. Historical control design.
   C. Group-sequential interim analysis design.
   D. Equivalent study design.
   E. None of these.

9. Interim analyses are for:
   A. ethical reasons,
   B. financial reasons,
   C. scientific reasons,
   D. all of the three above reasons.
   What answer is the most adequate one?

10. Interim analyses should include:
    A. more than one variable,
    B. more than one interim analysis per trial,
    C. analysis by a dependent group,
    D. a priori defined stopping rules.

11. Flexible alpha and beta: for fatal disease and non-toxic compound choose:
    A. 1-beta = 70%,
    B. 1-beta = 80%,
    C. 1-beta = 90%.

12. Flexible alpha and beta: for non-fatal disease and toxic compound choose:
    A. 1-beta = 70%,
    B. 1-beta = 80%,
    C. 1-beta = 95%.

13. How can the art of medicine be changed into the science of medicine?
    A. By quality of life assessments.
    B. By flexible alphas.
    C. By including a safety factor in a sample size computation.
    D. By assessment of interaction between patients and their physicians.

14. The first randomized controlled trials in the 50ths were often negative because of:
    A. Unrepresentative samples.
    B. Wrong prior hypotheses.
    C. Biases.
    D. Inappropriate statistics.

    What statement is untrue.

CHAPTER 13

BIAS DUE TO CONFLICTS OF INTERESTS, SOME GUIDELINES

1. INTRODUCTION

The controlled clinical trial, the gold standard for drug development, is in jeopardy. The pharmaceutical industry rapidly expands its commend over clinical trials. Scientific rigor requires independence and objectivity. Safeguarding such criteria is hard with industrial sponsors, benefiting from favorable results, virtually completely in control. The recent Good Clinical Practice Criteria adopted by the European Community[1] were not helpful , and even confirmed the right of the pharmaceutical industry to keep everything under control. Except for the requirement that the trial protocol should be approved by an external protocol review board, little further external monitoring of the trial is required in Europe today. The present paper was written to review flawed procedures jeopardizing the credibility of current clinical trials, and to look for possible solutions to the dilemma between sponsored industry and scientific independence.

2. THE RANDOMIZED CONTROLLED CLINICAL TRIAL AS THE GOLD STANDARD

Controlled clinical trials began in the UK with James Lind, on H.M.S. Salisbury, a royal Frigate, by the end of the 18th century. However, in 1948 the first randomized controlled trial was actually published by the English Medical Research Council in the British Medical Journal.[2] Until then, published observations had been uncontrolled. Initially, trials frequently did not confirm hypotheses to be tested. This phenomenon was attributed to little sensitivity due to small samples, as well as inappropriate hypotheses based on biased prior trials. Additional flaws were being recognized and, subsequently were better accounted for: carryover effects due to insufficient washout from previous treatments, time effects due to external factors and the natural history of the condition under study, bias due to asymmetry between treatment groups, lack of sensitivity due to a negative correlation between treatment responses etc. Such flaws mainly of a technical nature have been largely implemented and lead to trials after 1970 being of significantly better quality than before. And so, the randomized clinical trial has gradually become accepted as the most effective way of determining the relative efficacy and toxicity of new drug therapies. High quality criteria for clinical trials include clearly defined hypotheses, explicit description of methods, uniform data analysis, but, most of all, a valid design. A valid design means that the trial should be made independent, objective, balanced, blinded, controlled, with objective measurements. Any research but, certainly, industrially-sponsored drug reseach where sponsors benefit from favorable results, benefits from valid designs.

3. NEED FOR CIRCUMSPECTION RECOGNIZED

The past decade focused, in addition to technical aspects, on the need for circumspection in planning and conducting clinical trials.[3] As a consequence, prior to approval, clinical trial protocols started to be routinely scrutinized by different circumstantial organs, including ethic committees, institutional and federal review boards, national and international scientific organizations, and monitoring committees charged with conducting interim analyses. And so things seems to be developing just fine until something else emerged, the rapidly expanding commend of the pharmaceutical industry over clinical trials. Scientific rigor requires independence and objectivity of clinical research, and safeguarding such principles is hard with sponsors virtually completely in control.

