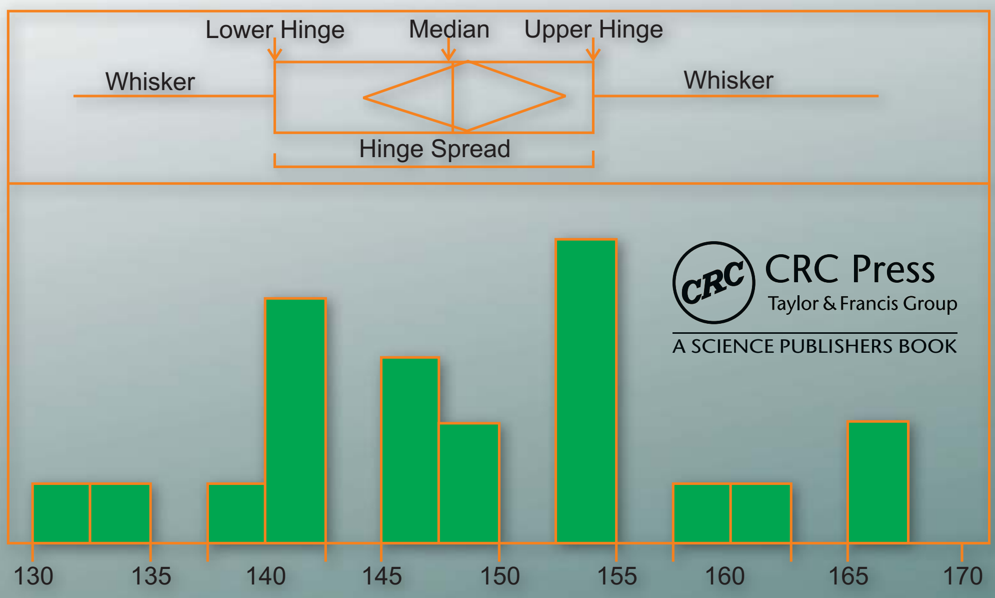
A Handbook of Applied Statistics in Pharmacology

Katsumi Kobayashi
K. Sadasivan Pillai



CRC Press
Taylor & Francis Group
A SCIENCE PUBLISHERS BOOK

A Handbook of Applied Statistics in Pharmacology

A Handbook of Applied Statistics in Pharmacology

Katsumi Kobayashi
Safety Assessment Division, Chemical Management Center
National Institute of Technology and Evaluation (NITE)
Tokyo, Japan

K. Sadasivan Pillai
Frontier Life Science Services
(A Unit of Frontier Lifeline Hospitals)
Thiruvallur District
Chennai, India

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A SCIENCE PUBLISHERS BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 2012919

International Standard Book Number-13: 978-1-4665-1540-6 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Foreword

Life expectancy has significantly increased in the last century, thanks to the discovery and development of new drugs by pharmaceutical industries. Search for new therapeutics is the primary activity of the R&D of pharmaceutical industries and it involves complex network of tasks such as synthetic chemistry, *in vitro*/*in vivo* efficacy, safety, preclinical and clinical research. Statistical analysis has always been the foundation to establish the safety and efficacy of drugs. The decision to- or not to- advance preclinical drug candidates to very expensive clinical development heavily relies on statistical analysis and the resulting significance of preclinical data. Recent reports attributed failure of certain drugs in clinical stages of development to improper conduct of preclinical studies and inappropriate application of statistical tools. Applying appropriate statistical tools is sagacious to analysis of data from any research activity. Though scientists expect computerized statistical packages to perform analyses of the data, he/she should be familiar with the underlying principles to choose the appropriate statistical tool.

'A Handbook of Applied Statistics in Pharmacology' by Katsumi Kobayashi and K. Sadasivan Pillai is a very useful book for scientists working in R&D of pharmaceuticals and contract research organizations. Most of the routine statistical tools used in pharmacology and toxicology are covered perspicuously in the book. The examples worked out in the book are from actual studies, hence do not push a reader having less or no exposure to statistics outside his/her comfort zone.

Dr. K.M. Cherian
M.S., F.R.A.C.S., Ph.D., D.Sc. (Hon.), D.Sc. (CHC), D.Sc. (HC)
Chairman & CEO
Frontier Lifeline Hospitals
Chennai, India

Preface

Scientists involved in pharmacology have always felt that statistics is a difficult subject to tackle. Thus they heavily rely on statisticians to analyse their experimental data. No doubt, statisticians with some scientific knowledge can analyse the data, but their interpretation of results often perplexes the scientists.

Statistics play an important role in pharmacology and related subjects like, toxicology, and drug discovery and development. Improper statistical tool selection to analyze the data obtained from studies conducted in these subjects may result in erroneous interpretation of the performance- or safety- of drugs. There have been several incidents in pharmaceutical industries, where failure of drugs in clinical trials is attributed to improper statistical analysis of the preclinical data. In pharmaceutical Research & Development settings, where a large number of new drug entities are subjected to high-throughput *in vitro* and *in vivo* studies, use of appropriate statistical tools is quintessential.

It is not prudent for the research scientists to totally depend on statisticians to interpret the findings of their hard work. Factually, scientists with basic statistical knowledge and understanding of the underlying principles of statistical tools selected for analysing the data have an advantage over others, who shy away from statistics. Underlying principle of a statistical tool does not mean that one should learn all complicated mathematical jargons. Here, the underlying principle means only ‘thinking logically’ or applying ‘common sense’.

The authors of this book, with decades of experience in contract research organizations and pharmaceutical industries, are fully cognizant of the extent of literacy in statistics that the research scientists working in pharmacology, toxicology, and drug discovery and development would be interested to learn. This book is written with an objective to communicate statistical tools in simple language. Utmost care has been taken to avoid complicated mathematical equations, which the readers may find difficult to assimilate. The examples used in the book are similar to those that the scientists encounter regularly in their research. The authors have provided cognitive clues for selection of an appropriate statistical tool to analyse the data obtained from the studies and also how to interpret the result of the statistical analysis.

Contents

1
Probability

Probability and Possibility

We all are familiar with the words, possibility and probability. Though these words seem to convey similar meanings, in reality they do not. Imagine, your greatest ambition is to climb Mount Everest. But you do not know the basics of mountaineering and have not climbed even a hill before. It may still be *possible* for you to climb Mount Everest, if you learn mountaineering techniques and undergo strenuous training in mountaineering. But the *probability* of accomplishing your ambition of climbing Mount Everest is remote. Possibility is the event that can happen in life, whereas the probability is the chance of that happening. In statistical terminology, an event is collection of results or outcomes of a procedure. Probability is the basic of statistics.

Mathematicians developed the 'principle of indifference' over 300 years ago to elucidate the 'science of gaming' (Murphy, 1985). According to Keynes (1921), the 'principle of indifference' asserts that "if there is no known reason for predicating of our subject one rather than another of several alternatives, then relatively to such knowledge the assertions of each of these alternatives have an equal probability." In other words, if you have no reason to believe the performance of drug A is better than B, then you should not believe that drug A is better than B.

The two approaches to probability are classical approach and relative frequency approach. In classical approach, the number of successful outcomes is divided by the total number of equally likely outcomes. Relative frequency is the frequency of an event occurring in large number of trials. For example, you flip a coin 1000 times and the number of occurrences of *head up* is 520. The probability of *head up* is 520/1000=0.52.

Both the classical and frequency approaches have some drawbacks. Because of these drawbacks, an axiomatic approach to probability has been suggested by mathematicians (Spiegel *et al.*, 2002).

However, in pharmacology and toxicology experiments, relative frequency approach proposed by Mises and Reichenbach (Carnap, 1995) works well.

We shall understand probability a bit more in detail by working out examples.

Probability—Examples

Let us try to define a probability with regard to frequency approach. The probability of an occurrence for an event labeled A is defined as the ratio of the number of events where event A occurs to the total number of possible events that could occur (Selvin, 2004).

Let us understand some basic notations of probability:

P denotes probability.

If you toss a coin, only two events can occur, either a *head up* or a *tail up*.

P(H) denotes probability of event head is up. You can calculate the probability of head coming up using the formula:

$$P(H) = \frac{\text{Number of times head is up}}{(\text{Number of times head is up+Number of times tail is up})}$$

Remember, a *head up* and a *tail up* have equal chance of occurring. Ideally you will get a value very close to 50% for *P(H),* if you toss the coin several times.

You roll an unbiased six-sided dice. The total number of outcomes is six, which are equally likely. This means the likelihood of ‘any number’ coming up is same as ‘any other number’. The probability of any number coming up is 1/6. The probability of any two numbers coming up is 2/6.

Let us come back to our example of tossing a coin. The probability of a *head up* is ½ (0.5 or 50%). Now you flip the coin twice. The probability of a *head up* both times is ½ x ½ = ¼.

Mutually exclusive events

While you toss a coin either a *head up* or a *tail up* occurs. When the event *head up* occurs, the event *tail up* cannot occur and *vice versa*; one event precludes the occurrence of the other. In this example, *head up* or *tail up* that occurs while tossing a coin is a mutually exclusive event.

Equally likely events

Occurrence of *head up* or *tail up* is an equally likely event when you toss a fair coin. This means *P(H) = P(T),* where *P(H)* denotes probability of event *head up* and *P(T)* denotes probability of event *tail up*.

Probability Distribution

Let us try to understand probability distribution with the help of an example. You flip a coin twice. In this example the variable, *H* is number of heads that results from flipping the coin. There are only 3 possibilities:

$H = 0$

$H = 1$

$H = 2$

Let us calculate the probabilities of the above occurrences of *head up*.

The probability of not occurring a *head up* in both the times (H=0) =0.25

The probability of occurring a *head up* in one time (H=1) = 0.5

The probability of occurring a *head up* in both times (H=2) = 0.25

0.25, 0.5 and 0.25 are the probability distribution of *H*.

Cumulative Probability

A cumulative probability is a sum of probabilities. It refers to the probability that the value of a random variable falls within a specified range.

You toss a dice. What is the probability that the dice will land on a number that is smaller than 4? The possible 6 outcomes, when a dice is tossed are 1, 2, 3, 4, 5 and 6.

The probability that the dice will land on a number smaller than 4:

$$\mathrm{P}(X < 4\ ) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) = 1/6 + 1/6 + 1/6 = 1/2$$

The probability that the dice will land on a number 4 or smaller than 4:

$$P(X \leq 4) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$$

Cumulative probability is commonly used in the analysis of data obtained from pharmacological (Kuo *et al.*, 2009; Rajasekaran *et al.*, 2009) and toxicological experiments.

Probability and Randomization

In order to evaluate the efficacy of an anti-diabetic drug in rats, twenty rats are administered streptozotocin to induce diabetes. The blood sugar of individual rats is measured to confirm induction of diabetes. You find that 13 rats have blood sugar >250 mg/dl and remaining 7 rats have blood sugar <200 mg/dl. The 20 rats are then distributed randomly in two equal groups (Group 1 and Group 2). You want to treat the Group 1 (control group) with the vehicle alone and the Group 2 (treatment group) with the drug.

Initiate randomization by picking up a rat without any bias and place it in Group 1.

The probability of picking up a rat having blood sugar >250 mg/dl = 13/20 = 65%

The probability of picking up a rat having blood sugar <200 mg/dl = 7/20 = 35%

Assign 10 rats to Group 1 and then the remaining to Group 2. It is most likely that you will have more rats with blood sugar >250 mg/dl in Group 1.

Remember that both the groups are physiologically and metabolically different, because it is most likely that more number of rats in Group 1 will have blood sugar >250 mg/dl and more number of rats in Group 2 will have blood sugar <200 mg/dl. It is unlikely that the experiment with these groups will yield a fruitful result. Randomization is very important in animal studies. We shall be discussing more on randomization of animals in pharmacological studies in later chapters.

References

Carnap, R. (1995): Introduction to the Philosophy of Science. Dover Publications, Inc., New York, USA.

Keynes, J.M. (1921): A Treatise on Probability. Macmillan, London, UK.

Kuo, S.P., Bradley, L.A. and Trussell, L.O. (2009): Heterogeneous kinetics and pharmacology of synaptic inhibition in the chick auditory brainstem. J. Neurosci., 29 (30), 9625–9634.

Rajasekaran, K., Sun, C. and Bertram, E.H. (2009): Altered pharmacology and GABA-A receptor subunit expression in dorsal midline thalamic neurons in limbic epilepsy. Neurobiol. Res., 33(1), 119–132.

Murphy, E.A. (1985): A Companion to Medical Statistics. Johns Hopkins University Press, Baltimore, USA.

Selvin, S. (2004): Biostatistics—How It Works. Pearson Education (Singapore) Pte. Ltd., India Branch, Delhi, India.

Spiegel, M.R., Schiller, J.J., Srinivsan, R.A. and LeVan, M. (2002): Probability and Statistics. The McGraw Hill Companies, Inc., USA.

2 Distribution

History

The most commonly used probability distribution is the normal distribution. The history of normal distribution goes way back to 1700s. Abraham DeMoivre, a French-born mathematician introduced the normal distribution in 1733. Another French astronomer and mathematician, Pierre-Simon Laplace dealt with normal distribution in 1778, when he derived 'central limit theorem'. In 1809 Johann Carl Friedrich Gauss (1777–1855), a German physicist and mathematician, studied normal distribution extensively and used it for analysing astronomical data. Normal distribution curve is also called as Gaussian distribution after Johann Carl Friedrich Gauss, who recognized that the errors of repeated measurements of an object are normally distributed (Black, 2009).

Variable

We need to understand a terminology very commonly used in statistics, *i.e.,* 'variable'. Variable is the fundamental element of statistical analysis. Variables are broadly classified into categorical (attribute) and quantitative variables. Categorical and quantitative variables are further classified into two subgroups each—Categorical variables into nominal and ordinal, and Quantitative variables into discrete and continuous.

Nominal variable: The key feature of nominal variables is that the observation is not a number but a word (example—male or female, blood types). Nominal variables cannot be ordered. It makes no difference if you write the blood types in the order A, B, O, AB or AB, O, B, A.

Ordinal variable: Here the variable can be ordered (ranked); the data can be arranged in a logical manner. For example, intensity of pain can be ordered as—mild, moderate and severe.

Discrete variable: Discrete variable results from counting. It can be 0 or a positive integer value. For example, the number of leucocytes in a µl of blood.

Continuous variable: Continuous variable results from measuring. For example, alkaline phosphatase activity in a dl of serum.

The variables can be independent and dependent. In a 90 day repeated dose administration study you measure body weight of rats at weekly intervals. In this situation week is the independent variable and the body weight of the rats is the dependent variable.

Stem-and-Leaf Plot

Stem- and Leaf-Plot (Tukey, 1977) is an elegant way of describing the data (Belle *et al.*, 2004). Let us construct a stem-and-leaf plot of the body weight of rats given in Table 2.1.

Table 2.1. Body weight of rats

Body weight (g)
132, 139, 134, 141, 145, 141, 140, 166, 154, 165, 145, 158, 162, 148, 154, 146, 154, 148, 140, 153, 154

Now arrange the data in an ascending order as given in Table 2.2:

Table 2.2. Body weight of rats arranged in an ascending order

Body weight (g)
132, 134, 139, 140, 140, 141, 141, 145, 145, 146, 148, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166

Stem-and-leaf plot of the above data is drawn in Figure 2.1:

Stem	Leaf
13	2 4 9
14	0 0 1 1 5 5 6 8 8
15	3 4 4 4 4 8
16	2 5 6

Figure 2.1. Stem- and- Leaf plot

Each data is split into a “leaf” (last digit) and a “stem” (the first two digits). For example, 132 is split into 13, which forms the ‘stem’ and 2, which forms the ‘leaf’. The stem values are listed down (in this example 13, 14, 15 and 16) and the leaf values are listed on the right side of the stem values.

The Stem-and-leaf plot provides valuable information on the distribution of the data. For example, the plot indicates that more number of the animals is having body weight in the 140 g range, followed by the 150 g range.

Box-and-Whisker Plot

Another way of describing the data is by constructing a box-and-whisker plot. The usefulness of box-and-whisker plot is better understood by learning how to construct it. For this purpose we shall use the same body weight data given in Table 2.1. As we have done for plotting the stem-and-leaf plot, arrange the data in an ascending order (Table 2.2). The first step in constructing a box-and-whisker plot is to find the median. You will learn more about median in Chapter 3.

The median of the data given in Table 2.2 is the 11th value, *i.e.,* 148 (see Table 2.3).

Table 2.3. Median value of the body weight data

Median

⇩

132, 134, 139, 140, 140, 141, 141, 145, 145, 146, **148**, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166

The median divides the data into 2 halves (a lower and an upper half). The lower half consists of a range of values from 132 to 146 and the upper half consists of a range of values from 148 to 166 (see Table 2.4).

Table 2.4. Median value of the lower and upper quartiles

Median

Lower half ◀—— ⇩ ——▶ Upper half

132, 134, 139, 140, 140, 141, 141, 145, 145, 146, **148**, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166

Next step is to find the median of lower half and upper half:
Median of the lower half = (140+141)/2 = 140.5
Median of the upper half = (154+154)/2 = 154.0

Median of the lower half is also called as 'lower hinge' or ' lower quartile' and the median of the upper half as 'upper hinge' or ' upper quartile'. The term, quartile was introduced by Galton in 1882 (Crow, 1993). About 25% of the data are at or below the 'lower hinge', about 50% of the data are at or below the median and about 75% of the data are at or below the 'upper hinge'.

Next step is calculation of 'hinge spread', the range between lower and upper quartiles:

Hinge spread = 154.0–140.5 = 13.5

Hinge spread is also called as inter-quartile range (IQR).

Now, we need to determine 'inner fence'. The limits of 'inner fence' are determined as given below:

Lower limit of 'inner fence' = Lower hinge–1.5 x hinge spread
= 140.5–(1.5x13.5)= 120.25
Upper limit of 'inner fence' = Upper hinge+1.5 x hinge spread
= 154.0+(1.5x13.5)= 174.25

We now have all the required information to construct the 'whiskers'. The lowest body weight data observed (see Table 2.4) between 140.5 g and 120.25 g is 132 g and the highest body weight data observed between 154.0 g and 174.25 g is 166 g. Hence, the whiskers are extended from the lower quartile to 132 g and from the upper quartile to 166 g.

Box-and-whisker plot of the data (Table 2.1) is given in Figure 2.2.

The box-and-whisker plot is based on five numbers: the least value, the lower quartile, the median, the upper quartile and the greater value in a data set.

If the data are normally distributed:

1. the median line will be in the centre of the box dividing the box into two equal halves
2. the whiskers will have similar lengths
3. observed values will scarcely be outside the 'inner fence'.

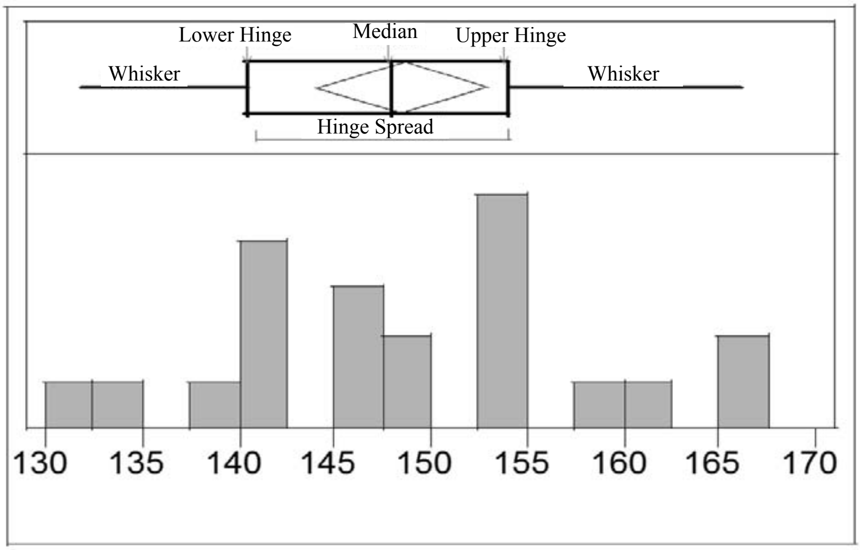


Figure 2.2. Box-and-whisker plot of the data

It is important to examine whether the data are normally distributed before applying a statistical tool. We shall learn more about this in later chapters.

References

Belle, G,V., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics-A Method for the Health Sciences. 2nd Edition, Wiley Interscience, New Jersey, USA.

Black, K. (2009): Business Statistics: Contemporary Decision Making. 6th Edition. John Wiley and Sons, Inc., USA.

Crow, J.F. (1993): Francis Galton: Count and measure, measure and count. Genetics, 135, 1–4.

Tukey, J.W. (1977): Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts, USA.

3
Mean, Mode, Median

Average and Mean

Average and mean are interchangeably used in everyday life. Average is the synonym for the central tendency. There are various types of central tendencies, such as mean, mode and median.

Mean

The procedure for calculating mean is very simple; sum of all individual observations divided by the sum of number of observations. There are several types of means, such as arithmetic mean, geometric mean and harmonic mean. Let us work out the example given in Table 3.1 to familiarise the reader with the calculation procedure of these means.

Table 3.1. Calculation of arithmetic mean of body weight of rats

Body weight (g)	Sum
132, 139, 134, 141, 145, 141, 140, 166, 186, 183	1507

In statistics, the number of observations is denoted by the letter *N* or *n* (both cases). Number of observations is also called the sample size. The Greek letter Σ (uppercase only) is used to denote sum. Mean is denoted as $\overline{X}$ (X bar).

Mean body weight of above example:

$$\overline{X} = \frac{\sum X}{N} = \frac{1507}{10} = 150.7 \text{ g}$$

Mean in the above example is called the arithmetic mean. Arithmetic mean is sensitive to extreme values in data set. There is a condition for calculating arithmetic mean—the data should fit a normal distribution.

Geometric Mean

Mathematically geometric mean is defined as the nth root of the product of n numbers. An easy way to calculate the geometric mean is to find the mean of logarithmic values of the data and then to find the antilog of the mean. Steps involved in the calculation of the geometric mean of the body weight data of the rats (Table 3.1) are given in Table 3.2.

Table 3.2. Calculation of geometric mean of body weight of rats

Body weight (g)		Σ	*N*	$\overline{X}$
Linear scale	132, 139, 134, 141, 145, 141, 140, 166, 186, 183	1507	10	150.7
Log scale	2.12, 2.14, 2.13, 2.15, 2.16, 2.15, 2.15, 2.22, 2.27, 2.26	21.7	10	2.17

Geometric mean is the antilog of 2.17 = 147.9

If any observed value is 0 or negative, geometric mean cannot be calculated. Geometric mean is very rarely used in pharmacology. However, use of it is witnessed in some pharmacokinetic studies (Schuirmann, 1987).

Harmonic Mean

Harmonic mean is calculated by finding the mean of the reciprocals of the values and then finding the reciprocal of the mean.

Calculation procedure of the harmonic mean of the data given in Table 3.1 is described in Table 3.3:

Table 3.3. Calculation of harmonic mean of body weight of rats

Body weight (g)		Σ	*N*	$\overline{X}$
Linear scale	132, 139, 134, 141, 145, 141, 140, 166, 186, 183	1507	10	150.7
Reciprocal	0.0076, 0.0072, 0.0075, 0.0071, 0.0069, 0.0071, 0.0071, 0.0060, 0.0054, 0.0055	0.0673	10	0.0067

Harmonic mean = 1/0.0067 = 148.5 g

Unlike arithmetic mean, the harmonic mean is not influenced by the extreme values. Harmonic mean has limited use in pharmacology. A pharmacokinetic study carried out with cyclosporine-A revealed that there was little use of harmonic mean to describe the central tendency (Lum *et al.*, 1992). However, Iwamoto *et al.* (2008) used harmonic mean to evaluate the central tendency of pharmacokinetics in a clinical study conducted with Raltegravir in healthy subjects.

Weighted Mean

In an experiment designed to administer a drug to rats, 15 rats were randomly assigned to 3 cages (Cage 1, Cage 2 and Cage 3), each cage consisting of 5 rats. At the end of 2 weeks of the drug administration in Cages 1 and 2, two rats each survived, whereas in Cage 3 all the five rats survived. The body weight of the survived rats is given in Table 3.4.

Table 3.4. Body weight (g) of rats in 3 Cages at the end of 2 weeks following a drug administration

Cage	N	Mean (g)
1	2	119
2	2	125
3	5	134

Let us calculate the grand mean:

(119+125+134)/3 = 126 g. But there is a problem with this grand mean. It is very close to the mean value of 2 animals of Cage 2, but does not seem to represent the body weight of animals in Cages 1 and 3. Hence calculating grand mean by the above method is not advisable. In such situation, calculating weighted mean value may be the right approach.

Weighted Mean = [(2x119)+(2x125)+(5x134)]/(2+2+5) =1158/9 = 128.7 g

Mode

The mode is the value which appears the most in the data. It is usually calculated for discrete data (Belle *et al*., 2004). There can be more than one mode, if there is more than one value which appears the most.

In the following data,

130, 140, 140, 150, 140, 160, 140, 110, 120

The mode is 140 (140 appears 4 times in the data).

In the following data,

130, 140, 140, 150, 140, 160, 140, 110, 120, 130, 130

There are two modes, 140 and 130 (140 appears 4 times in the data, whereas 130 appears 3 times).

Median

To measure the central tendency, median is second in popularity to mean (Rosner, 2006). Median is also termed as 0.50 quantile. Another term for the median is the 50th percentile.

The first step to calculate the median is to rank the values from lowest to the highest. If the number of data values is odd, add 1 to the number of data values and divide that by 2. For example, if there are 9 sample values, divide (9+1) by 2. The median is the 5th ranked value. If the number of data values is even, again add 1 to the number of data values and divide that by 2. For example, if there are 10 sample values, divide (10+1) by 2 to get 5.5. Median is the mean of the 5th and 6th ranked values.

The second situation where the median is useful is when it is impractical to measure all of the values, such as when you are measuring the time until something happens. Survival time is a good example of this; in order to determine the mean survival time, you have to wait until every individual is dead, whereas to determine the median survival time you do not need to wait until every individual is dead; you need to wait only until half the individuals are dead.

Mean, mode and median are theoretically the same for the data collected from a symmetrical distribution (Lemma, 2008). Median and mode are not affected by the extreme values (outliers). One disadvantage of mode is that it does not include all the data for the analysis. Though mean and median are commonly used in statistical analysis of pharmacological and toxicological data, the use of mode is not very common.

References

Belle, V.G., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics—A Methodology for the Health Sciences. John Wiley & Sons, Inc., New Jersey, USA.

Iwamoto, M., Wenning, L.A., Petry, A.S., Laethem, M., De Smet, M., Kost, J.T., Merschman, S.A., Strohmaier, K.M., Ramael, S., Lasseter, K.C., Stone, J.A., Gottesdiener, K.M. and Wagner, J.A. (2008): Safety, tolerability, and pharmacokinetics of Raltegravir after aingle and multiple doses in healthy subjects. Clin. Pharmacol. Therapeutics, 83, 293–299.

Lemma, A. (2008): Introduction to the Practice of Psychoanalytic Psychotherapy. John Wiley & Sons Ltd., Chichester, UK.

Lum, B.L., Tam, J., Kaubisch, S. and Flechner, S.M. (1992): Arithmetic versus harmonic mean values for cyclosporin-A pharmacokinetic parameters. J. Clin. Pharmacol., 32 (10), 911–014.

Rosner, B. Fundamentals of Biostatistics. 6th Edition, Thomson Brooks/Cole, Belmont, USA.

Schuirmann, D.J. (1987): A comparison of the two one-sided tests procedure and the power approach for assessing the bioequivalence of average bioavailability. J. Pharmacokinetics Biopharm., 15, 657–680.

4

Variance, Standard Deviation, Standard Error, Coefficient of Variation

Variance

Even the inbred animals maintained under well controlled animal house conditions may show some variations among the individuals in responding to a treatment in a pharmacology or toxicology study. Though majority of the individual animals respond to the treatment in a similar manner or magnitude, few of them will be too sensitive or resistant to the treatment. There are several factors that may affect the outcome of an animal experimentation, for example factors related to the experimenter. In a nutshell, even a well designed animal experimentation is bound to show some variations in the result and it is important to understand these variations for interpreting the experimental data. We shall work out an example, to make it very clear.

For a pharmacology experiment 5 rats are randomly picked up and placed them in a cage. As all the rats are of similar age and maintained in identical animal house conditions, one would assume that all the animals will have comparable body weight. The body weight of the rats is given in Table 4.1.

It is evident from the Table that the assumption of 'all animals having comparable body weight' is incorrect. In animal experiments, one can seldom get identical animals. There could be several differences (for example difference in water and feed consumption, difference in activity, difference in certain clinical chemistry parameters, etc.) among them. These differences have an important role in determining the outcome of an

Table 4.1. Body weight of rats (g)

Column 1	**Column 2**	**Column 3**	**Column 4**
Rat No.	**Body Weight** (X)	(X-$\bar{X}$)	(X-$\bar{X}$)^2
1	245	–7.6	57.76
2	254	+1.4	1.96
3	239	–13.6	184.96
4	266	+13.4	179.56
5	259	+6.4	40.96
Number of observations (n)	5	-	-
Sum (Σ)	1263	0	465.2
Mean ($\bar{X}$)	252.6	-	-

animal experimentation. Let us try to find an estimate for these differences. In the example given in Table 4.1, the mean body weight is calculated as 252.6 g. Now, calculate the difference of each observation from the mean value (X–$\bar{X}$). A better statistical terminology for the difference is deviation, which is given in column 3 of Table 4.1. One may think that an estimate of the deviations can be obtained easily by summing up (X–$\bar{X}$). By doing so what you get is a zero. You cannot go further with this zero. When (X–$\bar{X}$) given in column 3 is closely examined, one would realize that the sum of the values bearing plus (+) sign is equal to the sum of the values bearing minus (–) sign. That is why a zero is obtained for the sum of (X–$\bar{X}$). This can be easily solved by squaring (X–$\bar{X}$). Squares of (X–$\bar{X}$) are given in column 4 of Table 4.1. Summing up (X-$\bar{X}$)^2, a value 465.2, is called as sum of the squares (SS) of deviations is obtained. By dividing 465.2, i.e., the sum of the squares of deviations by n–1, a very important statistical parameter called 'variance' is derived.

Variance = 465.2/(5–1) = 116.3

One may ask why the SS is divided by 4 (n–1), instead of 5 (n). The denominator to calculate the variance is called as 'degrees of freedom'. Degrees of freedom is one less than the total number of observations. Let us try to explain this logically. Five different coloured boxes, say Black, Blue, Green, Red and Yellow are placed on a table. You have the 'freedom' to pick up the boxes in an unbiased manner, one by one. You may think that there are 5 boxes and the number of the 'freedoms' that you can exercise in picking up the boxes is also 5. You exercised your 'freedoms' to pick up the boxes as given in Table 4.2.

Table 4.2. Degrees of freedom exercised in picking up coloured boxes

Boxes picked up	Degrees of freedom exercised	Degrees of freedom left
Red	1	5–1 = 4
Yellow	2	5–2 = 3
Black	3	5–3 = 2
Green	4	5–4 = 1
Blue	This is the last box left out. You cannot exercise any degree of freedom for picking up this box.	

Initially, you thought that you would have had 5 degrees of freedom before picking up any box. Firstly, you picked up the red box and your degrees of freedom is reduced by 1 (5–1). The next time you picked up the yellow box, and now your degrees of freedom is reduced by 2 (5–2). When you picked up the black box, you have only 2 degrees of freedom left. After picking the green box, you have only 1 degree of freedom left. But you cannot exercise any freedom to pick up the blue box. Blue box is the last box left out and you have to pick up this without any choice. Therefore, the actual degrees of freedom that one can exercise is not equal to the total number of observations, but 1 less than the total number of observations.

Standard Deviation (SD)

Standard deviation is the square root of variation:

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{116.3} = \pm 10.78$$

A ± sign should always be added as a prefix to SD.

Some statisticians are of the opinion that the ± symbol is superfluous (Everett and Benos, 2004). According to them, a standard deviation is a single positive number, the notation of the SD should be: Mean (SD X), where X is the value for SD (for example, body weight of rats = 252.6 g (SD 10.78). We are in favor of prefixing a ± sign to SD as it gives an easily perceivable indication about the lowest and highest values of the sample observations.

Standard deviation is a useful measure to explain the distribution of the sample observations around the mean. SD can also be used to see whether a single observation falls within the normal range (Cumming, 2007). If the observations follow a normal distribution, mean ± 1 SD covers a range of 68% of the observations. About 95% of individuals will have values within 2 standard deviations of the mean (mean ± 2 SD), the other 5%

being equally scattered above and below these limits (Altman and Bland, 1995). Mean ± 3 SD covers a range of 99.7% of the observations.

Standard Error (SE)

SE is the SD of the mean. SE is considered as a measure of the precision of the sample mean (Altman and Bland, 2005). It provides an estimate of the uncertainty of the true value of the population mean (Everett, 2008). In simple words, SE measures the variation in the means of the samples. It can be calculated using the formula:

$$\mathrm{SE} = \mathrm{SD}/\sqrt{n} = 10.78/\sqrt{5} = \pm 4.82$$

Always prefix ± sign to SE.

Coefficient of Variation (CV)

CV is a numerical value where the proportion of the standard deviation in the mean value is shown as a percentage:

$$\mathrm{CV} = \frac{\mathrm{SD}}{\mathrm{Mean}} \times 100 = \frac{10.78}{252.6} \times 100 = 4.27\%$$

CV is an excellent statistical tool that can be used to compare different analytical methods and performance of equipments. Since CV is independent of the scale of measurement, it can be used to compare variables measured on different scales (Daniel, 2007). In a clinical chemistry laboratory, biochemists routinely use the commercially available reagent kits for analyzing clinical chemistry parameters in blood. It is difficult to choose from the plenty of kits available in the market. In such cases, kit with the lowest CV given in the packet insert should be chosen.

CV plays a very important role in determining the significant difference in pharmacology and toxicology experiments. Kobayashi *et al.* (2011) examined 59 parameters from 153 numbers of 28-day repeated dose administration studies conducted in 12 test facilities in order to understand the influence of CV in determining significant difference of quantitative values. CV of electrolytes was comparatively small, whereas enzymes had large CV. A significant difference between the sexes was observed in the CVs of feed consumption, reticulocyte, platelet and leucocyte counts, cholesterol, total protein, albumin, albumin/globulin ratio, alkaline phosphatase, inorganic phosphorus, and pituitary and adrenals weights.

Large differences in CV were observed for major parameters among 7 test facilities. The authors inferred that a statistically significant difference is usually detected if there is a difference of 7% in mean values between the groups and the groups have a CV of about 7%. A parameter with a CV as high as 30% in two groups can be significantly different from each other, if the difference between the two mean values of the groups is about 30% and the number of observation (n) in each group is 10. The authors suggested that it would be ideal to use median value to assess the treatment-related effect, rather than mean, when the CV is very high.

Matsuzawa *et al.* (1993) analyzed historical control data pertaining to clinical pathology of study population covering 14000 rats, 10000 dogs and 1400 monkeys. The authors stated that the serum assay values showed greater variation than the plasma values. Aoyama (2005) suggested that when the number of animals is adjusted, the decentralization of data, like body weight and the organ weight, become comparatively smaller, and a CV of about 10% is obtained. CV for blood levels of various hormones, even in control animals is large. Often, the standard deviation exceeds the mean value by more than 50% for these parameters.

There is a misconception that the variability in the experimental data occurs only in animal experiments. One may think that the instruments used in bioanalytical laboratories are highly sophisticated and automated, hence the results obtained from these instruments show minimum to no variation. This is not true. There is variability in analytical chemistry and the measured values differ from the actual values and 'if the variability of a measurement is not characterized and stated along with the result of the measurement, then the data can only be interpreted in a limited sense' (USP, 2008).

When to Use a Standard Deviation (SD)/Standard Error (SE)?

Pharmacologists and toxicologists ambiguously use SD and SE in their study reports. A confusion in the use of SD and SE is evident in scientific articles published in various journals (Herxheimer, 1988; Nagele, 2003).

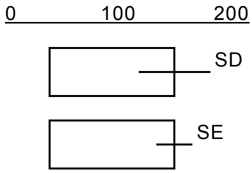


Figure 4.1. SD and SE calculated for human γ-GTP data[a]

[a]Data—42, 60, 26, 48, 56, 31, 30, 80, 79, 93 γ-GTP (IU/l)

Since SE is smaller than the SD (see Figure 4.1), some authors use SE, perhaps intentionally, in order to reduce the variability of their samples (Streiner, 1996; Lang, 1997; Fisher, 2000).

Although SD and SE are related, they give two very different types of information (Carlin and Doyle, 2000). In animal experiments, generally SD is 8–20% of the mean of the measured values, hence, the bar presented by the SD in a graph seems to be well balanced against the mean value. It is not permitted to use SE intentionally just to show a small width of the bar (Matsumoto, 1990). The next question is how precisely mean and SD should be specified? Mean should not be specified with more than one extra decimal place over the raw data but for SD greater precision can be given (Altman and Bland, 1996).

In conclusion, SD gives a fairly good indication about the distribution of the observed values around the mean. SE gives an indication about the variability of the mean. In toxicology experiments, especially with rodents, where the number of animals in a group is usually 10, it would be more ideal to use SD and in pharmacology experiments, where the number of animals in a group is usually <5 it would be more ideal to use SE, though there is no hard and fast rule for these.

References

Altman, D.G. and Bland, J.M. (1995): Statistics notes: The normal distribution. BMJ, 310, 298.

Altman, D.G. and Bland, J.M. (1996): Presentation of numerical data. BMJ, 312, 572.

Altman, D.G. and Bland, J.M. (2005): Standard deviations and standard errors. BMJ, 331, 903.

Aoyama, H. (2005): Applications and limitations of *in vivo* bioassays for detecting endocrine disrupting effects of chemicals on mammalian species of animals. J. Natl. Inst. Public Health, 54(1), 29–34.

Carlin, J.B. and Doyle, L.W. (2000): Basic concepts of statistical reasoning: standard errors and confidence intervals. J. Paediatr. Child Health, 36, 502–505.

Cumming, G. (2007): Error bars in experimental biology. JCB*,* 177(1), 7–11.

Daniel, W.W. (2007): Biostatistics-A Foundation of Analysis in the Health Sciences. 7th Edition, John Wiley & Sons (Asia) Pte. Ltd., Singapore.

Everett, D.C. (2008): Explorations in statistics: standard deviations and standard errors. Adv. Physiol. Educ., 32, 203–208.

Everett, D.C. and Benos, D.J. (2004): Guidelines for reporting statistics in journals published by the American Physiological Society. Adv. Physiol. Educ., 28, 85–87.

Fisher, D.M. (2000): Research Design and Statistics in Anesthesia. In: Anesthesia, 5th Edition, Vol. 1., Edited by Miller, R.D., Churchill Livingston, Philadelphia, USA.

Herxheimer, A. (1988): Misuse of standard error of the mean. Br. J. Clin. Pharmacol., 26, 197.

Kobayashi, K., Sakuratani, Y., Abe, T., Yamazaki, K., Nishikawa, S., Yamada, J., Hirose, A., Kamata, E. and Hayashi, M. (2011): Influence of coefficient of variation in determining significant difference of quantitative values obtained from 28-day repeated-dose toxicity studies in rats. J. Toxicol. Sci., 36(1), 63–71.

Lang, T.A.S.M. (1997): How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. American College of Physicians, Philadelphia, USA.

Matsuzawa, T., Nomura, M. and Unno, T. (1993): Clinical pathology reference ranges of laboratory animals. J. Vet. Med. Sci., 55(3), 351–362.

Matsumoto, K. (1990): Japanese Laboratory Animal Engineer Society, No. 6.

Nagele, P. (2003): Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br. J. Anaesthesiol., 90, 514–516.

Streiner, D.L. (1996): Maintaining standards: differences between the standard deviation and standard error, and when to use each. Can. J. Psychiatry, 41, 498–502.

USP (2008): The United States Pharmacopeia, The National Formularly, USP 31, NF 26, Asian Edition, Volume1, Port City Press, Baltimore, USA.

5

Analysis of Normality and Homogeneity of Variance

Distribution of Data in Toxicology and Pharmacology Experiments

It is important to know how the data are distributed for selecting a statistical tool for the analysis of the data (Bradlee, 1968). In toxicology and pharmacology experiments, data could be distributed in various patterns. The three commonly seen patterns of data distribution are given in Figure 5.1.

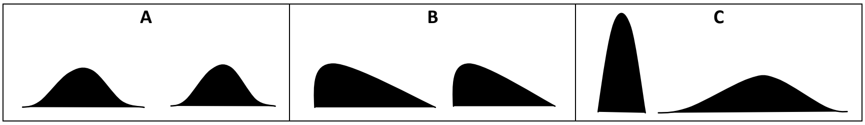


Figure 5.1. Three patterns of data distribution in toxicology and pharmacology experiments

A. Normal distribution and homogeneity of variance, B. Non-normal distribution and homogeneity of variance, C. Non-normal distribution in and heterogeneity in variance.

Analysis of Normality

The two types of non-normal distributions that are generally encountered in statistical analysis are skewness and kurtosis. The mean and median are different in a skewed distribution. Skewness can be positive or negative. The data are positively skewed, when the tail of the distribution curve is extended towards more positive values and the data are negatively skewed, when the tail of the distribution curve is extended towards more negative values (Čisar and Čisar, 2010).

Peakedness of a distribution is depicted by kurtosis. A distribution can be 'platykurtic' or 'leptokurtic'. Platykurtic is more flat-topped and leptokurtic is less flat-topped. Usually platykurtic has long tails, whereas leptokurtic has short tails. In a leptokurtic distribution, the individual measures are concentrated near the mean, whereas in a platykurtic distribution, the individual measures are spread out across their range.

Most of the results obtained from toxicity studies do not follow a normal distribution. When Weil (1982) examined the distribution pattern of hematological and blood chemistry parameters of toxicological studies, skewness and kurtosis were observed in many cases. Kobayashi (2005) examined the measured items of a carcinogenicity/chronic toxicity study in rats. He reported that majority of hematological and biochemical parameters presented a non-normal distribution—mean corpuscular volume, mean corpuscular hemoglobin, platelets, protein, alanine aminotransferase, aspartate aminotransferase, gamma-glutamyl transpeptidase, creatinine phosphokinase, cholesterol and potassium were skewed positively, whereas hematocrit, hemoglobin, red blood cells and mean corpuscular hemoglobin concentration were negatively skewed.

Tests for Analyzing Normal Distribution

Several tests are available for analyzing normal distribution of the data, for example, Kolmogorov-Smirnov (Chakravarti *et al.*, 1967; Park, 2008), Lilliefors (1967), Shapiro-Wilk's *W* (Shapiro and Wilk, 1965) and Chi-distribution using goodness of fit tests (Snedecor and Cochran, 1989).

The Kolmogorov-Smirnov test is used to analyse continuous distributions. The Lilliefors test is a modified Kolmogorov-Smirnov test. The Shapiro-Wilk *W* test is capable of detecting non-normality for a wide variety of statistical distributions. Owing to this, a lot of attention has been paid to this test in the literature (Sen *et al.*, 2003). The power of Shapiro-Wilk's *W* test for detecting a non-normal distribution is better than other normality tests (Chen, 1971; Liang *et al.*, 2009). The chi-square test is an excellent test to examine whether the data are normally distributed. The major advantage of the chi-square test is that it can be applied to discrete distributions and its disadvantage is that it requires a larger sample size.