4. THE EXPANDING COMMEND OF THE PHARMACEUTICAL INDUSTRY OVER CLINICAL TRIALS

Today megatrials are being performed costing billions of dollars paid by the industry. Clinical research has become fragmented among many sites, and the control of clinical data often lies exclusively in the trial sponsor's hands.[4] A serious issue to consider here are adherence to scientific criteria like objectivity, and validity criteria like blindness during the analysis phase. In the USA, the FDA audits ongoing registered trials for scientific validity. However, even on-site-audits can hardly be considered capable of controlling each stage of the trial. Not any audits are provided by the FDA′s European counterparts. Instead, in 1991, the European Community endorsed the Good Clinical Practice (GCP) criteria developed[1] as a collaborative efforts of governments, industries, and the profession. For each of the contributing parties benefits are different. Governments are interested in uniform guidelines and uniform legislation. For the profession the main incentives are scientific progress, and the adherence to scientific and validity criteria. In contrast, for the pharmaceutical industry a major incentive is its commercial interest. And so, the criteria are, obviously, a compromise. Scientific criteria like clearly defined prior hypotheses, explicit description of methods, uniform data analyses are broadly stated in the guidelines given.[1] However, scientific criteria like instruments to control independence and objectivity of research are not included. Validity criteria like control groups and blinding are recognized, but requirements like specialized monitoring teams consistent of a group of external independent investigators guiding such criteria, and charged with interim analysis and stopping rules are not mentioned. And so, the implementation of the Good Clinical Practice Criteria are not helpful for the purpose of safeguarding scientific independence. Instead, they confirmed the right of the pharmaceutical industry to keep everything under control.

5. FLAWED PROCEDURES JEOPARDIZING CURRENT CLINICAL TRIALS

Flawed procedures jeopardizing current clinical trials can be listed as follows. Industries, at least in Europe, are allowed to choose their own independent protocol review board prior to approval. Frequently, a pharmaceutical company chooses one-and-the-same-board for all of its (multicenter) studies. The independent protocol review board may approve protocols, even if the research is beyond its scope of expertise, for example, specialized protocols like oncology-protocols without an oncologist among its members. Once the protocol is approved, little further external review is required in Europe today. Due to recent European Community Regulations, health facilities hosting multicenter trials are requested to refrain from scientific or ethic assessment. Their local committees may assess local logistic aspects of the trial but no more than that. And so, the once so important role of local committees to improve the objectivity of sponsored research is minimized. Another problem with the objectivity of industrially-sponsored clinical trials is the fact that the trial monitors are often employees of the pharmaceutical industry. Furthermore, data control is predominantly in the hands of the sponsor. Interim analyses are rarely performed by independent groups. The scientific committee of the trial consists largely of prominent but otherwise uninvolved physicians attached to the study, the socalled *guests*. Analysis and report of the trial is generally produced by clinical associates at the pharmaceutical companies, the socalled *ghosts*, and, after a brief review, co-signed by prominent physicians attached to the study the socalled *graphters*.

Table Flawed procedures jeopardizing current clinical trials.

1. Pharmaceutical industries, at least in Europe, are allowed to choose their own independent review board prior to approval.
2. the independent protocol review board approves protocol even if the research is beyond the scope of its expertise.
3. Health facilities hosting multicenter research are requested to refrain from ethic or scientific assessment after approval by the independent review board.
4. Trial monitors are often employees of pharmaceutical industry.
5. Data control is predominantly in the hands of the sponsor.
6. Interim analyses are rarely peformed by independent groups.
7. The scientific committee of a trial consists largely of guests (names of prominent physicians attached to the study) and graphters (for the purpose of giving the work more impact).
8. The analysis and report is produced by *ghosts* (clinical associates at the pharmaceutical companies) and is after a brief review co-signed by the *guests* and *graphters*.

6. THE GOOD NEWS

The Helsinki guidelines rewritten in the year 2000 have been criticized[5] for its incompleteness regarding several ethical issues, e.g., those involving developing countries. However, these independently written guidelines also included important improvements. For the first time the issue of conflict of interests has been assessed in at least 5 paragraphs. Good news is also the American FDA's initiative to start auditing sponsored trials on site. In May 1998 editors of 70 major journals have endorsed the Consolidated Standards of Reporting Trials Statement (CONSORT) in an attempt to standardize the way trials are conducted, analyzed and reported. The same year, the Cochrane Collaborators together with the British journals The Lancet and The British Medical Journal have launched the "Unpublished Paper Amnesty Movement", in an attempt to reduce publication bias.There is also good news from the basis. E.g., in 30 hospitals in the Netherlands local ethic committees, endorsed by the Netherlands Association of Hospitals, have declared that they will not give up scrutinizing sponsored research despite approval by the independent protocol review board.

In our educational hospital house officers are particularly critical of the results of industrially-sponsored research even if it is in the Lancet or the New England Journal of Medicine, and they are more reluctant to accept results not fitting in their prior concept of pathophysiology, if the results are from industrially-sponsored research. Examples include: ACE-inhibitors for normotensive subjects at risk for cardiovascular disease (HOPE Study[6]), antihypertensive drugs for secondary secondary prevention of stroke in elderly subjects (PROGRESS Study[7]), beta-blockers for heart failure (many sponsored studies, but none of them demonstrating an unequivocal improvement of cardiac performance[8]), cholesterol-lowering treatment for patients at risk of cardiovascular disease but normal LDL-cholesterol levels (Heart Protection Study), hypoglycemic drugs for prediabetics (NAVIGATOR Study). As a matter of fact, all of the above studies are based on not so sensitive univariate analyses. When we recently performed a multivariate analysis of a secondary prevention study with statins, we could demonstrate that patients with normal LDL-cholesterol levels did not benefit.[9]

7. FURTHER SOLUTIONS TO THE DILEMMA BETWEEN SPONSORED RESEARCH AND THE INDEPENDENCE OF SCIENCE

After more than 50 years of continuous improvement, the controlled clinical trial has become the most effective way of determining the relative efficacy and toxicity of new drug therapies. This gold standard is, however, in jeopardy due to the exanding commend of the pharmaceutical industry. Mega-trials are not only paid for by the industry but also designed, carried-out, and analyzed by the industry. Because objectivity is at stake when industrial money mixes with the profession[9] it has been recently suggested to separate scientific research and the pharmaceutical industry. However, separation may not be necessary, and might be counterproductive to the progress of medicine. After all, pharmaceutical industry has deserved substantial credits for developing important medicines, while other bodies including governments have not been able to develop medicines in the past 40 years, with the exception of one or two vaccines. Also, separation would mean that economic incentives are lost not only on the part of the industry but also on the part of the profession while both are currently doing well in the progress of medicine. Money *was* and *is* a major motive to stimulate scientific progress. Without economic incentives from industry there may soon be few clinical trials. Circumspection from independent observers during each stage of the trial has been recognized as an alternative for increasing objectivity of research and preventing bias.[3] In addition, tight control of study data, analysis, and interpretation by the commercial sponsor is undesirable. It not only raises the risk of biased interpretation, but also limits the opportunities for the scientific community to use the data for secondary analyses needed for future research.[4] If the pharmaceutical industry allows the profession to more actively participate in different stages of the trial, scientific research will be better served, and reasonable biological questions will be better answered. First on the agenda will have to be the criteria for adequate circumspection (Table underneath). Because the profession will be more convinced of its objective character, this allowance will not be counterproductive to the sales. Scientific research will be exciting again, confirming prior hypotheses, and giving new and sound ideas for further research.

Table Criteria for adequate circumspection.

1. Disclosure of conflict of interests and the nature of it from each party involved
2. Independent ethical and scientific assessment of the protocol
3. Independent monitoring of the conduct of the trial
4. Independent monitoring of data management
5. Independent monitoring of statistical analysis including the cleaning-up of the data
6. The requirement to publish even if data do not fit in the commercial interest of the sponsor.
7. Requirement that interim analyses be performed by an independent group.