Shapiro-Wilk's W test

Let us understand Shapiro-Wilk's *W* test in detail by working out an example given in Table 5.1, body weight of F344 male rats. The data are arranged in an orderly fashion.

Table 5.1. Body weight of F344 male rats

Animal No.	1	2	3	4	5	6	7	8	9	10
Body weight (g)	71	86	92	95	100	102	105	108	118	123
Observation	1	1	2		4				1	1

The data in Table 5.1. is analysed using SAS-JMP and the statistics are given in Tables 5.2. and 5.3. The body weight distribution is given in Figure 5.2.

Table 5.2. Quantiles

100%	**Maximum**	**123.0**
99.5%		123.0
97.5%		123.0
90.0%		122.5
75.0%	**Quartile**	**110.5**
50.0%	**Median**	**101.0**
25.0%	**Quartile**	**90.5**
10.0%		72.5
2.5%		71.0
0.5%		71.0
0.0%	**Minimum**	**71.0**

Note: The term, quantile was introduced by Kendall (1940). Quantiles divide the distributions such that there is a given proportion of observations below the quantile. Quartiles and percentiles are quantiles. Quartile divides the quantile into four equal parts (0–25%, 25–50%, 50–75% and 75–100%). A percentile is the value of a variable below which a certain percent of observations fall. For example, the 10th percentile is that position in a data set which has 90% of data points above it, and 10% below it.

Table 5.3. Estimates

N	10
Sum (Σ)	1000
Mean ($\overline{X}$)	100
Standard error (SE)	4.7981478
Upper 95% mean	110.85416
Lower 95% mean	89.145836
Sum of squares (X-$\overline{X}$)^2	2072
Standard deviation (SD)	15.173076
Variance	230.22222
Coefficient of variation (CV)	15.173076
Skewness	–0.36285
Kurtosis	0.3549171

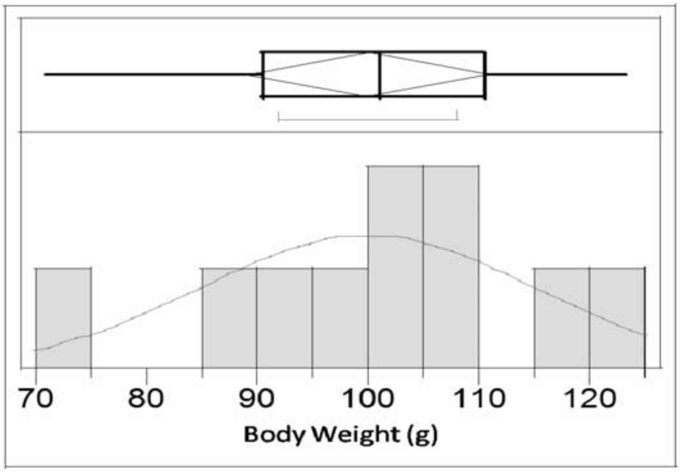


Figure 5.2. Body weight of F344 male rats

Shapiro-Wilk's W test-calculation steps

Step 1: Find the difference between the first set of extreme values (123 and 71 g from Table 5.1). Then find the difference between the second set of extreme values (118 and 86 g). In such a manner find the difference between the extreme values of remaining sets sequentially. If the number of samples is an odd number, ignore the remaining value.

Step 2: Find the Shapiro-Wilk *W* coefficients corresponding to the difference between the extreme values from the Appendix 1. In this example, the number of samples, N=10. The Shapiro-Wilk *W* coefficients corresponding to the difference between the 1st, 2nd, 3rd, 4th and 5th sets of extreme values are 0.5739, 0.3291, 0.2141, 0.1224 and 0.0399, respectively. Calculate the product of difference between extreme values and Shapiro-Wilk *W* coefficients (Table 5.4).

Step 3: Calculate the statistic, *W*, as given below:

$$W = \frac{45.10^2}{2072} = 0.98166$$

Compare *W* (0.98166) with the quantiles of the Shapiro-Wilk *W* test statistic given in Appendix 2. At 10 degrees of freedom the quantiles at 0.95 and 0.98 are 0.978 and 0.983, respectively. Since, the calculated *W* (0.98166) falls between 0.978 and 0.983, it could be concluded that the body weight of all the 10 animals follow a normal distribution. The same is confirmed by Test for goodness of fit:

Test for goodness of fit by Shapiro-Wilk test

W	Prob < W
0.981120	0.9673

Since the probability 0.9673<0.981120 (*W*), it is confirmed that the body weight of all the 10 animals follow a normal distribution pattern.

Table 5.4. Product of difference between extreme values and Shapiro-Wilk *W* coefficients

Animal No.	Body weight (g)	Difference between extreme values (D)		Shapiro-Wilk *W* coefficients (C)	Product (DxC)
1	71	First set	123–71=52	0.5739	29.8428
2	86	Second set	118–86=32	0.3291	10.5312
3	92	Third set	108–92=16	0.2141	3.4256
4	95	Fourth set	105–95=10	0.1224	1.2240
5	100	Fifth set	102–100=2	0.0399	0.0798
6	102	-	-	-	-
7	105	-	-	-	-
8	108	-	-	-	-
9	118	-	-	-	-
10	123	-	-	-	-
Sum					45.10

Power of Shapiro-Wilk's W test

Shapiro-Wilk's *W* test can be used in small as well as large sample sizes (Singh, 2009).

However, the power of this test varies with the number of animals in the group. This can be demonstrated with the help of an example of weight of rats on week 13, in a repeated dose administration study. Four situations are simulated in the example:

Situation 1 (Seventeen observations): 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, and 131. The distribution of the observations is given in Figure 5.3a.

Statistics

Mean	103.05882
SD	16.009648
SE	3.8829099
Upper 95% mean	111.29022
Lower 95% mean	94.827422
N	17

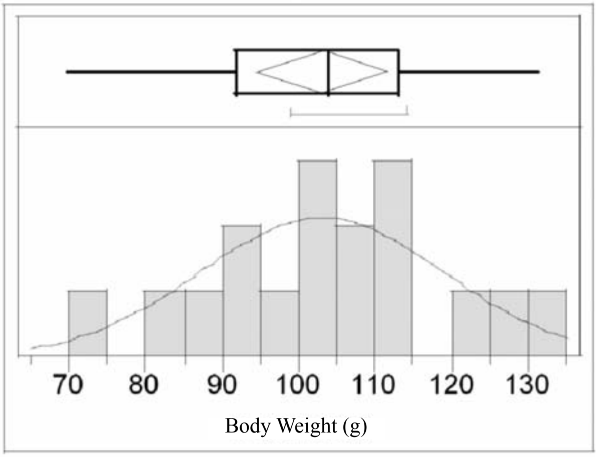


Figure 5.3a. Distribution pattern of body weight (g) of rats—17 observations

Shapiro-Wilk's *W* test

W	Prob <*W*
0.987278	0.9891

Situation 2 (Thirty four observations, the observations of situation 1 are used twice): 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, and 131. The distribution of the observations is given in Figure 5.3b.

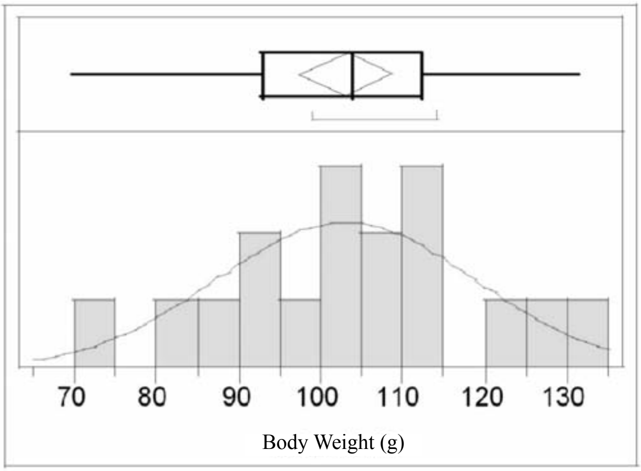


Figure 5.3b. Distribution pattern of body weight (g) of rats—34 observations

Statistics

Mean	103.05882
SD	15.765211
SE	2.7037114
Upper 95% mean	108.55957
Lower 95% mean	97.558081
N	34

Shapiro-Wilk's *W* test

W	Prob <*W*
0.968746	0.5017

Situation 3 (Fifty one observations, the observations of situation 1 are used thrice): 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, and 131. The distribution of the observations is given in Figure 5.3c.

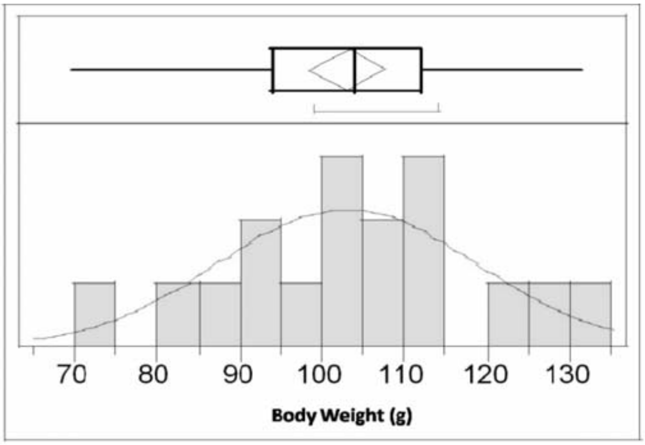


Figure 5.3c. Distribution pattern of body weight (g) of rats—51 observations

Statistics

Mean	103.05882
SD	15.686187
SE	2.1965056
Upper 95% mean	107.47063
Lower 95% mean	98.647012
N	51

Test for goodness of fit, Shapiro-Wilk's *W* test

W	Prob <*W*
0.959888	0.1486

Situation 4 (Sixty eight observations, the observations of situation 1 are used four times): 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, 131, 70, 80, 85, 90, 94, 99, 101, 102, 104, 105, 108, 111, 112, 114, 121, 125, and 131. The distribution of the observations is given in Figure 5.3d.

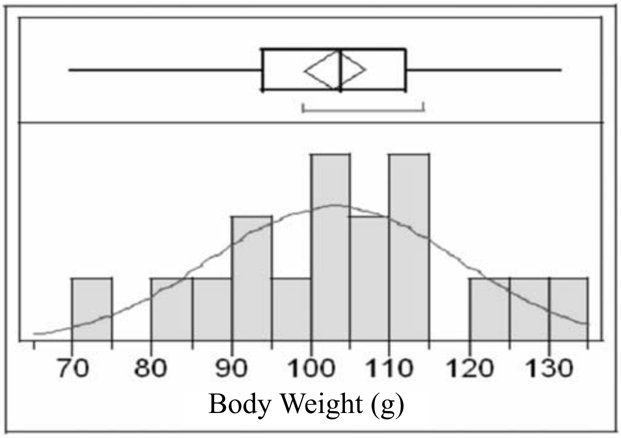


Figure 5.3d. Distribution pattern of body weight (g) of rats—68 observations

Statistics

Mean	103.05882
SD	15.647118
SE	1.8974918
Upper 95% mean	106.84623
Lower 95% mean	99.271414
N	68

Shapiro-Wilk's *W* test

W	Prob <*W*
0.954862	0.0383

The statistics given in Figure 5.3a–Figure 5.3d are consolidated in Table 5.5. Shapiro-Wilk's *W* test revealed a significant *P*, when the number of animals was 68, indicating a non-normal distribution.

Table 5.5. Change in power of Shapiro-Wilk's *W* test with the change in number of animals

N	Mean	Coefficient of variance (%)	Shapiro-Wilk's *W* test	
			W	*P*
17	103	15.5	0.987278	0.9891 (NS)
34		15.3	0.968746	0.5017 (NS)
51		15.2	0.959888	0.1486 (NS)
68		15.2	0.954862	**0.0383** (S)

NS-Not significant (normal distribution); S-Significant (non-normal distribution)

Parametric and Non-parametric Analyses

The two basic assumptions for any statistical analysis are the distribution of the data (normal or non-normal) and homogeneity of variance (homogeneous or heterogeneous). If the variances of the groups are heterogeneous and or the data are non-normally distributed, the choice of the statistical tools is non-parametric (Kobayashi *et al.*, 2011a). Non-parametric tests are also called as 'distribution-free tests'. A parametric test is always based on the assumption that the data follow a normal distribution and variances of the groups are homogeneous.

Analysis of Homogeneity of Variance

One of the assumptions of parametric analysis is that variances of the observations in the individual groups are equal (the other assumption is that the data are normally distributed). When the variances of the groups are equal, the situation is referred to as homogeneity of variance (also called as homoscedasticity of variance). When the variances of the groups are different (not homogeneous), the situation is called as heteroscedasticity.

Bartlett's homogeneity test

In most of the pharmacological and toxicological studies, Bartlett's test is commonly used to examine the data for homogeneity of variance (Bartlett, 1937). However, according to Finney (1995) "Bartlett's test is notorious for its unwanted sensitivity to non-normality of error distribution, and is an untrustworthy instrument for classifying some data sets as homogeneous in variance, other as heterogeneous."

Homogeneity of variance by Bartlett's test is calculated using the below given formula:

$$X^2 cal = 2.3026 \times \frac{\left\{ (\text{Sum of N} - \text{Number of group}) \times \log V - \text{N of each group} - 1 \times \log \text{Sum of Variance} \right\}}{1 + \frac{\frac{1}{\text{Sum of (N of each group} - 1)} - \frac{1}{\text{Sum of total number} - \text{Number of group}}}{3 \times (\text{Numbe of group} - 3)}}$$

where,

$$V = \frac{\left(\text{Variance of each group} \times \text{Sum of } (N - 1)\right)}{(\text{Sum of } N) - \text{Number of group}}$$

X^2 *cal* (chi square calculated) is compared with the value given in chi square Table (N=number of groups-1) at 5% probability level. If the computed value is less than the table value, it is interpreted that the variances of the groups are similar (no heterogeneity). It may be noted that Bartlett's test is not suitable for detecting a heterogeneity when the number of animals in a group very small.

Levene's homogeneity test

Another test used to examine the data for homogeneity of variance is Levene's test (Levene, 1960; Nichols, 1994), which has less sensitivity to non-normality of error distribution. Interestingly, compared to Bartlett's test, Levene's test is less commonly used to analyse the data obtained from toxicological and pharmacological experiments.

Power of Bartlett's and Levene's homogeneity tests

Bartlett's test is used for testing the homogeneity of variance of the data that follow a normal distribution. Bartlett's test is very sensitive to the data that are non-normal to the slightest extent. According to Finney (1995), Bartlett's test is not necessarily to be carried out for examining homogeneity of variance before ANOVA (Analysis of variance, an important statistical tool for comparing more than two groups; you will learn more about ANOVA in Chapter 11). The reason for this is that the power of the Bartlett's homogeneity test is too strong for examining homogeneity of variance, as mentioned above. Toxicity studies using Bartlett's test for testing homogeneity of variance at 1% probability level, which is not so conventional, have been reported (Hayashi *et al.*, 1994; Katsumi *et al.*, 1999; Kudo *et al.*, 2000; Mochizuki *et al.*, 2009; Ishii *et al.*, 2009). The reason for setting a 1% probability level for detecting a significant difference probably could be: if a significant difference is detected by Bartlett's test at the conventional 5% probability level, then the data should be analysed using the non-parametric Dunnett type rank sum test (joint type) (Yamazaki *et al.*, 1981) and/or Dunn test (Hollander and Wolfe, 1973), which have low detection power. Therefore, when the probability level is set at 1%, it is unlikely that the data show a heteroscedacity in variance by Bartlett's test. The reason for this is that to detect a significant difference at 1% probability level, the chi square value has to be larger than that of the 5% probability level.

Do We Need to Examine the Data for Both Normality and Homogeneity?

Kobayashi *et al.* (2011b) made an attempt to compare the statistical tools used to analyse the data of repeated dose administration studies with rodents conducted in 45 countries, with that of Japan. They found that the statistical techniques used for testing the above data for homogeneity of variance are similar in Japan and other countries. In most of the countries, including Japan, the data are generally not tested for normality.

Kobayashi *et al.* (2008; 2011b) suggested that the data may be examined for both homogeneity of variance and normal distribution. However, in bioequivalence clinical trials, because of the limited sample size a reliable determination of the distribution of the data set is not required (EMEA, 2006).

Which Test to be Used for Examining Homogeneity of Variance?

In pharmacological and toxicological experiments, treatments that lower mean values often decrease variance in the treated groups, substantially (Colquhoun, 1971). In these cases, statistical analyses based on the assumption of normal distribution and homogeneity of variance are inappropriate (Spector and Vesell, 2006).

Water consumption of B6C3F1 female mice during the week 13 of a repeated dose administration study is given in Table 5.6. There were four groups and each group consisted of 10 mice. Homogeneity of variances among the groups was analysed using Brown-Forsythe's (Brown and Forsythe, 1974), O'Brien's, Levene's and Bartlett's tests.

Table 5.6. Water consumption (g/week)of B6C3F1 female mice during the week 13 of a repeated dose administration study

Groups	N	Mean± S.D.	*P*			
			O'Brien	Brown-Foresythe	Levene	Bartlett
1	10	43.8 ± 9.0	0.0459	0.0340	0.0014	<0.0001
2	10	35.4 ± 3.4				
3	10	31.9 ± 1.5				
4	10	30.7 ± 2.1				

It is clear from the table that the sensitivity of Bartlett's test is higher, followed by Levene's test. O'Brien's and Brown-Forsythe's tests have very low sensitivity.

Brown-Forsythe's test is a modified Levene's test. Both Brown-Forsythe's and Levene's tests use transformed values (Maxwell and Delaney, 2004). It is more appropriate to use the Levene's, Brown-Forsythe's or O'Brien's tests (O'Brien, 1979; 1981) for testing the homogeneity of variance of the data that follow a non-normal distribution (SAS, 1996). Kobayashi *et al.* (1999) suggested Levene's test for examining homogeneity of variance of the data obtained from toxicity studies.

References

Bartlett, M.S. (1937): Properties of sufficiency and statistical tests. Proceedings of the Royal Statistical Society Series A, 160, 268–282.

Bradlee, J.V. (1968): Distribution-Free Statistical Tests. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Brown, M.B. and Forsythe, A.B. (1974): Robust tests for equality of variances. J. Am. Stat. Assoc., 69, 364–367.

Chakravarti, I.M, Laha, R.G. and Roy, J. (1967): Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, New York, USA.

Chen, E.H. (1971): The power of Shapiro-Wilk *W* test for normality in samples from contaminated normal distribution. J. Am. Stat. Assoc., 66(336), 760–762.

Čisar, P. and Čisar, S.M. (2010): Skewness and kurtosis in function of selection of network traffic distribution. Acta Polytech. Hung., 7(2), 95–106.

Colquhoun, D. (1971): Lecture on Biostatistics. Clarendon Press, Oxford, UK.

EMEA (2006): European Medicines Agency. Biostatistical Methodology in Clinical Trials. ICH Topic E 9—Statistical Principles for Clinical Trials, CPMP/ICH/363/96, London, UK.

Finney, D.J. (1995): Thoughts suggested by a recent paper: Questions on non-parametric analysis of quantitative data (letter to editor). J. Toxicol. Sci., 20(2), 165–170.

Hayashi, T., Yada, H., Auletta, C.S., Daly, I.W., Knezevich, A.L. and Cockrell, B.Y. (1994): A six-month interperitoneal repeated dose toxicity study of tazobactam/piperacillin and tazobactam in rats. J. Toxicol. Sci., 19, Suppl. II, 155–176.

Hollander, M. and Wolf, D.A. (1973): Nonparametric Statistical Methods, John Wiley and Sons, New York, USA.

Ishii, S., Ube, M., Okada, M., Adachi, T., Sugimoto, J., Inoue, Y., Uno, Y. and Mutai, M. (2009): Collaborative work on evaluation of ovarian toxicity (17). J. Toxicol. Sci., 34, SP175–SP188.

Kendall, M.G. (1940): Note on the distribution of quantiles for large samples. Supp. J. Royal Stat. Soc., 7(1), 83–85.

Kobayashi, K. (2005): Analysis of quantitative data obtained from toxicity studies showing non-normal distribution. J. Toxicol. Sci., 30(2), 127–134.

Kobayashi, K., Kitajima, S., Miura, D., Inoue, H., Ohori, K., Takeuchi, H. and Takasaki, K. (1999): Characteristics of quantitative data obtained in toxicity rodents—The necessity of Bartlett's test for homogeneity of variance to introduce a rank test. J. Environ. Biol., 20, 3748.

Kobayashi, K., Pillai, K.S., Suzuki, M. and Wang, J. (2008): Do we need to examine the quantitative data obtained from toxicity studies for both normality and homogeneity of variance? J. Environ. Biol., 29, 47–52.

Kobayashi, K., Sadasivan Pillai, K., Soma Guhatakurta, Cherian, K.M., and Ohnishi, M. (2011b): Statistical tools for analysing the data obtained from repeated dose toxicity studies with rodents: A comparison of the statistical tools used in Japan with that of used in other countries. J. Environ. Biol., 32: 11–16.

Kobayashi, K., Sakuratani, Y., Abe, T., Yamazaki, K., Nishikawa, S., Yamada, J., Hirose, A., Kamata, E. and Hayashi, M. (2011a): Influence of coefficient of variation in determining significant difference of quantitative values obtained from 28-day repeated-dose toxicity studies in rats. J.Toxicol. Sci., 36(1), 63–71.

Kudo, S., Tanase, H., Yamasaki, M., Nakao, M., Miyata, Y., Tsuru, K. and Imai, S. (2000): Collaborative work to evaluate toxicity on male reproductive organs by repeated dose studies in rats (23). J. Toxicol. Sci., 25, SP223–SP232.

Levene, H. (1960): Robust tests for the equality of variances. In: Contributions to Probability and Statistics, Edited Olkin, I. Stanford University Press, USA.

Liang, J., Tang, M.L. and Chan, P.S. (2009): A generalized Shapiro-Wilk's *W* statistic for testing high dimensional normality. Comp. Stat. Data Anal., 53(11), 3883–3891.

Lilliefors, H. (1967): On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc., 62, 399–402.

Maxwell, S.E. and Delaney, H.D. (2004): Designing Experiments and Analysing Data—A Model Comparison Perspective. 2nd Ed., Lawrence Erlbaum Associates, Inc., New Jersey, USA.

Mochizuki, M., Shimizu, S., Urasoko, Y., Umeshita, K., Kamata, T., Kitazawa, T. Nakamura, D., Nishihata, Y., Ohishi, T. and Edamoto, H. (2009): Carbon tetrachloride-induced hepatotoxicity in pregnant and lactating rats. J. Toxicol. Sci., 34(2), 175–181.

Nichols, D. (1994): Levene test, SPSS Inc., nichols@spss.com

O'Brien, R.G. (1979): A general ANOVA method for robust test of additive models for variance. J. American Stat. Asso., 74, 877–880.

O'Brien, R.G. (1981): A simple test for variance effects in experimental designs. Psych. Bull., 89, 570–574.

Park, H.M. (2008): Univariate analysis and normality test using SAS, Stata and SPSS. Univ. Inf. Tech. Serv., Centre Stat. Math. Comp., Indiana Univ., Bloomington, USA.

SAS (1996): JMP Start Statistics. SAS Institute, USA.

Sen, P.K., Jureckov, J. and Picek, J. (2003): Goodness-of-Fit Test of Shapiro-Wilk Type with Nuisance Regression and Scale. Aust. J. Stat., 32(1&2), 163–167.

Shapiro, S.S. and Wilk, M.B. (1965): An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591–611.

Singh, K. (2009): Quantitative Social Research Methods. Sage Publication Pvt. Ltd., New Delhi, India.

Snedecor, G.W. and Cochran, W.G. (1989): Statistical Methods, 8th Edition, Iowa State University Press, Ames, USA.

Spector, R. and Vesell, E.S. (2006): Pharmacology and statistics: Recommendations to strengthen a productive partnership. Pharmacology, 78, 113–122.

Weil, C.S. (1982): Statistical analysis and normality of selected hematologic and clinical chemistry measurements used in toxicologic studies. Arch. Toxicol. Suppl., 5, 237–253.

Yamazaki, M., Noguchi, Y., Tanda, M. and Shintani, S. (1981): Statistical method appropriate for general toxicological studies in rats. J. Takeda Res. Lab., 40(3/4), 163–187.

6

Transformation of Data and Outliers

Transformation of Data

There are situations in pharmacological and toxicological experiments that the data show heterogeneous variance across the groups of animals. Using parametric tests to analyse such data may give rise to Type I error. One way to overcome this situation is to transform the data (Wallenstein *et al.,* 1980). It is most likely that the variance of the transformed data show homogeneity.

In Table 6.1, transformed values of alanine aminotransferase activity of Wistar rats of the control group in a 14-day repeated dose administration study is given.

Table 6.1. Alanine aminotransferase activity (U/L) of Wistar rats of the control group in a 14-day repeated dose administration study

45.3, 63.8, 82, 42, 40.8, 38.2, 35.9, 37.9, 39.1, 35.5 (N=10)		
Transformation	Mean±SD	CV (%)
None	46 ± 15	32.7
Logarithm	1.6 ± 0.12	7.2
Square root	6.7 ± 1.0	15.0
Reciprocal	0.02 ± 0.005	22.8

For the non-transformed data, the CV was 32.7%, which substantially decreased, when the data were transformed to logarithms. CV also decreased when the data were transformed to square roots and reciprocals, but in a lesser magnitude than the logarithmic-transformed data.

Concentrations of blood constituents usually show a non-normal distribution (Flynn *et al*., 1974). Therefore, statistical analysis is usually carried out with the transformed values of blood constituents (Niewczas *et al*., 2009). According to Lew (2007), in pharmacology, the data may be transformed to their logarithms in order to eliminate heterogeneity in variation. For example, plasma/serum concentration of drug and/or its metabolites in drug metabolism and pharmacokinetic studies (DMPK) in laboratory animals (Girard *et al*., 1992; Steinke *et al*., 2000; Zheng *et al*., 2010) and bioavailability/bioequivalence (BA/BE) studies in volunteers (Dubey *et al*., 2009) are usually analysed in their logarithmic-transformed values. FDA (2003) and EMEA (2006) recommend logarithmic-transformation of exposure measures before statistical analysis in BA/BE studies. It should be borne in mind that the data showing a non-normal distribution may also display other patterns of uneven variation that cannot be easily eliminated (Keppel and Wickens, 2004).

Statistical analysis using transformed values are not the same as using measured values. Therefore, interpreting the transformed values may be difficult (Jenifer, 2010). In the words of Finney (1995), "When a scientist measures a quantity such as concentration of a chemical compound in a body fluid, his interest usually lies in the scale, perhaps mg/ml, that he has used; he is less likely to be interested in a summary of results relating to a transformed quantity such as the logarithm of blood concentration. If he analyzes in terms of logarithm, encouraged perhaps by an elementary but uncritical statistical textbook or by a convenient software package, he may find significant differences but to express his conclusions in meaningful numbers may be impossible. I do not assert that a scientist should never transform data before analysis; I urge that data should be transformed only after careful consideration of all consequences. Textbook implications that; 'In certain specified circumstance, data must be transformed' should not be unthinkingly accepted. Remember that any transformation is likely to increase the difficulty of interpreting results in relation to the original measurements." Therefore, when a significant difference is obtained for transformed values, following a statistical analysis, it is necessary to describe that the significant difference obtained is for the transformed values.

Outliers

Data obtained from pharmacological and toxicological studies are not free from outliers. An outlier can be defined as 'an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism' (Hawkins, 1980). Outliers can have deleterious effects on statistical analyses (Rasmussen, 1988; Schwager and Margolin, 1982). Outliers increase error rates and distort statistical estimates when using either parametric or nonparametric tests (Zimmerman, 1995; 1998). Outliers arise from two sources—from errors in the data and from the inherent variability of the data (Anscombe, 1960). According to Barnett and Lewis (1994), 'not all outliers are illegitimate contaminants, and not all illegitimate scores show up as outliers'.

Hypoglycemic property of a drug was evaluated in alloxan-induced hyperglycemic rats. These rats were divided into two groups (5 rats/group), Groups 1 and 2. Group 1 (control) was treated with vehicle and Group 2 was treated with the drug. Following the administration of vehicle or drug, blood glucose was determined in individual rat (Table 6.2.).

Table 6.2. Blood glucose (mg/dl) in alloxan-treated rats following administration of drug

Group 1 (Vehicle treated)	Group 2 (Drug treated)
189, 195, 169, 206, 175	138, 161, 156, 171, ***259***
Mean ± SD = 186.8 ± 14.9 (n=5)	Mean ± SD = 177.0 ± 47.4 (n=5)
	Mean ± SD = 156.5 ± 13.4 (n=4)

The blood glucose level of the vehicle treated group was 186.8 ± 14.9 mg/dl (mean ± SD), whereas the drug treated group was 177.0 ± 47.4 mg/dl (mean ± SD). Though a decrease in blood glucose level was observed in the drug treated animals, it was statistically insignificant by Aspin Welch's *t*-test using one-sided (we used Aspin Welch's *t*-test because the variance of the groups is different. You will read more about this test in Chapter 8). The SD of drug treated group exploded considerably, indicating a large variance. Close observation of the individual values of the drug treated animals shows that all the values in this group are close to each other, except the value, 259. Let us recompute the mean and SD of this group, after removing 259 from the data. The revised mean ± SD is 156.5 ± 13.4 (n=4). We are comfortable with this SD, as this is very close to the SD of the vehicle treated group, indicating a homogeneity of variance between the vehicle treated and drug treated animals. The blood glucose of drug treated animals (after removing the value, 259) is statistically different from the vehicle treated animals by Student's *t*-test (we used the Student's *t*-test because the variance of the groups is not different. You will read more about this test in Chapter 8). In this example, the value 259 is an outlier, as it clearly stands out of other values, but in many pharmacological and toxicological experiments it is not easy to spot an outlier. A simple method to identify an outlier mentioned in several books on statistics is given below (Hogan and Evalenko, 2006):

Lower outlier = 25th percentile – (1.5 x IQR)
Upper outlier = 75th percentile + (1.5 x IQR)

Readers may go back to Chapter 2 and refresh their memory on box-and-whisker plot and IQR (inter-quartile range or hinge spread).

There are several statistical tools available for detecting an outlier. Among them, the Dixon test and Grubb test are widely used (Verma and Ruiz, 2006) and these tests are suggested by ASTM (2008). Outlier tests suggested in USP (2008) are ESD test, Dixon-Type test and Hampel's rule.

We shall discuss 3 outlier tests in detail:

1. **Masuyama's Rejection Limit Test** (Shibata, 1970)

Let us examine whether the value 259 of the example given in Table 6.2 is an outlier. Masuyama's rejection limit test is calculated using the following equation:

$$\overline{X} \pm \left(Sx \cdot \sqrt{\frac{n+1}{n}} \cdot t_{(n-1)\,0.05} \right), \text{ where}$$

Sx: Standard deviation; $t_{(n\text{-}1)0.05}$ is *t* value at 5% probability level (n–1 degrees of freedom).

The mean and SD of the data (138, 161, 156, 171, ***259***) given in Table 6.2 are;

Mean = 177.0; SD = 47.4 (n=5)

$t_{(5-1)0.05} = 2.776$ [from *t* Table by two-tailed test]

$$\text{Rejection limits} = 177 \pm (47.4 \times \sqrt{\frac{5+1}{5}} \times 2.776) = 177 \pm 157.89$$

$\therefore 19.11 \sim 334.89$

As indicated above, Masuyama's rejection limit test gives the rejection limits in a wider range. Masuyama's rejection test is not sensitive in detecting an outlier. Hence, use of this test should be done in toxicology/pharmacology with a little caution.

2. Thompson's Rejection Test (Thompson, 1935)

Let us again work out the example of blood glucose levels of drug-treated rats given in Table 6.2. The values are 138, 161, 156, 171, ***259*** mg/dl. We shall apply Thompson's rejection test to examine whether the value, ***259*** mg/dl is an outlier.

$\Sigma X = 885$, $\overline{X} = 177$, Sum of squares (SS), $\Sigma(X-\overline{X})^2 = 8978$

$\therefore \delta = 177–259 = –82$

$$Sn = \sqrt{\frac{8978}{5}} = \sqrt{1795.6} = 42.37$$

$$\therefore \tau = \frac{-82}{42.37} = -1.94$$

When you substitute these calculations for the expression of *t*:

$$t_{(5-2)} = \frac{-1.94\sqrt{5-2}}{\sqrt{5-1-\left(-1.94^2\right)}}$$

$$\therefore t_{(3)} = 14.2$$

The Table value for *t* at 0.001 probability level (Table 6.3) for three degrees of freedom, is 12.923. Since the calculated *t* value is greater than the table value, we consider the blood glucose value, ***259*** mg/ dl is an outlier.

Table 6.3. *t* test critical values (Yoshimura, 1987)

df\2α	0.2	0.1	0.05	0.02	0.01	0.002	0.001
df\α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
2	1.885	2.919	4.302	6.964	9.924	22.327	31.59
3	1.637	2.353	3.182	4.540	5.840	10.214	**12.923**
4	1.533	2.131	2.776	3.746	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	2.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587

α=One-sided, 2α=Two-sided test.

3. Smirnov-Grubbs' Rejection Test (Grubbs, 1969)

Smirnov-Grubbs' rejection test is one of the tests for outliers used widely in various fields of biology (Sunaga *et al.*, 2006; Kawano *et al.*, 2007; Ishikawa *et al.*, 2010; Okubo *et al.*, 2010).

In animal experiments, the Smirnov-Grubbs' test is used more frequently than the Thompson's rejection test. Smirnov-Grubbs' test has a high power when the outlier is only one observation. However, when outliers are two or more observations, power of this test decreases due to the masking effect of one outlier to the other.

The calculation procedure of Smirnov- Grubbs' test is very simple. We can use the same example that we used for Thompson's rejection test.

First, calculate T_n.

$$T_n = \frac{(X_1 - \overline{X})}{\sqrt{V}},$$

Where n = Number of samples; X_1= The outlier.

Blood glucose level of drug treated rats are 138, 161, 156, 171, ***259*** mg/dl.

$\Sigma X = 885$, $\overline{X} = 177$, Sum of squares (SS), $\Sigma(X - \overline{X})^2 = 8978$, Variance ($V$) = 1795.6.

$$T_5 = \frac{|259 - 177\ |}{\sqrt{1795.6}} = \frac{82}{42.37} = 1.94$$

The Table value for Smirnov- Grubbs at 0.01 probability level (Table 6.4) for 5 degrees of freedom, is 1.749. Since the calculated value (1.94) is greater than the table value (1.749), the test confirms that the blood glucose value ***259*** mg/dl is an outlier.

A Cautionary Note

Though human and other errors are major contributing factors for outliers, a positive outcome from an outlier test should be investigated (Ellison *et al.*, 2009). Before discarding an outlier, one has to confirm that the value discarded as an outlier is not a genuine data point. Hubrecht and Kirkwood (2010) suggested that one way to deal with an outlier is to carry out the statistical analysis with and without it. If the analytical results provide

Table 6.4. Smirnov-Grubbs' Table[a] (Aoki, 2002; 2006)

N	0.1	0.05	0.025	0.01
3	1.148	1.153	1.154	1.155
4	1.425	1.462	1.481	1.493
5	1.602	1.671	1.715	1.749
6	1.729	1.822	1.887	1.944
7	1.828	1.938	2.020	2.097
8	1.909	2.032	2.127	2.221
9	1.977	2.110	2.215	2.323
10	2.036	2.176	2.290	2.410
11	2.088	2.234	2.355	2.484
12	2.134	2.285	2.412	2.549
13	2.176	2.331	2.462	2.607
14	2.213	2.372	2.507	2.658
15	2.248	2.409	2.548	2.705
16	2.279	2.443	2.586	2.747
17	2.309	2.475	2.620	2.785
18	2.336	2.504	2.652	2.821
19	2.361	2.531	2.681	2.853
20	2.385	2.557	2.708	2.884

[a]One-sided table.

similar interpretation, the outlier should not be discarded. By merely not falling in the 'expected' range should not be the only reason for considering a data point as an outlier and discarding it (Petrie and Sabin, 2009). Let us examine the data on hemoglobin concentration of F344 male rats on week 104 in a repeated dose administration study given in Figure 6.1.

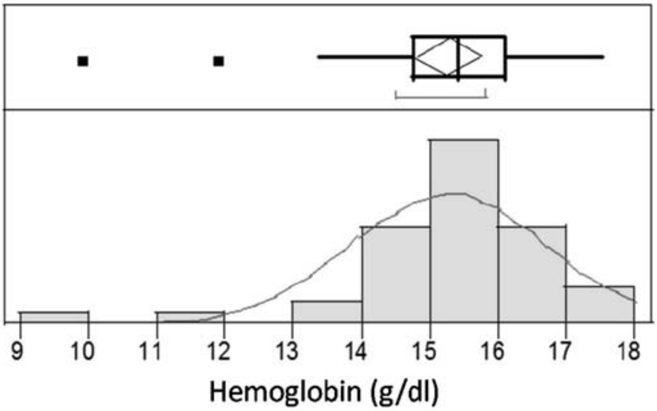


Figure 6.1. Hemoglobin concentration (g/dl) of F344 male rats on week 104

The data between 9 and 13 g/dl, appear to be outliers. Box-and-Whisker plot given in the upper section of the Figure provides useful information on the spread of the data and two outlier data points. It may be also possible that an outlier test done on the data of the Figure 6.1 confirms this view. But the values lower than 13 g/dl should not be considered as outliers, since this is how hemoglobin is distributed in the rat population of the study, which is non-normal. However, according to Ye (2003), an outlier is valid if it represents an accurate measurement and still falls well outside range of majority of values.

Non-normal distribution of several parameters is normally seen in biological experiments. In a non-normal distribution, the data points that fall outside the range of majority of the values should not be considered as outliers. It is worth mentioning here that in bioequivalence trials the regulatory agencies may permit exclusion of outliers from the statistical analysis if they are caused by product or process failure but the regulatory agencies may not permit exclusion of outliers from the statistical analysis if they are caused by subject-by-treatment interaction (Schall *et al.*, 2010).

References

Anscombe, F.J. (1960): Rejection of outliers. Technometrics, 2, 123–147.

Aoki, S. (2002): http://aoki2.si.gunma-u.ac.jp/lecture/Grubbs/Grubbs-table.html

Aoki, S. (2006): http://aoki2.si.gunma-u.ac.jp/lecture/Grubbs/Grubbs.html

ASTM (2008): American Society for Testing Materials. Standard Practice for Dealing With Outlying Observations, ASTM E178-08, ASTM International Philadelphia, USA.

Barnett, V. and Lewis, T. (1994): Outliers in Statistical Data, 3rd Edition. Wiley, New York, USA.

Dubey, S.K., Patni, A., Khuroo, A., Thudi, N.R., Reyar, S., Arun Kumar, Tomar, M.S., Jain, R., Nand Kumar and Monif, T. (2009): A quantitative analysis of memantine in human plasma using ultra performance liquid chromatography/Tandem mass spectrometry. E- J. Chem., 6(4), 1063–1070.

Ellison, S.L.R., Barwick, V.J. and Farrant, T.J.D. (2009): Practical Statistics for the Analytical Scientist—A Bench Guide. 2nd Edition, The Royal Society of Chemistry, Cambridge, U.K.

EMEA (2006): European Medicines Agency. Biostatistical Methodology in Clinical Trials. ICH Topic E 9—Statistical Principles for Clinical Trials, CPMP/ICH/363/96, London, UK.

FDA (2003): Food and Drug Administration. Guidance for Industry Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Rockville, USA.

Finney, D.J. (1995): Thoughts suggested by a recent paper: Questions on non-parametric analysis of quantitative data (letter to editor). J. Toxicol. Sci., 20(2), 165–170.

Flynn, F.V., Piper, K.A.J., Garcia-Webb, P., McPherson, K. and Healy, M.J.R. (1974): The frequency distributions of commonly determined blood constituents in healthy blood donors. Clin. Chim. Acta, 52,163–171.

Girard, D., Gootz, T.D. and Mcguirk, P.R. (1992): Studies of CP-74667, a new quinolone, in laboratory animals. Antimicrobial Agents and Chemotherapy, 36(8), 11671–1676.

Grubbs, F.E. (1969): Procedures for detecting outlying observations in samples. Technometrics, 11, 1–21.

Hawkins, D.M. (1980): Identification of Outliers. Chapman and Hall Ltd., New York, USA.

Hogan, T.P. and Evalenko, K. (2006): The elusive definition of outliers in introductory statistics text books. Teaching of Psychology, 33, 252–256.

Hubrecht, R. and Kirkwood, J. (2010): The UFAW Handbook on the Care and Management of Laboratory and Other Research Animals. John Wiley and Sons, West Sussex, UK.

Ishikawa, Y., Kiyoi, H., Watanabe, K., Miyamura, K., Nakano, Y., Kitamura, K., Kohno, A., Sugiura, I., Yokozawa, T., Hanamura, A., Yamamoto, K., Iida, H., Emi, N., Suzuki, R., Ohnishi, K. and Naoe, T. (2010): Trough plasma concentration of imatinib reflects BCR-ABL kinase inhibitory activity and clinical response in chronic-phase chronic myeloid leukemia: a report from the BINGO study. Cancer Sci., 101(10), 2186–2192.

Jenifer, L.H. (2010): S Guide to Doing Statistics in Second Language Research Using SPSS. Taylor & Francis, New York, USA.

Kawano, N., Egashira, Y. and Sanada, H. (2007): Effect of dietary fiber in edible seaweeds on the development of D-galactosamine-induced hepatopathy in rats. J. Nutr. Sci. Vitaminol., 53(5), 446–450.

Keppel, G. and Wickens, T.D. (2004): Design and Analysis, a Researcher's Handbook. 4th Edition, Pearson Prentice Hall, New Jersey, USA.

Lew, M. (2007): Good statistical practice in pharmacology Problem 1. Br. J. Pharmacol., 152(3), 295–298.

Niewczas, M.A., Ficociello, L.H., Johnson, A.C., Walker, W., Rosolowsky, E.T., Roshan, B., Warram, J.H. and Krolewski, A.S. (2009): Pathways and renal function in nonproteinuric patients with type 1 diabetes. Clin. J. Am. Soc. Nephrol., 4, 62–70.

Okubo, Y., Kaneoka, K., Imai, A., Shiina, I., Tatsumura, M., Izumi, S. and Miyakawa, S. (2010): Comparison of the activities of the deep trunk muscles measured using intramuscular and surface electromyography. J. Mech. Med. Biol., 10(4), 611–620.

Petrie, A. and Sabin, C. (2009): Medical Statistics at a Glance. 3rd Edition. Wiley-Blackwell, Chichester, UK.

Rasmussen, J.L. (1988): Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. Multivariate Behavioral Res., 23(2), 189–202.

Schall, R., Endrenyi, L. and Ring, A. (2010): Residuals and outliers in replicate design crossover studies J. Biopharm. Stat., 20(4), 835–849.

Schwager, S.J. and Margolin, B.H. (1982): Detection of multivariate outliers. Ann. Stat., 10, 943–954.

Shibata, K. (1970): Biostatistics, Tokyo University of Agriculture, Tokyo, Japan.