8. REFERENCES

1. Anonymous 1997 International guidelines for good clinical practice. Ed: NEFARMA (Netherlands Association of Pharmaceutical Industries), Utrecht, Netherlands
2. Medical Research Council 1948 Streptomycin treatment of pulmonary tuberculosis. Br Med J 2: 769-782
3. Cleophas TJ, Zwinderman AH, Cleophas TF 2002 Statistics applied to clinical trials, second edition. Kluwer Academic Publishers, Boston, USA
4. Montaner JS, O'Shaughnessy MV, Schechter MT 2001 Industry-sponsored clinical research: a double-edged sword. Lancet 358: 1893-1895
5. Diamant JC 2002 The revised Declaration of Helsinki - is justice served. Int J Clin Pharmacol Ther 40: 76-83
6. Sleight P, Yusuf S, Pogue J, Tsuyuki R, Diaz R, Probsfield J 2001 Blood pressure reduction and cardiovascular risk in HOPE Study. Lancet 358: 2130-2131
7. 7.PROGRESS Collaborative Group 2001 Randomised trial of a perindopril-based blood-pressure lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. Lancet 358: 1033-1041
8. Meyer FP, Cleophas TJ 2001 Meta-analysis of beta-blockers in heart failure. Int J Clin Pharmacol Ther 39: 561-563 and 39: 383-388
9. Cleophas TJ, Zwinderman AH 2002 Efficacy of HMG-CoA reductase inhibitors dependent on baseline cholesterol levels, secundary analysis of the Regression Growth Evaluation Statin Study (REGRESS). Br J Clin Pharmacol: accepted for publication
10. Relman AJ, Cleophas TJ, Cleophas GI 2001 The pharmaceutical industry and continuing medical education. JAMA 286: 302-304

Statistical Tables

T-Table: v = degrees of freedom fot t-variable, Q = area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

v	*Q* = **0.4**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
	I2Q = **0.8**	**0.5**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.002**
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

Chi-square distribution

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

F-distribution

df of denominator	2-tailed *P*-value	1-tailed *P*-value	Degrees of freedom (*df*) of the numerator												
			1	2	3	4	5	6	7	8	9	10	15	25	500
1	0.05	0.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3	968.6	984.9	998.1	1017.0
1	0.10	0.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	249.3	254.1
2	0.05	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.46	39.50
2	0.10	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.49
3	0.05	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.12	13.91
3	0.10	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.53
4	0.05	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.50	8.27
4	0.10	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.64
5	0.05	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.27	6.03
5	0.10	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.52	4.37
6	0.05	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.11	4.86
6	0.10	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.68
7	0.05	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.40	4.16
7	0.10	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.24
8	0.05	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	3.94	3.68
8	0.10	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	2.94
9	0.05	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.60	3.35
9	0.10	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.72
10	0.05	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.35	3.09
10	0.10	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.55
15	0.05	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.69	2.41
15	0.10	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.28	2.08
20	0.05	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.40	2.10
20	0.10	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.86
30	0.05	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.12	1.81
30	0.10	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.64
50	0.05	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.92	1.57
50	0.10	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.46
100	0.05	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.77	1.38
100	0.10	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.31
1000	0.05	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.85	1.64	1.16
1000	0.10	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.52	1.13

Paired non-parametric test: Mann-Whitney test, the table uses smaller of the two ranknumbers

N pairs	P<0.05	P<0.01
7	2	0
8	2	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19

Unpaired non-parametric test: Wilcoxon rank sum test. Table uses difference of added up rank numbers between group 1 and group 2

P<0.01 levels

n_1→ n_2 ↓	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	.81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7													

Unpaired non-parametric test: Wilcoxon rank sum test. Table uses difference of added up rank numbers between group 1 and group 2

P<0.05 levels

n_1 → n_2 ↓	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

ANSWERS TO QUESTIONS AND EXERCISES

CHAPTER 1 / INTRODUCTION TO THE STATISTICAL ANALYSIS OF CLINICAL TRIALS, CONTINUOUS DATA ANALYSIS

1. B
2. A
3. D
4. B
5. A
6. C
7. d
8. A
9. C
10. A
11. D

CHAPTER 2 / EQUIVALENCE TESTING

1. B
2. A
3. A
4. C
5. B
6. A
7. A

CHAPTER 3 / POWER, SAMPLE SIZE

1. B
2. D
3. A
4. B
5. B
6. C
7. A
8. C
9. B
10. A
11. 1. p<0.01
    2. 75-90%
    3. yes
12. 1. n = 16
    2. n = 62

CHAPTER 4 / PROPORTIONAL DATA ANALYSIS, PART I

1. D

2. $SE_1^2 = [4/16\ \ (1 - 4/16)] / 16 = 48 / 16^3$
$SE_2^2 = [12/16\ (1 - 12/16)] / 16 = 48 / 16^3$
$z = d/SE = \text{proportion}_1 - \text{proportion}_2\ /\ \sqrt{}(SE_1^2 + SE_2^2) =$
$= 8/16\ /\ \sqrt{}(6/16^2) = 0.5\ / 0.153 = 3.26$
$p<0.002$ answer C is correct.

3.

	yes	no
Group 1	4(a)	12 (c)
Group 2	12(b)	4 (d)

$\chi^2 = (4x4 - 12x12)^2\,(32)\ /\ \ 16x16x16x16 = 524{,}288 / 65{,}536 = 8.$
With 1 dfs $p< 0.01$ answer C is correct.

4.

	burn out	no burn out
Group 1	3	7
Group 2	0	10

$\chi^2 = (0 - 3.10)^2\,(20)\ /\ \ 3.17.10.10 = 18{,}000 / 5{,}100 = 3.6.$
With 1 df $0.05< p< 0.01$ answer A is correct.

5. A.

6. C.

7.

	yes	no
Group 1	4	2
Group 2	2	4

$P = 6!.6!.6!.6!\ /\ 12!.4!.2!.4!.2! = 13{,}095 / 133{,}065 = 0.0984,$
A is the correct answer.

8. $(a-\alpha)^2 / \alpha = 4^2\ /8 = 2$
$(b-\beta)^2 / \beta = 4^2\ /4 = \underline{4\ +}$
$\chi^2 = 6$ 1 df $p< 0.05$, B is correct answer.

9. $\chi^2 = (28\text{-}12)^2 / (12 + 28) = 256 / 40 = 6.4$ 1 df $p < 0.05$,
B is correct answer.

10. Odds ratio = 28/12= 2.33
95 % confidence interval (CI):
ln 2.33 - 1.96 √ (1/12 + 1/28) to ln 2.33 + 1.96 √ (1/12 + 1/28)=
(0.847-0.690) to (0.847+0.690)=
0.157 to 1.537,
to find 95% confidence intervals convert back using anti ln:
95% confidence intervals = 1.16 to 4.65 (p<0.05).

11. Odds ratio= 103/46 / 77/62 = 2.239 / 1.242= 1.803
95 % confidence interval:
ln 1.803 - 1.96 √ (1/103+1/46+1/77+1/62)
to ln 1.803 + 1.96 √ (1/103+1/46+1/77+1/62)=
0.589±0.482= 0.107 to 1.071,
to find 95% confidence intervals convert back using anti ln:
95% confidence intervals = 1.11 to 2.92 (p<0.05).

12. Are the two Kaplan Meier curves significantly different from one another?

```
survivors 2x 15 ___________
                   I____    I____
                        I____    I___________
                             I____
                                  I_____
                                        I_____

________________________________________
           1    2    3    4    5    6   period
```

Answer to 12:	a b	15 0	14 1	14 1	13 2	13 2	13 2
	c d	14 1	13 2	12 3	11 4	11 4	10 5
	period	1	2	3	4	5	6

Mantel Haenszl χ^2 summary test (= log rank test):
$\chi^2 = 1.42$, 1df, ns.

CHAPTER 5 / PROPORTIONAL DATA ANALYSIS, PART II

1. B
2. B
3. A
4. C
5. B
6. C
7. A
8. B
9. B
10. C

CHAPTER 6 / META-ANALYSIS

1. A
2. A
3. B
4. 4
5. D
6. A
7. A
8. B

CHAPTER 7 / INTERIM-ANALYSES

1. C
2. A
3. C
4. B
5. B
6. B

CHAPTER 8 / MULTIPLE TESTING

1. A
2. C
3. C
4. B
5. C

CHAPTER 9 / PRINCIPLES OF LINEAR REGRESSION

1. WRONG
2. A IS CORRECT
3. C IS CORRECT
4. D IS CORRECT
5. 2 IS CORRECT
6. F-TEST
7. C IS CORRECT, A IS CORRECT ONLY IF $P<0.05$
8. C IS CORRECT
9. 2ND DRUG = -0.41 + 0.52 1ST DRUG
10. POSITIVE
11. A IS CORRECT
12. YES. EXPLANATION: $102.77 = (10.14)^2$
13. YES

CHAPTER 10 / SUBGROUP ANALYSIS USING REGRESSION MODELING

1. C
2. B
3. B
4. B
5. B

CHAPTER 11 / RELATIONSHIP AMONG STATISTICAL DISTRIBUTIONS

1. D IS WRONG
2. BECAUSE WITH NORMAL DISTRIBUTIONS THE ADD-UP SUM EQUALS ZERO
3. B IS CORRECT
4. B IS CORRECT
5. C IS THE CORRECT ANSWER

Answer to 6.

$$\begin{array}{lrll} (a-\alpha)^2 / \alpha = (60-40)^2 / 40 = & 10.0 & & \\ (b-\beta)^2 / \beta = (100-120)^2 / 120 = & \underline{\ 3.5\ } & + & \\ & 13.5 & & p < 0.001 \end{array}$$

Answer to 7.