Steinke, W., Archimbaud, Y., Becka, M., Binder, R., Busch, U., Dupont, P. and Maas, J. (2000): Quantitative distribution studies in animals: Cross-validation of radioluminography versus liquid-scintillation measurement. Reg. Toxicol. Pharmacol., 31, S33–S43.

Sunaga, H., Kaneko, M. and Amaki, Y. (2006): The efficacy of intratracheal administration of vecuronium in rats, compared with intravenous and intramuscular administration. Int. Anesthesia Res. Soc., 103(3), 601–607.

Thompson, W.R. (1935): On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation, Ann. Math. Stat., 6, 215–219.

USP (2008): The United States Pharmacopeia, The National Formularly, USP 31, NF 26, Asian Edition, Volume1, Port City Press, Baltimore, USA.

Verma, S.P. and Ruiz, A.Q. (2006): Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. Revista Mexicana de Ciencias Geológicas, 23(2), 133–161.

Wallenstein, S., Zucker, C.L. and Fleiss, J.L. (1980): Some statistical methods useful in circulation research. Circ. Res., 47, 1–9.

Ye, N. (2003): The Handbook of Data Mining. Lawrence Elbaum Associate Inc., New Jersey, USA.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Press, Tokyo, Japan.

Zheng, Y., Liu, H., Ma, G., Yang, P., Zhang, L., Gu, Y., Zhu, Q., Shao, T., Zhang, P., Zhu, Y., and Cai, W. (2010): Determination of S-propargyl-cysteine in rat plasma by mixed-mode reversed-phase and cation-exchange HPLC–MS/MS method and its application to pharmacokinetic studies. J. Pharm. Biomed. Anal., 54(5), 1187–1191.

Zimmerman, D.W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. J. Exp. Edu., 64(1), 71–78.

Zimmerman, D.W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. J. Exp. Edu., 67(1), 55–68.

7

Tests for Significant Differences

Null Hypothesis

The main objective of conducting an animal experiment is to know whether the treatment with a test item causes any effect compared to the control group. The comparison between the treatment group/s and the control group is made using various statistical tools. The selection of an appropriate statistical tool is based on certain assumptions. Before we go further, we need to understand a hypothesis called 'null hypothesis'.

In the statistical context, a hypothesis is a statement about a distribution (example, normal distribution), or its underlying parameter (example, mean value, μ) or a statement about the relationship between probability distribution (example, there is no statistical difference between the treated and the control groups) or its parameter ($\mu_1 = \mu_2$) (Le, 2009). Why is it called as 'null hypothesis'? Let us try to understand 'null hypothesis' using the explanation proposed by Yoshida (1980). No pharmaceutical company will venture in developing a new drug, A1, if it is not superior to the drug currently in use, A2. In a statistical analysis, we first hypothesize that drugs A1 and A2 have the same therapeutic value. That is, we hypothesize A1 = A2, which is contrary to our assumption A1 > A2. When the experimental data fail to show A1 = A2, we judge that A1 differs from A2 and reject the hypothesis. Thus, in a statistical test, we first hypothesize A1 = A2 in contrast to our assumption A1 > A2, and then show that it is not true (A1 ≠ A2). The original hypothesis A1 = A2, which is desirably rejected, is called the null hypothesis. In most of the statistical books null hypothesis is notated as:

$H_0=\mu_1= \mu_2$, and the alternate hypothesis is notated as:

$H_1=\mu_1\neq\mu_2$, where μ_1 and μ_2 are the mean values of two groups.

Generally, a statistical process starts from the null hypothesis, which assumes no difference between the control group and the treated group or among the groups, and if a significant difference is detected at 5% significant level ($P<0.05$), the null hypothesis is rejected.

Significant Level, Type I and Type II Errors

In the publications of pharmacological and toxicological experiments one would have come across authors using $P<0.05$, usually as footnotes of the Tables to denote a significant difference. *P* stands for probability. In order to detect a significant difference we have to challenge the null hypothesis. When $P <0.05$, the null hypothesis is rejected. It means there is only a 5% chance of rejecting null hypothesis, when it is true. We are not supposed to reject null hypothesis, when it is true, if we reject it, it is called as Type I error. In a pharmacological experiment, if you reject the null hypothesis, when actually it is true, *i.e.*, $H_0=\mu_1= \mu_2$ (there is no difference between the treated and control groups), you would report that the drug that you tested had an effect, causing a Type I error. Hence, this Type I error is also called as 'false positive'. Experimental design in pharmacology should be proper so that misleading claims concerning the effectiveness of a treatment (Type I error) are not made (Spina, 2007). Type II error is opposite to Type I error, also called as 'false negative', occurs when you falsely accept the null hypothesis.

Why at 5% Significant Level?

In statistical analysis, the smallest probability for rejecting a null hypothesis, when it is true, is considered as 5% (Madsen, 2011). The same is used in most of the pharmacological and toxicological studies, where a significant difference between the treated and the control groups is judged at 5% probability level. Why the statisticians look upon 5% probability as the cut-off point for assessing a significant difference? Let us try to explain it with an example: A tennis player loses several matches against an opponent of supposedly equal skill level. How many losses will be required for the player to regard the opponent as a better player than him? It is not odd for a player to lose three consecutive games to his opponent with equal ability, but the fourth consecutive loss leads the player to believe that his opponent to be a better player. After losing five consecutive games, the player may abandon the null hypothesis (null hypothesis in this case is that the player and his opponent have equal skill level) and consider that his opponent is a better player than him. If the player and his opponent have equal ability, the probability of losing the game once by the player is 1/2, but the probabilities of losing four and five games consecutively by the player are $(1/2)^4 = 6.3\%$ and $(1/2)^5 = 3.2\%$, respectively. The mid-point of these probabilities is about 5% [(6.3+3.2)/2=4.8%)].

The five percent significant level which implies 1 mistake in 20 observations (1/20) is normally unavoidable in biological experiments and has been used for more than half a century in bioassays including toxicity tests (Dunnett, 1955; Kornegay *et al.*, 1961). Hence, according to Bailey (1995), the five percent significant level can be generally used for flagging a significant difference. Conventionally, a *P* value of <0.05 indicates statistical significance (Doll and Carney, 2005).

However, strictly adhering to a 5% significant level to delineate a significant difference has been questioned by few statisticians. Fisher (1955) recommended a 5% significant level based on a single hypothesis, H_0. Neyman and Pearson (1928, 1936) proposed a decision process which seeks to confirm or reject *a priori* hypothesis and rejected Fisher's idea that only the null hypothesis needs to be tested. Statisticians posed questions against Fisher's 5% probability level; the question was 'what should be the smallest *P* value that warrants rejection of the null hypothesis?' In later years, Fisher (1971) stated that the *Q* value can be significant at a 'higher standard, if *P* is 1%' and at a 'lower standard if *P* is 5%'. It again states, though indirectly, that a significant difference can be obtained only when the *P* is between 1 and 5%. (Note: *Q* value is the 'false discovery rate' analogue of *P*).

Many statisticians do not favor strictly characterizing the result of a statistical analysis into a positive or negative finding on the basis of a *P* value. They suggest, when reporting the results of significance tests, precise *P* values (example, $P<0.049$ or $P<0.051$) should be reported rather than referring to specific critical values. Interpretation of the results of a statistical analysis should not be made solely on the basis of null hypothesis. The hypothesis testing has been challenged and there has been suggestion to report confidence intervals rather than *P* (Krantz, 1999). According to Gelman and Stern (2006) 'dichotomization into significant and non-significant results encourages the dismissal of observed differences in favor of the usually less interesting null hypothesis of no difference'. In the case of experiments conducted in pharmacology and toxicology, biological relevance of the results also should be considered for interpreting the data. Declaring a result non-significant does not mean that the effect is not biologically relevant; it only means that there is not sufficient evidence to reject the null hypothesis. In a nutshell, statistical analysis should not override the experience of the experimenter in interpreting the results of the experiments.

How to Express *P*?

The published articles express the *P* in two ways: P <0.05 or $P \leq 0.05$. The question is how the *P* should be expressed—P <0.05 or $P \leq 0.05$? Though, technically, it may be better to express $P \leq 0.05$, P<0.05 also conveys similar information on statistical significance. We conducted a small investigation on the expression of *P* in toxicological/pharmacological articles published in few journals. In most of the journals investigated, we observed that $P \leq 0.05$ and P <0.05 were used at similar frequencies. In the toxicological/ pharmacological experiments conducted in Japan, P<0.05 tended to be used slightly more frequently than $P \leq 0.05$. In the technological report of the National Toxicology Program of NIH, USA, P <0.05 is more widely used.

One-sided and Two-sided Tests

Generally, it has been stated that a one-sided test is used in the following cases: 1) the difference, large or small is questioned and 2) the inter-group difference (plus or minus) is known in advance. On the other hand, a two-sided test is used in the following cases: 1) only the presence or absence of an inter-group difference is questioned and 2) it is not certain whether the inter-group difference is plus (positive) or minus (negative). The detection rate of a significant difference differs depending on the selection of a one-sided or a two-sided test. Let us work out an interesting example: A customer went to a grocery shop to buy a loaf of bread. The weight of a loaf of bread printed on the bread wrappers was 450 g. On a hunch, the customer purchased one loaf of bread from the shop daily for seven days and weighed the loaves. The weights were 444, 434, 450, 430, 458, 446 and 422 g. He informed the grocer that the weight printed on the bread wrapper did not match with the actual weight of the bread. The grocer offered to analyse the data provided by the customer using a two-sided test. The calculated *t* value (2.14) was less than the value of *t*-distribution Table

(2.447), hence the null hypothesis was not rejected (Note: Normally we analyse the data using a statistical formula to obtain a 'calculated value'. Then, we compare this 'calculated value' with the value (critical value) given in the appropriate statistical Table. If the calculated value is greater than the Table value (critical value), we consider the null hypothesis is rejected. In this particular example we have analysed the data using a *t*-test and got a *t* value. This *t* value was compared with the value given in the *t* Table. You shall learn about various statistical tools and their applications in later chapters). Not-rejection of the null hypothesis means there is no statistical significant difference among the weights of seven loaves of the bread that the customer purchased. The customer was not convinced with the result of the two-sided test provided by the grocer. The customer decided to analyse the data using a one-sided test, with the assumption that the weight of the loaf of the bread is less than 450 g. When the customer analysed the data using the one-sided test, he found that the calculated *t* value (2.14) was greater than the value of *t*-distribution Table (1.943). Therefore, "Null hypothesis" is rejected, which means that there is a statistical significant difference among the weights of seven loaves of the bread that he purchased.

Which Test to Use: One-sided or Two-sided?

It is interesting to note that scientists have different views in choosing between one-sided test and two-sided test. Kobayashi *et al.* (2008) examined whether a one-sided or a two-sided test was used in the analysis of the data obtained from 122 numbers of 28-day repeated dose administration studies in rats. The studies were conducted as per Chemical Substances Control Law, Japan (CSCL, 1986) or OECD test guideline (OECD, 2008). Out of 122 studies examined, quantitative data of 22 studies were analysed by the one-sided test, 87 studies were analysed by two-sided test, whereas there was no mention about whether the one-sided or two-sided test was used in 13 studies. With regard to qualitative data, in 34 and 22 studies the data were analysed by the one-sided and two-sided tests, respectively, whereas there was no mention about whether the one-sided or two-sided test was used in 66 studies.

Drewitt *et al.* (1993) used a two-sided *t*-test for preliminary studies and one-sided test for the main studies. Shertzer and Sainsbury (1991) used a one-sided *t*-test for the detection of a significant difference between two groups. Yoshimura and Ohashi (1992) recommended the one-sided test because the results of a toxicity study are evaluated by the presence or absence of an increase in the mean value of the treated groups in comparison with the control group. Shirley (1997) used a two-sided test for Student's *t*-test and Cochran's *t*-test, and if a significant difference is observed in the ANOVA, used the one-side test in Dunnett's multiple comparison test. Dunnett (1955) recommended the use of a two-sided test to determine simultaneously the upper and lower limits to the difference between the control group and each treated group and a one-sided test to determine either the upper or lower limit to the difference between the control group and each treated group. Gad and Weil (1988) explained the significant difference between the control and treated groups in body weight by using the two-sided test. Sakuma (1977) suggested to select either a one- or a two-sided test referring to the reports of similar studies. Nakamura (1986) stated that selection of one- or two-sided test may depend on the objective of the study, and he suggested that the statistical significance of the data should not be foreseen. Kobayashi (1997) recommended a one-sided test for the analysis of data obtained from toxicological studies.

A significant difference is more apt to be observed in a one-sided test than in a two-sided test. According to a survey, the detectability of a significant difference by the two-sided test was 71–95% of that by a one-sided test in the Dunnett's *t*-test (Table 7.1) (Kobayashi, 1997).

Table 7.1. Difference in number of significant differences ($P < 0.05$) by one- and two-sided test by Dunnett's *t*-test in a chronic toxicity and carcinogenicity study

Items	No. of statistical analyses	Dunnett's *t*-test	
		One-sided	Two-sided
Body weight (b.w.)	528	223	212 (95)
Feed consumption	832	235	189 (80)
Hematology	352	123	105 (85)
Blood chemistry	576	215	181 (84)
Urinalysis	64	7	5 (71)
Organ weight	224	47	42 (89)
Organ weight/b.w.	224	82	67 (81)
Total	2800	932	801 (86)

Note: Values in parentheses show the percent significant difference by two-sided test with regard to one-sided test.

Overall significant difference shown by the two-sided test is 86% of the one-sided test. The reason for this is that one-sided test requires less strength of evidence than the two-sided test for a significant difference. It is likely that an item shown as insignificant by a two-tailed test can be significant by a one-sided test. One-sided test should never be used to make a conventionally non-significant difference significant (Bland and

Bland, 1994). Therefore, it is important to decide to use a one-sided or a two-sided test before the data collection (Rosner, 2010).

The rejection limit value at 5% probability level of *t*-test and Dunnett's multiple comparison test was excerpted and is shown in Table 7.2. The rejection limit value of the one-sided test does not become 1/2 the value of the Table of the two-sided test, but it becomes 78% of two-sided test in *t*-test, and it becomes 85% of two-sided test at four groups setting in Dunnett's multiple comparison test.

Table 7.2. Rejection limits of *t*-test and Dunnett's multiple comparison test with one- and two-sided (Yoshimura, 1987)

DF	Rejection limit at 5% level			
	t-Table		Dunnett's Table[a]	
	Two-sided	One-sided	Two-sided	One-sided
1	12.706	6.314	–	–
2	4.303	2.920	–	–
3	3.182	2.353	3.867	2.912
4	2.776	2.132	3.310	2.598
5	2.571	2.015	3.030	2.433
6	2.447	1.943	2.863	2.332
7	2.365	1.895	2.752	2.264
8	2.306	1.860	2.673	2.215
9	2.262	1.833	2.614	2.178
10	2.228	1.812	2.268	2.149
•	•	•	•	•
•	•	•	•	•
21	2.080	1.721	2.370	2.021
22	2.074	1.717	2.363	2.016
•	•	•	•	•
•	•	•	•	•
31	2.040	1.696	2.317	1.986
32	2.037	1.694	2.314	1.984
•	•	•	•	•
•	•	•	•	•
41	2.020	1.683	2.291	1.969
42	2.018	1.682	2.289	1.968
•	•	•	•	•
•	•	•	•	•
60	2.000	1.671	2.265	1.952
120	1.980	1.657	2.238	1.934
240	1.970	1.651	–	–
∽	1.960	1.645	2.212	1.916
Rate[b]	1:0.78		1:0.85	

[a] Four groups setting.

[b]Value when total of two-sided is assumed to be one.

The decision to use a one-sided or a two-sided test should be made carefully, as it has an impact on sample size calculation. Minimum sample size required for one-sided test is less, because it focuses on only tail of the probability distribution (Moye and Tita, 2002). The decision to use a one-sided or a two-sided test also has an impact on assessment of study results by regulatory authorities (Freedman, 2008). When you carry out initial pharmacological or toxicological tests with an unknown molecule, it would be appropriate to use a two-sided test. In subsequent tests, for confirming the findings of the initial tests, one-sided test may be used.

References

Bailey, N.T.J. (1995): Statistical Methods in Biology, Cambridge University Press, New York, USA.

Bland, J.M. and Bland, D.G. (1994): Statistics notes: One and two sided tests of significance. BMJ, 309, 248.

CSCL (1986): Chemical Substance Control Law. http://www.safe.nite.go.jp/ kasinn/genkou/kasinhou02.html.

Doll, H. and Carney, S. (2005): Statistical approaches to uncertainty: p values and confidence intervals unpacked. Evid. Based Med., 10, 133–134.

Drewitt, P.N., Butterworth, C.D., Springall, C.D. and Moorhouse, S.R. (1993): Plasma levels of aluminum after tea ingestion in healthy volunteers. Food Chem. Toxic., 31, 19–23.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. Am. Stat. Assoc., 50, 1096–1211.

Fisher, R.A. (1955): Statistical methods and scientific induction. J. Royal Stat. Soc. B., 17, 69–78.

Fisher, R.A. (1971): The Design of Experiments. 9th Edition. Hafner Press, New York, USA.

Freedman, L. (2008): An analysis of the controversy over classical one-sided tests. Clin. Trials, 5(6), 635–640.

Gad, S.C. and Weil, C.S. (1988): Statistics and Experimental Design for Toxicologists. Telford Press, New Jersey, USA.

Gelman, A. and Stern, H. (2006): The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat., 60(4), 328–331.

Kobayashi, K. (1997): A comparison of one- and two-sided tests for judging significant differences in quantitative data obtained in toxicological bioassay of laboratory animals. J. Occup Health, 39, 29–35.

Kobayashi, K., Pillai, K.S., Sakuratani, Y., Abe, T., Kamata, E. and Hayashi, M. (2008): Evaluation of statistical tools used in short-term repeated dose administration toxicity studies with rodents. J. Toxicol. Sci., 33(1), 97–104.

Kornegay, E.T., Clawson, A.J., Smith, F.H. and Barrick, E.R. (1961): Influence of protein source on toxicity of gossypol in swine ration. J. Anim. Sci., 20, 597–602.

Le, C.T. (2009). Health and Numbers-A Problems-Based Introduction to Biostatistics, 3rd Edition, John Wiley & Sons Inc., New Jersey, USA.

Krantz, D.H. (1999): The null hypothesis testing controversy in psychology. J. Am. Stat. Assoc., 94, 1372–1381.

Madsen, B. (2011): Statistics for Non-Statisticians. Springer-Verlag, Berlin, Germany.

Moye, L.A. and Tita, A.T.N. (2002): Defending the rationale for the two-tailed test in clinical research. Circulation, 150, 3062–3065.

Nakamura, G. (1986): Practice, Statistical Analyses. Kaiumeisha, Tokyo, Japan.

Neyman, J. and Pearson, E.S. (1928): On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika, 20A, 263–94.

Neyman, J. and Pearson, E.S. (1936): Sufficient statistics and uniformly most powerful tests of statistical hypotheses. Stat. Res. Mem., 1, 113–137.

OECD (2008): Organization for Economic Cooperation and Development. OECD Guidelines for the Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents., No. 407. OECD, Geneva, France.

Rosner, B. (2010): Fundamentals of Biostatistics. 7th Edition. Brooks/cole, Cengage Learning, Boston, USA.

Sakuma, A. (1977): Statistical Methods in Pharmacometrics I. 56, Tokyodaigaku Shupankai, Tokyo, Japan.

Shertzer, H.G. and Sainsbury, M. (1991): Chemoprotective and hepatic enzyme induction properties of indol and indenoindol antioxidants in rats, Food Chem. Toxic., 29, 391–400.

Shirley, E.A. (1977): Non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, 33, 386–389.

Spina, D. (2007): Statistics in Pharmacology. Br J. Pharmacol., 152(3), 291–293.

Yoshida, M. (1980): Design of Experiments for Animal Husbandry. Yokendo Press, Tokyo, Japan.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Press, Tokyo, Japan.

Yoshimura, I. and Ohashi, S. (1992): Statistical Analysis for Toxicology Data. Chijin-Shokan, Tokyo, Japan.

8
t-Tests

Student's *t*-Test—History

The history of statistical significance tests dates back 17th century. Perhaps the earliest statistical analysis published was by John Arbuthnot on London birth rates with regards to gender in 1710 (Hacking, 1965). One of the most popular significance tests is the Student's *t*-test, which has wide scientific applications (Papana and Ishwaran, 2006). The Student's *t*-test is a parametric test for comparing two groups. Readers may be interested to know why it is called as Student's *t*-test. 'Student' was the pseudonym of W.S. Gossett (1876–1937) (Box, 1987). He worked as a chemist at the Guinness brewery, Ireland. He chose this pseudonym because his company did not allow its scientists to publish confidential data (Raju, 2005). His company regarded use of statistics in quality control as a trade secret. In an article published in Biometrika, Gossett described a procedure to assess population means by using small samples under the pseudonym, "Student" (Student, 1908).

t-Test for One Group

The temperature of an animal room was set at 22°C. The temperature of the room measured everyday at 9.00 am for seven days is given in Table 8.1. The temperature measured was not the same as the temperature set (22°C) in any of these days. Let us examine whether the temperature measured during the seven days is statistically similar to the temperature set (22°C).

Table 8.1. Temperature of the animal room

Day	1	2	3	4	5	6	7
Temperature (°C)	22.3	22.6	22.4	22.4	22.6	22.5	22.4

N = 7; Mean = 22.46; SD = 0.1134; SE =0.0429

$$tcal = \frac{22.46 - 22.0}{0.0429} = 10.723$$

The *t*-distribution Table value (Table 8.2.) at 0.05 probability, for 6 (7–1) degrees of freedom is 2.447 (two-sided). Since calculated value (10.723) is greater than the Table value (2.447), it is considered that the temperature measured in the animal room during the seven days differed from the temperature set (22°C).

Table 8.2. *t*-distribution Table (Yoshimura, 1987)

DF\2α*	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF\α**	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	**2.447**	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408

DF, Degrees of freedom; *One-sided; **Two-sided

t-Test for Two Groups

The use of a repeated *t*-test for comparison of three or more groups might cause the error of the first kind (Type I error). Three kinds of *t*-tests are commonly used (Figure 8.1). Depending on the variance ratio (*F*) and the number of samples in the group, a *t*-test is selected.

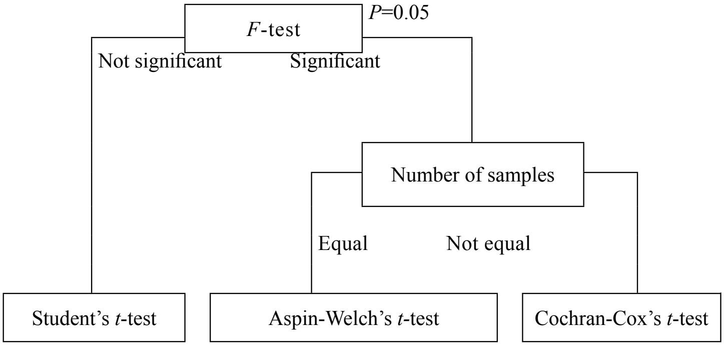


Figure 8.1. Selection of a *t*-test

F-value is the variance ratio. It is calculated by dividing the larger variance by the smaller variance. If the calculated *F*-value is smaller than the value given in *F*-distribution Table at 5% probability level, the two groups are regarded to have the same distribution and the data are analysed using Student's *t*-test. On the contrary, if the calculated *F*-value is greater than the value given in *F*-distribution Table at 5% probability level, the two groups are regarded to have different distributions and the data are analysed using either Aspin-Welch's *t*-test (if the number of samples in the two groups is equal) or Cochran-Cox's *t*-test (if the number of samples in the two groups is not equal). Cochran-Cox's test has a low power to detect a significant difference. This may be the reason why Aspin-Welch's *t*-test is often used regardless of the number of samples in the two groups.

Student's t-test

The height of male and female students in a class room is given in Table 8.3. We would like to examine whether the male and female students have similar heights.

Table 8.3. Height (cm) of male and female students

Male (Group 1)	Female (Group 2)
170	160
168	154
170	162
169	160
179	151
162	159
172	148
169	159
169	150
179	162

Statistics

Estimates	Male (Group 1)	Female (Group 2)
N	10	10
Sum	1707	1565
Mean	170.7	156.5
SD	5.0783	5.2546
Variance	25.79	27.61
Sum of squares	232.10	248.50

Let us examine the distribution of the data of males and females by calculating *F*-value:

$$F_9^9 = \frac{27.6}{25.8} = 1.07$$

Note: F_9^9—The superscript and subscript to *F* indicate the degrees of freedom of the numerator and denominator, respectively.

Compare the calculated *F*-value with the *F*-distribution Table value (Table 8.4). *F*-distribution Table value is the value, where the degrees of freedom of numerator and denominator intercept.

(Note: The *F*-distribution is named after Sir Ronald A. Fisher (1890 –1962), who is known to be the father of modern statistics (Kennedy, 2003). *F*-test is a ratio of the sample variances. However, *F*-test is not suitable for the data showing a non-normal distribution.).

Table 8.4. *F*-distribution values at 5% probability level (Yoshimura, 1987)

$N_1 \diagdown N_2$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30
7	5.59	4.73	4.34	4.12	3.97	3.86	3.78	3.72	3.67	3.63	3.57	3.52	3.49	3.46	3.44	3.37
8	5.31	4.45	4.06	3.83	3.68	3.58	3.50	3.43	3.38	3.34	3.28	3.23	3.20	3.17	3.15	3.07
9	5.11	4.25	3.96	3.63	3.48	3.37	3.29	3.22	**3.17**	3.13	3.07	3.02	2.98	2.96	2.93	2.86
10	4.96	4.10	3.70	3.47	3.32	3.21	3.13	3.07	3.02	2.97	2.91	2.86	2.82	2.79	2.77	2.69
11	4.84	3.98	3.58	3.35	3.20	3.09	3.01	2.94	2.89	2.85	2.78	2.73	2.70	2.67	2.64	2.57

N_1= Degrees of freedom of numerator, N_2= Degrees of freedom of denominator

The calculated *F* value (1.07) is less than the Table value (3.17). Hence, F_9^9 is not considered significant, indicating that the variances of both the groups having a similar distribution. Therefore, as given in Figure 8.1, the data can now be analysed using Student's *t*-test.

The *t* value is calculated using the equation,

$$tcal = \frac{\left| \bar{X}_1 - \bar{X}_2 \right|}{\sqrt{SS_1 + SS_2}} \times \sqrt{\frac{N_1 \times N_2}{N_1 + N_2}(N_1 + N_2 - 2)}$$

Where,

$\bar{X}_1$ = Mean of Group 1; $\bar{X}_2$ = Mean of Group 2; SS_1 = Sum of squares of Group 1; SS_2= Sum of squares of Group 2; N_1 = Degrees of freedom of Group 1; N_2 = Degrees of freedom of Group 2.

$$tcal = \frac{\left| 170.7 - 156.5 \right|}{\sqrt{232.1 + 248.5}} \times \sqrt{\frac{10 \times 10}{10 + 10}(10 + 10 - 2)} = 6.145$$

Compare the calculated *t* value with the *t*-test critical value given in Table 8.5.

Table 8.5. *t*-test critical values (Yoshimura, 1987)

P= 2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
P= α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF							
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	**1.734**	2.101	2.552	2.878	2.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850

Note: α=one-sided, 2α=two-sided.

The *t*-test critical value at 5% probability level for N_1+N_2-2 (10+10–2=18) degrees of freedom is 1.734. Since calculated *t*-value (6.145) is greater that the *t*-test critical value, it is considered that the height of male and female students is different.

Aspin-Welch's t-test

This test is used to compare the means of two groups having different distributions, but number of samples (observations) is the same.

A study was conducted in volunteers to find the effect of high fat content. Diet containing high fat content was given to 10 individuals (Group 1). Concurrently, normal diet was given to another 10 individuals for comparison (Group 2). At the end of the 7 days treatment, alanine aminotransferase (ALT) activity was measured in the individuals of both the Groups. The ALT determined in the individuals is given in Table 8.6.

Table 8.6. Alanine aminotransferase activity (IU/l) of individuals

Diet containing high fat content (Group1)	Normal diet (Group 2)
42	30
60	34
26	35
48	32
56	36
31	41
30	42
80	28
79	71
93	35

Statistics

Estimates	Diet containing high fat content (Group 1)	Normal diet (Group 2)
N	10	10
Sum	545	384
Mean	55	38
SD	23.4011	12.2493
Variance (Sx^2)	548	150

F-ratio =

$$F_9^9 = \frac{548}{150} = 3.65$$

Compare the calculated F-value with the Table value (Table 8.4). The derived F value (3.65) is greater than the Table value (3.17). Hence, F_9^9 is considered significant, indicating that the variances of both the groups are distributed differently. According to Figure 8.1, Aspin-Welch's t-test is the appropriate statistical tool for the analysis of this data. The t is calculated using the following formula:

$$tcal = \frac{\left|\overline{X}_1 - \overline{X}_2\ \right|}{\sqrt{\frac{Sx_1}{N_1} + \frac{Sx_2}{N_2}}}$$

Where,

$\overline{X}_1$ = Mean of Group 1; $\overline{X}_2$ = Mean of Group 2; Sx_1= Variance of Group 1; Sx_2= Variance of Group 2; N_1= Degrees of freedom of Group 1; N_2 = Degrees of freedom of Group 2.

$$tcal = \frac{\left|55 - 38\ \right|}{\sqrt{\frac{548 + 150}{10}}} = 2.03$$

Unlike Student's t-test, where the degrees of freedom is $N_1+N_2–2$, degrees of freedom needs to be calculated for Aspin-Welch's t-test. The degrees of freedom for Aspin-Welch's t-test is calculated as given below:

$$N = \frac{1}{\frac{C^2}{N_1 - 1} + \frac{(1-C)^2}{N_2 - 1}}$$

Where,

$$C = \frac{\frac{Sx_1{}^2}{N_1}}{\frac{Sx_1{}^2}{N_1} + \frac{Sx_2{}^2}{N_2}}$$

$$C = \frac{54.8}{54.8 + 15.0} = 0.79$$

$$N = \frac{1}{\frac{0.79^2}{9} + \frac{(1-0.79)^2}{9}} = 13.5$$

Compare the derived *t* value with the *t*-test critical value given in Table 8.7 at 5% probability level for fourteen degrees of freedom (14 degrees of freedom is obtained by rounding up the calculated *N*, 13.5). Since the calculated *t*-value, 2.03 is greater than the *t*-test critical value given in the Table 8.7 (1.761), it can be stated that there is a difference in ALT between the high fat diet-treated and normal diet treated-individuals.

Table 8.7. *t*-test critical values (Yoshimura, 1987)

2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF=14	1.345	**1.761**	2.145	2.624	2.977	3.787	4.140

α=one-sided, 2α=two-sided.

Cochran-Cox's t-test

Cochran-Cox's *t*-test is used to compare the means of two samples having different distributions and different number of observations. We shall modify the data given in Table 8.6 and analyse it using Cochran-Cox's *t*-test. The values modified are given in Table 8.8. We have not made any change in the ALT values of Group 1. But, the values of Group 2 are changed and only nine individuals of this group are used for the analysis.

Table 8.8. Alanine aminotransferase activity (IU/l) of individuals

Diet containing high fat content (Group1)	Normal diet (Group 2)
42	57
60	45
26	55
48	46
56	26
31	33
30	41
80	35
79	43
93	-

Statistics

Estimates	Diet containing high fat content (Group 1)	Normal diet (Group 2)
N	10	9
Sum	545	381
Mean	55	42
SD	23.4011	10.0374
Variance (Sx^2)	548	101

F-ratio =

$$F_8^9 = \frac{548}{101} = 5.43$$

Compare the derived *F*-value with the Table value (Table 8.4). The calculated *F*-value (5.43) is greater than the Table value (3.38). Hence, F_8^9 is considered significant, indicating that the variances of both the groups are distributed differently. According to Figure 8.1, Cochran-Cox's *t*-test is the appropriate statistical tool for the analysis of the data given in Table 8.8.

In Cochran-Cox's *t*-test, we need to calculate two *t* values (*t* calculated and *t'* calculated).

$$tcal = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{Sx_1{}^2}{N_1} + \frac{Sx_2{}^2}{N_2}}}$$

$$tcal = \frac{|\ 55 - 42\ |}{\sqrt{\frac{548}{10} + \frac{101}{9}}} = 1.21$$

$$t'cal = \frac{\frac{t_1 \times Sx^2{}_1}{N_1} + \frac{t_2 \times Sx^2{}_2}{N_2}}{\frac{Sx^2{}_1}{N_1} + \frac{Sx^2{}_2}{N_2}}$$

$$t'cal = \frac{\frac{1.833 \times 548}{10} + \frac{1.860 \times 101}{9}}{\frac{548}{10} + \frac{101}{9}} = 1.83$$

Since the *t* calculated (*tcal* = 1.21) is smaller than the *t*' calculated (t'cal=1.83), it is concluded from the analysis that there is no significant difference in ALT between the high fat diet-treated and normal diet treated-individuals.

Paired *t*-Test

Let us assume one needs to test an antidiabetic drug in diabetic rats. One way to do is to measure the blood sugar before and after treatment with the drug and calculate the respective mean values, and compare the mean values using an appropriate *t*-test (select the appropriate *t*-test as per Figure 8.1). Another way is to analyse the data using paired *t*-test.

Blood sugar values of individual rats before and after the drug treatment is given Table 8.9.

Table 8.9. Blood sugar values (mg/dl) of individual rats

Rat Number	1	2	3	4	5	Mean	Variance	SD	SE
Before treatment	274	287	277	259	237	-	-	-	-
After treatment	165	142	215	209	198	-	-	-	-
Difference between before and after treatments	109	145	62	50	39	81	1992	44.6	19.9

$$tcal = \frac{Mean}{SE}$$

$$tcal = \frac{81}{19.9} = 4.07$$

Compare the calculated *t*-value with the *t*-test critical value given in Table 8.10 at 5% probability level for N-1 degrees of freedom. N is number of pairs, hence N–1=4. Since the calculated *t* value, 4.07 is greater than the *t*-test critical value given in the Table 8.10 (2.132), it can be stated that treatment with the drug significantly decreased the blood sugar in rats.

Table 8.10. *t*-test critical values (Yoshimura, 1987)

2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF=4	1.533	**2.132**	2.776	3.747	4.604	7.173	8.610

α=one-sided, 2α=two-sided.

A Note of Caution

It is well known that with Student's two-independent-sample *t*-test, the actual level of significance can be well above or below the nominal level, confidence intervals can have inaccurate probability coverage, and power can be low relative to other methods.

In Student's two-independent-sample *t*-test, the variance heterogeneity can distort rates of Type I error (Kaselman *et al.*, 2004). Therefore, when the variance of the two populations is different, Student's *t*-test is not suitable (Ruxton, 2006). When the number of the groups is more than two, multiple comparison with Student's *t*-test can cause Type I error.

References

Box, J.F. (1987): Guinness, Gosset, Fisher, and Small Samples. Stat. Sci., 2(1), 45–52.

Hacking, I. (1965): Logic of Statistical Inference. Cambridge University Press, New York.

Kaselman, H.J., Othman, A.R., Wilcox, R.R. and Fradette, K. (2004): The new and improved two-sample *t*-test. Psych. Sci., 15(1), 47–51.

Kennedy, P. (2003): A Guide to Econometrics. 5th Edition. MIT Press, UK.

Papana, A. and Ishwaran, H. (2006): CART variance stabilization and regularization for high-throughput genomic data. Bioinformatics, 22(18), 2254–2261.

Raju, T.N.K. (2005): William Sealy Gosset and William A. Silverman: Two "Students" of Science. Pediatrics, 116(3), 732–735.

Ruxton, G. (2006): The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney U test. Behavioral Ecol., 17(4), 688–690.

Student (1908): The Probable Error of a Mean. Biometrika, 6(1), 1–25.

Yoshimura, I. (1987): Statistical Analysis of Toxicity and Drug Efficacy Data. Scientist Inc., Tokyo, Japan.

9 Correlation Analysis

Correlation and Association

Correlations are relationships between two or more variables or sets of variables (Cohen and Cohen, 1983). In statistics there is a distinction between an association and a correlation, though these terms are often used interchangably. Two variables are associated if one of them provides information about the likely value of the other. If the association between two variables is linear, there is a correlation. Therefore, strictly speaking, "non-linear correlation" is an incorrect terminology, a better term is "non-linear association".

Statisticians' definition of correlation is that it is 'a parameter of the bivariate normal distribution'. The variables are random variables when one variable is not depended on the other. In statistics, correlation is referred to as coefficient of correlation (Paler-Calmorin and Calmorin-Piedad, 2008). The correlation coefficient is denoted by the letter *r* which might have originated from the letter, *r* of the word, relation. A number between –1 and +1 is used to 'quantify' the correlation of the variables (Glantz, 2005). The closer the absolute value of *r* to 1 or –1, the higher the degree of correlation. When one variable increases as the other variable increases, it is called a 'positive correlation', and when one variable decreases as the other variable increases, it is called a 'negative correlation'. When $r = -1$, there is a 100% negative correlation, when $r = +1$, there is a 100% positive correlation and when $r = 0$, there is a 100% no correlation. But, if $r = 0.5$, it does not mean that there is a 50% correlation. Therefore, *r* does not indicate the percent of correlation (Gurumani, 2005).

Pearson's Product Moment Correlation Coefficient

A commonly used measure of correlation is Pearson's product moment correlation coefficient. Correlation coefficient is a standardised covariance (Field, 2009; Berkman and Reise, 2011). Covariance is a measure of joint variances of two variables; the deviation of each variable is computed and multiplied. Since there are two variables, there are two standard deviations. Multiply these standard deviations and divide joint variances by it.

Standardised covariance, $r = \frac{1}{n-1} \frac{\sum (x-\bar{x})(y-\bar{y})}{(s_x)(s_y)}$, where

s_x and s_y are the standard deviations of variable x and variable y, respectively. Above equation can be rewritten as follows:

$$r = \frac{1}{n-1} \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left(\frac{\sum (x-\bar{x})^2}{n-1}\right)\left(\frac{\sum (y-\bar{y})^2}{n-1}\right)}}$$

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

The above equation was formulated by Karl Perason, hence called Pearson's correlation coefficient.

Let us compute correlation coefficient, r for the variables x and y given in Table 9.1.

Table 9.1. Calculation of correlation coefficient

x	y	x^2	y^2	xy
1	93	1	8649	93
2	87	4	7569	174
3	76	9	5776	228
4	70	16	4900	280
5	62	25	3844	310
6	45	36	2025	270
7	40	49	1600	280
8	32	64	1024	256
9	25	81	625	225
10	10	100	100	100
Σx= 55	Σy=540	Σx^2=385	Σy^2=36112	Σxy= 2216

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

$$\sum(x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\sum x \times \sum y}{n} = 2216 - \frac{55\times 540}{10} = -754$$

$$\sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 385 - \frac{(55)^2}{10} = 82.5$$

$$\sum(y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 36112 - \frac{(540)^2}{10} = 6952$$

$$r = \frac{-754}{\sqrt{82.5\times 6952}} = \frac{-754}{757.32} = -0.996$$

Significance of *r*

When the sample size is not too large, the significance of a correlation coefficient can be tested using a *t*-test:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.996\sqrt{10-2}}{\sqrt{1-(-0.996)^2}} = \frac{-2.8171}{0.0894} = -31.51$$

Above is Students *t*-test with n–2 degrees of freedom.

Alternatively, significance of a correlation coefficient can be tested as given below, which involves no calculation procedure:

Compare the correlation coefficient, *r* with the value given in correlation coefficient table (Table 9.2) for eight degrees of freedom. The computed correlation coefficient, *r* (–0.996) is less than the correlation coefficient Table value (–0.765) at 1% probability level. Hence the correlation coefficient is considered to be significant. The negative sign of the correlation coefficient indicates that the variables *x* and *y* are negatively correlated. Had the *r* been 0.996 (positively correlated), we would have compared it with 0.765 (without a minus sign). In this case, in order to consider the *r* to be significant, it has to be greater than 0.765.

Table 9.2. Correlation coefficient Table (Shibata, 1970)

DF	5%	1%	DF	5%	1%	DF	5%	1%
1	0.997	1.000	17	0.456	0.575	45	0.288	0.372
2	0.950	0.990	18	0.444	0.561	50	0.273	0.354
3	0.878	0.959	19	0.433	0.549	60	0.250	0.325
4	0.811	0.917	20	0.423	0.537	70	0.232	0.302
5	0.754	0.874	21	0.413	0.526	80	0.217	0.283
6	0.707	0.834	22	0.404	0.515	90	0.205	0.267
7	0.666	0.798	23	0.396	0.505	100	0.195	0.254
8	0.632	0.765	24	0.388	0.496	125	0.174	0.228
9	0.602	0.735	25	0.381	0.487	150	0.159	0.208
10	0.576	0.708	26	0.374	0.478	200	0.138	0.181
11	0.553	0.684	27	0.367	0.470	300	0.113	0.148
12	0.532	0.661	28	0.361	0.463	400	0.098	0.128
13	0.514	0.641	29	0.355	0.456	500	0.088	0.115
14	0.497	0.623	30	0.349	0.449	1000	0.062	0.081
15	0.482	0.606	35	0.325	0.418			
16	0.468	0.590	40	0.304	0.393			

Confidence Interval of Correlation Coefficient

A confidence interval of correlation coefficient, *r* can be determined by using a transformation of *r* to a quantity *z*, which has an approximately normal distribution. This transformed *z* is calculated using the equation:

$$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

For the example given in Table 9.1, $r = 0.996$. The transformed Z is:

$$Z = \frac{1}{2}\ln\left[\frac{1+(-0.996)}{1-(-0.996)}\right] = \frac{1}{2}\ln\left[\frac{0.004}{1.996}\right] = -3.1063$$

Now, we need to calculate an estimate called Error of Estimate:

Error of Estimate = $1/\sqrt{n-3} = 1/\sqrt{10-3} = 0.3780$

Using the Error of Estimate we can calculate Z_1 and Z_2 with 95% confidence level:

$$Z_1 = -3.1063 - (1.96\text{x}0.3780) = -3.8472$$
$$Z_2 = -3.1063 + (1.96\text{x}0.3780) = -2.3654$$

Next step is to transform the Z_1 and Z_2 back to original scale. Confidence interval of r is:

$$\frac{e^{2z_1}-1}{e^{2z_1}+1} \text{ to } \frac{e^{2z_2}-1}{e^{2z_2}+1} =$$

$$\frac{e^{2\times -3.8472}-1}{e^{2\times -3.8472}+1} \text{ to } \frac{e^{2\times -2.3654}-1}{e^{2\times -2.3654}+1} =$$

$$\frac{0.9995}{1.0005} \text{ to } \frac{0.9912}{1.0088}$$

Confidence interval of r is calculated as 0.983–0.999.

Coefficient of Determination

The coefficient of determination is the square of r (R^2; coefficient of determination is usually denoted by the capital letter R^2), which expresses the strength of the relationship between the x and y variables (McDonald, 2009). This is reviewed in Chapter 10, in greater detail.

Rank Correlation

When the variables are not linearly associated, Pearson's product moment correlation analysis does not work well. In this situation the association is transformed into linear by ranking the variables. Rank correlation is a nonparametric alternative to the linear correlation coefficient (Ruby, 2008). There are several rank correlation analyses available, amongst them, Spearman's rank correlation is more commonly employed (Hassard, 1991).