$(ad-bc)^2\ (a+b+c+d) / (a+b)(\ c+d)(\ b+d\ )(a+c) = 5.81818$ for 1df $p<0.02$

Alternative approach

$$\begin{array}{lll}
\alpha = [(a+c) / (a+b+c+d)]\ \ x\ (a+b) & & =50 \\
\beta = & & 50 \\
\gamma = & & 110 \\
\delta = & & 110 \\
(a-\alpha)^2 /\alpha = (60-50)^2 / 50 & = & 2 \\
(b-\beta)^2 /\beta = (40-50)^2 / 50 & = & 2 \\
(c-\gamma)^2 /\gamma = (100-110)^2 /110 & = & 0.9 \\
(d-\delta)^2/\delta\ = \ldots\ldots. & = & 0.9\ + \\
\hline
 & & 5.8
\end{array}$$

5.8 for 1 df $p<0.02$

Answer to 8.

$$\begin{array}{ll}
\alpha = [(a+b+c) / (a+b+c+d+e+f)]\ \ x\ (a+d) & = 52.17 \\
\beta\ \ldots.. & =114.78 \\
\gamma & =73.04 \\
\delta\ = [(d+e+f))/(a+b+c+d+e+f)]\ x\ (a+d) & = 47.83 \\
\varepsilon\ \ldots. & = 57.39 \\
\xi\ \ldots. & = 66.96
\end{array}$$

$$\begin{array}{ll}
(a-\ \alpha)^2 /\alpha & = 1.175 \\
(b-\ \ldots & =1.903 \\
(c-\ \ldots & = 0.663 \\
(d-\ \ldots & = 1.282 \\
(e-\ \ldots & =68.305 \\
(f-\ \ldots & = 0.723\ + \\
\hline
 & = 72.769
\end{array}$$

= 72.769 for 3-1= 2 dfs
(2 columns and 3 rows= (2-1)x(3-1)= 2dfs)
$p< 0.001$

Answer to 9.

$\alpha = [(a+b) / (a+b+c+d+e+f)] \ x \ (a+c+e) = 55.000$
β .... $= 55.000$
$\gamma = [(c+d) / (a+b+c+d+e+f)] \ x \ (a+c+e) = 51.613$
$\delta =$ ... $= 51.613$
ε ... $= 55$
ξ ... $= 55$
$(a-\alpha)^2 / \alpha = 0.45$
(b ... $= 0.45$
(c .. $= 0.847$
(d .. $= 1.363$
(e .. $= 0.45$
(f .. $= 0.45$ +

$=2.21$ for 3-1= 2 dfs (2 rows and 3 colums = (2-1)x(3-1)= 2dfs) $0.05 < p < 0.1$

Answer to 10.

$\frac{\text{SE data}}{\text{SE required}} = 0.9 / 0.5 = 1.8.$ Our SE is thus 1.8 times the SE required and so our variance is $1.8^2 = 3.24$ times the required variance. with 1df $\chi^2 = 3.24$ $0.05<p<0.1$

Answer to 11.

pooled OR (95% CI) = 0.63 (0.51-0.79) p<0.0001, heterogeneity is ns.

Answer to 12.

pooled OR (95% CI) = 0.87 (0.70-1.09) ns, heterogeneity p= 0.025.

CHAPTER 12 / STATISTICS IS NOT BLOODLESS ALGEBRA

1. D
2. C
3. A
4. A
5. A
6. A
7. E
8. E
9. D
10. D
11. A, BEING FALSELY NEGATIVE (TYPE II ERROR) IS NOT SO SERIOUS HERE, AS LONG AS SOME PATIENTS BENEFIT.
12. C, BEING FALSELY NEGATIVE IS SERIOUS HERE BECAUSE OF THE RISK OF SERIOUS ADVERSE EFFECTS.
13. A
14. D

INDEX