Spearman's Rank Correlation

As stated, in Spearman correlation analysis, the variables are converted to ranks. Spearman rank correlation analysis is also used, when there are two measurement variables and one "hidden" nominal variable. If you measure body weight and body surface area of rats with the rat identification number, the identification number of the rat is the nominal variable. The major advantages of Spearman's rank correlation are that it is not affected by the distribution of the population and it can be applied to small samples (Gauthier, 2001).

Canonical Correlation

Canonical correlation analysis developed by Hotelling (1936), is the study of the linear relationships between two sets of variables, and is considered as a fundamental statistical tool (Bulut *et al.*, 2010). It is the multivariate extension of correlation analysis and it measures the interrelationships among sets of multiple dependent variables and multiple independent variables (Green, 1978). Canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables. It is a very useful tool in pharmacology and toxicology (Kelder, 1982; Hu *et al.*, 2003; Tanaka, 2010), where interrelationships between several dependent and independent variables need to be assessed.

An elaborative discussion on canonical correlation is beyond the scope of this book. Several books are available that cover the subject in depth (Green and Carroll, 1978; Das and Sen, 1994).

Misuse of Correlation Analysis

There are a several situations in which the correlation coefficient can be misinterpreted. Fifteen errors related to correlation and regression were identified in articles published in three leading medical journals in the year, 1997 (Porter, 1999). Perhaps the most important error committed in these articles was, not presenting confidence intervals of correlation coefficient (this error could be seen in many of the scientific articles, even today). Another error in interpreting the correlation coefficient is, the failure to consider that there may be a third variable related to both of the variables being investigated, which is responsible for the apparent correlation. Often the correlation coefficient fails to detect the existence of a nonlinear association between two variables (Bewick *et al.*, 2003).

A high correlation coefficient (for example, $r = > 0.997$) is not always a useful indicator of linearity in method validation; other statistical tests like Lack-of-fit and Mendel's fitting test may be used for evaluating the linearity (Loco *et al.*, 2002).

A correlation coefficient will have limited use as a stand-alone quantity without reference to the number of observations, the pattern of the data and the slope of the regression line (Sonnergaard, 2006). It is recommended to plot the variables and understand the pattern of the data before interpreting the correlation analysis.

References

Berkman, E.T and Reise, S.P. (2011): A Conceptual Guide to Statistics Using SPSS. SAGE Publications Inc., California, USA.

Bewick, V., Cheek, L. and Ball, J. (2003): Statistics review 7: Correlation and regression. Critical Care, 7, 451–459.

Bulut, M., Gultepe, N., Mendes, M., Guroy, D. and Palaz, M. (2010): According to Canonical correlation, the evaluation of bluefish (*Pomatomus saltatrix*) blood chemistry. J. Animal Vet. Adv., 9(4), 666–670.

Cohen, J. and Cohen, P. (1983): Multiple Regression/Correlation for the Behavioral Sciences. 2nd Edition. Erlbaum Associates, Hillsdale, New Jersey, USA.

Das, S. and Sen, P.K. (1994): Restricted canonical correlations. Linear algebra and its applications, 210, 29–47.

Field, A. (2009): Discovering Statistics Using SPSS. 3rd Edition. SAGE Publications Ltd., London, UK.

Gauthier, T.D. (2001): Detecting trends using Spearman’s rank correlation coefficient. Exp. Forensics, 2, 359–362.

Glantz, S.A. (2005): Primer of Biostatistics. Mc Graw-Hill Companies Inc., USA.

Green, P.E. (1978): Analyzing Multivariate Data. Holt, Rinehart & Winston, Illinois, USA.

Green, P.E., and Carroll, J.D. (1978): Mathematical Tools for Applied Multivariate Analysis. Academic Press, New York, USA.

Gurumani, N. (2005): An Introduction to Biostatistics. 2nd Edition. MJP Publishers, Chennai, India.

Hassard, T.H. (1991): Understanding Biostatistics. Mosby-Year Book Inc., St. Loius, Missouri, USA.

Hotelling, H. (1936): Relations between two sets of variates. Biometrika, 28, 321–377.

Hu, Q.N., Liang, Y.Z., Peng, X.L., Hong, Y. and Zhu, L. (2003): Application of orthogonal block variables and canonical correlation analysis in modeling pharmacological activity of alkaloids from plant medicines. J. Data Sci., 1, 405–423.

Kelder, J. (1982): Prediction of the Bobon clinical profile of neuroleptics from animal pharmacological data. Psychopharmacol., 77(2), 140–145.

Loco. J.V., Elskens, M., Croux, C. and Beernaert, H. (2002): Linearity of calibration curves: use and misuse of the correlation coefficient. Accred. Qual. Assur., 7, 281–285.

McDonald, J.H. (2009): Handbook of Biological Statistics. 2nd Edition. Sparky House Publishing Baltimore, Maryland, USA.

Paler-Calmorin, L. and Calmorin-Piedad, M.L.P. (2008): Nursing Biostatistics with Computer. Rex Printing Co. Inc., Florentino St., Quezon City, Philippines.

Porter, A.M.W. (1999): Misuse of correlation and regression in three medical journals. J. Royal Soc. Med., 92, 123–128.

Ruby, J. (2008): Elementary Statistics. Thompson Brooks. Cole, Belmont, USA.

Shibata, K. (1970): Biostatistics, Tokyo University of Agriculture, Tokyo, Japan.

Sonnergaard, J.M. (2006): On the misinterpretation of the correlation coefficient in pharmaceutical sciences. Int. J. Pharm., 321(1-2), 12–17.

Tanaka, T. (2010): Biological factors influencing exploratory behavior in laboratory mice, *Mus musculus*. Mammal Study, 35(2), 139–144.

10

Regression Analysis

History

The origin of the term 'regression' in statistics has an interesting history. Francis Galton (1822–1911) had deep interest in heredity, biometrics and eugenics (Crow, 1993). He found that sons of tall men to be shorter than their fathers. He called this phenomenon regression towards the mean, and thus the term 'regression' originated (Dupont, 2002).

Unlike correlation, where there is no 'dependence relationship', there are dependent and independent variables in regression analysis. In regression analysis, y is assumed to be a random variable and x is assumed to be a fixed variable. The underlying assumption of regression analysis is that the dependent variable follows a normal distribution and scatter about the regression line.

In animal experiments regression analysis is used to evaluate cause (variable x) and effect (variable y) relationships; for example in a repeated dose administration study, the rate of decrease in body weight (y) as the exposure period (x) increases can be determined using regression analysis.

Linear Regression Analysis

The regression equation is:

$y = a + bx,$ where y = Dependent variable, x = Independent variable, a = Intercept and b = slope.

The intercept represents the estimated average value of y when x equals zero and the slope represents the estimated average change in y when x increases/decreases by one unit. Slope and intercept are derived using the least-square method.

If the underlying assumptions of the least-square model are not met, the regression slope and intercept may be incorrect. Two factors that cause incorrect regression coefficients are: (i) imprecision in the measurement of the independent (x) variable and (ii) inclusion of outliers in the data analysis (Cornbleet and Gochman, 1979). Outliers have profound effect on the slope (Farnsworth, 1990; Glaister, 2005).

The slope, *b* is calculated using the formula:

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

The intercept a can be calculated from the equation:

$$\bar{y} = a + b\,\bar{x}$$

Let us work out an example for calculating *b* and *a*. Body weight of babies measured in different months is given in Table 10.1. Month is the independent variable (x) and the body weight is the dependent variable (y).

Table 10.1. Body weight of babies measured in different months

Age (Month) (x)	Body weight (kg) (y)	x^2	y^2	xy
1	3.8	1	14.44	3.8
2	4.2	4	17.64	8.4
3	4.8	9	23.04	14.4
5	5.7	25	32.49	28.5
6	6.4	36	40.96	38.4
7	6.9	49	47.61	48.3
8	7.1	64	50.41	56.8
9	7.8	81	60.84	70.2
10	8.6	100	73.96	86
12	10.4	144	108.16	124.8
$\Sigma x = 63$ $\bar{X} = 6.3$	$\Sigma y = 65.7$ $\bar{Y} = 6.57$	$\Sigma x^2 = 513$	$\Sigma y^2 = 469.55$	$\Sigma xy = 479.6$

We shall calculate the slope, *b* first:

$$\sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\sum x \times \sum y}{n} = 479.6 - \frac{63 \times 65.7}{10} = 65.69$$

$$\sum (x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 513 - \frac{(63)^2}{10} = 116.1$$

$$\sum(y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} == 469.55 - \frac{(65.7)^2}{10} = 37.90$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{65.69}{116.1} =0.5658$$

Once the slope, b is calculated, it is easy to calculate the intercept, a:

$$\bar{y} = a + b\,\bar{x}$$

$$6.57 = a + 0.5658 \times 6.3$$

$$a = 6.57 - (0.5658 \times 6.3) = 3.005$$

Regression equation:

$$y = a + bx$$
$$y = 3.005 + 0.5658\,x$$

Significance of regression line can be determined by ANOVA (Table 10.2).

We wish to test the hypothesis:

H_0: b = 0 *vs* H_1: b ≠ 0, where b is the slope.

Table 10.2. Significance of regression line by ANOVA

Source of variation	Degrees of freedom	SS	Mean SS	F
Total SS for $y = \sum y^2 - \frac{(\sum y)^2}{n}$	9	37.90	4.21	-
Reduction due to regression (Residual SS) = $\frac{\left[\sum xy - \frac{\sum x \times \sum y}{N}\right]^2}{\sum\left[x-\bar{x}\right]^2}$	1	37.17	37.17	407
Error	8	0.73	0.0913	-

SS—Sum of squares

Since the calculated F-value is greater than the F-Table value (Table 10.3), the null hypothesis is rejected and the alternative hypothesis (H_1: b ≠ 0) is accepted. This means the slope of the regression line is significantly different from 0, which implies that there is a significant relationship between age and body weight of the babies.

Table 10.3. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1\backslash N_2$	1	2	3	4	5	6	7	8	9	10
8	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.05	11.77	11.54

The test of significance is based on the assumption that the distribution of the deviation from the regression line (residual values) of all the values of dependent variable, *y* is the same for all the independent varable, *x*. The residue of each observation is given by the difference between the observed value and the fitted value of the regression line (Chan, 2004). Let us understand the terminology the residue of *y,* by plotting the data given in the Table 10.1. Figure 10.1 is the body weight *vs* age plot.

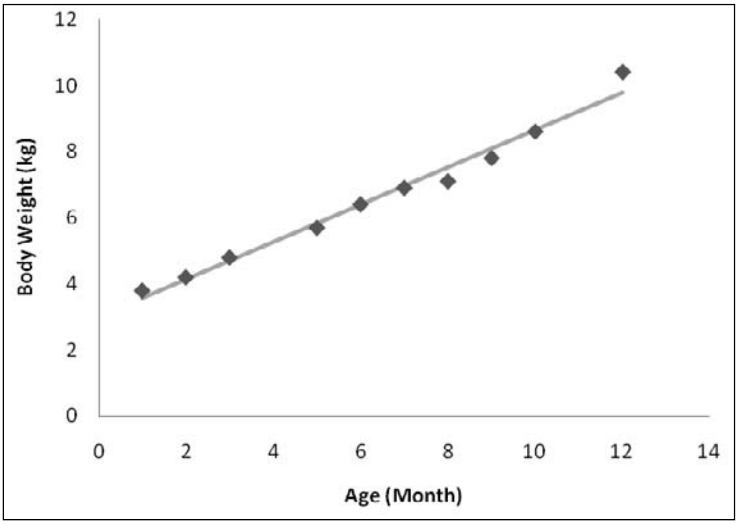


Figure 10.1. Body weight of babies measured in different months

Solid squares are the actual values. The line passing through the actual values is the regression line. For each value of *x* variable, the predicted *y* value is computed using the regression equation, $y' = 3.005 + 0.5658\ x$ (predicted *y* is denoted as *y'* in order to differentiate it from the actual *y*). Thus, *y'* is derived for each *x*, and the predicted *y*'s are joined together to obtain the regression line. By closely observing the plot, one can find that all the actual values do not fall on the regression line, though they are very close to the regression line. Linear regression line is called a ‘best fit line’, since it best fits the data points. The “best” fit line minimizes the squared vertical distances between the actual values and the line. An estimate of the squared vertical distances between the actual values and the line (in other words, variation of the actual values from the predicted values) can easily be arrived at (*vide* Table 10.4). You would have noticed that this estimate is the sum of squares for error component given in the ANOVA Table (Table 10.2).

Table 10.4. Calculation of variation of the actual *y* values from the predicted *y*' values

Age (Month) (x)	Body weight (kg) (y)	y' ($y' = 3.005 + 0.5658\,x$)	$y - y'$	$(y - y')^2$
1	3.8	3.5708	0.2292	0.052533
2	4.2	4.1366	0.0634	0.00402
3	4.8	4.7024	0.0976	0.009526
5	5.7	5.834	–0.134	0.017956
6	6.4	6.3998	0.0002	0.00000004
7	6.9	6.9656	–0.0656	0.004303
8	7.1	7.5314	–0.4314	0.186106
9	7.8	8.0972	–0.2972	0.088328
10	8.6	8.663	–0.063	0.003969
12	10.4	9.7946	0.6054	0.366509
-	-	-	-	$\sum (y - y')^2 =$ 0.733249

Confidence Limits for Slope

95% confidence limits for the slope (b) can be derived by using the formula:

$b \pm t_{0.05.n-2}$ SE (b), where b is the slope (0.5658); $t_{0.05.n-2}$ is the critical value for t at 5 % probability level for n–2 degrees of freedom (2.306);

SE (b) is the standard error of $b = \sqrt{\dfrac{ErrorMeanSS}{\sum\left[x-\bar{x}\right]^2}} = \sqrt{\dfrac{0.0913}{116.1}} = 0.0280$

95% confidence limits for the slope (b) = 0.5658 ± (2.306 x 0.0280) = 0.5658 ±0.0646.

The significance of slope can be tested using the t-test, when the number of samples is smaller than about 30 (Bailey, 1995):

$t_{0.05.n-2} = \dfrac{b-\beta}{s/\sqrt{\sum\left[x-\bar{x}\right]^2}}$ where $t_{0.05.n-2}$ is the critical value for t at 5% probability level for n–2 degrees of freedom; b is the slope (b=0.5658);

β is the hypothetical value (*β* =0) (we are testing whether the observed *b* value is different from the hypothetical value); *s* is the square root of error mean sum of squares

$$s = \sqrt{0.0913} = 0.3022 \ ; \ \sum\left[x - \overline{x}\right]^2 = 116.1.$$

$$t_{0.05.\text{n-2}} = \frac{0.5658 - 0}{0.3022/\sqrt{116.1}} = \frac{0.5658}{0.0280} = 20.17$$

The derived *t* value (20.17) is greater than the Table *t*-value (2.228) at 5% probability level and 10 degrees of freedom; hence the slope is significant.

Comparison of Two Regression Coefficients

The regression coefficient, *b* measures how much the dependent variable, *y* changes (increases or decreases), for each unit change in the independent variable, *x*. The slopes of two similar studies can be compared using the formula:

$$\text{d} = \frac{b_1 - b_2}{\sqrt{\left[\frac{s_1^{\ 2}}{\sum\left[X_1 - \overline{X_1}\right]^2} + \frac{s_2^{\ 2}}{\sum\left[X_2 - \overline{X_2}\right]^2}\right]}}$$

Suffix 1 refers to independent variable x_1, and 2 independent variable x_2. Since d is normally distributed, the difference between b_1 and b_2 can be examined for statistical significance using *t*-test:

$$t = \frac{b_1 - b_2}{s\sqrt{\left[\frac{1}{\sum\left[X_1 - \overline{X_1}\right]^2} + \frac{1}{\sum\left[X_2 - \overline{X_2}\right]^2}\right]}}, \text{ where}$$

$$s = \sqrt{\frac{(n_1 - 2)s_1^{\ 2} + (n_2 - 2)s_2^{\ 2}}{n_1 + n_2 - 4}}$$

The calculated *t* value is compared with the Table *t*-value at $n_1 + n_2 - 4$ degrees of freedom.

R^2

R^2 is interpreted as the proportion of total variability of the outcome that is accounted by the model (Vittinghoff *et al.*, 2005). In other words, it is the proportion of the variation in the *y* variable that is "explained" by the variation in the *x* variable. R^2 is called as the 'coefficient of determination'. R^2 can vary from 0 to1. An R^2 close to 1 indicates that the actual *y* values fall almost right on the regression line. An R^2 close to 0 indicates that there is little or no relationship between *x* and *y*.

Multiple Linear Regression Analysis

In most situations, the dependent variable is associated with more than one independent variable. For example, the body weight of rats measured in a repeated dose administration study is associated with several independent variables like, age, sex and feed consumption of the animals. Multiple regression analysis is a very useful tool for finding out which independent variable/s has/have genuine relationship with the dependent variable. Multiple linear regression model is an extension of the simple linear regression model (Ambrosius, 2007).

The regression equation for two independent variables is:

$y = a + b_1x_1 + b_2x_2$, where y = Dependent variable, x_1 and x_2 are the independent variables, a = Intercept and b_1 = Slope of x_1 and b_2 = Slope of x_2.

We shall examine the steps involved in calculating multiple linear regression coefficient:

$$\sum\left(x_1-\overline{x_1}\right)^2 \qquad = \text{A}$$

$$\sum\left(x_1-\overline{x_1}\right)\left(x_2-\overline{x_2}\right) \qquad = \text{B}$$

$$\sum\left(x_2-\overline{x_2}\right)^2 \qquad = \text{C}$$

$$\sum\left(x_1-\overline{x_1}\right)\left(y-\overline{y}\right) \qquad = \text{D}$$

$$\sum\left(x_2-\overline{x_2}\right)\left(y-\overline{y}\right) \qquad = \text{E}$$

$$\sum\left(y-\overline{y}\right)^2 \qquad = \text{F}$$

$$b_1 = \frac{CD - BE}{AC - B^2}$$

$$\mathrm{b}_2 = \frac{AE - BD}{AC - B^2}$$

Once the slopes are derived, *a* can be calculated using the formula:

$$y = a + b_1\bar{x}_1 + b_2\bar{x}_2$$

Multiple correlation coefficient can be computed using the formula:

$$\mathrm{R}= \frac{\Sigma yy'}{\sqrt{\sum y^2 y'^2}}\text{, where}$$

R = Multiple correlation coefficient; y = Actual value; y'= Predicted y (calculated using the regression equation, $\underline{y} = a + b_1\bar{x}_1 + b_2\bar{x}_2$;

$$\Sigma yy' = \left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]$$

$$\sum y^2 = \sum y^2 - \frac{(\sum y)^2}{n};\ \sum y'^2 = \sum y'^2 - \frac{(\sum y')^2}{n}$$

Significance of the multiple regression equation can be checked by ANOVA (Table 10.5).

Polynomial Regression

Linear regression does not hold good, when the data of your dependent variable follows a curved line, rather than a straight line. Transforming the *y* or *x* or both the variables to their logarithms, reciprocals, square roots etc., may straighten certain curves, but not all. Another way to solve this issue is to use a curvilinear regression equation. Polynomial regression equation is an example of curvilinear regression equation, which is used to predict toxicological variables (Vogt, 1989). Given the complexity of the calculations in polynomial regression analysis, it is not being included in the coverage of this book. The purpose of touching upon polynomial

Table 10.5. Significance of multiple regression equation by ANOVA

Source of variation	Degrees of freedom	SS	Mean SS
Total SS for *Y*	n–1	$\sum y^2 - \frac{(\sum y)^2}{n}$	-
Reduction due to regression (Residual SS)	k	$\sum y'^2 - \frac{(\sum y')^2}{n}$	$\sum Y'^2 - \frac{(\sum Y')^2}{n}$/k
Error	n–k–1	$\left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]^2$	$\left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]^2$ /n–k–1

k is the number of independent variables.

F value is calculated by dividing Reduction due to regression (Residual SS) with error.

regression analysis, is to create awareness that before carrying out linear regression analysis one should ensure that the trend of the association between the two variables is linear.

Misuse of Regression Analysis

Use of a regression equation is considered to be inappropriate for estimating an independent variable, rather than a dependent variable (Williams, 1983). It is important to understand the nature of the data before choosing a regression model. This can be easily done by plotting the data, which will help understanding the nature of the data and selecting appropriate regression model. One should not fit a straight line using a linear regression equation for a 'non-linear data'.

References

Ambrosius, W.T. (2007): Topics in Biostatistics. Humana Press Inc., New Jersey, USA.

Bailey, N.T.J. (1995): Statistical Methods in Biology. Cambridge University Press, Cambridge, UK.

Chan, Y.H. (2004): Biostatistics 201: Linear regression analysis. Singapore Med. J., 45 (2), 55–61.

Cornbleet, P.J. and Gochman, N. (1979): Incorrect least-squares regression coefficients in method-comparison analysis. Clin. Chem., 25, 432–438.

Crow, J.F. (1993): Francis Galton: Count and measure, measure and count. Genetics, 135, 1–4.

DuPont, W.D. (2002): Statistical Modeling for Biomedical Researchers. Cambridge Univ. Press, Cambridge, U.K.

Farnsworth, D.L. (1990): The effect of a single point on correlation and slope. Internat. J. Math. Math. Sci., 13(4), 799–806.

Glaister, P. (2005): Robust linear regression using Theil's method. J. Chem. Educ., 82(10), 1472–1473.

Vittinghoff, E., Glidden, D.N., Shiboski, S.C. and McCulloch, C.E. (2005): Statistics for Biology and Health. Springer Science+Business Media, Inc., New York, USA.

Vogt, N.B. (1989): Polynomial principal component regression: An approach to analysis and interpretation of complex mixture relationships in multivariate environmental data. Chemometrics Intelligent Lab Systems, 7(1-2), 119–130.

Williams, G.P. (1983): Improper use of regression equations in earth sciences. Geology, 11(4), 195–197.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Inc., Tokyo, Japan.

11

Multivariate Analysis

Analysis of More than Two Groups

Student's *t*-test is used to test the equality of the means from two different populations (Rothmann, 2005). Use of Student's *t*-test for comparing more than two groups can cause Type I error. This can be better understood from the example below:

Absolute weight of the liver of female mice in a 13-week repeated dose administration study is given in Table 11.1.

Table 11.1. Liver weight (g) of female mice in a 13-week repeated dose administration study

Group	N	Mean ± SD	Tukey's multiple range test			Repeated comparison with Student's *t*-test		
			A	B	C	A	B	C
A	10	1.083±0.057	-	-	-	-	-	-
B	10	1.098±0.077	NS	-	-	NS	-	-
C	10	1.154±0.050	NS	NS	-	S	NS	-
D	10	1.273±0.062	S	S	S	S	S	S

NS—Not significant; S—Significant (P<0.05).

Repeated analysis by Student's *t*-test revealed a significant difference between Groups A and C. Actual increase in liver weight in Group C compared to Group A is only 6.6%. In this case, the significant difference between Groups A and C detected by repeated comparison with the *t*-test is caused by Type I error. When the groups were compared using Tukey's multiple range test, no significant difference was observed between Groups A and C (Tukey's multiple range test is the ideal test in this situation, since the number of groups to be compared is more than two).

There are several methods available for multiple comparison of means, but most of them have often been misused (Gill, 1990). An appropriate tool for analyzing more than two groups is analysis of variance (Wallenstein *et al.*, 1980). One advantage of ANOVA (Analysis of Variance is abbreviated as ANOVA) is that it is easy to execute (Muir *et al.*, 2006) and it has great utility and flexibility (Armstrong *et al.*, 2000). Like Student's *t*-test, for carrying out ANOVA, it is a prerequisite that homogeneity of variance prevails across all the groups (Moder, 2007) and the data has normal distribution. However, normality is rarely tested in ANOVA, because, a slight departure from normality does not affect the conclusion drawn from the analysis (Norman and Streiner, 2008).

ANOVA is also an excellent tool for analysing data obtained from factorial experiments. In a factorial experiment, there can be several factors at several levels. For example, to test a drug against hypercholesterolemia in rats, we may use a standard drug for comparison. The test drug and the standard drug are called factors. We may test these drugs at different dose levels. Depending upon the number and levels of factors, an ANOVA can be one-way, two-way or multi-way.

One-way ANOVA

One-way ANOVA is used to find if the given factor has significant effect on the expected outcome of the experiment. Jaundice index (x) of a newborn baby measured in weeks 36, 38 and 40 is presented in Table 11.2. We want to examine if the factor (week) has any significant effect on the jaundice index.

Table 11.2. Jaundice index (x) of newborn baby

Week					
36 (Group 1)		38 (Group 2)		40 (Group 3)	
x_1	13	x_{11}	9	x_{21}	5
x_2	6	x_{12}	11	x_{22}	5
x_3	11	x_{13}	11	x_{23}	4
x_4	12	x_{14}	10	x_{24}	7
x_5	14	x_{15}	7	x_{25}	7
x_6	10	x_{16}	7	x_{26}	3
x_7	9	x_{17}	5	x_{27}	3
x_8	11	x_{18}	8	x_{28}	4
x_9	11	x_{19}	7	x_{29}	5
x_{10}	10	x_{20}	10	x_{30}	3

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	10	10	10
Mean ± SD	10.7 ± 2.2	8.5 ± 2.0	4.6 ± 1.5
Sum	107	85	46
Grand sum		238	

Total sum of squares = $(x_1{}^2 + x_2{}^2 \cdots\cdots x_{29}{}^2 + x_{30}{}^2) - \dfrac{(\sum x)^2}{\sum N} =$

$$= (13^2 + 6^2 \cdots\cdots 5^2 + 3^2) - \frac{(238)^2}{30} = 291.9$$

Sum of squares of among the groups

$$= \frac{107^2}{10} + \frac{85^2}{10} + \frac{46^2}{10} - \frac{(238)^2}{30} = 190.9$$

Total sum of squares for error = Total sum of squares—Sum of squares of among the groups
= 291.9–190.9 = 101

We have all the estimates required for constructing the ANOVA Table. See Table 11.3 given below:

Table 11.3. ANOVA Table

Source of variation	SS	DF	Variance (MS)	*F*-value	*P*
Total	291.9	29	-	-	-
Groups	190.9	2	95.5	25.5	*P*<0.001
Error	101	27	3.74		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.
Note: There are 30 observations, hence the DF for SS total is 30–1 = 29; Total number of groups are three, hence the DF for SS groups is 3–1 = 2; DF for error SS = DF for SS total—DF for groups SS (29–2 = 27).

$$F_{27}^{2} calc = \frac{95.5}{3.74} = 25.5$$

Compare the derived *F* value with the value given in the *F* distribution Table (Table 11.4):

Table 11.4. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
27	13.613	**9.019**	7.272	6.326	5.726	5.308	4.998	4.759	4.568	4.412

N_1—DF associated with the numerator (in this example, the DF associated with 95.5); N_2—DF associated with the denominator (in this example, the DF associated with 3.74). Since the calculated *F*-value is greater than the Table value, it is considered that the jaundice index of the newborn baby is significantly different among the weeks at 0.1% probability level.

post hoc Comparison

ANOVA indicates that the jaundice index of the newborn baby is significantly different among the groups. The question is, which group is different from the other group or groups? Are all the groups are different from each other? The possible comparisons that we can make in this particular example are:

Group 1 *vs* Group 2

Group 1 *vs* Group 3

Group 2 *vs* Group 3

There are several tests available in the literature for *post hoc* comparison. Few tests that are commonly used in pharmacology and toxicology are explained below:

Dunnett's multiple comparison test

Dunnett's multiple comparison test (Dunnett, 1955) is a widely used approach for comparing all groups with the control (Cheung and Holland, 1991).

To compare the Jaundice indices of weeks 36 and 38 with that of week 40 (*i.e.*, Group 1 *vs* Group 3 and Group 2 *vs* Group 3), Dunnett's multiple comparison test is the most appropriate statistical tool. Here, we are considering Group 3 as some sort of 'standard' or 'control'. Dunnett's multiple comparison test should not be used for other comparison, such as, comparison between Group 1 and Group 2.

Comparison between Group 1 and Group 3:

$$= \frac{10.7 - 4.6}{\sqrt{3.74} \times \sqrt{\frac{2}{10}}} = \frac{6.1}{0.8648} = 7.05 \qquad \therefore p < 0.001$$

Comparison between Group 2 and Group 3:

$$= \frac{8.5-4.6}{\sqrt{3.74}\times\sqrt{\frac{2}{10}}} = \frac{3.9}{0.8648} = 4.51 \qquad \therefore p < 0.001$$

The calculated values (7.05 and 4.51) are greater than the Dunnett's *t*-test critical value given in Table 11.5. Dunnett's *t*-test critical value at 3 (numerator)/27 (denominator) degrees of freedom is 3.674

Hence, it is considered that Jaundice indices of weeks 36 and 38 are different from that of week 40.

Table 11.5. Dunnett's *t*-test critical values (one-sided test at 0.1% probability level) (Yoshimura, 1987)

DF	2	3	4	5	6	7	8
27	3.422	**3.674**	3.821	3.922	3.999	4.061	4.114

Note: One-sided *t*-test is more appropriate in this example as it is an established fact that the jaundice index decreases in newborn babies as their age increases.

The power of the Dunnett's test decreases as the number of groups increases. This could be better understood from the data given in Table 11.6.

Table 11.6. Change in the power of the Dunnett's test when the number of groups increases

Data and tests	Control	Low dose	Mid dose	High dose	Top dose
Hemoglobin level (g/dl) of B6C3F1 male mice at Week 78	13.9, 14.3	14.0, 13.3	14.0, 13.8	14.1, 13.9	14.2, 14.2
	13.7, 13.8	15.0, 13.8	13.7, 13.8	14.3, 14.0	14.7, 13.9
	14.0, 14.3	14.1, 13.3	13.5, 14.1	14.2, 14.1	14.3, 13.7
	13.9, 13.7	14.1, 13.9	14.2, 13.8	14.3, 14.4	14.3, 14.4
	13.9, 13.5	13.8, 13.4	14.1, 14.0	14.4, 14.4	14.0, 14.3
N	10	10	10	10	10
Mean ± SD	13.9 ± 0.3	13.9 ± 0.4	13.9 ± 0.2	14.2 ± 0.2	14.2 ± 0.3
Rejection value in Dunnett's Table at 0.05 (two-sided)	2.45				
Statistical result		NS	NS	S	
Rejection value in Dunnett's Table at 0.05 (two-sided)	2.53				
Statistical result		NS	NS	NS	NS

NS-Not significant; S-Significant ($P < 0.05$)

In the four-group setting (control, low dose, mid dose and high dose), the high dose group showed a significant difference from the control group, whereas in the in the five-group setting (control, low dose, mid dose, high dose and top dose), no significant difference was seen in the high dose group compared to the control group, indicating a decrease in the power of Dunnett's test to detect a significant difference as the number of groups increases.

Tukey's multiple range test (Yoshida, 1980)

Tukey's multiple range test, also known as Tukey range test, Tukey's honest significance test (Tukey's HST) or the Tukey–Kramer test (Mathews, 2005), is used to compare all possible pairs of means.

This is exemplified by reviewing the example given in Table 11.2.

The variance of the error is 3.74 (Table 11.3).

$$S\bar{x} = \sqrt{\frac{3.74}{10}} = 0.6116$$

Find the *Q* (critical) value from the Table of Tukey (Table 11.7). In this example, *Q* at 5% probability level is 2.8882 [Number of groups = 2 ; Degrees of freedom for error = 30. Actual degrees of freedom of error is 27 (Table 11.3); since this value is not given in Table 11.7, the value 30 is considered].

Table 11.7. Tukey's critical value at 5% probability level (Yoshida, 1980)

Degrees of freedom for error	Number of Groups						
	2	3	4	5	6	8	10
24	2.9188	3.5317	3.9013	4.1663	4.3727	4.6838	4.9152
30	**2.8882**	3.4864	3.8454	4.1021	4.3015	4.6014	4.8241

Next step is the calculation of significant difference *D*. It is the product of Sx and *Q* (Sx×Q):

$$D = S\bar{x} \times Q = 0.6116 \times 2.8882 = 1.7664$$

If the difference between any two means is greater than *D*, the difference is considered significant.

The difference between the means is given in Table 11.8. All means are different from each other.

Table 11.8. Jaundice index of newborn baby-Difference between mean values

Estimates	Week		
	36 (Group 1)	**38 (Group 2)**	**40 (Group 3)**
Mean	10.7	8.5	4.6
Difference of Means	Group 1 and Group 2	2.2^{a}	Significant (P<0.05)
	Group 1 and Group 3	6.1^{b}	Significant (P<0.05)
	Group 2 and Group 3	3.9^{c}	Significant (P<0.05)

Note: The superscripts of the mean values can be explained as—"Values bearing similar superscripts are statistically the same". Since the superscripts of the mean values are different, it can be stated that each mean value is different from the other.

Williams's test

Most of the regulatory guidelines prescribe that the repeated-dose administration studies with rodents should be conducted with a minimum of three levels of doses (low, mid and high doses) and a control group (OECD, 1995). The high dose is chosen with the aim to induce toxicity but not death or severe suffering (OECD, 1998; EPA, 2000), whereas the low dose is chosen with the assumption that animals exposed to this dose level will not show any effect of the treatment compared to the control group (Kobayashi *et al.*, 2010). However, these guidelines do not state how to determine the mid dose. It only indicates that this dose is required to examine dose dependency. According to Gupta (2007), the mid dose selection should consider threshold in toxic response and mechanism of toxicity. Choosing the mid dose is as important as choosing the high and low doses in repeated dose administration studies, since mid dose plays a determining role in establishing the dose dependency. It is not uncommon to encounter situations where mid dose alone shows an insignificant difference compared to the control group, whereas low and high doses show a significant difference. In this situation the data are examined for a dose-related trend. Williams' test is generally carried out to test dose-related trend (Bretz, 2006).

For the data that show a dose-related trend and a significant difference by Dunnett's test (Dunnett, 1955), the interpretation of the data analysis can be done in a straight forward manner. In a four group-setting repeated dose administration study, seven different situations can be expected (Table 11.9). Interpretation is relatively easier in situations 1–3, whereas it is difficult in situations 4–7, where further investigation on dose-related trend is required.

Table 11.9. Significant difference shown by the treatment groups by Dunnett's test—Possible situations

Test Group	●: Significant difference, ○: No significant difference from the control group						
	Situation 1	Situation 2	Situation 3	Situation 4	Situation 5	Situation 6	Situation 7
Control	○	○	○	○	○	○	○
Low dose	●	○	○	●	●	○	●
Mid dose	●	●	○	○	●	●	○
High dose	●	●	●	○	○	○	●
Investigation	Not required	Not required	Not required	Required	Required	Required	Required
Visual dose-related trends	Yes	Yes	Yes	No	No	No	No

Absolute kidney weight of rats from a repeated dose administration study is given in Table 11.10. These data were analysed using Dunnett's and Williams' tests. Dunnett's test showed a significant difference in low and high dose groups, whereas Williams' test showed a significant difference in all the groups.

Table 11.10. Absolute kidney weights of rats

Absolute kidney weights	Dose group			
	Control	Low	Mid	High
Individual data, (g)	2.558	3.269	3.116	2.706
	2.789	3.428	2.791	3.293
	2.764	3.083	2.981	3.535
	2.707	3.532	3.337	3.387
	2.793	3.546	2.432	3.064
	3.041	2.677	2.934	3.102
	3.000	2.822	3.388	3.279
	-	3.656	2.911	-
	-	3.271	2.798	-
	-	3.348	3.208	-
	-	3.031	2.876	-
	-	3.742	2.703	-
Number of animal	7	12	12	7
Mean ± Standard deviation	2.807±0.167	3.284±0.329	2.956±0.273	3.195±0.269
Bartlett's homogeneity test	*P* = 0.4130 (No heterogeneity)			
Dunnett's test		*P* = 0.0026*	*P* = 0.5190	*P* = 0.0332*
Mean value used for Williams' test	2.807	3.284	3.120	3.195
Williams' test		*P*<0.05*	*P*<0.05*	*P*<0.05*
Jonckheere's trend test	No significant difference			

*Significantly different from control group.

Use of Williams' test is not recommended when the number of animals in the groups is different (Williams, 1972) and extremely less (Williams, 1971; 1972). But, Sakaki *et al.* (2000) stated that Williams' test can be used even if number of the animals in a group differs about 2 times compared to other group/s.

Williams' test analyzes the difference of the mean values between each treated group and the control, like Dunnett's test, when the mean value of the treated groups changes in one direction. The example given in Table 11.11 does not show a dose-dependence as the mid dose showed an insignificant liver weight compared to control (by Dunnett's test). When the data were analysed by Williams' test, significance in the liver weight is observed in the mid dose group. The reason for this may be better explained by elucidating the calculation procedure of Williams' test as given below:

Table 11.11. Liver weight of rats in a 4-week repeated dose administration study

Group	Liver weight (g), N=5, (Sum)	Mean ± SD (% change with respect to control)	Results of Dunnett's test	Mean for Williams' test (% change with respect to control)	Results of Williams' test
Control	10.7, 11.5, 11.6, 12.0, 11.0 (56.8)	11.36 ± 0.51 (100)		11.36 (100)	
Low dose	11.6, 12.3, 12.5, 12.3, 12.7 (61.4)	12.28 ± 0.41 (108.1)	*P*<0.05	12.28 (108.1)	*P*<0.05
Mid dose	11.2, 11.5, 11.6, 11.5, 11.5 (57.3)	11.46 ± 0.15 (100.9)	Not significant	11.87 (104.5)	*P*<0.05
High dose	12.2, 12.5, 12.0, 11.9, 13.0 (61.6)	12.32 ± 0.44 (108.5)	*P*<0.05	12.32 (108.5)	*P*<0.05

Calculation procedure of Williams' test:

(1) Control *vs* High dose

$$\frac{61.4+57.3+61.6}{5+5+5}=12.02$$

(Note: Numerator—sums of low dose + mid dose + high dose; denominator—number of observations of low dose + mid dose + high dose).

$$\frac{57.3+61.6}{5+5}=11.89$$

(Note: Numerator—sums of mid dose + high dose; denominator—number of observations of mid dose + high dose).

$\frac{61.6}{5} = 12.32$ ←This largest value is used for the calculation of *t* value.

(Note: Numerator—sum of high dose; denominator—number of observations of high dose).

We have all estimates for calculating the *t* value, except the mean SS of error variance. Let us analyse the data using ANOVA:

Liver weight of rats in a 4-week repeated dose administration study

Statistics

Estimates	Liver weight (g)			
	Control	Low dose	Mid dose	High dose
N	5	5	5	5
Mean ± SD	11.36 ± 0.51	12.28 ±0.41	11.46±0.15	12.32 ± 0.44
Sum	56.8	61.4	57.3	61.6
Grand sum	237.1			

Total sum of squares

$$= \ (x_1^{\ 2} + x_2^{\ 2} \cdots\cdots x_{29}^{\ 2} + x_{30}^{\ 2}) - \frac{(\sum x)^2}{\sum N} =$$

$$= \ (10.7^2 + 11.5^2 \cdots\cdots 11.9^2 + 13^2) - \frac{(237.1)^2}{20} = 6.6095$$

Sum of squares of among the groups

$$= \frac{56.8^2}{5} + \frac{61.4^2}{5} + \frac{57.3^2}{5} + \frac{61.6^2}{5} - \frac{(237.1)^2}{20} = 3.9895$$

Total sum of squares for error = Total sum of squares—Sum of squares of among the groups

$= \ 6.6095 - 3.9895 = 2.62$

The ANOVA Table constructed is given below (Table 11.12).

Table 11.12. ANOVA Table

Source of variation	SS	DF	MS	*F* value	*P*
Total	6.6095	19	-	-	-
Groups	3.9895	3	1.32983	8.12	*P*<0.001
Error	2.62	16	0.16375		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Mean SS for error is 0.16375. Now we have all the required estimates for calculating *t*:

$$t = \frac{11.36 - 12.32}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 3.751$$

t-value is significant at 5% level (Table 11.13, Number of groups-4; DF-16).

(2) Control *vs* Mid dose

$\frac{61.4 + 57.3}{5 + 5} = 11.87$←This largest value is used for the calculation of *t*-value.

(Note: Numerator—sums of low dose + mid dose; denominator—number of observations of low dose + mid dose).

$$\frac{57.3}{5} = 11.46$$

(Note: Numerator- sum of mid dose; denominator- number of observations of mid dose).

$$t = \frac{11.36 - 11.87}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 1.993$$

t value is significant at 5% level (Table 11.13, Number of groups-3; DF-16).

(3) Control *vs* Low dose

$$\frac{61.4}{5} = 12.28$$

(Note: Numerator- sum of low dose; denominator- number of observations of low dose).

$$t = \frac{11.36 - 12.28}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 3.595$$

t-value is significant at 5% level (Table 11.13, Number of groups-2; DF-16).

Table 11.13. Williams' Table

DF	Number of groups							
	2	3	4	5	6	7	8	9
15	1.753	1.839	1.868	1.882	1.891	1.896	1.900	1.903
16	**1.746**	**1.831**	**1.860**	1.873	1.882	1.887	1.891	1.893
17	1.740	1.824	1.852	1.866	1.874	1.879	1.883	1.885

The reason for Williams' test showing a significant difference in the weight of the liver of the mid dose group, when compared with the control group, is that the test used 11.87 as the mean value of the mid dose group for the comparison instead of the actual value (11.46).

Williams' test is a useful statistical tool in toxicology as it provides information on evidence of toxicity and also the dose level that causes the toxicity (Shirley, 1977). Williams' test is similar to Dunnett, Tukey and Duncan multiple comparison (range) tests as it uses the error variance of the ANOVA (Nagata and Yoshida, 1997) in the calculation procedure. Williams' test is a closed procedure. If no significant difference between control group and highest dose group is seen, all the other treated groups are considered to have no significant difference compared to the control group and no further analysis is carried out. If there is a significant difference in the highest dose group, then the next highest dose level is examined for the significant difference from the control. If this dose group does not show a significant difference, no further analysis is carried out. But if it shows a significant difference, the next highest dose level is examined for the significant difference from the control group. Thus all the dose groups are sequentially examined.

Williams' test is effective in monotonic and non-monotonic dose-response relationships (Dmitrienko *et al.*, 2007). Since estimated mean values are used in the calculation procedure of Williams' test, it is likely that this test might show a dose-related trend, where it actually does not exist. It also may be noted in this context that, according to Gad and Weil (1988) dose-related trend is necessarily not evident in all the parameters.

Duncan's multiple range test (Shibata, 1970)

Duncan's multiple range test is generally used for comparison of more than 2 groups, when the number of observations of the groups is different. We shall work on the example given in Table 11.2. The data is slightly modified by changing the number of observations of Groups 1 and 2. The changed data are given in Table 11.14.

Table 11.14. Jaundice index of newborn baby. Reproduced from Table 11.2. Number of observations of Groups 1 and 2 was changed

Week		
36 (Group 1)	38 (Group 2)	40 (Group 3)
13	9	5
6	11	5
11	11	4
12	10	7
14	7	7
10	7	3
9	5	3
	8	4
		5
		3

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	7	8	10
Mean ± SD	10.7±2.7	8.5±2.1	4.6±1.5
Sum	75	68	46
Grand sum		189	

Calculation steps:

Total sum of squares =

$$(x_1{}^2 + x_2{}^2 \cdots\cdots x_{24}{}^2 + x_{25}{}^2) - \frac{(\sum x)^2}{\sum N} =$$

$$(13^2 + 6^2 \cdots\cdots 5^2 + 3^2) - \frac{(189)^2}{25} = 260.2$$

Sum of squares of among the groups

$$= \frac{75^2}{7} + \frac{68^2}{8} + \frac{46^2}{10} - \frac{(189)^2}{25} = 164.3$$

Total sum of squares for error = Total sum of squares—Sum of squares among the groups

= 260.2 – 164.3 = 95.9

Let us construct the ANOVA Table (Table 11.15).

Table 11.15. ANOVA Table

Source of variation	SS	DF	MS	*F* value	*P*
Total	260.2	24	-	-	-
Groups	164.3	2	82.2	19.1	*P*<0.001
Error	95.9	22	4.3		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Note: There are 25 observations, hence the DF for total SS is 25–1 = 24; Total number of groups are three, hence the DF for SS groups is 3–1 = 2; DF for error SS = DF for total SS—DF for SS among groups (24 – 2 = 22).

$$F_{22}^{2}calc = \frac{82.2}{4.3} = 19.1$$

Compare the derived *F* values with that of the value given in the *F* Distribution Table (Table 11.16.)

Table 11.16. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1\backslash N_2$	1	2	3	4	5	6	7	8	9	10
22	14.380	**9.612**	7.796	6.814	6.191	5.758	5.438	5.190	4.993	4.832

N_1—DF for the numerator; N_2–DF for the denominator.

Since the derived *F*-value is greater than the Table value, it is considered that the jaundice index of the newborn baby is significantly different among the weeks at 0.1% probability level.

Let us carry out *post hoc* comparison using Duncan's multiple range test. The first step is calculation of 'least significant range', *Rp*:

R_p = Sm×Q, where

$$Sm = \sqrt{\frac{\text{MS for error variance}}{\sum N / \text{Number of groups}}}$$

Q = Critical value from Duncan's table

$$Sm = \sqrt{\frac{4.3}{25/3}} = 0.72$$

Note: 4.3 is variance of error (see Table 11.15); Total number of observation = 25; Total number of groups = 3).

Critical *Q* values are obtained from Duncan's Table (Table 11.17). *Q* values at 22 degrees of freedom (Degrees of freedom of the error component; see Table 10.15) for 2 and 3 Groups are 2.93 and 3.08, respectively.

Table 11.17. Duncan's critical values at 5% probability level (Shibata, 1970)

DF	Group								
	2	3	4	5	6	7	8	9	10
22	**2.93**	**3.08**	3.17	3.24	3.29	3.32	3.35	3.37	3.39

$$R_{2(0..05)} = 0.72 \times 2.93 = 2.11$$

$$R_{3(0..05)} = 0.72 \times 3.08 = 2.22$$

Arrange the mean values orderly:

Group 1 (Week 36) = 10.7

Group 2 (Week 38) = 8.5

Group 3 (Week 40) = 4.6

Let us compare the largest sample means range, *i.e.,* 10.7 and 4.6. The difference between these two mean values is 6.1, which is greater than the 'least significant range', R_3. Hence, the difference between these two mean values (Group 1 and Group 3) is considered significant. Let us compare the next set of mean values, 10.7 and 8.5. The difference between these two mean values is 2.2, which is greater than the 'least significant range', R_2. Hence the difference between the mean values of Group 1 and Group 2 is also considered significant.

Scheffé's multiple comparison test (Scheffe, 1953)

We shall use the data given in Table 11.14 for demonstrating Scheffé's multiple comparison test.

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	7	8	10
Mean ± SD	10.7±2.7	8.5±2.1	4.6±1.5
Sum	75	68	46
Grand sum		189	

Comparisons:

Group 3 *vs* Group 2

$$F = \frac{(4.6 - 8.5)^2}{(3-1) \times 4.3 \times (\frac{1}{10} + \frac{1}{8})} = 7.86 \qquad \therefore p < 0.05$$

Group 3 *vs* Group 1

$$F = \frac{(4.6 - 10.7)^2}{(3-1) \times 4.3 \times (\frac{1}{10} + \frac{1}{7})} = 17.82 \qquad \therefore p < 0.05$$

Group 2 *vs* Group 1

$$F = \frac{(8.5 - 10.7)^2}{(3-1) \times 4.3 \times (\frac{1}{8} + \frac{1}{7})} = 2.10 \qquad \therefore p > 0.05(NS)$$

Note: 4.3 is the variance of error (*vide* Table 11.15).

These derived *F*-values are compared with the values given in *F* distribution Table (Table 11.18) given below:

Table 11.18. *F*-distribution values at 5% probability level (Yoshimura, 1987)

N_1\ N_2	1	2	3	4	5	6	7	8	9	10
22	4.301	**3.443**	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297

N_1-DF for the numerator; N_2-DF for the denominator.

All the derived *F* values, except the one computed for the comparison between Group 2 and Group 1, are significant at 5% probability level.

The Scheffé's multiple comparison test is used for all-pair comparisons, like the Duncan's multiple comparison test. However, the power to detect a significant difference is low with the Scheffé's multiple comparison test compared to that of the Duncan's multiple comparison test (*vide* Table 11.19).

Duncan's multiple comparison test showed a significant difference in the mid dose and high dose groups, whereas the Scheffé's multiple comparison test did not show a significant difference in these groups, indicating it's low power to detect a significant difference. Therefore, use of Scheffé's multiple comparison test should be done with little caution in the safety evaluation studies with animals.

Table 11.19. Comparison of the power to detect a significant difference between Scheffés and Duncan's multiple comparison tests. LDH activity (U/l) of F344 female rats at week 78 in a repeated dose administration study is given.

Estimates	Control	Low dose	Mid dose	High dose
-	168, 188, 181, 250, 122, 89, 125, 135, 211, 204	112, 168, 175, 241, 218, 49, 49, 76, 66, 30	69, 86, 145, 244, 135, 46, 105, 40, 53, 73	43, 59, 73, 99, 129, 181, 49, 69
N	10	10	10	8
Mean ± SD	167 ± 49	118 ± 76	100 ± 62	88 ± 47
In % of control	-	71	60	53
ANOVA	$P < 0.05$			
Duncan's test		N.S.	S	S
Scheffé's test		N.S.	N.S.	N.S.

N.S.—Not significant ($P > 0.05$); S—Significant ($P < 0.05$)

Two-way ANOVA

It is an extension of one-way ANOVA. The difference in 2-way ANOVA is that it has 2 independent factors. The data is arranged in tabular fashion in such a way that the column represents one factor and the row, the other factor (Belle *et al.*, 2004).

An example is provided to illustrate the computations required in two-way ANOVA (Kibune and Sakuma, 1999).The diameter of the head of the three human embryos was measured by four observers. Each observer measured the diameter of three embryos. The data is arranged in a tabular fashion as given in Table 11.20. We are interested to know: 1. Among the observers, is there any difference in the diameter of embryos measured 2. Among the embryos, is there any difference in the diameter of embryos measured and 3. Is there any simultaneous influence of observer and embryo in the diameter measured (interaction)

Calculation steps:

1) Correction factor (CF)
   $=(\text{Grand sum})^2/N = 558.1^2/36=8652.1$
2) Total sum of squares
   $= (14.3^2+14.0^2+......+12.9^2+13.8^2)\text{-CF}=8979.7–8652.1=$**327.6**
3) Sum of squares of among the observers
   $=1/9\ (141.0^2+137.6^2+138.2^2+141.3^2)\text{-CF}=8653.2–8652.1=$**1.199**

4) Sum of squares of among the embryos
=1/12 ($167.9^2+236.3^2+153.9^2$)-CF=8976.1–8652.1=**324**
5) Embryo × Observer (Interaction)
=1/3 ($43.1^2+59.4^2+38.5^2+41.0^2+58.9^2+37.7^2+41.4^2+58.8^2$
$+38.0^2+42.4^2+59.2^2+39.7^2$)-CF=8977.8–8652.1=**325.7**
Sum of squares of interaction is calculated as given below:
325.7–1.199–324= **0.501**. The DF for interaction is (3–1) (4–1)=6.
6) Sum of squares of error
327.6–1.199–324–0.501=**1.9**. The DF for error is 35–2–3–6=24

Table 11.20. Diameter of three human embryos (cm) measured by four observers

	Observer 1	Observer 2	Observer 3	Observer 4	Sum
Embryo 1	14.3	13.6	13.9	13.8	167.9 (109)
	14.0	13.6	13.7	14.7	
	14.8	13.8	13.8	13.9	
Sum	43.1	41.0	41.4	42.4	
Embryo 2	19.7	19.8	19.5	19.8	236.3 (154)
	19.9	19.3	19.8	19.6	
	19.8	19.8	19.5	19.8	
Sum	59.4	58.9	58.8	59.2	
Embryo 3	13.0	12.4	12.8	13.0	153.9 (100)
	12.6	12.8	12.7	12.9	
	12.9	12.5	12.5	13.8	
Sum	38.5	37.7	38.0	39.7	
Total sum	141.0 (99.8)	137.6 (97.4)	138.2 (97.8)	141.3 (100)	558.1

Let us construct the ANOVA Table (Table 11.21):

Table 11.21. Two-way layout ANOVA

Source of variation	SS	DF	MS	*F* value	*P*
Embryo*	324	2	162	2051	*P*<0.001
Observer*	1.199	3	0.399	5.05	*P*<0.01
Embryo×Observer**	0.501	6	0.084	1.06	NS
Error	1.9	24	0.079		
Total sum	327.6	35			

*Main effects **Interaction, SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

The computed *F* values are compared with the *F* distribution values given in *F* distribution Table (Table 11.22). For the comparison of all the sources of variation (embryo, observer and embryo × observer interaction), the denominator remains the same (DF of error, which is 24), but the numerator differs. The *F* values should be compared with *F* distribution values at 2/24 (numerator/denominator) for embryo, 3/24 for observer and 6/24 for embryo × observer interaction.

Table 11.22. *F* distribution values at 1% probability level (Yoshimura, 1987)

$N_1\backslash N_2$	1	2	3	4	5	6	7	8	9	10
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168

N_1- DF for the numerator; N_2 –DF for the denominator.

Discussion:

1. Embryo: The *F*-value is greater than the Table *F*-value (2051>5.614); hence there is a significant difference among embryos.
2. Observer: The *F*-value is greater than the Table *F*-value (5.05>4.718); hence there is a significant difference among observers.
3. The embryo × observer interaction: The *F*-value is less than the Table *F* value (1.06<3.667); hence embryo×observer interaction is not significant.

Since the interaction is not significant, the ANOVA Table can be reconstructed excluding interaction as a source of variation. The SS of interaction is added to the SS of error and the DF of the interaction is added to the DF of error. The Table thus reconstructed after excluding interaction as a source of variation is given below (Table 11.23):

Table 11.23. ANOVA Table excluding the interaction

Source of variation	SS	DF	MS	*F* value	*P*
Embryo	324	2	162	2024	*P*<0.001
Observer	1.199	3	0.399	4.99	*P*<0.001
Error	2.40	30	0.080		
Total sum	327.6	35	-		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Dunnett's Multiple Comparison Test and Student's *t* Test—A Comparison

In pharmacological and toxicological experiments the number of groups usually employed is more than two. If the data obtained from such studies are anlaysed by Student's *t*-test (picking up any two groups and analyzing by Student's *t*-test), it may cause Type I error.

We analysed data obtained from several repeated dose administration studies in rats using Dunnett's multiple comparison test and Student's *t*-test to know to what extent repeated analysis by Student's *t*-test shows a Type I error. Our finding is given in Table 11.24.

Table 11.24. Analysis of data obtained from repeated dose administration studies in rats by Dunnett's multiple comparison test and Student's *t*-test

Item	Number of analyses	Dunnett's multiple comparison test	Student's *t*-test[a]
Body weight	528	223	246 (10)
Feed consumption	832	235	349 (49)
Hematology	352	123	159 (29)
Blood chemistry	576	215	272 (27)
Urinalysis	64	7	11 (57)
Organ weight	224	47	80 (70)
Organ weight/ body weight ratio	224	82	104 (27)
Total	2800	932	1221 (31)

[a] Values given in parentheses are percent increase compared to Dunnett's multiple comparison test.

The number of items showing a significant difference by Student's *t*-test increased, compared to those showing a significant difference by Dunnett's multiple comparison test. Overall, there was an increase by 31% in the items, when they were analysed by Student's *t*-test. This increase is due to the Type I error. Yoshimura and Tsubaki (1993) suggested that to assess the toxicity, Dunnett's multiple comparison test is the appropriate statistical approach; on the contrary, from a consumer point of view, Student's *t*-test, may be more appropriate.

References

Armstrong, R.A., Slave, S.V. and Eperjesi, F. (2000): An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. Ophthalmic Physiol. Opt., 20(3), 235–241.

Belle, G.V., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics—A Method for Health Sciences. John Wiley & Sons, Inc., New Jersey, USA.

Bretz, F. (2006): An extension of the Williams' trend test to general unbalanced linear models. Comp. Stat. Data Anal., 50(7), 1735–1748.

Cheung, S.H. and Holland, B. (1991): Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics, 47, 21–32.

Dmitrienko, A., Chaung-Stein, C. and D'Agostino, R. (2007): Pharmaceutical Statistics. Using SAS—A Practical Guide. SAS Institute, NC, USA.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. Am. Stat. Assoc., 50, 1096–1211.

EPA (2000): United States Environmental Protection Agency. Health Effects Test Guidelines, OPPTS 870.3050, Repeated Dose 28—Day Oral Toxicity Study in Rodents, EPA 712–C–00–366 2000. EPA, USA.

Gad, S. and Weil, C.S. (1988): Statistics and Experimental Design for Toxicologists. Telford Press, New Jersey, USA.

Gill, L. (!990): Uses and abuses of statistical methods in research in parasitology. Vet. Parasitol., 36(3-4), 189–209.

Gupta, R.C. (2007): Veterinary Toxicology—Basic and Clinical Principles. Academic Press, New York, USA.

Kibune, Y. and Sakuma, A. (1999): Practical Statistics for Medical Research, Scientist Press, Tokyo, Japan.

Kobayashi, K., Pillai, K.S., Michael, M., Cherian, KM., Ohnishi, M. (2010): Determining NOEL/NOAEL in repeated-dose toxicity studies, when the low dose group shows significant difference in quantitative data. Lab. Anim. Res., 26(2),133–137.

Mathews, P.G. (2005): Design of Experiments with MINITAB. American Society for Quality, Milwaukee, USA.

Moder, K. (2007): How to keep the Type I error rate in ANOVA if variances are heteroscedastic. Aust. J. Stat., 6(3), 179–188.

Muir, W.M., Romero-Severson, J., Rider Jr., S.D., Simons, A. and Ogas, J. (2006): Application of one sided *t*-tests and a generalized experiment-wise error rate to high-density oligonucleotide microarray experiments: An example using Arabidopsis. J. Data Sci., 4, 323–341.

Nagata, Y. and Yoshida, M. (1997): Tokeiteki-tajuhikakuho-no-kiso. Scientist Co. Ltd, Tokyo, Japan.

Norman, G.R. and Streiner, D.L. (2008): Biostatistics—The Bare Essentials. 3rd Edition. BC Decker Inc., Ontario, Canada.

OECD (1995): Organization for Economic Cooperation and Development. OECD Guidelines for Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents. No. 407, OECD, France.

OECD (1998): Organization for Economic Co-operation and Development. OECD Guideline for the Testing of Chemicals. Repeated Dose 90 Day Oral Toxicity Study in Rodents, No. 408. OECD, France.

Rothmann, M. (2005): Type I error probabilities based on design-stage strategies with applications to noninferiority trials. J. Biopharm. Stat., 15(1), 109–127.

Sakaki, H., Igarashi, S., Ikeda, T., Imamizo, K., Omichi, T., Kadota, M., Kawaguchi, T., Takizawa, T., Tsukamoto, O., Terai, K., Tozuka, K., Hirata, J., Handa, J., Mizuma, H., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Scheffé, H. (1953): A method for judging all contrasts in the analysis of variance. Biometrica, 40, 87–104.

Shibata, K. (1970): Biostatistics.Tokyo University of Agriculture, Tokyo, Japan.

Shirley, E. (1997): A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, 33(2), 386–389.

Wallenstein , S., Zucker, C.L. and Fleiss, J.L. (1980): Some statistical methods useful in circulation research. Circ. Res., 47, 1–9.

Williams, D.A. (1971): A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, 27,103–117.

Williams, D.A. (1972): The comparison of several dose levels with zero dose control. Biometrics, 28, 519–531.

Yoshida, M. (1980): Design of Experiments for Animal Husbandry, Yokendo Press, Tokyo, Japan.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Inc., Tokyo, Japan.

Yoshimura, I and Tsubaki, H. (1993): Multiple comparison test for more than three dosed groups settings (debate). Japanese Society for Biopharmaceutical Statistics, August 7, 1993, Sohyo Kaikan, Tokyo, Japan.

12

Non-Parametric Tests

Non-parametric and Parametric Tests—Assumptions

Statistical methods are based on certain assumptions. For applying parametric statistical tools, the assumptions made are that data follow a normal distribution pattern and are homogeneous. In many situations, the data obtained from animal studies contradict these assumptions, and are not suitable to be analysed with the parametric statistical methods. Non-parametric tests do not require the assumption of normality or the assumption of homogeneity of variance. Hence, these tests are referred to as distribution-free tests. Non-parametric tests usually compare medians rather than means, therefore influence of one or two outliers in the data is annulled. We shall deal with some of the most commonly used non-parametric tests in toxicology/pharmacology.

Sign Tests

Perhaps, the sign test is the oldest distribution-free test which can be used either in the one-sample or in the paired sample contexts (Sawilowsky, 2005). Sign test is probably the simplest of all the non-parametric methods (Whitley and Ball, 2002; Crawley, 2005). The null hypothesis of the sign test is that given a pair of measurements (xi, yi), then xi and yi are equally likely to be larger than each other (Surhone *et al.*, 2010). Though the sign test is rarely used in toxicology, it can be used in certain pharmacological *in vivo* experiments to evaluate whether a treatment is superior to the other. The sign test may be used in clinical trials to know whether either of the two treatments that are provided to study subjects is favored over the other (Nietert and Dooley, 2011).

The calculation procedure of sign test for small sample size ($n \leq 25$) is different from that of large sample size ($n>25$):

Calculation procedure of sign test for small sample size

A study was conducted to evaluate the hypoglycemic effect of an herbal preparation in rats. Hyperglycemia was induced in rats by administering streptozotozin. Following the administration of streptozotozin, the blood sugar was measured in individual rats to confirm hyperglycemia. Then the hyperglycemic rats were given the herbal preparation daily for 14 consecutive days. On day 15, again blood sugar was measured in these rats. The blood sugar measured in hyperglycemic rats before and after the administration of the herbal preparation is given in Table 12.1.

Table 12.1. Blood sugar level (mg/dl) in hyperglycemic rats

Rat No.	1	2	3	4	5	6	7	8
Blood sugar level before administration of herbal preparation (X_a)	236	223	211	229	205	245	243	231
Blood sugar level after administration of herbal preparation (X_b)	155	156	172	198	209	181	231	231
Difference (X_b- X_a)	–81	–67	–39	–31	+4	–64	–12	0
Sign	- (–1)	- (–1)	- (–1)	- (–1)	+ (+1)	- (–1)	- (–1)	± (0)

$$p={}_7C_1\left(\frac{1}{2}\right)^6\frac{1}{2}+{}_7C_0\left(\frac{1}{2}\right)^7$$

$$=\left({}_7C_1+{}_7C_0\right)\left(\frac{1}{2}\right)^7$$

$$=0.0546+0.0078=0.0624$$

Note: ${}_nC_r=\dfrac{n\,!}{r\,!(n-r)!}$; Rat No. 8, which did not show any change in the blood sugar is not included in the analysis.

Since P=0.0624 is >0.05, it is considered that the decrease in blood sugar in rats administered with herbal preparation is insignificant.

Calculation procedure of sign test for large sample size

The effect of two analgesics, drugs A and B was evaluated five times by 32 doctors and their findings are given in Table 12.2. The objective of the

Table 12.2. Analgesic effect of drugs A and B evaluated by 32 doctors

Doctor No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Drug A (X_a)	4	5	4	5	5	3	5	4	3	4	5	3	3	4	5	4	3	4	4	5	4	5	4	3	3	5	2	1	4	5	3	5
Drug B (X_b)	2	3	3	4	2	3	2	5	3	5	3	4	2	4	3	5	4	3	5	3	4	3	5	4	4	2	4	3	2	4	2	2
Sign (X_b- X_a)	-	-	-	-	-	±	-	+	±	+	-	+	-	±	-	+	+	-	+	-	±	-	+	+	+	-	+	+	-	-	-	-

study was to know whether the analgesic effect of drugs A and B is similar or different.

The pairs, which showed a difference of 0 (± sign) are excluded from the calculation procedure. In this example four pairs showed a difference of 0 (± sign). Therefore, number (n) of data becomes 32–4=28. Number of + sign, which indicates that the effect of drug B is better than drug A, is 11. *Z* is obtained from the equation given below:

$$z = \frac{(r + 0.5) - \mu_r}{\sigma_r} = \frac{11.5 - 14}{2.65} = -0.94$$

$$\text{Mean } \mu_r = \frac{28}{2} = 14 \qquad \sigma_r(SD) = \frac{\sqrt{28}}{2} = 2.65$$

r = Total number of + sign = 11

The $p(|z| > 0.94 = 0.9) = 0.36812$ from normal distribution Table (Table 12.3) is greater than 0.05 (two-sided test). Therefore, it can be concluded that both the drugs have similar effect.

Table 12.3. Normal distribution table (Yoshimura, 1987)

%	Two-sided *P*	Upper *P*
Z	2α	α
0.8	0.423711	0.211855
0.9	**0.368120**	0.184060
1.0	0.317311	0.158655

Signed Rank Sum Tests

The major disadvantage of the sign test is that it considers only the direction of difference between pairs of observations, not the size of the difference (Mc Donald, 2009). Ranking the observations and then carrying out the statistical analysis can solve this issue. Signed rank sum test is more powerful than the sign test (Elston and Johnson, 1994).

Wilcoxon Rank-Sum test (Wilcoxon, 1945)

The Wilcoxon rank-sum test is one of the most commonly used non-parametric procedures (Le, 2003). This is the non-parametric analogue to the paired *t*-test. The null hypothesis of Wilcoxon rank-sum test is that the median difference between pairs of observations is zero.

The performance of six classes of two schools expressed in average scores is given in Table 12.4. We shall analyse this data using Wilcoxon rank-sum test.

Table 12.4. Average scores of six classes of two schools

School	Average score					
School A	79.5	85.5	83.5	93.5	91.5	77.5
School B	95.5	87.5	89.5	98.0	97.5	81.5

Step 1: Combine the scores of both the schools and arrange them from the smallest to the largest. Then assign a rank from 1 to 12 to the scores as given in Table 12.5. (Note: if there are tied observations, assign average rank to each of them).

Table 12.5. Ranks assigned to the combined scores of two schools

Scores arranged from smallest to largest	Rank
77.5	1
79.5	2
81.5	3
83.5	4
85.5	5
87.5	6
89.5	7
91.5	8
93.5	9
95.5	10
97.5	11
98	12

Step 2: Arrange the rank corresponding to the original scores as given in Table 12.6 and calculate the sum of the ranks.

Table 12.6. Ranks arranged to the original scores

School	Ranks						Sum of rank
School A	2	5	4	9	8	1	29
School B	10	6	7	12	11	3	49

Calculation Procedure:

The number of samples (classes) in each group = 6

Sum of rank of School B, R_2=10+6+7+12+11+3=49

$$V = \frac{\begin{bmatrix}(2-6.5)^2+(5-6.5)^2+(4-6.5)^2+(9-6.5)^2+(8-6.5)^2+(1-6.5)^2\\ +(10-6.5)^2+(6-6.5)^2+(7-6.5)^2+(12-6.5)^2+(11-6.5)^2+(3-6.5)^2\end{bmatrix}\times 6\times 6}{12\times 11}$$

$=39$

Where,

$$6.5 = \frac{29+49}{12}$$

12 = Sum of number of samples (classes) of School A and School B

11 = (Sum of number of samples (classes) of School A and School B) – 1

Let us calculate T

$$T = \frac{49-6\times\frac{13}{2}}{\sqrt{39}} = 1.601$$

Where,

13 = (Sum of number of samples (classes) of School A and School B) + 1

2 = Constant

Calculated T value (T=1.601) is smaller than the $U(\alpha)$ = 1.644854 at P= 0.05 (see Table 12.7). Hence, it is considered that there is no significant difference in scores between the schools.

Table 12.7. Standard normal distribution Table (Yoshimura, 1987)

Two tailed *P*	Upper *P*	% point
2α	α	U(α)
0.05000	0.025000	1.959964
0.06000	0.030000	1.880791
0.07000	0.035000	1.811911
0.08000	0.040000	1.750686
0.09000	0.045000	1.695398
0.10000	0.050000	**1.644854**

Fisher's exact test

Fisher's exact test is used in the analysis of contingency tables with small sample sizes (Fisher, 1922; 1954). It is similar to χ^2 test, since both Fisher's exact test and χ^2 test deal with nominal variables. In Fisher's exact test, it is assumed that the value of the first unit sampled has no effect on the value of the second unit. It is interesting to learn how the Fisher's exact test was originated. Dr Muriel Bristol of Rothamsted Research Station, UK claimed that she could tell whether milk or tea had been added first to a cup of tea. Fisher designed an experiment to verify the claim of Dr Muriel

Bristol. Eight cups of tea were made. In four cups, milk was added first and in the other four cups tea was added first. Thus, the column totals were fixed. Dr. Bristol was asked to identify the four to 'tea first', and the four to 'milk first' cups. Thus, the row totals were also fixed in advance. Fisher proceeded to analyse the resulting 2 × 2 table, thus giving birth to Fisher's exact test (Clarke, 1991; Ludbrook, 2008).

Manual analysis of data using Fisher's exact test is beyond the scope of this book, hence not covered. The power to detect a significant difference is more with Fisher's exact test than the χ^2 test as seen in Table 12.8.

Table 12.8. Power to detect a significant difference—Comparison between χ^2 test and Fisher's exact test

Incidence of pathological lesions (Control *vs* dosed group)	*P*-value	
	Chi-square test*	Fisher's test (α)
0/5 *vs* 1/5	1.00000	0.50000
0/5 *vs* 2/5	0.42920	0.22222
0/5 *vs* 3/5	0.16755	0.08333
0/5 *vs* 4/5	0.05281	0.02381
0/5 *vs* 5/5	0.01141	0.00397
1/5 *vs* 2/5	1.00000	0.50000
1/5 *vs* 3/5	0.51861	0.26190
1/5 *vs* 4/5	0.20590	0.10317
1/5 *vs* 5/5	0.05281	0.02381
2/5 *vs* 3/5	1.00000	0.50000
2/5 *vs* 4/5	0.51861	0.26190
2/5 *vs* 5/5	0.16755	0.08333

*Yetes's correction (Note on Yetes's correction: χ^2 slightly overestimates the 'difference between expected and observed' results. This overestimation can be corrected by decreasing the 'difference between expected and observed' by 0.5).

McKinney *et al.* (1989) reviewed the use of Fisher's exact test in 71 articles published between 1983 and 1987 in six medical journals. Nearly 60% of articles did not specify use of a one- or two-sided test. The authors concluded that the use of Fisher's exact test without specification as a one- or two-sided version may misrepresent the statistical significance of data.

Mann-Whitney's *U* ***test***

Mann-Whitney's *U* test, a test equivalent of Student's *t*-test for comparing two groups, was independently developed by Mann and Whitney (1947) and Wilcoxan (1945). The calculation procedure of Mann-Whitney's *U* test is very much similar to Wilcoxan signed rank sum test. For understanding Mann-Whitney's *U* test in a detailed manner, let us analyse the data given in Table 12.9. Our objective of the analysis is to find whether there is a significant difference in hemoglobin content between Group A and Group B.

Table 12.9. Hemoglobin content (g/dl) in two experimental groups of rats following the administration of a drug at 10 mg/kg b.w. (Group A) and at 20 mg/kg b.w. (Group B)

Group A	9.3	6.4	10.8	5.6
Group B	5.9	9.7	9.9	6.7

Let us pool the data and arrange them from the smallest to the largest, ignoring the Group to which they belong and rank them. Then, tag them with the identity of the Group to which they belong (Table 12.10).

Table 12.10. Ranking the data

Pooled data	5.6	5.9	6.4	6.7	9.3	9.7	9.9	10.8
Ranked data	1	2	3	4	5	6	7	8
Tagged data with respective group	A	B	A	B	A	B	B	A

(Note for tied observations: Assign mean score for the tied observations. For example, if the value of ranks 2nd and 3rd is 5.9, give each value a rank of 2.5).

Let n_a = Number of observations in Group A, n_b = Number of observations in Group B, T_a = Rank sum for Group A, T_b = Rank sum for Group B:

$T_a = 1+3+5+8 = 17$

$T_b = 2+4+6+7 = 19$

Let us calculate U_1 and U_2:

$$U_1 = T_a - \frac{n_a(n_a+1)}{2} = 17 - \frac{4(4+1)}{2} = 7$$

$$U_2 = T_b - \frac{n_b(n_b+1)}{2} = 19 - \frac{4(4+1)}{2} = 9$$

The smallest value 7 is the *U* value.

The smallest *U* value, 7 is compared with the Mann-Whitney *U* Table value at n_1=4 and n_2=4. Relevant part of the *U* Table is reproduced in Table 12.11.

Table 12.11. Mann-Whitney *U* Table

n_1	n_2	Two-sided		One-sided	
		α=0.05	α=0.01	α=0.05	α=0.01
2	2	---	---	---	---
3	3	---	---	---	---
4	4	0	---	1	---
5	5	2	0	4	1
6	6	5	2	7	3

Since the computed *U* value is greater than the values given in the Mann-Whitney *U* Table, it is not significant at 5% level by two-sided and one-sided tests (at 5 % significant level the *U* Table values are 0 and 1 for two-sided and one-sided tests, respectively).

When the size of either of the groups exceeds 20, the significance of *U* can be tested using the *Z* statistic:

$$Z = \frac{U - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}}$$

Z score for normal distribution is shown in Appendix 3.

(A note on *Z* statistic: *Z* is designated to a standard normal variate. It is computed by subtracting the measured value from the population mean, then dividing by the population SD(σ). A standard normal variate has a normal distribution with mean 0 and variance 1. The total area under a normal distribution curve is unity (or 100%). The notation, Prð($-1 < z < 1$) = 0.6826, indicates that about 68% of the area is contained within ± 1 SD).

Mann-Whitney's *U* test works well in the analysis of data obtained from toxicity studies, where the number of animals in each group is 27 or less. By Mann-Whitney's *U* test, a significant difference (one-sided test) can be detected even with three animals in each group. Therefore, this test can be used in experiments with dogs, where each group usually consists of three animals/sex. This test seems to be extensively used for analyzing urinalyses data and pathological findings in repeated dose administration studies in rodents.

The power to detect a significant difference is more with Mann-Whiney's *U* test than the Fisher's test. Analysis of pathological findings of a repeated dose administration study by Mann-Whitney's *U* and Fisher's tests is given in Table 12.12.

Table 12.12. Analysis of pathological findings of a repeated dose administration study by Mann-Whitney's *U* and Fisher's tests

Groups	Lesions grades and number of animals with lesions grade				Mann-Whitney's *U* test	Lesions grades and number of animals with lesions grade		Fisher's test
	-	±	+	++	*P*=0.0032 (One-sided)	-	> ±	*P*=0.0238 (One-sided)
Control	4	1	0	0		4	1	
High dose	0	0	3	2		0	5	

The computed *P* value for Mann-Whitney's *U* test (*P*=0.0032) is considerably less than that of the Fisher's test (*P* = 0.0238), indicating that the power to detect a significant difference is more with Mann-Whitney's *U* test than the Fisher's test.

The power of the Mann-Whitney's *U* test decreases when the groups to be compared have the same order of rank. There is a possibility in having the same order of rank, when the number of digits after decimal of the raw data is truncated. This can be better understood from the data given in Table 12.13.

Table 12.13. Change in the pattern of significant difference detection as the number of digits after decimal of the raw data decreases. Absolute liver weight (g) of male rats from a 28-day repeated dose administration study is given in the Table.

Number of digits after decimal	Items	Groups		P
		Control (N = 6)	High dose (N = 6)	Mann-Whitney's *U* test
3	Raw data	10.391, 11.442, 13.653, 10.224, 10.783, 10.414	13.194, 11.444, 13.701, 11.572, 12.683, 12.661	< 0.05
	Mean ± SD	11.151 ± 1.301	12.543 ± 0.889	
	Mean rank	4.3	8.6	
2	Raw data	10.39, 11.44, 13.65, 10.22, 10.78, 10.41	13.19, 11.44, 13.70, 11.57, 12.68, 12.66	Not significant
	Mean ± SD	11.15 ± 1.30	12.54 ± 0.89	
	Mean rank	4.4	8.5	
1	Raw data	10.4, 11.4, **13.7**, 10.2, 10.8, 10.4	13.2, 11.4, **13.7**, 11.6, 12.7, 12.7	Not significant
	Mean ± SD	11.2 ± 1.3	12.6 ± 0.9	
	Mean rank	4.5	8.5	

The high dose group is significantly different from the control group as per Mann-Whitney's *U* test, when the data of both the groups have three digits after decimal and no data from the control group is repeated in the high dose group and *vice versa*. When the number of digits after the decimal of the data was truncated to two decimals, the value 11.44 was repeated in both the groups, resulting in an insignificant difference between the control and high dose groups. When the number of digits after the decimal of the data was restricted to one decimal, the values 11.4 and 13.7 were repeated in both the groups, resulting in an insignificant difference between the control and high dose groups.

There are two methods for calculating the Mann-Whitney's *U* test. When the number of observations in each group is small (N= <27), the Mann-Whitney's *U* test can be calculated by using a ready reckoner (*http://aoki2.si.gunma-u.ac.jp/lecture/Average/u-tab.html*). When the number of observations in each group is large (N= >27), it is calculated using the *Z* distribution Table method. Table 12.14 demonstrates the analysis of a simulated data with a strong dose-related pattern by Mann-Whitney's *U* test using the *Z* distribution Table method. Table 12.15 demonstrates the analysis of a simulated data with strong dose-related pattern by Mann-Whitney's *U* test using the ready reckoner.

Table 12.14. Power of Mann-Whitney's *U* test for three and four samples with a strong dose-related pattern (calculated by using *Z* distribution Table)

Number of samples	Group	Raw data (ranked)	Mean rank	*Z* value	*P* value	
					Two-sided	One-sided
3	Control	1, 2, 3	2	1.96	0.04953	0.02500
	Dose	4, 5, 6	5			
4	Control	1, 2, 3, 4	2.5	2.30	0.0209	0.010
	Dose	5, 6, 7, 8	6.5			

Table 12.15. Power of Mann-Whitney's *U* test for three and four samples with a strong dose-related pattern (calculated by using the ready reckoner—*http://aoki2.si.gunma-u.ac.jp/lecture/Average/u-tab.html*)

Number of samples	Group	Raw data (ranked)	Mean rank	*U* value	*P* value	
					Two-sided	One-sided
3	Control	1, 2, 3	2	0.0	Not significant	*P*<0.05.
	Dose	4, 5, 6	5			
4	Control	1, 2, 3, 4	2.5	0.0	*P*=0.05	*P*<0.05.
	Dose	5, 6, 7, 8	6.5			

The Tables 12.14 and 12.15 indicate that there is not much difference in *P* values between *Z* distribution Table and ready reckoner methods, when the number of samples is as small as 3 to 4. However, we recommend a ready reckoner when the number of observations in each group is small (N= <27) and a *Z* distribution Table when the number of observations in each group is large (N= >27).

Kruskal-Wallis Nonparametric ANOVA by Ranks
(Kruskal and Wallis, 1952)

The Kruskal–Wallis test is identical to one-way ANOVA with the data replaced by their ranks. It has also been stated that this test is an extension of the two-group Mann-Whitney's *U* (Wilcoxon rank) test (Mc Kight and Najab, 2010). It assumes that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (for example, one is skewed to the right and another is skewed to the left or they have different variances), the Kruskal–Wallis test may give inaccurate results (Fagerland and Sandvik, 2009).

Calculation Procedure:

The data is ranked and the sum of the ranks is calculated. Then the test statistic, *H*, is calculated (hence this test is also called as Kruskal-Wallis *H* test). *H* is approximately chi-square distributed. Kruskal-Wallis test is not suitable if the sample size is small, say less than 5.

The formula for the calculation of chi-square value is given below (Equation 1):

$$X^2 = \frac{12 \times \left(\frac{r_1{}^2}{N_1} + \frac{r_2{}^2}{N_2} + \cdots\cdots + \frac{r_a{}^2}{N_a} \right)}{N(N-1)} - 3(N+1)$$

If the groups have data with same ranks, the chi-square value is calculated as given below (Equation 2):

$$X^2 = \frac{(N-1)S}{\left\{ \left(r_{11} - \frac{N+1}{2} \right)^2 + \cdots\cdots + \left(r_{ana} - \frac{N+1}{2} \right)^2 \right\}}$$

$$S = \frac{\left(\frac{r_1 - N_1(N+1)}{2} \right)^2}{N_1} + \frac{\left(\frac{r_2 - N_2(N+1)}{2} \right)^2}{N_2} + \cdots\cdots + \frac{\left(\frac{r_a - N_a(N+1)}{2} \right)^2}{N_a}$$

If the derived chi-square value is larger than the chi distribution Table value, then it indicates a significant difference.

Let us work out an example. Lymphocyte count determined in four groups in a clinical study is given in Table 12.16.

Table 12.16. Lymphocyte counts (%) determined in a clinical study

	Group A	Group B	Group C	Group D
	40.6	31.9	32.7	30.6
	38.0	36.8	31.3	35.9
	41.1	32.4	32.9	29.6
	52.7	34.8	31.9	29.2
	48.8	43.1	**28.5**	**28.5**
	41.1	39.0	31.2	30.8
	39.9	33.6	33.1	30.5
	43.1	34.3	34.1	29.4
	32.7	34.0	31.2	30.8
	30.1	33.8	31.7	32.0
Mean	40.8	35.4	31.9	30.7
N	10	10	10	10

Number group = 4; Total number of samples = 40.

Combine the lymphocytes counts of all the four groups, and arrange them from the smallest to the largest. Then assign a rank from 1 to 40 to them as given in Table 12.17. (Note: we have done a similar exercise while working out the example of scores for performance of six classes of two schools for explaining Wilcoxon rank-sum test; *vide* Tables 12.4 and 12.5).

Table 12.17. Ranks assigned to the lymphocyte counts (%) of four groups

	Group A	Group B	Group C	Group D
	34	15.5	19.5	8
	31	30	13	29
	35.5	18	21	5
	40	28	15.5	3
	39	37.5	**1.5**	**1.5**
	35.5	32	11.5	9.5
	33	23	22	7
	37.5	27	26	4
	19.5	25	11.5	9.5
	6	24	14	17
Mean rank	31.1	26	15.55	9.35

Equation 2 (page 117) is used to calculate the chi-square value.

Let us calculate r_1, $r_{2.}$ r_3 and r_4:

$$r_1 = 34+31+\cdots\cdots\cdots+19.5+6 = 311$$

$$r_2 = 15.5+30+\cdots\cdots\cdots+25+24=260$$

$$r_3 = 19.5+13+\cdots\cdots\cdots+11.5+14=155.5$$

$$r_4 = 8+29+\cdots\cdots\cdots+9.5+17=93.5$$

S is calculated as 2914.35 (see below):

$$S = \frac{\left(311-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(260-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(155.5-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(93.5-\frac{10\times 41}{2}\right)^2}{10} = 2914.35$$

X^2 is calculated as 21.3 (see below):

$$X^2 = \frac{(40-1)\times 2914.35}{\left(34-\frac{(40+1)}{2}\right)^2 + \left(31-\frac{(40+1)}{2}\right)^2 + \cdots + \left(9.5-\frac{(40+1)}{2}\right)^2 + \left(17-\frac{(40+1)}{2}\right)^2}$$

$$= \frac{113659.7}{5326.5} = 21.3$$

The computed X^2 value is compared with the X^2 Table value (Table 12.18) at 4–1=3 degrees freedom. Since the computed X^2 value (21.3) is greater than the X^2 Table value (16.266), it is considered that there is a significant difference in lymphocyte counts among the groups (P<0.001).

Table 12.18. Chi square Table (Yoshimura, 1987)

DF\α	0.1	0.05	0.01	0.001
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	**16.266**
4	7.779	9.488	13.277	18.467
5	9.236	11.070	15.086	20.515

Comparison of Group Means

Wilcoxon Rank-Sum test or Kruskal-Wallis test provides the information, whether a significant difference exists among the group means. If these tests reveal a significant difference, it does not indicate that every group means are significantly different from each other. One of the robust tests used to find out which group means are significantly different from each other is the Dunn's multiple comparison test. Dunn's multiple comparison test can be used to find the difference of 3 or more groups (Israel, 2008).

Dunn's multiple comparison test for more than three groups (Gad and Weil, 1986; Hollander and Wolf, 1973)

Let us review the example given in Table12.17. The mean rank values are reproduced in Table 12.19.

Table 12.19. Mean rank of lymphocyte (%)

	Group A	Group B	Group C	Group D	
Mean rank	31.1	26	15.6	9.4	
N	10	10	10	10	Sum=40

Calculation procedure

Group A** vs **Group B:

Difference of mean rank: 31.1–26=5.1

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

Group A vs Group C:

Difference of mean rank: 31.1–15.6=15.5

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

Group A vs Group D:

Difference of mean rank: 31.1-9.4=21.7

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

4 (3) = Number of group × Number of group – 1; The value 2.63 is obtained from Table 12.20 **(**the value, 0.00417 can be rounded to 0.0042. This value lies between 0.0043 and 0.0041 of *Z* value. In this case, 0.0043 was considered. The *Z* value corresponding to 0.0043 is 2.63).

The numerator (40) is total number of samples, (41) is total number of sample + 1; The denominator 12 is a constant, whereas 10 is number of samples in the groups.

Table 12.20. *Z* score for normal distribution (Gad and Weil, 1986)

Z	Proportional parts									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.6	0.0047	0.0045	0.0044	**0.0043**	**0.0041**	0.0040	0.0039	0.0038	0.0037	0.0036

The difference between the two mean scores is compared with the Probability (critical) value (13.7). If the difference between the two mean scores is greater than the Probability (critical) value, then the difference is considered significant (see below given Table 12.21).

Table 12.21. Significant difference between the groups

Analysis	Difference	Critical value	P
Group A *vs* Group B	31.1–26=**5.1**	13.7	Not significant (P>0.05)
Group A *vs* Group C	31.1–15.6=**15.5**		Significant (P<0.05)
Group A *vs* Group D	31.1–9.4=**21.7**		Significant (P<0.05)

Steel's multiple comparison test for more than three groups
(Steel, 1961)

The power of Steel's test is higher than the other multiple comparison tests. Usually the number of groups employed is four (three treatment groups + one control group) in most of the animal studies. For a parameter which shows a strong dose-related pattern, a significant difference can be detected by Steels's test, even if the number of animals in a group is as low as four (Yoshimura and Ohashi, 1992; Inaba, 1994). Let us work out an example (Table 12.22).

Calculation procedure:

Control group *vs* Low dose group

1) Sum of rank of low dose group, R_2=5+6+7+8=26

Table 12.22. Quantitative data from a toxicity study

Group	Control	Low dose	Mid dose	High dose
	1	5	9	13
	2	6	10	14
	3	7	11	15
	4	8	12	16
Mean rank	2.5	6.5	10.5	14.5

Note: Ranked values are given.

2) Calculation of SS(S_2) and Variance (V_2)
$S_2 = (1–4.5)^2 + (2–4.5)^2 + (3–4.5)^2 + (4–4.5)^2 + (5–4.5)^2 + (6–4.5)^2 + (7–4.5)^2 + (8–4.5)^2 = 42$, where

4.5 = Sum of number of samples of control group and number of samples of low dose group + 1 divided by number of groups [(4+4+1)/2]= 4.5).

$$V_2 = \frac{4\times 42}{4\times 8\times 7} = 0.75\text{, where}$$

4×42 = Number of sample in control group × S_2value, 42; 4×8×7 = Number of sample in low dose × Sum of number of samples of control and low dose groups ×Sum of number of samples of control and low dose groups – 1.

3) Calculation of t_2

$$t_2 = \frac{\frac{26}{4} - \frac{4+4+1}{2}}{\sqrt{0.75}} = \frac{2}{0.866} = 2.309\text{, where}$$

26/4 =R_2/4 (4=Number of sample in low dose), (4+4+1)/2 =(Number of samples in control group + Number of samples in low dose group + 1)/2; 0.75 = V_2.

4) Calculated t_2 value, 2.309 is compared with the critical value given in Table 12.23. As the size of each group is similar, the critical value becomes (∞, 4) =2.062.

5) Since computed t_2 value, 2.309 is greater than the Table value,2.062, it is considered that the low dose group is significantly different from the control.

Table 12.23. Dunnett's *t* test critical values, one-sided at 0.05 probability level (Yoshimura, 1987)

Number of group	2	3	4	5	6	7	8
∞	1.645	1.916	**2.062**	2.160	2.234	2.292	2.340

Using the calculation procedure mentioned above for comparing control group *vs* low dose group, comparison between other groups (control group *vs* mid dose group and control group *vs* high dose group) can be made.

Rank Sum Tests—Some Points

An interesting example of a rank sum test analysis is given in Table 12.24. Creatinine value of F344 rats on week 52 in a repeated dose administration study is given in the Table.

Table 12.24. Creatinine value (mg/dl) of F344 rats at 52 weeks after dosing

Group	Individual value (20 animals/group)	Mean ± SD
Control	0.70 0.68 0.70 0.74 0.60 0.65 0.65 0.72 0.63 0.78 0.67 0.64 0.63 0.66 0.88 0.73 0.57 0.79 0.78 0.65	0.69 ± 0.07
Low dose	0.72 0.64 0.66 0.66 0.88 0.68 dead 0.51 0.65 0.63 0.79 0.60 0.69 0.68 0.62 0.57 dead 0.66 0.59 0.54	0.65 ± 0.09
Middle dose	0.56 0.59 0.66 0.68 0.57 0.67 0.70 0.83 0.86 0.68 0.60 0.68 0.57 0.67 0.53 0.57 0.64 0.61 0.86 0.67	0.66 ± 0.10
High dose	0.51 0.59 0.49 0.60 0.58 0.62 0.51 0.57 0.60 2.96 0.56 0.65 0.71 0.55 0.54 0.41 0.52 0.62 0.59 0.59	0.69 ± 0.54**

**Significantly different from control by rank sum test (P<0.01).

Bartlett's test for homogeneity of variance showed a significant difference, therefore Dunnett type rank test was used for the analysis of the data. The Dunnett type rank test revealed a significant difference between the high dose group and the control group (P<0.01), though the mean values of these groups are the same (0.69). Close examination of the individual values of the high dose group revealed that one of the values among them (2.96) is extremely high compared with the other values. If a number slightly higher than 0.88, which is the next highest value among the high dose and control groups, replaces 2.96 of the high dose group, the mean value of this group becomes lower than that in the control group, but the rank is not changed, *i.e.,* the result of the rank sum test will not be changed. Thus, the significant difference between the control group and high dose group detected by the rank sum test is understandable, though the mean values of these groups are the same.

Another important point in rank sum test analysis is that one should know the minimum number of animals required in each group to detect a significant difference. Table 12.25 shows the minimum number of animals required in four-group and five-group settings to detect a significant difference.

Table 12.25. Minimum number of animals in four-group and five-group settings necessary to show a significant difference

Test	Four groups	Five groups
Scheffé type	22	40
Hollander-Wolfe*	19	30
Tukey type	18	32
Dunnett type	15	26
Wilcoxon	8	12
Steel type	4	6
Mann-Whitney *U***	3	

*Dunn's test. **Test for 2 group alone.

The power also depends on the number of treatment groups, which implies that inclusion of further non-significant treatment group/s can result in overlooking significant effects (Hothorn, 1990).

As mentioned earlier, the power to detect a significant difference is high with Steel's test. A comparison of the power to detect a significant difference between Dunnett type rank test and Steel's test is given in Table 12.26.

Table 12.26. Comparison of the power to detect a significant difference between Dunnett type rank test and Steel's test

Parameter analysed and tests	Control (N=5)	Low dose (N=5)	Mid dose (N=5)	High dose (N=4)	Top dose (N=4)
Urine volume (ml)	2.4, 2.8, 2.4, 2.4, 2.4	43, 45, 40, 41, 46	62, 48, 68, 52, 55	73, 72, 102, 104	52, 97, 99, 103
Mean ± SD	2.5 ± 0.18	43 ± 2.55	57 ± 8.0	87.8 ± 17.6	87.8 ± 24
Bartlett's homogeneity test	P = 0.0001				
Kruskal-Wallis's test	P = 0.0006				
Dunnett type rank test		NS	S	S	S
Steel's test		S	S	S	S

NS-Not significant (P>0.05); S-Significant (P<0.05)

The low dose group was not significantly different, when analysed using Dunnett type rank test, whereas, this dose group was significantly different, when analysed using Steel's test.

Most of the pharmacologists and toxicologists express their concern about use of non-parametric tests like rank sum tests, because of their low sensitivity in detecting a significant difference. However, some biostatisticians are of the opinion that the rank sum tests are more useful for analyzing the biological data than the parametric tests.

References

Clarke, S.C. (1991): Invited commentary on R. A. Fisher. Am. J. Epidemiol., 134(12), 1371–1374.

Crawley, M.J. (2005): Statistics: An Introduction Using R. John Wiley and Sons Ltd., Chichester, UK.

Elston, R.C. and Johnson, W.D. (1994): Essentials of Biostatistics. F.A. Davis & Co., Philadelphia, USA.

Fagerland, M.W. and Sandvik, L. (2009): The Wilcoxon-Mann-Whitney test under scrutiny. Statist. Med., 28, 1487–1497.

Fisher, R.A. (1922): On the interpretation of χ^2 from contingency tables, and the calculation of *P*. J. Royal Stat. Soc., 85(1), 87–94.

Fisher, R.A. (1954): Statistical Methods for Research Workers. Oliver and Boyd, London, UK.

Gad, S. and Weil, C.S. (1986): Statistics and Experimental Design for Toxicologists. The Telford Press, New Jersey, USA.

Hollander, M. and Wolf, D.A. (1973): Non-Parametric Statistical Methods.John Wiley, New York, USA.

Hothorn, L. (1990): Biometrische Analyse spezieller Untersuchungen der regulatorischen Toxikologie. In: Aktuelle Probleme der Tbxikologie, Vol. 5 Grundlagen der Statistik fuer Toxikologen (M. Horn and L. Hothorn, Eds.) Verlag Gesundheit Gmbh, Berlin, Germany.

Inaba, T. (1994): Problem of multiple comparison method used to evaluate medicine of enzyme inhibitor X1, Japanese Society for Biopharmaceutical Statistic, 40, 33–36.

Israel, D. (2008): Data Analysis in Business Research-A Step by Step Non-Parametric Approach. SAGE Publications India Pvt. Ltd., New Delhi, India.

Kruskal, W.H. and Wallis, A.W (1952): Use of ranks in one criterion analysis of variance. J. Am. Stat. Assoc., 47(260), 583–621.

Le, C.T. (2003): Introductory Biostatistics. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Ludbrook, J. (2008): Analysis of 2 × 2 tables of frequencies: Matching test to experimental design. Int. J. Epidemiol., 37(6), 1430–1435.

Mann, H.B. and Whitney, D.R. (1947): On a test of whether one of 2 random variables is stochastically larger than the other. Ann. Math. Stat., 18, 50–60.

Mc Donald, J.H. 2009: Handbook of Biological Statistics, 2nd Edition. Sparky House Publishing, Baltimore, USA.

Mc Kight, P.E. and Najab, J. (2010): Kruskal-Wallis Test. In: Corsini Encyclopedia of Psychology. Editors, Weiner, I.B. and Craighead, W.E., Wiley Online Library, DOI: 10.1002/9780470479216.

Mc Kinney, W.P., Young, M.J., Hartz, A. and Lee, M.B. (1989): The inexact use of Fisher's exact test in six major medical journals. JAMA, 16, 261(23), 3430–3433.

Nietert, P.J. and Dooley, M.J. (2011): The power of the sign test given uncertainty in the proportion of tied observations, 32(1), 147–150.

Sawilowsky, S. (2005): Encyclopedia of Statistics in Behavioral Science. Wiley Online Library, DOI: 10.1002/0470013192.bsa615.

Steel, R.G.D. (1961): Some rank sum multiple comparison tests. Biometrics, 17(4), 539–552.

Surhone, L.M., Timpledon, M.T. and Marseken, S.F. (2010): Sign Test. VDM Verlag Dr Mueller AG&Co., KG, Germany.

Whitley, E. and Ball, J. (2002): Statistics review 6: Nonparametric methods, Crit. Care, 6(6), 509–513.

Wilcoxan, F. (1945): Individual comparisons by ranking methods. Biometrics Bull., 1(6), 80–83.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Press, Tokyo, Japan.

Yoshimura, I. and Ohashi, Y. (1992): Statistical Analysis for Toxicology Data. Chijin-Shokan, Tokyo, Japan.

13
Cluster Analysis

What is Cluster Analysis?

Cluster analysis is used to classify observations into a finite and small number of groups based upon two or more variables (Finch, 2005). The term cluster analysis was first used in 1939 by Tryon (Tryon,1939). 'Numerical taxonomy' is another term used for cluster analysis in some areas of biology (Romesburg, 2004). There is no *a priori* hypothesis in cluster analysis, unlike other statistical analysis. In cluster analysis the variables are arranged in a natural system of groups (Kirkwood, 1989). The heterogeneous data collected are sorted into series of sets. Data in a cluster are considered to be 'similar' or highly correlated to each other. Clusters can be exclusive (a particular variable is included in only one cluster) and overlapping (a particular variable is included in more than one cluster). Cluster analysis method is used in a variety of research problems (Hartigan, 1975; Scoltock, 1982; Moore *et al.*, 2010). It is applied extensively in the fields of toxicogenomics (Hamadeh *et al*., 2002), genetics (Shannon *et al*., 2003; Makretsov *et al.*, 2004) and molecular biology (Furlan *et al.*, 2011). Cluster analysis only discovers structures in data, but does not explain why such structures exist.

Cluster analysis can be carried out using several methods. Three commonly used methods are described below:

Hierarchical cluster analysis

As the name indicates, hierarchical cluster analysis produces a hierarchy of clusters. The clusters thus produced are graphically presented. This graphical output is known as a dendrogram (from Greek *dendron* 'tree', *gramma* 'drawing'). The dendrogram can be used to examine how clusters are formed in hierarchical cluster analysis (Schonlau, 2002). Hierarchical clustering can be of two types. One type is agglomerative clustering, where grouping of clusters is done small clusters to large ones. The other type is divisive clustering, where grouping of clusters is done large clusters to small ones. For illustrative purpose a dendrogram is given in Figure 13.1.

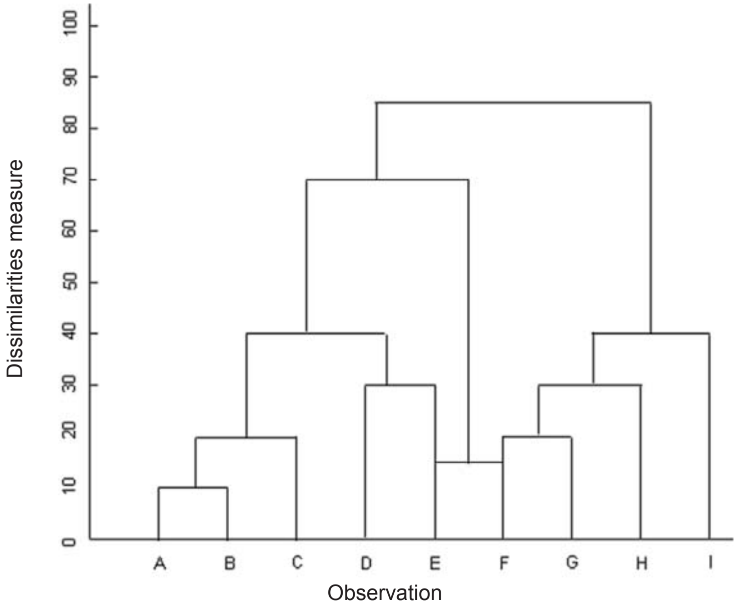


Figure 13.1. Dendrogram

The individual observations (A–I) are arranged evenly along the X axis of the dendrogram. They are called as leaf nodes. The vertical axis indicates a distance or dissimilarity measure. The height of a leaf node represents the distance of the two clusters that the node joins. In this dendrogram, the similarity of samples A and B is better than the other samples, and the first cluster is formed by these two samples.

Ward's method of cluster analysis (Ward, 1963; Ward and Hook, 1963)

This method is more efficient than hierarchical cluster analysis. Ward's method uses the squared distances between-clusters and within-clusters (Rencher, 2002). Hence, Ward's method is also called as the 'incremental sum of squares' method.

k-means cluster analysis

This method of clustering is used when *a priori* hypothesis concerning the number of clusters in variables are available. *k* is the number of clusters that we desire.

Data collected in repeated dose administration toxicity studies is enormous and are either qualitative or quantitative in nature. No observed adverse effect level (NOAEL) of the test substance is judged based on these data. Sometimes the toxicity effects manifested are not dose-dependent, which makes judging an NOAEL difficult. In such situations, cluster analysis is extremely useful for judging an NOAEL. Now the question is whether to consider only those data which show a significant difference compared to control for the cluster analysis or all data collected in the study, irrespective of their difference from the control is significant or not.

We shall try to understand cluster analysis with the help of an example. Groups (10/sex/dose) of seven-week-old Crj: CD rats were administered the test substance at low, mid, high and top doses by gastric intubation daily for 28 days. A concurrent control group was also maintained. Rats were daily examined for general behavior. During the dosing period, body weight, food and water consumption of the animals were measured. Animals were sacrificed on day 29 after overnight starvation for assessment of hematology, blood biochemistry, serum protein electrophoresis, urinalysis, myelogram and ophthalmologic and pathological (organ weight measurement and gross and histopathology) examinations (Kobayashi, 2004).

Salivation in both sexes in the high dose group, staggering gait in the top dose group, slight suppression of the body weight gain in males in the top dose group, slight anemic trend in both sexes in the top dose group, higher values in alkaline phosphatase in both sexes in the high dose and top dose groups, lower values in albumin in males in the top dose group and in females in the high dose and top dose groups, bone fractures, mobilization of the sinusoidal cell and extramedullary hematopoiesis in the liver in both sexes in the top dose group and squamous hyperplasia, and erosion of the fore-stomach in both sexes in the high and top dose groups were observed as the main changes attributable to the repeated oral administration of the test substance. Based on above observations and determinations, the NOAEL was considered to be the mid dose for both males and females.

The data obtained in the study was analyzed statistically. Continuous data was subjected to Bartlett's test for examining homogeneity of variance and was analysed (two-sided analysis) using the statistical techniques as given in the decision tree proposed by Kobayashi *et al*. (2000) (Figure 13.2). Gross and histopathological findings were analyzed by the Fisher's exact test (Gad and Weil, 1986). The level of significance for the above mentioned statistical analysis was set at P<0.05.

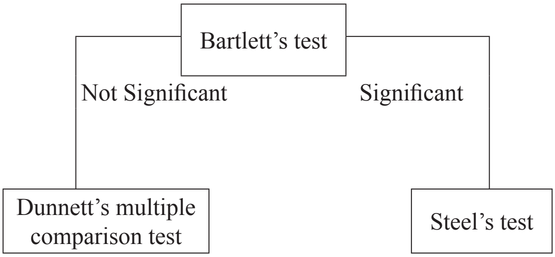


Figure 13.2. Analytical methods by a decision tree

We shall analyse the data of the study described above using Ward's method of cluster analysis (Milligan 1980). The software used for the analysis was JMP (version 5) of the SAS (SAS Institute, Japan).

Cluster-1

The items in the dosed groups that showed a significant difference compared to the control group were—body weight gain, food efficiency, hematocrit, hemoglobin, red blood cell count, platelet count, neutrophil (%), lymphocytes (%), blood urea nitrogen, total protein, alanine aminotranferase, alkaline phosphatase, glucose, prothrombin time, albumin, albumin/globulin ratio, inorganic phosphorus in urine, lung weight, relative weights of the lung, liver, kidneys and testes, gross pathology findings, and microscopic findings. These items were grouped in Cluster 1.

Each dosed group was divided into Group 1 and Group 2. Group 1 was further divided into Subgroup 1 and Subgroup 2 (Table 13.1).

Table 13.1. Results of cluster analysis: Cluster 1—Items showing a significant difference (P<0.05) compared to control

Dose Group	Number of animals		
	Group 1		Group 2
	Subgroup-1	Subgroup-2	
Control	10	0	0
Low	10	0	0
Mid	10	0	0
High	2	8	0
Top	0	4	6

The dendrogram obtained from the above data is given in Figure 13.3.

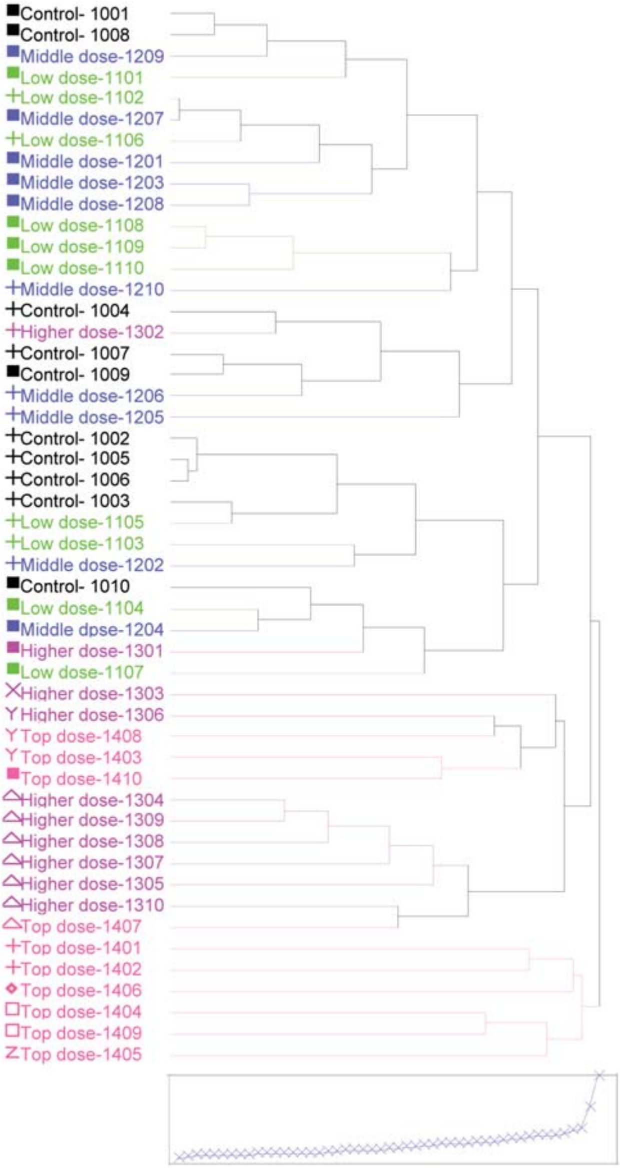


Figure 13.3. Dendrogram of items that are significantly different from control (Ward's method)

Note: Animal identification mark, dose group and animal number are given on the left side of the dendrogram.

Cluster 2

The items which did not show a significant difference compared to control were—food and water consumption, leucocyte count, lymphocyte count, reticulocyte count, activated partial thromboplastin time, total cholesterol, free cholesterol, triglyceride, phospholipid, non esterified fatty acid, creatinine, total bilirubin, sodium, potassium, chloride, calcium, inorganic phosphorus, alanine aminotransferase, lactate dehydrogenase, alpha-1 (%), gamma (%), urine volume, urine specific gravity, and sodium, potassium, chloride, calcium and inorganic phosphorus in urine, and weights of the brain, heart, liver, kidneys, spleen, adrenals, testes, thyroid and thymus, and relative weights of the brain, heart, spleen, adrenals, thyroid and thymus. These items were grouped in Cluster 2.

Each dosed group was divided into Group 1 and Group 2. Groups 1 and 2 were further divided into two Subgroups each (Table 13.2).

Table 13.2. Results of cluster analysis: Cluster 2—Items showing no significant difference (P>0.05) compared to control

Dose group	Number of animal			
	Group 1		Group 2	
	Subgroup-1	Subgroup-2	Subgroup-1	Subgroup-2
Control	8	2	0	0
Low	6	4	0	0
Mid	7	3	0	0
High	5	0	5	0
Top	0	0	5	5

The dendrogram obtained from the above data is given in Figure 13.4.

As you would have observed from the dendrograms, when the number of observations are more, it is very difficult to distinguish each observation. Dendrograms are only suitable for hierarchical cluster analysis. Schonlau (2002) proposed a clustergram, which is suitable for non-hierarchical cluster analysis. For hierarchical cluster analysis, a radial clustergram was proposed by Agrafiotis *et al*. (2007). In radial clustergram, clusters are arranged into a series of layers, each representing a different level of the tree. However, for small set of data, a dendrogram is still preferable to a clustergram.

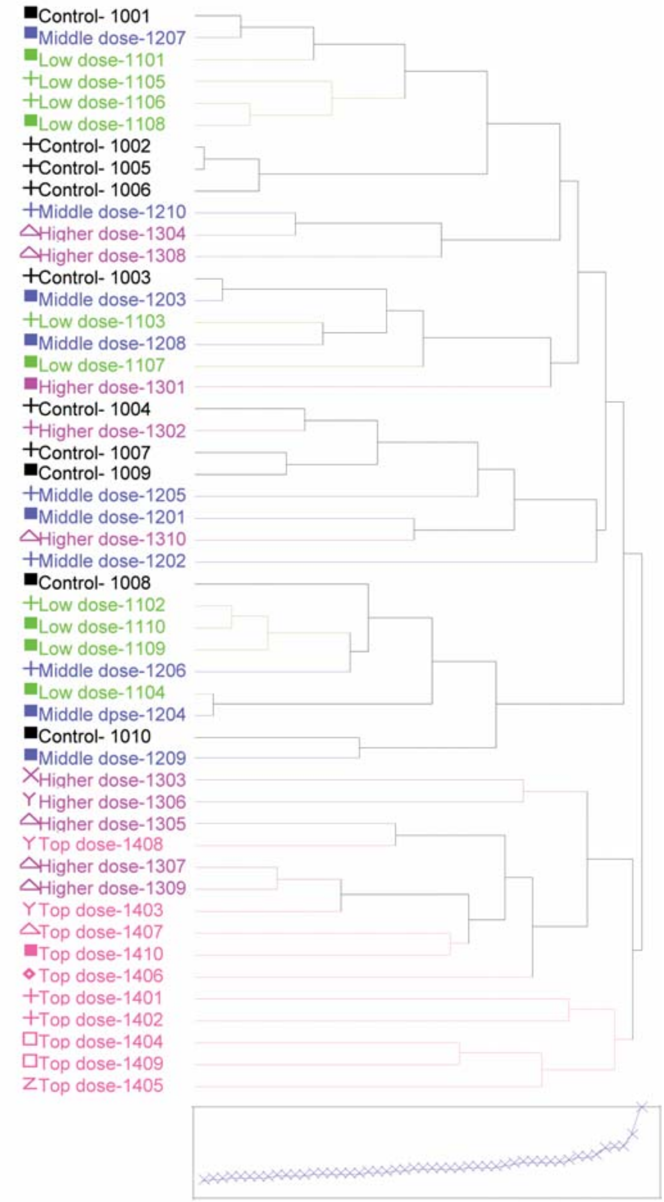


Figure 13.4. Dendrogram of items that are not significantly different from control (Ward's method)

Note: Animal identification mark, dose group and animal number are given on the left side of the dendrogram.

References

Agrafiotis, D.K., Bandyopadhyay, D. and Farnum, M. (2007): Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. J. Chem. Inf. Model, 47, 69–75.

Finch, H. (2005): Comparison of distance measures in cluster analysis with dichotomous data. J. Data Sci., 3, 85–100.

Furlan, D., Carnevali, I.W., Bernasconi, B., Sahnane, N., Milani, K., Cerutti, R., Bertolini, V., Chiaravalli, A.M., Bertoni, F., Kwee, I., Pastorino, R. and Carlo, C. (2011): Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. Modern Path., 24, 126–137.

Gad, S. and Weil, C.S. (1986): Statistics and Experimental Design for Toxicologists, The Telford Press, New Jersey, USA.

Hamadeh, H.K., Bushel, P.R., Jayadev, S., DiSorbo, O., Bennett, L., Li, L., Tennant, R., Stoll, R., Barrett, C., Paules, R.S., Blanchard, K. and Afshari, C.A. (2002): Prediction of compound signature using high density gene expression profiling. Toxicol. Sci., 67, 232–240.

Hartigan, J.A. (1975): Clustering Algorithms. John Wiley & Sons, Inc., New York, USA.

Kikwood, B. (1989): Medical Statistics, Blackwell Scientific Publications, London, UK.

Kobayashi, K. (2004): Evaluation of toxicity dose levels by cluster analysis. J. Toxicol. Sci., 29(2), 125–129.

Kobayashi, K., Kanamori, M., Ohori, K. and Takeuchi, H. (2000): A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies on rodent. San Ei Shi, 42, 125–129.

Makretsov, N.A., Huntsman, D.G., Nielsen, T.O., Yorida, E., Peacock, M., Cheang, M.C.U., Dunn, S.E., Hayes, M., van de Rijn, M., Bajdik, C. and Gilks, C.B. (2004): Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. Clin. Cancer Res., 10, 6143–6151.

Milligan, G.W. (1980): An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325–342.

Moore, C.W., Meyers, D.A., Wenzel, S.E., Teague, G.W., Li, H., Li, X., D'Agostino, Jr., R., Castro, M., Curran-Everett, D., Fitzpatrick, A.M., Gaston, B., Jarjour, N.N., Sorkness, R., Calhoun, W.J., Chung, K.F., Comhair, S.A.A., Dweik, R.A., Israel, E., Peters, S.P., Busse, W.W., Erzurum, S.C. and Bleecker, E.R. (2010): Identification of asthma phenotypes using cluster analysis in the severe asthma research program. Am. J. Resp. Crit. Care Med., 181, 315–323.

Rencher, A.C. (2002): Methods of Multivariate Analysis. 2nd Edition, Wiley-Interscience, New York, USA.

Romesburg, H. (2004): Cluster Analysis for Researchers. Lulu Press, North Carolina, USA.

Schonlau, M. (2002): The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses. The Stata J., 3, 316–327.

Scoltock, J. (1982): A survey of the literature of cluster analysis. Computer J., 25(1), 130–134.

Shannon, W., Culverhouse, R. and Duncan, J. (2003): Analyzing microarray data using cluster analysis. Pharmacogenomics, 4(1), 41–52.

Tryon, R.C. (1939): Cluster Analysis. Edward Brothers, Ann. Arbor., MI, USA.

Ward, J.H., Jr. (1963): Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc., 58(301), 235–244.

Ward, J.H., Jr. and Hook, M.E. (1963): Application of an hierarchical grouping procedure to a problem of grouping profiles. Edu. Psych. Measurement, 23, 69–81.

14
Trend Tests

Introduction

In pharmacology and toxicology experiments three or more than three treatment groups are usually used. One of the objectives for carrying out the experiment with three or more than three groups is to assess the dose-dependency of the test substance. Dose-dependency is an important concept for evaluating toxicological data (Hamada *et al.*, 1997). In order to examine whether the change in a parameter observed in a study is dose-dependent, a trend test is used. A trend test examines whether the results in all dose groups together increase as the dose increases (EPA, 2005). Trend tests have been recommended as a customary method for analyzing data from subchronic and chronic animal studies (Selwyn, 1995). For examining quantitative data, Jonckheere's trend test (Jonckheere, 1954) is generally used. The frequency data are examined by Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955).

Jonckheere's trend test

Jonckheere's test is a frequently used nonparametric trend test for the evaluation of preclinical studies and clinical dose-finding trials (Neuhäuser *et al.*, 1999). Predicted trend can be evaluated using this test (Cohen and Holliday, 2001). Since it does not require specification of a covariate, it has generated a continued interest (Jones, 2001). Jonckheere's test is based on the idea of taking a score in a particular condition and counting how many scores in subsequent conditions are smaller than that score (Field, 2004). In order to use the Jonckheere's test, the number of groups should be 3 or more than 3 and each group should have equal number of observations.

Water consumption of B6C3F1 mice fed on a diet containing a test substance at week eight is given in Table 14.1. There are three dose groups and one control group. Let us examine whether there is a trend in the water consumption across the groups.

Table 14.1. Water consumption (g/week) of B6C3F1 mice fed on a diet containing a test substance at week eight

Group	Group 1 (Control)	Group 2 (Low dose)	Group 3 (Mid dose)	Group 4 (High dose)
	40.6	31.9	32.7	30.6
	38.0	36.8	31.3	35.9
	41.1	32.4	32.9	29.6
	52.7	34.8	31.9	29.2
	48.8	43.1	28.5	28.5
	41.1	39.0	31.2	30.8
	39.9	33.6	33.1	30.5
	43.1	34.3	34.1	29.4
	32.7	34.0	31.2	30.8
	30.1	33.8	31.7	32.0
Mean	40.8	35.4	31.9	30.7
SD	6.7	3.4	1.5	2.1
N	10	10	10	10

Formula:

$$J = \frac{\left[\sum ijTij + \frac{\sum ijSij}{2} - \frac{N^2 \sum iN_1{}^2}{4}\right] - 0.5}{\sqrt{V}}$$

$$V = \frac{N(N-1)(2N+5) - \sum iN_1(N_i-1)(2Ni+5)}{72}$$

$$+ \frac{\left\{\sum iNi(ni-1)(Ni-2)\right\}\left\{\sum i\tau(\tau i-1)(\tau i-2)\right\}}{36N(N-1)(N-2)}$$

$$+ \frac{\left\{\sum iNi(Ni-1)\ \right\}\left\{\sum i\tau i(\tau i-1)\right\}}{8N(N-1)}$$

If the computed J value is greater than the Z value given in the standard normal distribution Table, it is considered to be significantly different.

Calculation of T values:

We need this information for the calculation of J. Arrange the data in each group in the order of prediction. Let us calculate T_{12} (Control Group *vs* Low Dose Group). For each control value the number of values that are lesser than it in the low dose group are counted, and their total is calculated:

$T_{12} = 9+ 8+ 9+10+10+9+ 9+ 9+ 2+ 0 = 75$

The first value of the control group is 40.6 and there are 9 values of the low dose group, which are lesser than 40.6. The second value of the control group is 38.0 and there are 8 values of the low dose group, which are lesser than 38.0, and so on.

Similarly, values are counted for other trends.

$T_{13} =10+10+10+10+10+10+10+10+ 6+ 1 = 87$

$T_{14} =10+10+10+10+10+10+10+10+ 9+ 4 = 93$

$T_{23} = 5+10+ 6+10+10+10+ 9+10+ 9+ 9 = 88$

$T_{24} = 8+10+ 9+ 9+10+10+ 9+ 9+ 9+ 9 = 92$

$T_{34} = 9+ 8+ 9+ 8+ 0+ 8+ 9+ 9+ 8+ 8 = 76$

where, T_{13} is Control Group *vs* Mid Dose Group, T_{14} is Control Group *vs* High Dose Group, T_{23} is Low Dose Group *vs* Mid Dose Group, T_{24} is Low Dose Group *vs* High Dose Group and T_{34} is Mid Dose Group *vs* High Dose Group.

τ Values: We also need to know how many times a value repeated within a group and across the groups. 43.1 is repeated twice-one each in Groups 1 and 2 (τ_1), 41.1is repeated twice within the Group 1 (τ_2), 32.7 is repeated twice-one each in Groups 1 and 3 (τ_3), 31.9 is repeated twice-one each in Groups 2 and 3 (τ_4), 30.8 is repeated twice within Group 4 (τ_5), 31.2 is repeated twice within Group 3 (τ_6)and 28.5 is repeated twice-one each in Groups 3 and 4 (τ_7).

$$V = \frac{40(40-1)(2\times 40+5)-10(10-1)(20+5)\times 4}{72}$$

$$+\frac{4\times 10(10-1)(10-2)\times\{\,2(2-1)(2-2)+2(2-1)(2-2)+2(2-1)(2-2)+2(2-1)(2-2)+2(2-1)(2-2)+ 2(2-1)(2-2)+2(2-1)(2-2)\,\}}{36\times(40-1)(40-2)}$$

$$+\frac{10(10-1)\times 4\{2(2-1)+2(2-1)+2(2-1)+2(2-1)+2(2-1)+2(2-1)+2(2-1)\}}{8\times 40(40-1)}=1717.003$$

$$J = \frac{\left\{75+87+93+88+92+76+\frac{1+1+0+1+0+1}{2}-\frac{40^2-10^2\times 4}{4}\right\}-0.5}{\sqrt{1717.003}} = \frac{212.5}{41.4} = 5.13$$

Note: 40 is total number of observations, 10 is number of observations in each group, 4 is total number of groups and the denominators 2 and 4, and 0.5 are the constants.

1+1+0+1+0+1= S_{12}+S_{13}+S_{14}+S_{23}+S_{24}+S_{34}; Number of values repeated across the groups (not within the groups)—the value 43.1 repeated in Groups 1 and 2 (S_{12}=1), 32.7 is repeated in Groups 1 and 3 (S_{13}=1), no value is repeated in Groups 1 and 4 (S_{14}=0), 31.9 is repeated in Groups 2 and 3 (S_{23}=1), no value is repeated in Groups 2 and 4 (S_{24}=0), and 28.5 is repeated in Groups 3 and 4 (S_{34}=1).

Computed value for *J*=5.13 is greater than the point and (α) = 3.290 (Table 14.2). Therefore, it could be stated that there is a dose-related trend in the decrease of water consumption of B6C3F1 mice fed on diet containing the test substance at week eight.

Table 14.2. Standard normal distribution Table (Yoshimura, 1987)

Two tailed *P*	Upper *P*	% point
2α	α	U(α)
0.00100	0.000500	**3.290527**
0.00200	0.0010000	3.090232

The Cochran-Armitage test

The Cochran-Armitage trend test is commonly used to examine whether a dose-response relationship exists in toxicological risk assessment, carcinogenicity studies and several other biomedical experiments (Mehta *et al*., 1998) including mutagenicity studies (Kim *et al*., 2000). It is also widely used in genetics and epidemiology to test linear trend (Buonaccorsi *et al*., 2011). The Cochran-Armitage test for trend is used in categorical data analysis. It can be used to test for linear correlation between a binomial response and an ordinal group variable (Walker and Shostak, 2010). In 1985, the US Federal Register recommended that the analysis of tumour incidence data is carried out with a Cochran-Armitage's trend test (Gad, 2009).

The presence of the antibody to the house dust was investigated in individuals of different age groups (see Table 14.3). Let us examine whether there is a tendency to increase the antibodies to the house dust as the age of the individuals increases.

Table 14.3. Individuals of different age groups expressing antibodies to house dust

Age	Conversion value	Independent variable (log transformed)	Number of investigations	Number of antibody positives
One's sixties	2.5	0.398	10	2
One's fifties	5	0.699	10	4
One's forties	10	1.000	10	6
One's thirties	20	1.301	10	8

A value of 10 is assigned to the age forties. Half of the value of the age forties (10/2=5) is assigned to the age fifties and half of the value of age fifties (5/2=2.5) is assigned to the age sixties. The value assigned for the age thirties is 20 (10x2).

Number of group = 4, Sum of number of sample = 40, rate of positive in total = (2+4+6+8)/40= 20/40= 0.5

$$\text{Mean} = \frac{(10\times 0.398+10\times 0.699+10\times 1.000+10\times 1.301}{40} = 0.8495$$

$$X^2 = \frac{\left\{(2\times 0.398+4\times 0.699+6\times 1.000+8\times 1.301)-40\times 0.5\times 0.8495\right\}^2}{0.5\times(1-0.5)\times 10\times\left\{(0.398-0.8495)^2+(0.699-0.8495)^2+(1.000-0.8495)^2+(1.301-0.8495)^2\right\}}$$

$$=\frac{9.0601}{1.1325}=8.000$$

From the chi-square Table (Table 14.4), for one degree of freedom, we find that the calculated value (8.000) is greater than the chi-square Table value (6.635) at 0.01 probability level. Hence, we conclude that there is a tendency to increase the antibodies to the house dust as the age of the person increases.

Table 14.4. Chi-square **(**Yoshimura, 1987)

DF	α			
	0.100	0.050	0.010	0.001
1	2.705	**3.841**	**6.635**	10.82
2	4.605	5.991	9.210	13.81
3	6.251	7.814	11.34	16.26
4	7.779	9.487	13.27	18.46
5	9.236	11.07	15.08	20.51
6	10.64	12.59	16.81	22.45
7	12.01	14.06	18.47	24.32
8	13.36	15.50	20.09	26.12
9	14.68	16.91	21.66	27.87
10	15.98	18.30	23.20	29.58

Armitage (1955) recommended the Cochran-Armitage test in case there is no *a priori* knowledge of the type of the trend. The Cochran-Armitage test is asymptotically efficient for all monotone alternatives (Tarone and Gart, 1980). But, this test should not be used for the data showing an extra-Poisson variability (Astuti and Yanagawa, 2002), where estimated variance exceeds estimated means. Antonello *et al.* (1993) stated that Tukey trend test is more powerful for monotonic dose-response toxicologic effects than the pair-wise comparison tests. But dichotomous endpoints are frequently observed in several toxicologic effects. For analysing dichotomous endpoints, Neuhauser and Hothorn (1997) proposed a trend test analogous to the nonparametric Jonckheere's trend test.

We propose Jonckheere's trend test for the analysis of quantitative data, such as body weight, erythrocyte count, alkaline phosphatase and organ weights. For qualitative data, such as a macroscopic-, microscopic-pathological findings and urinalysis (color, pH, protein, glucose, ketone, bilirubin and urobilinogen) we propose Cochran-Armitage test.

References

Antonello, J.M., Clark, R.L. and Heyse, J.F. (1993): Application of the Tukey trend test procedure to assess developmental and reproductive toxicity I. Measurement data. Tox. Sci., 21(1), 52–58.

Armitage, P. (1955): Tests for linear trends in proportions and frequencies. Biometrics, 11 (3), 375–386.

Astuti, E.T. and Yanagawa, T. (2002): Testing trend for count data with extra-Poisson variability. Biometrics, 58(2), 398–402.

Buonaccorsi, J.P., Laake, P. and Veierød, M.B. (2011): On the power of the Cochran-Armitage test for trend in the presence of misclassification. Stat. Methods Med. Res., August 2011; doi:10.1177/0962280211406424.

Cochran, W.G. (1954): Some methods for strengthening the common chi-square tests. Biometrics, 10(4), 417–451.

Cohen, L. and Holliday, M. (2001): Practical Statistics for Students. SAGE Publications Inc., California, USA.

EPA (2005): United States Environmental Protection Agency. Guidelines for Carcinogen Risk Assessment. U.S. Environmental Protection Agency, EPA/630/P-03/001F. USEPA, Washington D.C., USA.

Field, A.P. (2004): Discovering Statistics Using SPSS, 2nd Edition, SAGE, London, UK.

Gad, S.C. (2009): Drug Safety Evaluation, 2nd Edition. John Wiley & Sons, Inc., New Jersey.

Hamada, C., Yoshino, K., Matsumoto, K., Ikumi Abe, I., Yoshimura, I. and Nomura, M. (1997): A study on the consistency between statistical evaluation and toxicological judgment. Drug Inf. J., 31, 413–421.

Jonckheere A.R. (1954): A distribution-free *k*-sample test against ordered alternatives. Biometrika, 41, 133–145.

Jones, M.P. (2001): Unmasking the trend sought by Jonckheere-type tests for right censored data. Scand. J. Stat., 28(3), 527–535.

Kim, B.S., Zhao, B., Kim, H.J. and Cho, M.H. (2000): The statistical analysis of the *in vitro* chromosome aberration assay using Chinese hamster ovary cells. Mut. Res., 469(2), 243–252.

Mehta, C.R., Patel, N.R. and Senchaudhuri, P. (1988): Exact power and sample size computations for the Cochran-Armitage trend test. Biometrics, 54, 1615–1621.

Neuhauser, M. and Hothorn, L.A. (1997): Trend tests for dichotomous end points with application to carcinogenicity studies. Drug Inf. J., 31, 463–469.

Neuhäuser, M., Liu, P.Y. and Hothorn, L.A. (1999): Nonparametric tests for trend: Jonckheere's test, a modification and a maximum test. Biometrical J., 40(8), 899–909.

Selwyn, M.R. (1995): The use of trend tests to determine a no-observable-effect level in animal safety studies. Int. J. Toxicol., 14(2), 158–168.

Tarone, R.E. and Gart, J.J. (1980): On the robustness of combined tests for trends in proportions. J. Am. Stat. Assoc., 75, 110–116.

Walker, G.A. and Shostak, J. (2010): Common Statistical Methods for Clinical Research with SAS Examples, 3rd Edition. SAS Inst. Inc., North Carolina.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data, Scientist Press, Tokyo, Japan.

15
Survival Analysis

Introduction

Survival analysis is one of the oldest fields of statistics, going back to the 17th century. The first life-table was presented by John Graunt in 1662 (Kreager, 1988). Life-tables are used extensively in analysing the mortality data obtained from toxicology studies, especially carcinogenicity and long-term repeated dose administration studies (Portier, 1988; FDA, 2007) and ecotoxicology studies (Gentile *et al.*, 1982; Van Leeuwen *et al.*, 1985; Bechmann, 1994). A major advancement in the survival analysis took place in 1958, when Kaplan and Meier proposed their 'estimator of the survival curve' (Kaplan and Meier, 1958). Since then, the field of survival analysis progressed significantly with the contributions from several statisticians (Mantel and Haenszel, 1959; Cox, 1972; Aalen, 1976; Aalen, 1980; Diggle, *et al.*, 2007; Aalen *et al.*, 2008). The term "survival" is a bit misleading. Originally the analysis was concerned with time from treatment until death, hence the name, "survival analysis". Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs (Kleinbaum and Klein, 2005). According to Akritas (2004), survival analysis is a method for the analysis of data on an event observed over time and the study of factors associated with the occurrence rates of this event. The event could be the time until a generator's bearing seizes, the time until a patient dies or the time until a person finds employment (Cleves *et al.*, 2008). Survival analysis can be used in many fields, such as medicine, biology, public health and epidemiology (Kul, 2010). In pharmacology and toxicology survival analysis is used in analyzing the events like time to death, time to signs occurrence, disappearance and reoccurrence, time to recovery etc. of the experimental animals.

Another terminology that we need to understand in survival analysis is 'censored observation'. When animals do not have an event during the observation time, they are described as censored. Censored animals may or may not have an event after the end of observation time.

Hazard Rate

'Hazard rate' is an important concept in survival analysis. It provides information on the risk of event happening as a function of time, condition on not having happened previously (Aalen *et al.*, 2009), whereas survival curve provides information on how many have survived upto a certain time. Hazard function can be estimated using the equation:

H (t) = Number of individuals experiencing an event in interval beginning at t/(number of individuals surviving at time *t*) x (interval width)

The hazard function describes the risk of an outcome of an event in an interval after time *t*, conditional on the individual having experienced the event to time *t*. The hazard function is useful in determining whether toxicity is constant over time, or it increases or decreases as the exposure continues (Wright and Welbourn, 2002).

Kaplan-Meier Method

Survival analysis is normally carried out using Kaplan-Meier method or the log rank test. The log rank test is ideal for the analysis of two groups. The Kaplan–Meier estimator uses product-limit methods to estimate the survival ratio (Kaplan and Meier, 1958). This is a nonparametric maximum likelihood estimate of survival analysis and is used in animal experiments to measure the fraction of animals that lives after treatment.

Distribution of the survival time T from the start of the experiment (first dose administration) to the event of interest (for example mortality) is considered as a random variable. The survival rate, S_t, is defined as the probability that an animal survives longer than *t* units of time:

S_t=P (T> *t*); for example, if *t* is in years, S_2 is the two-year survival rate; if S_2=P (T> 2)=0.10, it indicates 10% is the probability the time from a treatment to death is greater than 2 years

Kaplan-Meier product-limit estimator

$$S_t = \prod \frac{r_i - d_i}{r_i},$$

r_i is the number of animals lived just before t_i; d_i is the number of animals which died in t_i. Π denotes the product (geometric sum) across all cases less than or equal to *t*. Kaplan-Meier product-limit estimator measures the fraction of animals living for a certain amount of time after treatment.

Let us review an example to understand Kaplan-Meier product-limit estimator. The survival rate of F344 rats in a 110-week chronic toxicity study is given in Table 15.1. The experimental group of rats (20 rats/group) was treated with 1000 ppm pesticide in diet. The control group of rats (20 rats/group) was given normal diet without the pesticide.

Table 15.1. Survival rate of F344 rats in a 110-week chronic toxicity (Funaki and Origasda, 2001)

Control group (Normal diet)				Treatment group (1000 ppm pesticide in diet)			
Animal ID-No.	Survival period (week)	Survival rate (s_t)	Size of effective sample (n')	Animal ID-No.	Survival period (week)	Survival rate (s_t)	Size of effective sample (n')
1001	85	0.950	20	1101	66	0.900	20
1002	87	0.900	19	1102	66		
1003	95	0.800	18	1103	62	0.850	18
1004	95			1104	63	0.800	17
1005	99	0.650	16	1105	68	0.750	16
1006	99			1106	70	0.650	15
1007	99			1107	70		
1008	101	0.550	13	1108	72	0.550	13
1009	101			1109	72		
1010	102	0.500	11	1110	75	0.400	11
1011	103	0.350	10	1111	75		
1012	103			1112	75		
1013	103			1113	77	0.300	8
1014	104	0.250	7	1114	77		
1015	104			1115	78	0.57	7
1016	106	0.150	5	1116	79	0.154	5
1017	106			1117	79		
1018	110	0.050	3	1118	80	0.051	3
1019	112	0.025	2	1119	80		
1020	120	-	1	1120	88	-	1

The survival rate is calculated using the equation:

$$S_t = \prod \frac{r_i - d_i}{r_i}$$

Calculation procedure of S_t is given below by working out few selected survival period of control group:

Week 85 (S_{85}): $0.950 = \frac{20-1}{20}$

Week 87 (S_{87}): $0.900 = 0.950 \times \frac{19-1}{19}$

Week 95 (S_{95}): $0.800 = 0.900 \times \frac{18-2}{18}$

Week 112 (S_{112}): $0.025 = 0.05 \times \frac{2-1}{2}$

Thus S_t is calculated for all survival periods of control group and given in Table 15.1.

Calculation of standard error of S_t:

$$SE = \sqrt{\frac{S_t(1-S_t)}{n'}}$$

Let us calculate SE for (S_{104}):

Week 104 (S_{104}): $0.25 = 0.35 \times \frac{7-2}{7}$

$$SE = \sqrt{\frac{0.25(1-0.25)}{7}} = 0.164$$

Survival rate at 95% confidence interval is:

$$0.25-(1.96\times0.164) \sim 0.25+(1.96\times0.164) = 0–0.57$$

The 95% confidence interval exploded in a wide range, because of the small sample size (N=7).

Similarly survival rate of the treatment group is computed and given in Table 15.1. Plot of survival curves is an important part of survival analysis (Freeman *et al*., 2008). A plot of the survival curves of data (Table 15.1) is given in Figure 15.1.

Though the survival curves provide a good information on the mortality rates in two groups, the comparison of the curves should be made using a statistical tool (Altman, 1991). Log-rank test is the common method used to compare survival curves (Cox, 1972). This test assigns equal weight to each event at whatever time it occurs (Tinazzi *et al*., 2008). The null

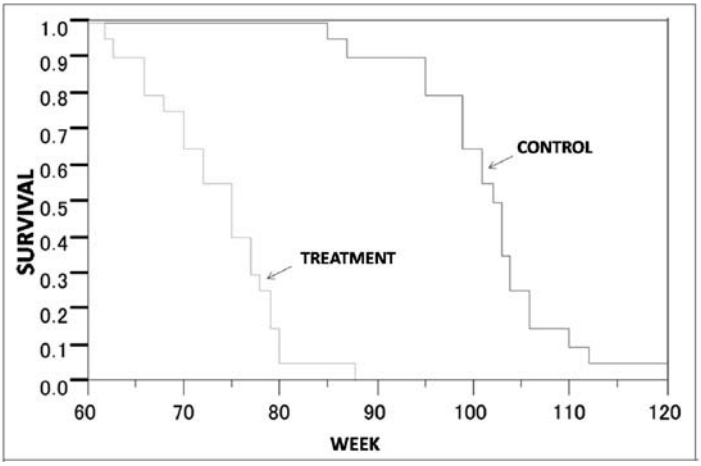


Figure 15.1. Survival rate of F344 rats in a 110-week chronic toxicity

hypothesis for the log-rank test is that there is no difference between the survivals of two or more populations that are being compared. The comparison is based on the difference between the observed number of events in each group and the expected number of events in case of non-difference between the two groups. The χ^2 equation is:

$$\chi^2_{\log-rank} = \sum_g \frac{(O_g - E_g)^2}{E_g}$$

where O is the number of observed events in each group g, and E is the total number of expected events in each group g. O and E are computed each time an event happens; if a survival time is censored, then the subject is considered to be at risk during the interval of censoring, but not anymore for the subsequent intervals. The test statistic is then compared with a χ^2 with g-1 degrees of freedom. The limitation of log rank test and Cox's proportional hazards model is that they are based on the assumption that the hazard ratio is constant over time (Bewick *et al.*, 2004).

Both the life-table and the Kaplan-Meier methods have advantages and disadvantages. In small data sets in which the time of occurrence event is measured precisely the Kaplan-Meier method is best used, whereas the life-table methods works well with large data sets and when the time of occurrence of an event cannot be measured precisely. The Kaplan-Meier method handles censored data better than life-table method.

References

Aalen, O.O. (1976): Nonparametric inference in connection with multiple decrement models. Scand. J. Stat., 3, 15–27.

Aalen, O.O. (1980): A model for non-parametric regression analysis of lifetimes. In: Mathematical Statistics and Probability Theory. Editors, Klonecki, W., Kozek, A. and Rosinski, J. Lecture Notes in Statistics, Vol. 2, Springer-Verlag, NewYork.

Aalen, O.O., Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (2009): History of applications of martingales in survival analysis. Electronic J. History Probability Stat., 5(1), 1–28.

Aalen, O.O., Borgan, Ø. and Gjessing, H.K. (2008): Survival and Event History Analysis: A Process Point of View. Springer-Verlag, NewYork.

Akritas, M.G. (2004): Nonparametric survival analysis. Stat. Sci., 19(4), 615–623.

Altman, D.G. (1991): Practical Statistics for Medical Research. Chapman & Hall/CRC, London.

Bechmann, R.K. (1994): Use of life tables and lC_{50} tests to evaluate chronic and acute toxicity effects of copper on the marine copepod *Tisbe furcata* (baird), Environmental Toxicology and Chemistry. 13(9), 1381–1548.

Bewick, V., Cheek, L. and Ball, J. (2004): Statistics review 12: Survival analysis. Crit Care, 8(5), 389–394.

Cleves, M., Gutierrez, R., Gould, W. and Marchenko, Y. (2008): An Introduction to Survival Analysis Using Data. 2nd Edition, Stata Press, Texas, USA.

Cox, D.R. (1972): Regression models and life tables (with discussion). J. Royal Stat. Soc. Ser., B34, 187–220.

Diggle, P., Farewell, D.M. and Henderson, R. (2007): Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. J. Royal Stat. Soc. Ser. C (Applied Statistics), 56, 499–550.

FDA (2007): United States Food and Drug Administration. Redbook 2000: IV.B.4 Statistical Considerations in Toxicity Studies. USFDA, MD, USA.

Freeman, J.V., Walters, S.J. and Campbell, M.J. (2008): How to display data. Blackwell BMJ Books, Oxford, UK.

Funaki, K. and Origasda, H. (2001): Statistics with Confidence, Scientist Press, Tokyo, Japan.

Gentile, J.H., Gentile, S.M., Hairston, N.G. and Sullivan, B.K. (1982): The use of life-tables for evaluating the chronic toxicity of pollutants to *Mysidopsis bahia*. Hydrobiologia, 93(1-2), 179–187.

Kaplan, E. L. and Meier, P. (1958): Nonparametric estimation from incomplete observations. J. Am. Stat. Assn., 53, 457–481.

Kleinbaum, D.G and Klein, M. (2005): Survival Analysis-A Self Learning Text. 2nd Edition, Springer+Business Media, Inc., New York, USA.

Kreager, P. (1988): Newlighton Graunt. Population Studies, 42,129–140.

Kul, S. (2010): The use of survival analysis for clinical pathways. Intl. J. Care Path Ways, 14, 23–26.

Mantel, N. and Haenszel, W. (1959): Statistical aspects of the analysis of data from retrospective studies of disease. J. National Cancer Inst., 22, 719–748.

Portier, C.J. (1988): Life table analysis of carcinogenicity experiments. Int. J.Tox., 7(5), 575–582.

Tinazzi, A., Scott, M. and Compagnoni, A. (2008): A gentle introduction to survival analysis. SAS Conference Proceedings: PhUSE 2008, October 12–15, 2008, Manchester, UK.

Van Leeuwen, C.J., Moberts, F. and Niebeek, G. (1985): Aquatic toxicological aspects of dithiocarbamates and related compounds. II. Effects on survival, reproduction and growth of *Daphnia magna.* Aquatic Toxicol., 7(3), 165–175.

Wright, D.A. and Welbourn, P. (2002): Environmental Toxicity. Cambridge Environmental Chemistry Series 11. Cambridge University Press, Cambridge, UK.

16 Dose Response Relationships

Dose and Dosage

Dose-response relationship is the association between the dose administered and the response/s that is/are exhibited. Response/s and dose are causally related (Eaton and Klaassen, 1996). Establishing a cause–response relationship is very important in the analysis/assessment of a risk (Christensen *et al.*, 2003). Though the terms 'dose' and 'dosage' refer to more or less a same thing, there is a difference between these two terms. Dose refers to a stated quantity or concentration of a substance to which an organism is exposed and is expressed as the amount of test substance per unit weight of test animal (example, mg/kg body weight), whereas dosage is a general term comprising the dose, its frequency and the duration of dosing. Dosages often involve the dimension of time (example, mg/kg body weight/day) (Hayes, 1991).

Margin of Exposure, NOAEL, NOEL

Determining the presence or absence of a dose-response relationship is one of the primary criteria of a risk assessment (IPCS, 2009). In drug development, assessment of dose-response should be an integral part in the study design. The studies should be designed to assess dose-response an inherent part of establishing the safety and effectiveness of the drug (EMEA, 2006). Once a dose-response relationship is established for a test substance, the margin of exposure is determined. The margin of exposure lies between a defined point on the dose-response relationship and the human exposure level. In animal experiments, NOAEL (No-observed-adverse-effect-level) and NOEL (No-observed-effect-level) on the dose-response curve are usually considered as this defined point. Though in reality, both NOAEL and NOEL have similar meaning, JECFA (Joint

FAO/WHO Expert Committee on Food Additives) differentiated between the terms NOEL and NOAEL in risk assessments with the following definitions (WHO, 2007):

NOEL: Greatest concentration or amount of a substance, found by experiment or observation, that causes no alteration of morphology, functional capacity, growth, development, or lifespan of the target organism distinguishable from those observed in normal (control) organisms of the same species and strain under the same defined conditions of exposure.

NOAEL: Greatest concentration or amount of a substance, found by experiment or observation, which causes no detectable adverse alteration of morphology, functional capacity, growth, development, or lifespan of the target organism under defined conditions of exposure.

An adverse response is defined as 'change in morphology, physiology, growth, development or life-span of an organism which results in impairment of the functional capacity or impairment of the capacity to compensate for additional stress or increase in susceptibility to the harmful effects of other environmental influences'. Decisions on whether or not any effect is adverse requires expert judgment (WHO, 1994). This definition shows that the environmental standard setting in general is adjusted to subtle effects which represent early steps in biological effect chains or can be interpreted as first signs of a pathological process (Neus and Boikat, 2000). An alternative approach is to classify dose-related effects in to physiological, toxic and pharmacological responses (OECD, 2000a). Physiological responses are not considered as adverse responses. For example, changes in pulse rate or respiration rate as long as it occurs within the normal functioning of the animal. Changes in physiological function as a result of interaction of a test substance with a cellular receptor site are considered as pharmacological responses. Pharmacological responses are reversible and of short duration, and can be adverse if they cause harm to the animals. Toxic responses are adverse and they can be reversible or irreversible. A chemical which causes a physiological or pharmacological effect may produce a toxic response if the exposure is prolonged and/or if the dose is increased beyond a certain level.

But, there is no consistent standard definition of NOAEL (Dorato and Engelhardt, 2005). In an FDA document (FDA, 2005) NOAEL is defined as the highest dose level that does not produce a significant increase in adverse effects in comparison to the control group. Any biologically significant effect is considered as an adverse effect, which may or may not statistically significant. NOEL refers to any effect, which may or may not be an adverse one. The definition of the NOAEL, in contrast to that of the NOEL, reflects the view that some effects observed in the animal may be acceptable pharmacodynamic actions of the therapeutic and may not raise a safety concern (FDA, 2005). Some other terminologies related to dose-response relationship are LOEL (Lowest-Observed-Effect Level), LOAEL (Lowest-Observed-Adverse-Effect Level) and threshold dose. LOEL is the lowest dose of a test substance which causes effects distinguishable from those observed in control animals and LOAEL is the lowest dose of a test substance which causes adverse changes distinguishable from those observed in control animals. Threshold dose is the minimum dose required to elicit a response. NOAEL has lot of importance in the clinical development of a drug. For example, the calculation of the first dose in man is based on NOAEL (EMEA, 2007). We may briefly explain some of the practical issues in determining NOEL/NOAEL.

Determining NOEL and NOAEL

One of the main objectives of conducting repeated-dose toxicity studies is to arrive at NOEL or NOAEL. Most of the regulatory guidelines prescribe that the repeated-dose toxicity studies with rodents should be conducted with a minimum of three treatment doses (low, mid and high doses) and a control group (OECD, 1995). The low dose level is carefully selected so that the animals exposed to this dose level will not show any effect of the treatment compared to the control dose. But, most of the repeated-dose toxicity studies show some effect of the treatment in few parameters of the low dose group. In such cases considering the low dose as an NOEL/NOAEL may be questionable. Kobayashi *et al.* (2010) investigated 109 numbers of 28-day repeated dose administration studies in rats and examined the measurable items (functional observational battery, urinalysis, hematology, blood chemistry and absolute and relative organ weights) of the low dose group. Their investigation revealed that, 205/12167 (1.6%) measurable items showed a significant difference ($P<0.05$) in the low dose groups compared to the respective controls. The authors concluded from the investigation that the low dose may be considered to be NOEL, if the significant difference of the measurable items showed by this dose group is about 2% (maximum <5%), compared to the control. However, due consideration may be given to the clinical relevance of the items that showed a significant difference.

It is not uncommon to encounter situations in repeated-dose toxicity studies where mid dose group alone shows an insignificant difference compared to control, whereas low and high dose groups show a significant difference. The guidelines do not mention how to determine the mid dose, except an indication that this dose is required to examine dose dependency. According to Gupta (2007), the mid dose selection should consider threshold in toxic response and mechanism of toxicity. Determining the mid dose is as important as determining the high and low doses in repeated-dose toxicity studies, since mid dose plays a determining role in establishing the dose dependency. For determining dose-related trend in repeated-dose toxicity studies, Williams' test is generally carried out (Bretz, 2006). The disadvantage of Williams' test is that it uses an estimated value for the mean rather than the original mean value for the analysis. Hence, it is likely that Williams' test may indicate a dose-related trend, when it actually does not exist (Williams' test is covered in detail in Chapter 11). Therefore, to analyse such data the use of Dunnett's multiple comparison test for comparing each dose group with the control, followed by Jonckheere's trend test for examining dose-related trend is recommended.

Benchmark Dose

NOAEL is based on a single data point and it does not consider the shape of the dose-response curve, the number of animals in the group, or the statistical variation in the response and its measurement (EPA, 1998). An alternative approach to NOAEL is the Benchmark dose approach (Kimmel and Gaylor, 1988). The Benchmark dose is defined as the dose of a chemical that is required to achieve a predetermined response of a toxicological effect (Sand *et al.*, 2006). The Benchmark dose method uses the full dose response data for the statistical analysis, hence the result obtained from the analysis is considered to be more reliable than the single data point based NOAEL. Unlike the NOAEL approach, the Benchmark dose method includes the determination of the response at a given dose, the magnitude of the dose at a given response and their confidence limits. According to EPA SAB (1998): "The [categorical regression] process makes use of every bit of data available. The underlying premise of the approach is that the severity of the effect, not the specific measurement or outcome incidence, is the information needed for assessing exposure-response relationships for non-cancer endpoints.... All the available data is plotted on a single chart and one can immediately see a rough picture of the level of the concentration multiplied by time values that can be expected to cause adverse effects of varying severity." The U.S. EPA's CatReg Program (Strickland, 2000) utilizes categorical regression to establish the relationship between concentration, time, and severity of the resulting effect. Response variability and uncertainty are addressed by confidence limits bounding the derived relationship curves. Three statistical models (Logit, Probit and Complementary Log-Log) are available in the CatReg program.

Probit Analysis

Probit analysis was originally published in Science by Bliss (Bliss, 1934). He was an entomologist and was involved in research to find a pesticide to control insects that fed on grape leaves (Greenberg, 1980). Bliss transformed the percentage mortality into a "probability units" (or "Probits") and plotted the 'Probits' against concentrations. But, he did not have a statistical tool to compare the effects among various pesticides. In 1952, Finney of the University of Edinburgh wrote a book, 'Probit Analysis' (Finney, 1952). Probit analysis, a preferred method for analyzing dose-response relationship even today described elaborately in Finney's book, is based on the idea developed by Bliss. One of the assumptions of Probit analysis is that the response *vs* dose data are normally distributed, if not, Finney suggested using the logit over the Probit transformation (Finney, 1952). Both Logit analysis (Muhammad *et al.*, 1990) and Probit analysis (Finney, 1978) are used in biological assays.

Performing Probit analysis manually is tedious. An example is provided below to show the steps involved in this statistical analysis. Most of the commercially available statistical software can perform Probit analysis.

Groups of rats (10 rats/group) were given a drug at different dose levels. The response shown by the number of animals at each dose level is given in Table 16.1.

Let us plot a graph with dose on X axis and percent response on Y axis (Figure 16.1).

The very purpose of carrying out the Probit analysis is to find out that dose which causes the response in 50% of the animals. If the response that we are looking at is mortality, the dose that causes mortality in 50% of animals is called as LD_{50}. Since the inception of the LD_{50} test by Trevan (1927), the test has gained wide acceptance as a measure of acute toxicity of all types of substances (DePass, 1989).

Table 16.1. Response shown by rats following the administration of a drug

Dose (mg/kg b.w.)	Response shown by number of animals	Percent Response
5	0	0
6	0	0
8.6	1	10
10.3	1	10
12.4	4	40
14.9	6	60
17.9	7	70
21.5	9	90
25.8	10	100
31	10	100

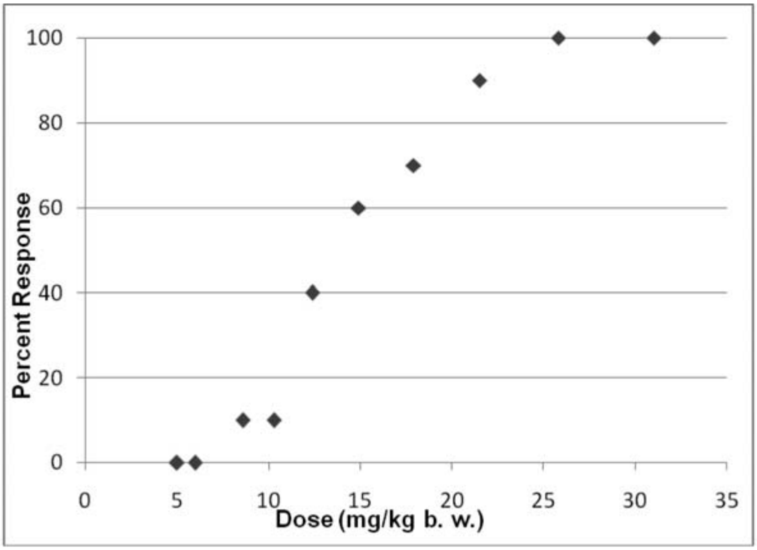


Figure 16.1. Dose *vs* response plot

We could have determined the dose which causes 50% response (for example, LD_{50}) straight away from the plot, had the plot been a straight line. In Finney's Probit analysis the dose response curve is converted to a straight line by transforming the doses to logarithmic values and percent mortality to Probit values (Finney, 1971). Let us try to understand what Probit values means. Percent response on Y axis can be converted to normal equivalent deviation (NED). What is an NED? We know that at one standard deviation below mean value (–1SD), 16% will show response and one standard deviation above mean value (+1SD) 84% will show the response. Such a relationship can be established between standard deviation and response. The response converted to the corresponding standard deviation is termed as NED. NEDs of below 50 percent response are negative numbers and above 50 percent response are positive numbers. To make the subsequent calculation steps easier, the negative numbers can be converted to positive numbers by simply adding 5 to all NEDs. Now these NEDs are called as probability units or Probits. Finding the Probits for percent response using the above steps is cumbersome. Probit value of a percent response can be directly read from the 'Probit Table' given in several statistical books. Such a Table is given hereunder in an abridged form (Table 16.2).

Table 16.2. Transformation of percentage response to Probit values

Percentage Response	0	10	20	30	40	50	60	70	80	90	100
Probits	-	3.72	4.16	4.48	4.75	5.00	5.25	5.52	5.84	6.28	-

Lets us now plot a graph with log dose on X axis and Probit on Y (Figure 16.2.).

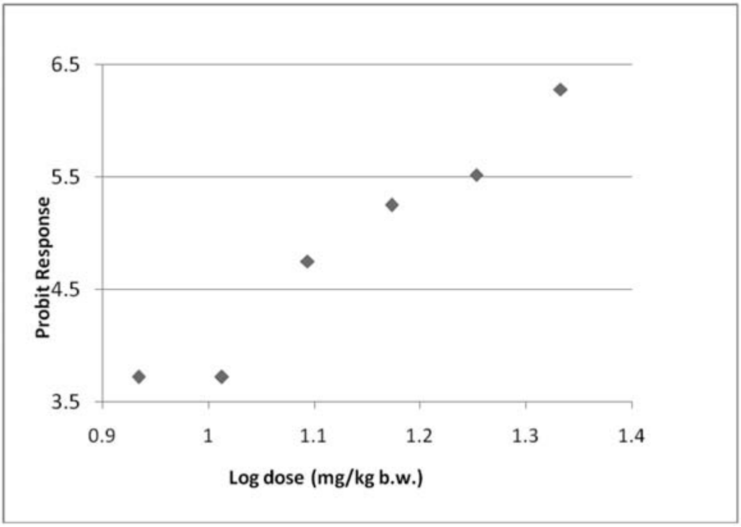


Figure 16.2. Log dose *vs* Probit response plot

You would have observed that Probit responses for 4 doses are missing in the Figure 16.2. The reason for this is that there are no Probit values for 0% and 100% responses. From the Figure one can find that the Probit values somewhat fall in a linear fashion. Let us closely observe the Probit values.

The middle region of the line (region of 50% response, *i.e.*, the region of Probit 5) is linear, hence this region is somewhat reliable for making a prediction. The two ends of the line, where the data are controlled by few animals, are not so linear in fashion, hence these regions are seldom used for making a prediction. The variation in the middle region of the line is less, whereas it is on the higher side in the 2 ends. This variation can be minimised by using weighting coefficients. Once a best-fit line is drawn using a regression equation, a 'statistically reliable median response dose' can be estimated:

$$\overline{Y} = \mathrm{a} + \mathrm{b}\overline{X}, \text{ where}$$

$\overline{Y}$ = 5 (Probit value corresponding 50% response)

$\overline{X}$= Log dose

a = Intercept

b = Slope

Mentioning the term 'statistically reliable median response dose', is intentional as several reports have stated that 'median response dose', for example, LD_{50} is notoriously variable. Usefulness of LD_{50} test has been criticized, as the test only expresses mortality; the test requires large number of animals and the outcome of the LD_{50} test is influenced by several factors associated with the animal (for example, species, age, sex, etc.), animal house condition (for example, temperature, humidity, light intensity, etc.) and human error; many times the findings of the test cannot be extrapolated to man. On the contrary, supporters of the LD_{50} test are of the opinion that a properly conducted LD_{50} test can yield information on the cause and time of death, symptomatology, nonlethal acute effects; slope of the mortality curve can provide information on the mode of action and metabolic detoxification; the results can be used for the basis for designing subsequent subchronic studies; the test is first approximation of hazards to workers (Hodgson, 2010).

This method for calculating LD_{50}, requires a large number of animals, thus, not desirable. Interested readers may refer to the Up and Down Procedure, which requires less number of animals (OECD, 2000b).

IC_{50} and EC_{50} Determination

IC_{50} and EC_{50} determinations are performed for assessing pharmacological affinity of new pharmaceutical compounds. IC_{50} is the concentration of the compound that provides 50% inhibition, whereas EC$_{50}$ is the concentration that provides 50% of compound's maximal response. IC$_{50}$ is determined for competition binding assays and functional antagonist assays, whereas EC$_{50}$ is determined for agonist/stimulator assays. The procedure for determining IC$_{50}$ and EC$_{50}$ is similar.

For fitting an IC$_{50}$/EC$_{50}$ curve, first convert the data into percentage inhibition/ percentage activity depending up on the assay performed. If the assay is carried out in replicates find the median percentage inhibition/percentage activity for each concentration. Plot a graph of log concentration *vs* percentage inhibition/percentage activity. The dose-response relationship can be derived using the Hill-slope model. It is also known as four parametric logistic model (4PL). The 4PL function is widely used in biological assays (Healy, 1972; Rodbard *et al*., 1978). The 4PL model equation is given below:

$$y = \text{Minimum Asymptote} + \frac{(\text{Maximum Asymptote} - \text{Minimum Asymptote})}{1 + (x / IC_{50} / EC_{50})^{\text{Hill Slope}}}$$

where y is the percentage activity/percentage inhibition and x is the corresponding concentration. The IC$_{50}$/EC$_{50}$ given in the equation is not the absolute IC$_{50}$/EC$_{50}$, but, relative IC$_{50}$/EC$_{50}$. Relative IC$_{50}$/EC$_{50}$ is the concentration giving a response half way between the fitted top and bottom of the curve. The relative IC$_{50}$/EC$_{50}$ serves the purpose for most of the assays.

Bioassays with a quantitative response showing a sigmoid log-dose relationship can be analysed by fitting a non-linear dose-response model directly to the data (Vølund, 1978). If the quantitative response shows a non-normal distribution, a five-parameter logistic (5PL) function is more ideal to fit dose-response data. The 5PL can dramatically improve the accuracy of asymmetric assays (Gottschalk and Dunn, 2005).

Usually, several concentrations of the compound are employed for the determination of IC$_{50}$. Turner and Charlton (2005) proposed a method for determining IC$_{50}$ using two concentrations. However, use of this method is not well accepted in drug discovery research.

Hormesis

All along we have been discussing about 'threshold dose-response curve'. It is widely believed that to initiate a biological effect some dose is required. This dose is called as the threshold dose. According to this belief a dose below the threshold dose level cannot initiate the effect. This concept has been disproven in recent years by introducing a hypothesis called 'hormesis'. The term hormesis was coined by Southam and Ehrlich

(1943). The hormesis hypothesis states that most of the chemical agents may stimulate or inhibit biological effects at doses lower than a threshold, while they are toxic at doses higher than the threshold. This hypothesis falls in line with Arndt-Schulz Law, which states that 'a weak stimulus increases physiologic activity, a moderate stimulus inhibits activity and a very strong stimulus abolish the activity (Schulz,1887). However, Arndt-Schulz Law is not widely known among the toxicologists and pharmacologists. One of the reasons for this is it was heavily criticised by earlier pharmacologists and toxicologists, hence did not find place in most books on toxicology and pharmacology. Alfred Clark, the renowned pharmacologist, in his book entitled 'The Mode of Action of Drugs on Cells' published in 1933 stated: "In 1885 Rudolf Arndt put forward the suggestion that if a weak stimulus excites an organism, then any drug in sufficiently weak dose ought to do this also. This suggestion was developed by Schulz, who had leanings to homeopathy" (Clark, 1933). Clark was well known among the statisticians like Fisher and Bliss, who contributed significantly to the threshold dose-response relationship. Another book by Clark, 'Handbook of Experimental Pharmacology" (Clark, 1937), which was very critical of the Arndt-Schulz Law, was published in seven editions, in 1970s, more than 30 years after his death. Holmstedt and Lijestrand in their book, 'Readings in Pharmacology, published in 1981 stated that Homoeopathic theories like the Ardnt-Schulz law and Weber-Fechner law were based on loose ideas around surface tension of the cell membranes but there was little physic-chemical basis to these ideas (Holmstedt and Lijestrand, 1981).

Brain-Cousens (1989) proposed a modified four-parameter logistic model in situations where hormesis is present. Several publications indicated that the hormetic dose-response is far more common and fundamental than the threshold dose-response models used in toxicology (Calabrese, 2005). According to Calabrese (2010), the hormetic dose-response model makes far more accurate predictions of responses in low dose zones than either the threshold or linear at low dose models.

References

Bliss, C.I. (1934): The method of Probits. Science, 79(2037), 38–39.

Brain, P. and Cousens, R. (1989): An equation to describe dose responses where there is stimulation of growth at low dose. Weed Res., 29, 93–96.

Bretz, F. (2006): An extension of the Williams trend test to general unbalanced linear models. Comp. Stat. Data Anal., 50(7), 1735–1748.

Calabrese, E.J. (2005): Toxicological awakenings: the rebirth of hormesis as a central pillar of toxicology. Toxicol. Appl. Pharmacol., 206(3):365–366.

Calabrese, E.J. (2010): Hormesis is central to toxicology, pharmacology and risk assessment. Hum. Exp. Tox., 29(4), 249–261.

Christensen, F.M., Andersen, O., Duijm, N.J. and Harremoës, P. (2003): Risk terminology—a platform for common understanding and better communication. J. Hazardous Materials, A103, 81–203.

Clark, A.J. (1933): The Mode of Action of Drugs on Cells. The Williams & Wilkins Company, Baltimore, USA.

Clark, A.J. (1937): Handbook of Experimental Pharmacology. Springer, Berlin, Germany.

DePass, L.R. (1989): Alternative approaches in median lethality (LD_{50}) and acute toxicity testing. Toxicol. Lett., 49(2-3), 159–170.

Dorato, M.A. and Engelhardt, J.A. (2005): The no-observed-adverse-effect-level in drug safety evaluations: Use, issues, and definition(s). Reg. Toxicol. Pharmacol., 42(3), 265–274.

Eaton, D.L. and Klaassen, D. (1996): Principles of Toxicology. In: Casarett and Doull's Toxicology; The Basic Science of Poisons, 5th Edition, McGraw-Hill, New York, USA.

EMEA (2006): European Medicines Agency. Note for Guidance on Dose Response Information to Support Drug Registration (CPMP/ICH/378/95). ICH Topic E4 Dose Response Information to Support Drug Registration. EMEA, London, UK.

EMEA (2007): European Medicines Agency. Guideline on Requirements for First-in-man Clinical Trials for Potential High-risk Medicinal Products. EMEA/CHMP/SWP/28367/2007. Committee for Medicinal Products for Human Use, EMEA, London, UK.

EPA (1998): United States Environmental Protection Agency. Methods for Exposure-Response Analysis for Acute Inhalation Exposure to Chemicals: Development of the Acute Reference Exposure (ARE). EPA/600/R-98/051. External Review Draft. April 1998. USEPA, Washington, DC., USA.

EPA SAB (1998):United States Environmental Protection Agency Science Advisory Board. A SAB Report: Development of the Acute Reference Exposure: Review of the Draft Document Methods for Exposure-Response Analysis for Acute Inhalation Exposure to Chemicals: Development of the Acute Reference Exposure (EPA/600/R-98/051) by the Environmental Health Committee of the Science Advisory Board (SAB), EPA-SAB-EHC-99-005. US EPA SAB. November 1998. USEPA, Washington, DC., USA.

FDA (2005): Food and Drug Administration. Guidance for industry—Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers. Centre for Drug Evaluation and Research, Food and Drug Administration, USFDA, Rockville, USA.

Finney, D.J. (1952): Probit Analysis. Cambridge University Press, Cambridge, UK.

Finney, D.J. (1971): Probit Analysis. 3rd Edition. Cambridge, London, UK.

Finney, D.J. (1978): Statistical Method in Biological Assay. 3rd Edition. Charles Griffin & Co., London, UK.

Gottschalk, P.G. and Dunn, J.R. (2005): The five-parameter logistic: A characterization and comparison with the four-parameter logistic. Anal. Biochem., 343, 54–65.

Greenberg, B.G. (1980): Chester I. Bliss, 1899–1979. International Statistical Review/ Revue Internationale de Statistique, 8(1), 135–136.

Gupta, R.C. (2007): Veterinary Toxicology—Basic and Clinical Principles. Academic Press, New York, USA.

Hayes, W.J. (1991): Dosage and other factors influencing toxicology. In: Hayes, W.J. & Laws, E.R. (Editors). Handbook of Toxicology, Vol. 1, General Principles. Academic Press, San Diego, USA.

Healy, M.J.R. (1972): Statistical analysis of radioimmunoassay data. Biochem. J., 130, 107–210.

Hodgson, E. (2010): A Text Book of Modern Toxicology. John Wiley & Sons Inc., New Jersey, USA.

Holmstedt, B. and Lijestrand, G. (1981): Readings in Pharmacology. Raven Press, New York, USA.

IPCS, (2009): International Programme for Chemical Safety. Principles and Methods for the Risk Assessment of Chemicals in Food, Chapter 5, Dose-response Assessment and Derivation of Health-based Guidance Values. Environmental Health Criteria 240, IPCS, Geneva, Switzerland.

Kobayashi, K., Pillai, K.S., Michael, M., Cherian, K.M. and Ohnishi, M. (2010): Determining NOEL/NOAEL in repeat-dose toxicity studies, when the low dose group shows significant difference in quantitative data. Lab. Animal Res., 26(2), 133-137.

Kimmel, C.A. and Gaylor, D.W. (1988): Issues in qualitative and quantitative risk analysis for development toxicology. Risk Anal., 8, 15–20.

Muhammad, F., Khan, A. and Ahmad, A. (1990): Logistic regression analyses in dose response studies. J. Islamic Acad. Sci., 3:2, 103–106.

Neus, H. and Boikat, U. (2000): Evaluation of traffic noise-related cardiovascular risk. Noise & Health, 2(7), 65–77.

OECD (1995): Organization for Economic Cooperation and Development. OECD Guidelines for Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents, No. 407. OECD, Paris, France.

OECD (2000a): Organization for Economic Cooperation and Development. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. OECD Environment, Health and Safety Publications Series on Pesticides, No. 10. OECD, Paris, France.

OECD (2000b). Organization for Economic Development and Co-operation. Acute Oral Toxicity: Up-and-Down Procedure. OECD Guideline for the Testing of Chemicals, Revised Draft Guideline, No. 425. OECD, Paris, France.

Rodbard, D., Munson, P.J. and DeLean, A. (1978): Improved curve fitting, parallelism testing, characterization of sensitivity and specificity, validation, and optimization for radioimmunoassays 1977. Radioimmunoassay and Related Procedures in Medicine 1, Vienna, Italy. Int. Atomic Energy Agency (1978) 469–504.

Sand, S., von Rosen, D., Victorin, K. and Filipsson, A.F. (2006): Identification of a critical dose level for risk assessment: Developments in Benchmark dose analysis of continuous endpoints. Tox. Sci., 90(1), 244–251.

Schulz, H. (1887): Zur lehre von der arzneiwirdung. Virchows Archiv fur Pathol. Anatom. und Physiol. fur Klinische Medizin,108, 423–445.

Strickland, J.A. (2000): CatReg Software User Manual. US Environmental Protection Agency. EPA 600/R-98/052.

Southam, C.M. and Ehrlich, J. (1943): Effects of extracts of western red-cedar heartwood on certain wood-decaying fungi in culture. Phytopath., 33, 517–524.

Trevan, J. (1927): The error of determination of toxicity. Proc. R. Soc., 101B, 483−514.

Turner, R.J and Charlton, S.J. (2005): Assessing the minimum number of data points required for accurate IC_{50} determination. Assay Drug Devp. Technol., 3(5), 525–531.

Vølund, A. (1978): Application of the four-parameter logistic model to bioassay: comparison with slope ratio and parallel line models. Biometrics, 34(3), 357–365.

WHO (1994): World Health Organisation. Assessing Human Health Risks of Chemicals: Derivation of Guidance Values for Health Based Exposure Limits. Environmental Health Criteria 170, WHO, Geneva, Switzerland.

WHO (2007): World Health Organisation. Evaluation of Certain Food Additives and Contaminants. Sixty-eighth Report of the Joint FAO/WHO Expert Committee on Food Additives. WHO Technical Report Series No. 947, WHO, Geneva, Switzerland.

17

Analysis of Pathology Data

Pathology in Toxicology

Pathology occupies a pivotal role in animal experiments. The toxicity of a compound can be assessed by linking compound-related changes in biochemical, haematological or urinalysis parameters with organ weight, gross pathology and/or histopathological changes (Tyson and Sawhney, 1985; Krinke *et al.*, 1991). All regulatory guidelines on animal experiments have given special emphasis to pathology. For example, in the long-term repeated dose administration studies, it is a regulatory requirement that all data relating to moribund or dead animals as well as the results of postmortem examinations is scrutinized and the analysis of the cause of individual deaths is done (OECD, 2000).

Pathologists usually make a biological judgment based on their experience, which differs from one pathologist to the other (Glaister, 1986). In a repeated dose administration study involving a large number of animals, the observation of tissue section slides may be completed over a substantial length of time. Thus it is not possible to maintain the consistency of grading the lesions, causing a ‘diagnostic drift’. It has been stated that even the nomenclature used to describe pathology findings in toxicology studies suffers from the lack of uniformity. Use of different nomenclature for describing the lesions causes difficulties while interpreting the observations (Haseman *et al.*, l984). Statistically and logically, blinding the slides is the best way to avoid the bias. But, several veterinary pathologists do not favor this, because they fear that blinded reading of slides of animal tissues/organs may result in loss of information critical to interpretation, such as the ability to relate observations in different tissues (Iatropoulos, 1984; Newberne and de la Lglesia, 1985; Prasse *et al.*, 1986; Goodman, 1988; House *et al.*, 1992; FDA, 2001). Mistakes can be easily made when assigning, opening codes, and recording results in blinded reading (Iatropoulos, 1988).

Microscopical data obtained from toxicity studies is usually classified into several grades. The grades of the control group is usually shown by minus (–) and those of the treated groups by (+1), (+2), (+3), so on. For statistical analysis, the difference of the grades between the control group and treatment groups is examined by Fisher's probability test or cumulative chi-square test. By these methods, only the presence or absence of a difference among several groups or between two groups can be ascertained and the degree of pathology lesions remains uncertain.

There is not enough specific statistical guidance available for the pathologists. Wade (2005) stated that most of the published statistical literature is not directly applicable to research in the field of pathology. In the toxicology studies with three or more groups, the relationship between the findings and the dose dependency should be examined. Dose dependency is often examined by the Cochran-Armitage trend test after Fisher's probability test or chi-square test. Kobayashi and Pillai (2003) proposed a method to examine both the degree of pathology lesions and the dose dependency. In this method, the pathology findings are scored in grades and analyzed by the rank sum test. For comparison between two groups, Mann-Whitney's or Wilcoxon's test, and for comparison among several groups, Dunnett's, Tukey's, Duncan's, Scheffe's, Wilcoxon's or Williams-Wilcoxon's non-parametric tests are proposed. However, the number of animals necessary to detect a significant difference between the low dose group and the control group greatly varies with these tests. Dunnett's multiple comparison test can detect a significant difference even with four animals per group when the dose dependency is very high. The authors suggested Jonckheere's trend test and Spearman's correlation coefficient (*r*) for examination of dose-dependency.

Analysis of Pathology Data of Carcinogenicity Studies

The objectives to be achieved as per the guidelines of OECD (2009) for rodent carcinogenicity studies are hazard characterization, describing the dose-response relationship and the derivation of an estimate of a point of departure such as the Benchmark dose or a no observed adverse effect level. Normally, carcinogenicity studies are conducted in rodents with a control group and 2 or 3 treatment groups, each group containing a minimum of 50 animals of each gender. Mice are normally exposed to the test compound for 18–24 months, whereas rats are exposed for 24–30 months. Animals are sacrificed at intervals or at the end of the experiment. The major observations carried out in a carcinogenicity study are the survival time and status (presence/absence) of specific tumour types.

National Toxicology Programme (NTP) and U.S. Food and Drug Administration (US FDA), reported that there were issues in the application of statistical methods to carcinogenicity studies (Gad and Rousseaux, 2002). Tumour incidence (tumour incidence is defined as the rate of tumour onset among the tumour-free population) is considered the most appropriate measure of tumourigenesis (Malani and Van Ryzin, 1988; Dinse, 1994). Tumours can be classified as 'incidental,' 'fatal,' and 'mortality-independent (or observable)' according to the contexts of observation described by Peto *et al.* (1980). Tumours that are not directly or indirectly responsible for the animal's death, but are merely seen at the autopsy of the animal after it has died of an unrelated cause, are said to have been observed in an incidental context. Tumours that kill the animal, either directly or indirectly, are said to have been observed in a fatal context. Tumours, such as skin tumours, whose detection occurs at times other than when the animal dies are said to have been observed in a mortality-independent context (Lin, 2000). Benign and malignant tumours should be analysed separately (Mc Connell *et al.,* 1986; EPA, 2005), if it is considered scientifically defensible, further statistical analysis may be performed on the combined benign and malignant tumours of the same histogenic origin, even when those tumours are in different tissues.

Peto test

While most pharmaceutical companies use the Peto test (Peto *et al.*, 1980), some do not categorize neoplasms as fatal or incidental. Generally, this test is considered to be useful for the groups with different survival rates. Before analysing, pathological findings should be examined (whether malignant of benign) and conclude whether the drug caused the death or not. Some categorize neoplasms as fatal or incidental based solely on the type of neoplasm rather than on an animal-by-animal basis. Others categorize neoplasms as fatal or incidental based on the gross and microscopic findings for each animal. Some controversies exist when relying on the Peto test for information on 'cause of death' (STP, 2002).

According to Lee *et al.* (2002), the 'fatal' definition is often misunderstood by the pathologists and there is a tendency for the over-designation of fatal tumours (Kodell *et al.*, 1982; Ahn *et al.*, 2000).

The US FDA recommends that both trend test and pair-wise comparison test be performed routinely for each study and that the results of both tests should be presented to regulatory officials (FDA, 2001). However, the Peto test is required for product registration in Europe. Based on current regulatory requirements, the STP recommends that the Peto test should be performed whenever the study pathologist and the peer review pathologist can consistently classify neoplasms as fatal or incidental (Morton *et al.*, 2002).

Decision rules

A distinguished characteristic of the Peto test is that it involves dosages in the calculation procedure. The power of the Peto test is very high, when the significance level is set at 5% probability level. However, the use of significance set at 5% and 1% probability levels in tests for positive trend in incidence rates of rare tumours and common tumours, respectively, will result in an overall false positive rate around 10% in a study in which only one 2-year rodent bioassay (plus the shorter rodent study) is conducted (Lin, 1998; Lin and Rahman, 1998). The power to detect a significant difference is greater with the trend tests than with the pair-wise comparisons in an animal experiment with a control group and more than two treatment groups. There are situations in which pair-wise comparisons between control and individual treated groups may be more appropriate than trend tests. However, both trend and pair-wise comparison tests are likely to cause false positive results. In order to control overall positive rates associated with trend tests and pair-wise comparisons certain statistical decision rules were developed (Haseman,1983). The decision rules were developed based on historical control data of Crl: CDÒ BR rats and Crl: CD-1Ò (ICR) BR mice to achieve an overall false positive rate of around 10% for the standard *in vivo* carcinogenicity studies in rodents. The decision rule tests the significance difference in tumour incidences between the control and the treatment groups at 5% probability level for rare tumours (tumours with background rate of 1% or less) and at 1% probability level for common tumours (frequent tumours). However, the decision rule described by Haseman (1983) to analyse the trend tests would lead to an excessive overall false positive error rate about twice as large as that associated with control-high dose pair-wise comparison tests. Statistical decision rules for controlling the overall false positive rates associated with tests for positive trend or with control *vs* high dose pair-wise comparison in tumour incidences in carcinogenicity studies were reported by FDA (2001). These decision rules test positive trend in tumour incidence at 2.5% probability level for rare tumours and at 0.5% probability level for common tumours. Although the overall false positive rate resulting from the use of the decision rule may vary from study to study, it is estimated that it will be around 10%.

The decision rules for testing positive trend or differences between control and individual treatment groups in incidence rates of tumours for standard studies using two species and two sexes as well as studies following ICH guidance and using only one 2-year rodent bioassay are summarized in Table 17.1.

Table 17.1. Statistical decision rules for controlling the overall false positive rates associated with tests for positive trend or with control *vs* high dose pair-wise comparisons in tumour incidences to around 10 percent in carcinogenicity studies of pharmaceuticals (FDA, 2001).

Study	Tests for positive trend	Control *vs* high dose pair-wise comparison
Standard 2-year studies with 2 species and 2 sexes	Common and rare tumours are tested at 0.5% and 2.5% probability levels, respectively	Common and rare tumours are tested at 1% and 5% probability levels, respectively
Alternative ICH studies (one two-year study in one species and one short- or medium-term study, two sexes)	Common and rare tumours are tested at 1% and 5% probability levels, respectively	Under development and not yet available.

Note: The decision rules were developed assuming the use of two-species and two-sex (or one-species and two-sex) for the standard design of a two-year study with 50 animals in each of the four treatment/sex/group.

Poly-k Type test

An alternative to the Peto-type is Poly-*k* type test (Bailer and Portier, 1988; Portier and Bailer, 1989; Piegorsch and Bailer, 1997). One advantage of this test is that it does not require the controversial 'cause of death' in the calculation procedure. NTP uses the Poly-*k* test to assess neoplasm and non-neoplastic lesion prevalence.

Analysis of Tumour Incidence—Comparison with Historical Control Data

Tumour incidence between the treatment group and control group is normally compared using Fisher's probability test. By this test, no significant difference in tumour incidence is observed between the treatment group and control group, if the incidence of tumour is 0/50 (number of animals in the group having tumour/total number of animals in the group) in the control group and 4/50 in the treatment group. However, a tumour incidence of 4/50 is considered to be significant from a pathological viewpoint. Comparison of the incidence of tumour in the treatment group with that of the historical control data may be useful, especially to assess the occurrence of rare tumours and marginally increased tumour incidences. But, certain requirements must be met before the use of historical control data, since the historical control data may change in time (Greim *et al.*, 2003). Several procedures have been proposed for incorporating historical control data into the analysis of data obtained from carcinogenicity studies (Sun, 1999). If the data of the treatment group is compared with the historical control data using *t*-test, it should be remembered that the number of animals used in these groups is different, being much larger in the historical control group, since the source of historical control data is several studies. Table 17.2 shows a comparison of incidence of tumour observed in 50 animals in the treatment group with several historical controls having differences in number of animals but with similar tumour incidence (%).

Table 17.2. Comparison of treatment group with historical control data using Kastenbaum and Bowman test (Kastenbaum and Bowman, 1966)

Incidence of tumour	Incidence of tumour in 50 animals (Treatment group)			
(Historical control data[a])	1 (2%)	2 (4%)	3 (6%)	4 (8%)
1/ 200 (0.5%)	NS	NS	NS	NS
2/ 500 (0.4%)	NS	NS	NS	*
3/ 700 (0.4%)	NS	NS	NS	**
4/1000 (0.4%)	NS	NS	*	**
5/1250 (0.4%)	NS	NS	*	**
7/1500 (0.5%)	NS	NS	*	**
7/1700 (0.4%)	NS	NS	*	**
8/2000 (0.4%)	NS	NS	*	**
10/2500 (0.4%)	NS	NS	*	**

[a] Number of animals in the historical controls showing tumour/total number of animals in the historical controls; NS-Not significance, *P<0.05, **P<0.01.

The incidence of tumour in 1 or 2 animals out of 50 animals in the treatment group is not significantly different compared with the historical control animals showing the tumour incidence in 1 animal out of 200 or 10 out of 2500 animals. However, the incidence of tumour seen in 3 animals out of 50 animals in the treatment group is significantly different from the historical control animals with incidences of tumour as 4/1000, 5/1250, 7/1500, 7/1700, 8/2000 and 10/2500 (number of animals showing incidence of tumour/total number of animals). The incidence of tumour 8% (4/50) in the treatment group is significantly different from the historical control data showing the incidence of tumour as 2/500, 3/700, 4/1000, 5/1250, 7/1500, 7/1700, 8/2000 and 10/2500. It is obvious from the Table 17.2 that the number of animals used in constructing the historical control data plays a crucial role in determining a significant difference between the historical control data and the treatment group.

The circumstances that prompted the use of historical control for the analysis of carcinogenicity data should be properly explained and justified. It must be remembered that the concurrent control group is the most relevant comparator for determining treatment-related effects in a study (FDA, 2001; EMEA, 2002; OECD, 2002). In evaluating the data from historical controls, statistically significant increases in tumours based on the concurrent control should not be discounted simply because incidence rates in the treatment groups are within the range of historical controls or because incidence rates in the concurrent controls are low (Keenan *et al.*, 2009). OECD guidelines (OECD, 2002) emphazise the historical control data should be generated by the same laboratory in animals of contemporaneous studies in the same species and strain, maintained under similar conditions, at which the study being assessed was performed. Furthermore, the historical control data should come from studies conducted within five years prior to, or within two to three years from the conclusion of the study. The guidelines recommend parameters that could affect the occurrence of spontaneous tumours in historical control data are identified. In studies exhibiting the lowest incidence (less than a few percent) of tumours, the Kastenbaum and Bowman test appears to be more relevant, since it takes into account the sample size of both the historical control data base and each treatment group in the study. In studies where a wider range of tumour incidence is exhibited, a statistical method which employs a rejection limits based on the range of incidence in the historical data is recommended. When malignant tumours are evident in treatment groups, no matter how low the incidence, the tumour should be analyzed statistically and compared with the incidence in the historical control data as well as those in the concurrent control group (Kobayashi and Inoue, 1994).

Analysis of Incidence of Tumour Using X^2 Test

Chi square test is an excellent tool to evaluate the significant difference in occurrence of tumours among the groups. An example is given in Table 17.3.

Table 17.3. Total number of occurrence of tumours in different organs in a two-year carcinogenicity study

Control	Low dose	Mid dose	High dose	Total
58	50	62	65	235

Note: Each group consists of 50 animals.

$$X^2 = \frac{58^2}{235 \times 0.25} + \frac{50^2}{235 \times 0.25} + \frac{62^2}{235 \times 0.25} + \frac{65^2}{235 \times 0.25} - 235 = 2.157$$

Note: 0.25=1/4: Assumed probability distribution (4=Number of groups).

The chi-squared Table value for 3 degrees of freedom is 7.82 at 5% probability level. The calculated value 2.157 is less than 7.82, which means that there is no significant differences in the occurrence of tumours among the groups. If a significant difference is observed, difference between control and each group is analyzed.

However, use of chi-square goodness-of-fit in multistage model to carcinogenicity has been questioned in recent years. According to Sielken (1988) "although the chi-square goodness-of-fit is a very widely used statistical test, it is also well documented (though not sufficiently widely known) that the test can have very little power to reject inaccurate models".

Comparison of Incidence of Tumours in Human, Rats, Mice and Dogs

Considerable debate about the need of conducting carcinogenicity studies in rats and mice has been taken place in recent years (Ennever and Lave, 2003; Billington *et al.*, 2010; Storer *et al.*, 2010). Most of the scientists are of the opinion that there is no need to conduct long-term rodent carcinogenicity studies in mice, since the use of the mice in carcinogenicity testing does not provide useful scientific information (Griffiths *et al.*, 1994;

Carmichael *et al.*, 1997; Meyer, 2003; Doe *et al.*, 2006). However, some current regulatory programmes require carcinogenicity testing in rats and mice.

Kobayashi *et al.* (1999) made an interesting comparison of incidence of spontaneous malignant tumours in human, rats, mice and dogs. The prevalence of each carcinoma in rodents was calculated as the population ratio *P*, at a 95% confidence interval, and compared with that in humans. The primary carcinomas according to sex in Japanese people who died of cancer were cited from the report of investigations on the population dynamics and economy in 1992, "Malignant neoplasm" published by the Welfare Statistics Association, Japan (Ministers' Secretariat, 1994). Data on spontaneous incidence of tumours in rats, mice and dogs were obtained from Biosafety Research Centre—Foods, Drugs and Pesticides, Japan. The incidence of spontaneous malignant tumours of various organs in humans, rodents and dogs is shown in Table 17.4.

Table 17.4. Incidence (%) of spontaneous malignant tumours in dead humans, rodents and dogs

Organ	Male			Female			Male+Female
	Human	Rat	Mouse	Human	Rat	Mouse	Dog
No. of deaths with cancer	139674	105	120	92243	117	100	5845
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Esophagus	4.7	0	0	1.4	0	0	0.3
Stomach	21.8	1.0	0	19.0	0.9	1.0	0.3
Intestine[a]	4.4	0	0.89	4.3	0	0	1.0
Liver	14.0	0.	52.5	8.1	1.7	24.0	0.7
Pancreas	5.6	0	0	6.9	0.9	2.0	0.5
Lung, trachea, bronchi	20.9	2.9	5.0	11.9	0	1.0	0.6
Mammary gland	<0.1	0	0.	7.1	2.6	8.0	9.1
Uterus	-	-	-	5.1	10.3	11.0	0.3
Leukemia	2.4	53.3	20.8	2.69	59.8	31.0	4.3
Other	26.1	42.9	20.8	33.9	23.9	22.0	82.9

[a]Including colon and anus in humans, small intestine, duodenum, large intestine and colon in rodents, and colon in dogs.

The incidence of tumours in the organs of humans who died of cancer differed considerably from that of mice, rats and dogs. For example, very low or no incidence of tumour was seen in esophagus, stomach and intestine of rats and mice. The incidence of hepatocellular carcinoma in mice and leukemia in rats and mice were higher than those in humans, while the incidence of malignant tumours in the lungs of rodents was lower than those in humans. The authors stated it is important to consider the spontaneous tumours and the probable target organ when selecting the appropriate species for a carcinogenicity study.

Analysis of Organ Weight Data

Organ weights (absolute and relative organ weights) are an important quantitative end point in the repeated dose administration studies. Many pathologists are of the opinion that it would be better to calculate organ weight relative to brain weight (organ-to-brain weight ratio) rather than to body weight (organ-to-body weight ratio). Animals are usually fasted before necropsy. The deprivation of food can affect the body weight of the animals, and also the physiological adaptability to fasting may vary significantly among the animals. When the body weight gain is affected, alterations of organ weight/body weight ratio may be due to the physiological response of the animal to decreased nutrient intake. Organ-to-body weight ratios are preferable for analysis of liver and thyroid weights, whereas organ-to-brain weight ratios are best for analysis of ovary and adrenal weights, and both organ-to-body weight ratios and organ-to-brain weight ratios do not accurately model brain, heart, kidney, pituitary, or testis weights (Bailey *et al*., 2004). Regardless of the study type or organs evaluated, organ weight changes must be evaluated within the context of the compound class, mechanism of action, and the entire data set for that study (Sellers *et al*., 2007).

Absolute weight of the mouse liver in a 13 week repeated dose administration study is given in Table 17.5.

Table 17.5. Absolute weight (g) of the mouse liver in a 13 week repeated dose administration study

Group	Control	Low Dose	Mid Dose	High Dose
Individual value	1.08, 1.09, 1.15, 1.09, 1.16, 1.00, 1.12, 1.01, 1.12, 1.02	1.09, 1.12, 1.15, 1.09, 1.04, 0.99, 1.24, 1.15, 0.99, 1.12	1.10, 1.20, 1.09, 1.02, 1.07, 1.12, 1.13, 1.06, 1.11, 1.20	1.16, 1.15, 1.24, 1.16, 1.22, 1.10, 1.18, 1.07, 1.18, 1.09
n	10	10	10	10
Mean ± SD	1.08 ± 0.06	1.10 ± 0.08	1.11± 0.06	1.16 ± 0.05
In % Control	100	102	103	107

Since there are four groups, the data is analysed using one-way ANOVA, which shows a non-significant *F* value, indicating that there is no significant difference in the absolute weight of the liver among the groups. Close examination of the mean value of the groups indicates that there is a dose-dependent increase in the absolute weight of the liver. When the data is analysed using Dunnett's multiple comparison test, absolute weight of the liver of the high dose group is found to be significantly different from the control group. It may be worth mentioning in this context that Dunnett (1964) did not recommend ANOVA prior to multiple comparison tests. Several authors are of the opinion that the error of second kind can be prevented by carrying out direct multiple comparison tests without subjecting the data to ANOVA (Hamada *et al.*, 1998; Sakaki *et al.*, 2000; Kobayashi *et al.*, 2000).

Interpretation of Pathology Observations

Interpretations made from the organ weight data should be used with caution. Indicating a significant or non-significant difference in organ weight alone by statistical analysis, particularly in studies with small size, has little use in evaluating the organ weight changes (Sellers *et al.*, 2007). According to Gad and Rousseaux (2002), treatment-related alterations in organ weight may not be statistically significant, similarly statistically significant alteration in organ weight may not be treatment related.

In the long-term toxicology studies, animals may show age-associated changes, which can have a significant effect on histopathology (Mohr *et al.*, 1992, 1994, 1996). Spontaneous degenerative lesions, especially when misinterpreted as toxic effects can cause major difficulty in hazard evaluation. In these situations, the data can be compared with historical control data. It has been stated that historical control tumour data is useful in the interpretation of long-term rodent carcinogenicity bioassays, especially to assess the occurrence of rare tumours and marginally increased tumour incidences (Deschl *et al.*, 2002). However, the advantage of a concurrent control as the comparator for treatment-related effects should not be overlooked, when historical control data are used as the comparator.

References

Ahn, H., Kodell, R.L. and Moon, H. (2000): Attribution of tumour lethality and estimation of time to onset of occult tumours in the absence of cause-of-death information. App. Stat., 49, 157–169.

Bailer, A.J. and Portier, C.J. (1988): Effects of treatment-induced mortality and tumour-induced mortality on tests for carcinogenicity in small samples. Biometrics, 44, 417–431.

Bailey, S.A., Zidell, R.H. and Perry, R.W. (2004): Relationships between organ weight and body/brain weight in the rat: what is the best analytic endpoint? Toxicol. Pathol., 32, 448–466.

Billington, R., Lewis, R., Mehta, J. and Dewhurst, I. (2010): The mouse carcinogenicity study is no longer a scientifically justifiable core data requirement for the safety assessment of pesticides. Crit. Rev.Toxicol., 40, 35–49.

Carmichael, N.G., Enzmann, H., Pate, I. and Waechter, F. (1997): The significance of mouse liver tumour formation for carcinogenic risk assessment: Results and conclusions from a survey of ten years of testing by the agrochemical industry. Environ. Health Perspect., 105, 1196–1203.

Deschl, U., Kittel, B., Rittinghausen, S., Morawietz, G., Kohler, M., Mohr, U. and Keenan, C. (2002): The value of historical control data—Scientific advantages for pathologists, industry and agencies. Toxicol. Pathol., 30, 80–87.

Dinse, G.E. (1994): A comparison of tumour incidence analyses applicable in single-sacrifice animal experiments. Stat. Med., 13, 689–708.

Doe, J.E., Boobis, A.R., Blacker, A., Dellarco, V., Doerrer, N.G., Franklin, C., Goodman, J.I., Kronenberg, J.M., Lewis, R., Mcconnell, E.E., Mercier, T., Moretto, A., Nolan, C., Padilla, S., Phang, W., Solecki, R., Tilbury, L., van Ravenzwaay, B. and Wolf, D.C. (2006): A tiered approach to systemic toxicity testing for agricultural chemical safety assessment. Cri. Rev. Toxicol., 36, 37–68.

Dunnett, C.W. (1964): New tables for multiple comparisons with a control. Biometrics, 20(3), 482–491.

EMEA (2002): European Medicines Agency. CPMP, Note for guidance on carcinogenic potential, EMEA, CPMP/SWP/2877/00, London, 25 July 2002. http://www.emea.europa.eu/pdfs/human/swp/287700en.pdf.

Ennever, F.K. and Lave, L.B. (2003): Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. Reg. Tox. Pharm., 38, 52–57.

EPA (2005): United States Environmental Protection Agency. Guidelines for Carcinogen Risk Assessment. U.S. Environmental Protection Agency (USEPA), Washington DC, USA.

FDA (2001): Food and Drug Administration. Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals. Draft Guidance, US FDA, Rockville, MD, USA.

Gad, S.C. and Rousseaux, C.G. (2002): Use and misuse of statistics in the design and interpretation of studies. In: Handbook of Toxicologic Pathology, 2nd Edition. Editors, Haschek, W.M., Rousseaux, C.G. and Wallig, M.A. Academic Press, San Diego, USA.

Glaister, J.R. (1986): Principles of Toxicological Pathology. Taylor & Francis, Philadelphia, USA.

Goodman, D.G. (1988): Factors Affecting Histopathologic Interpretation of Toxicity-Carcinogenicity Studies. Carcinogenicity: The Design, Analysis, and Interpretation of Long-Term Animal Studies. ILSI Monographs, Springer-Verlag, New York, USA.

Greim, H., Gelbke, H.P., Reuter, U., Thielmann, H.W. and Elder, L. (2003): Evaluation of historical control data in carcinogenicity studies. Hum. Exp. Toxicol., 22(10), 541–549.

Griffiths, S.A., Parkinson, C., McAuslane, J.A.N. and Lumley, C.E. (1994): The utility of the second rodent species in the carcinogenicity testing of pharmaceuticals. Toxicologist, 14(1), 214.

Hamada, C., Yoshino, K., Matsumoto, K., Nomura, M. and Yoshimura, I. (1998): Three-type algorithm for statistical analysis in chronic toxicity studies. J. Toxicol. Sci., 23 (3), 173–181.

Haseman, J.K. (1983): A reexamination of false-positive rates for carcinogenesis studies. Fund. Appl. Toxicol., 3, 334–339.

Haseman, J.K., Huff, J. and Boorman, G.A. (1984): Use of historical control data in carcinogenicity studies in rodents. Toxicol. Pathol*.,* 12, 126–135.

House, D.E., Berman, E., Seely, J.C. and Simmons, J.E. (1992): Comparison of open and blind histopathologic evaluation of hepatic lesions. Toxicol. Lett*.,* 63, 127–133.

Iatropoulos, M.J. (1984): Editorial : Toxicol. Pathol., 12(4), 305–306.

Iatropoulos, M.J. (1988): Society of Toxicologic Pathologists Position Paper: "Blinded" Microscopic Examination of Tissues from Toxicologic or Oncogenic Studies," In: Carcinogenicity, the Design, Analysis, and Interpretation of Long-Term Animal Studies, ILSI Monographs, Editros, Grice, H.C. and Ciminera, J.L., Spring-Verlag, New York, USA.

Kastenbaum, M.A. and Bowman, K.O. (1966): The minimum significant number of successes in a binominal sample. Oak Ridge National Laboratory (ORNL-3909), Oak, Tennessee, USA.

Keenan, C., Elmore, S., Francke-Carroll, S., Kemp, R., Kerlin, R., Peddada, S., Pletcher, J., Rinke, M., Schmidt, S.P., Taylor, I. and Wolf, D.C. (2009): Best practices for use of historical control data of proliferative rodent lesions. Toxicol. Pathol., 37, 679–693.

Kobayashi, K., Hagiwara, T., Miura, D., Ohori, K., Takeuchi, H., Kanamori, M. and Takasaki, K. (1999): A comparison of spontaneous malignant tumours in humans, rats, mice and dogs. J. Environ. Biol., 20(3), 189–193.

Kobayashi, K. and Inoue, H. (1994): Statistical analytical methods for comparing the incidence of tumours to the historical control data. J. Toxicol. Sci., 19(1), 1–6.

Kobayashi, K., Kanamori, M., Ohori, K., and Takeuchi, H. (2000): A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies on rodents. San Ei Shi, 42, 125–129.

Kobayashi, K. and Pillai, K.S. (2003): Applied Statistics in Toxicology and Pharamacology, Science Publishers, Enfield, USA.

Kodell, R.L, Farmer, J.H., Gaylor, D.W. and Cameron, A.M. (1982): Influence of cause-of-death assignment on time-to-tumour analyses in animal carcinogenesis studies. J. Natl. Cancer Inst., 69, 659–664.

Krinke, G.J., Perrin, L.P.A. and Hess, R. (1991): Assessment of toxicopathological effects in ageing laboratory rodents. Arch. Toxicol. Suppl., 14, 43–49.

Lee, P.N., Fry, J.S., Fairweather, W.R., Haseman, J.K., Kodell, R.L., Chen, J.J., Roth, A.J., Soper, K. and Morton, D. (2002): Current issues: statistical methods for carcinogenicity studies. Toxicol. Pathol., 30, 403–414.

Lin, K.K. (1998): CDER/FDA formats for submission of animal carcinogenicity study data. Drug Information J., 32, 43–52.

Lin, K.K. (2000): Progress report on the guidance for industry for statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. J. Biopharm. Stat., 10(4), 481–501.

Lin, K.K. and Rahman, M.A. (1998): False Positive Rates in Tests for Trend and Differences in Tumour incidence in Animal Carcinogenicity Studies of Pharmaceuticals under ICH Guidance S1B, Unpublished Report, Division of Biometrics 2, Center for Drug Evaluation and Research, Food and Drug Administration. USFDA, MD, USA.

Malani, H.M. and Van Ryzin, J. (1988): Comparison of two treatments in animal carcinogenicity experiments. J. Am. Stat. Assoc., 83, 1171–1177.

Mc Connell, E.E., Solleveld, H.A., Swenberg, J.A. and Boorman, G.A. (1986): Guidelines for combining neoplasms for evaluation of rodent carcinogenicity studies. J. Natl. Cancer Inst., 76, 283–289.

Meyer, O. (2003):Testing and assessment strategies, including alternative and new approaches. Toxicol. Lett., 140–141, 21–30.

Ministers' Secretariat (1994): The Report of Investigation on Population Dynamics and Social Economy in 1992 "Malignant Neoplasm", Welfare Statistics Association, Tokyo, Japan.

Mohr, U., Dungworth, D.L. and Capen, C.C. (1992): Pathobiology of the Aging Rat. Vol. 1. ILSI Press, Washington, DC, USA.

Mohr, U., Dungworth, D.L. and Capen, C.C. (1994): Pathobiology of the Aging Rat. Vol 2. ILSI Press, Washington, DC, USA.

Mohr, U., Dungworth, D.L., Ward, J., Capen, C.C., Carlton, W. and Sundberg, J. (1996): Pathobiology of the Aging Mouse. Vols. 1 & 2. ILSI Press, Washington, DC, USA.

Morton, D., Elwell, M., Fairweather, W., Fouillet, X., Keenan, K., Lin, K., Long, G., Mixson, L., Morton, D., Peters, T., Rousseaux, C. and Tuomari, D. (2002): The Society of Toxicologic Pathology's recommendations on statistical analysis of rodent carcinogenicity studies. Toxicol. Pathol*.,* 30(3): 415–418.

Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Newberne, P.M. and de la Lglesia, F.A. (1985): Editorial: Philosophy of blind slide reading. Toxicol. Pathol*.,* 13(4), 255.

OECD (2000): Organisation for Economic Cooperation and Development. Environment Directorate Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. OECD Series on Testing and Assessment, Number 32 and OECD Series on Pesticides, Number 10, ENV/JM/MONO(2000)18, Paris, France.

OECD (2002): Organisation for Economic Cooperation and Development. Environment, Health and Safety Publications. Series on Testing and Assessment No. 35 and Series on Pesticides No. 14. Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies. ENV/JM/MONO 19, Paris, France.

OECD (2009): Organization for Economic Cooperation and Development. Guidelines for Testing of Chemicals. Carcinogenicity Studies. Test Guideline 451. OECD, Paris, France.

Peto, R., Pike, M., Day, N.E., Gray, R.G., Lee, P.N., Parish, S., Peto, J., Richards, S. and Wahrendorf, J. (1980): Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-term Animal Experiments. In: IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans, Supplement of Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. International Agency for Research on Cancer, Lyon, France.

Piegorsch, W.W. and Bailer, A.J. (1997): Statistics for Environmental Biology and Toxicology, Chapman and Hall, London, UK.

Portier, C.J. and Bailer, A.J. (1989): Testing for increased carcinogenicity using a survival-adjusted quantal response test. Fundam. Appl. Toxicol.,12, 731–737.

Prasse, K., Hildebrandt, P. and Dodd, D. (1986): Letter to the Editor: Vet. Pathol*.,* 23, 540–541.

Sakaki, H., Igarashi, S., Ikeda, T., Imamizo, K., Omichi, T., Kadota, M., Kawaguchi, T., Takizawa, T., Tsukamoto, O., Terai, K., Tozuka, K., Hirata, J., Handa, J., Mizuma, H., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Sellers, R.S., Mortan, D., Michael, B., Roome, N., Johnson, J.K., Yano, B.L., Perry, R. and Schafer, K. (2007): Society of Toxicologic Pathology Position Paper: Organ weight recommendations for toxicology studies. Toxicol. Pathol., 35(5), 751–755.

Sielken, R.L. (1988): A critical evaluation of dose-response assessment of TCDD. Food Chem. Toxicol., 26(1), 79–83.

Storer, R.D., Sistare, F.D., Reddy, M.V. and DeGeorge, J.J. (2010): An industry perspective on the utility of short-term carcinogenicity testing in transgenic mice in pharmaceutical development. Toxicol. Pathol., 38, 51–61.

STP (2002): STP Peto Analysis Working Group (2002). The Society of Toxicological Pathology's recommendations on rodent carcinogenicity studies. Toxicol. Pathol., 30, 415–418.

Sun, J. (1999): On the use of historical control data for trend test in carcinogenicity studies. Biometrics, 55, 1273–1276.

Tyson, C.A. and Sawhney, D.S. (1985): Organ Function Tests in Toxicology Evaluation. Noyes Publications, Park Ridge, New Jersey, USA.

Wade, A. (2005): Fear or favour? Statistics in pathology. J. Clin. Pathol., 53,16–18.

18

Designing An Animal Experiment in Pharmacology and Toxicology—Randomization, Determining Sample Size

Designing Animal Experiments

The use of animals raises scientific and ethical challenges (Workman *et al.*, 2010). Therefore, an animal experiment should be designed with due consideration to ethics on a solid scientific platform. Animal experiment should have high precision, but should not waste resources or animals (Festing, 1997). It is important to select an appropriate study design to provide scientific evaluation of the research findings without bias (Lim and Hoffmann, 2007). Replication, randomization and blinding are the key components of the design of the animal experiment. But, these are less often used in animal research (Kilkenny *et al.*, 2009). Hess (2011) reviewed statistical design given in 100 articles on animal experiments published in Cancer Research in 2010. In 14 of the 100 articles, the number of animals used per group was not reported. In none of the 100 articles was the method employed to determine the number of animals used per group reported. Among the 74 articles in which randomization seemed feasible, only 21 reported that they had randomly allocated animals to various groups. None of these articles described how the randomization was carried out.

In animal experiments, bias could arise from lack of randomization, not blinding the groups, failure to report excluded animals, small sample sizes or use of statistical tools with low power (Dirnagl and Macleod, 2009). If there is a large difference between the treatment group and control of a well designed study, an experienced analyser can draw a conclusion without carrying out a statistical analysis of the data. But, if the difference is marginal, a mistaken or a biased conclusion could be avoided by subjecting the data to the statistical analysis (Lew, 2007).

It has been stated that several reports on animal experiments were biased or did not correctly model human disease and therefore were of little utility (Festing, 2003; Perel *et al.*, 2007). Though the findings of most of the animals studies cannot be directly extrapolated to man, a properly designed study may provide vital information on efficacy and toxicity of the test substance. Acclimation and randomization procedures of animals, and rationale for fixing the number of animals in a group should be explained in the study plan. There are additional issues such as rationale for selection of species, animal house conditions, bedding material, diet, drinking water, *etc.*, which need to be considered in the study plan, but beyond the scope of this book.

Acclimation

It should be ensured that the animals are not stressed at the start of the experiment. One way to ensure this is by acclimating the animals to the laboratory conditions. The acclimation period can be used for health-related quarantine and monitoring, and for behavioral conditioning. This period may include habituation to, desensitization to, and training for procedures that will be involved in experimental use (Bloomsmith *et al.*, 2006). Well-acclimated animals are able to deal appropriately with the challenges of the experimental environment. This ability is typically manifested in a transient divergence from equilibrium in response to a manipulation, followed by a gradual return to homeostatic balance (Schapiro and Everitt, 2006). Animals appearing to be behaviorally acclimated to a procedure may not necessarily physiologically acclimated to that procedure (Capitanio *et al.*, 2006). For example, acclimated animals may sometimes show change in metabolic profiles. Changes in nuclear magnetic resonance spectroscopic-based urinary metabolite profiles were observed in germ-free rats acclimated in standard laboratory animal facility conditions (Nicholls *et al.* 2003).

Randomization

Appropriate randomization and statistical procedures in the design of animal experimentation provide confidence that statistically significant results are not due to chance (EPA, 2005). Selection of an appropriate statistical tool is heavily depended on randomization, which is a fundamental element of good statistical design that acts to reduce potential bias during treatment allocation (Festing and Altman, 2002).

Infact the concept of randomization originated as early as 1935 (Fisher, 1935). Randomization transforms systematic errors into random errors and confirms comparability among experimental groups (Hamada and Ono, 2000). Though randomization is an important aspect in designing animal experiments, little consideration is given to it in most cases. This is evident from different terminologies that are used for randomization, like "animals were divided into four groups"; "animals were randomly divided"; "animals were sorted into groups"; "animals were randomly assigned"; and, "half of the animals were placed into one group and the other half in a second group" (Kozinetz, 2011). The key deficiencies that are seen in animal experiments are failure to randomly allocate animals to treatments and failure to blind observers to treatment assignment during outcome assessments (Hess, 2011). Failure of NXY-059, a neuroprotective agent for stroke patients, of Astra Zeneca, in Phase III has been attributed to improper randomization and bias in preclinical studies (Savitz, 2007). When comparing two treatments, analyser-related bias may occur. This bias can be avoided by blinding (Aguilar-Nascimento, 2005). In a clinical trial, blinding can take place at three levels: study units, researcher and data (Lim and Hoffmann, 2007). The same method can be applied to animal experiments also. In a blinded study, the researcher does not know which group of animals receives what treatment. According to Bebarta *et al.* (2003), "animal experiments where randomization and blind testing are not reported are five times more likely to report positive results". Therefore, effects of randomization have to be considered in planning and performing experiments as well as in the interpretation of experimental results (Vogt and Kloting, 1990).

In toxicological experiments, especially in repeated dose administration studies, young adult animals of an inbred strain are used. Though the animals of inbred strain are supposed to be isogenic, in reality it is not so. There could be some genetic variation between the individuals from one litter and the other. Let us work out an example. Body weight of rats from 3 litters is given in Table 18.1.

Let us randomly distribute the animals of litters 1, 2 and 3 into three groups. An unbiased randomization should distribute the variation of

Table 18.1. Body weight (g) of rats from 3 litters

Statistic	Litter 1	Litter 2	Litter 3
	180^{1}	195^{2}	210^{3}
	185^{1}	205^{2}	193^{3}
	189^{1}	215^{2}	190^{3}
	198^{1}	213^{2}	208^{3}
	203^{1}	211^{2}	201^{3}
Mean	191.00	207.80	200.40
CV (%)	4.93	3.89	4.42

Note: Superscripts indicate litter number.

animals of litter 1 more or less equally to the animals of litters 2 and 3. Similarly, an unbiased randomization should distribute the variation of animals of litter 2 more or less equally to the animals of litters 1 and 3 and so on. This can be achieved if the randomization results in an equal representation of animals from all the three litters to each group.

Just for academic interest the data (Table 18.1) was analysed using one-way ANOVA, and found that there is a significant difference in body weight among the groups.

Assign an arbitrary identification number to each animal and with the help of a random number table randomize the animals into three groups (Table 18.2).

One-way ANOVA of the above data (Table 18.2) resulted in a non-significant *F* value, indicating that the body weight of the rats did not differ among the groups. Strictly speaking, the randomization procedure is completed, but some researchers rearrange the animals among the groups, as explained below, to obtain a uniform mean value. On closely examining the mean values one should be satisfied with the mean values of Groups 1 and 2 since they are somewhat close to each other, but one should be concerned about the mean value of group 3, which deviates

Table 18.2. Body weight (g) of rats after randomization

Statistic	Group 1	Group 2	Group 3
	198^{1}	213^{2}	185^{1}
	205^{2}	189^{1}	193^{3}
	210^{3}	215^{2}	195^{2}
	201^{3}	208^{3}	190^{3}
	203^{1}	211^{2}	180^{1}
Mean	203.40	207.20	188.60
CV (%)	2.22	5.07	3.24

Note: Superscripts indicate litter number.

considerably from the mean values of groups 1 and 2, particularly of group 2. This can be overcome by selecting one or two animals based on their body weight from each group and distributing them in other groups in such a way that the mean values of all the groups are more or less similar.

One way to reduce the mean value of group 2 and increase the mean value of group 3 is to take out the rat with the largest body weight from group 2 (215 g) and place it in group 3 and take out the rat with the smallest body weight from group 3 (180 g) and place it to group 2. Now the animals are distributed as given in Table 18.3.

Table 18.3. Body weight (g) of rats after rearranging the animals (first time)

Statistic	Group 1	Group 2	Group 3
	198^1	213^2	185^1
	205^2	189^1	193^3
	210^3	180^1	195^2
	201^3	208^3	190^3
	203^1	211^2	215^2
Mean	203.40	200.20	195.60
CV (%)	2.22	7.39	5.87

Note: Superscripts indicate litter number.

One-way ANOVA of the data given in Table 18.3 indicates that there is no significant different in body weight of rats among the groups. This is still not satisfactory for few researchers. The difference of the body weight between groups 1 and 3 is about 8 g. In order to bring the mean body weight of these two groups closer, one more adjustment is required. A rat of 210 g is taken from group 1 and placed in group 3. Then a rat of 185 g is taken from group 3 and placed in group 1. Now the animals are distributed as given in Table 18.4.

Table 18.4. Body weight (g) of rats after rearranging the animals (second time)

Statistic	Group 1	Group 2	Group 3
	198^1	213^2	210^3
	205^2	189^1	193^3
	185^1	180^1	195^2
	201^3	208^3	190^3
	203^1	211^2	215^2
Mean	198.40	200.20	200.60
CV (%)	3.99	7.39	5.56

Note: Superscripts indicate litter number.

The mean values of the three groups are very close to each other, thus satisfactory. If you closely observe the individual values of the groups, you will realize that Group 3 represents animals from litters 2 and 3 and Groups 1 and 2 represent animals from all the three litters. Rearrangement increases variation within the groups, consequently, the animals respond to a treatment differently. This is evident from the Tables 18.2 and 18.4. The variations (CV%) of groups 1, 2 and 3 after randomization, but before rearrangement were 2.22, 5.07 and 3.24, respectively (Table 18.2). After the rearranging the animals a second time, the variations (CV%) of groups 1, 2 and 3 were 3.99, 7.39 and 5.56, respectively (Table 18.4). Such variations reduce the power of the experiment (Beynen *et al.*, 2001). In the first randomization (Table 18.2), each group represented animals from all the litters and the variation (CV%) among the groups are less and somewhat close to each other. Therefore, rearrangements of observations after the randomization to obtain desired mean values should be avoided as far as possible.

Determining Sample Size

In regulatory toxicology, the guidelines clearly indicate the number of animals to be used in a group for a study (Hauschke, 1997). In the research and development of a pharmaceutical company, where a large number of new chemical entities (NCEs) are synthesized, often the scientists carry out experiments with 'inadequate number' of animals. Results from such studies may not be reproducible and may fail to provide the desired information on the effectiveness of the molecule.

Using too few animals in experiments will result in a low power to detect a biologically meaningful results. Similarly, the use of too many animals is not ethical and drain organization's resources unnecessarily. The right number of animals (not too few and not too many) required for obtaining a biologically meaningful result should be an important component of any animal experimental design. In an *in vivo* efficacy study, the number of animals required to obtain the desired result is determined based on certain specifications: the desired magnitude of treatment effect, the chance of obtaining Type I and Type II errors and the inter-individual variability.

An *in vivo* efficacy study is a comparison-oriented study. The comparison of the NCE-treated animals is usually done with the control animals, using an appropriate statistical analysis. The two errors which can occur in such comparisons are Type I error (α error) and Type II error (β error). Though much attention is given to α error, β error is often overlooked. β error is a very potential error in animal experiments and in certain situations more potential than α error. For example, in an *in vivo* experiment you are confident that there is a treatment-related effect, but the statistical analysis does not show it because of random variation. This is a typical example of β error that commonly occurs in animal experiments. A large β error is a risk in detecting a genuine difference. Power of a study to detect a significant difference is explained by this risk:

Power = 1–β

In simple language, the power is the probability of obtaining a statistically significant result using a statistical test (Lenth, 2007). In other words, power of the test is the probability of correctly rejecting the null hypothesis, when false. A study with a high power is unlikely to fail in detecting a genuine significant difference, whereas a study with a weak power may fail in detecting a genuine significant difference. The power of the tests can be improved by increasing α, sample size, or limiting the statistical analysis to detection of large differences among samples (Hayes, 1987).

To design an experiment to investigate the effect of a hypoglycemic NCE in diabetic rats, the blood sugar in the individual diabetic rat is measured before and after the treatment with the NCE. Then the difference in blood sugar level of the individual rat is calculated. Another group of animals treated similarly, but with a placebo is also maintained. Let us work out number of animals required in each group to obtain the desired result. For that specifications of the study need to be defined:

1. The significance level (probability of α error). Usually it is set at 5% probability level.
2. Probability of β error is set at 10%. The statistical power (1–β) is 90%.
3. The desired treatment effect (difference between NCE treated group and placebo treated group. This is determined based on the factors like clinical, economical etc.)
4. Estimate of expected variation (variation between individual measurements with respect to difference of before and after treatments. This is estimated based on earlier experiments of similar nature or a pilot study)
5. Type of statistical analysis (since there are only two groups, the *t*-test would be better).

Number of animals in each group by two-sided test can be calculated using the formula,

$$n= 2\left[\frac{(Z_{a/2} - Z\pi)^2}{(\mu_1 - \mu_2/\sigma)^2}\right]$$

Number of animals in each group by one-sided test can be calculated using the formula,

$$n= 2\left[\frac{(Z_{a} - Z\pi)^2}{(\mu_1 - \mu_2/\sigma)^2}\right]$$

Let us work out an Example; $\alpha = 0.05$, $\pi = 0.9$, Desired effect = 25%; σ =15% (CV).

$Z_\alpha = 1.645$ (*vide* Appendix 3 for $Z_{0.05}$)
$Z_\pi = 1.282$ (*vide* Appendix 3 for $Z_{0.10}$)

$$n= 2\left[\frac{(1.645+1.282)^2}{(25/15)^2}\right]$$ = 6.2; Number of animals required in each group = 7

Animal Experimental Designs

Accuracy of an animal experiment depends on the design of the experiment. An animal experimental design should be unbiased, should have high precision, wide range of applicability and should be simple in design (Cox, 1958). An animal experiment can be designed in several ways, for example, completely randomized design, randomized block design, cross-over design, Latin square design etc. The commonly used design in pharmacology and toxicology is randomized design. Other designs may be adopted, especially for *in vivo* efficacy studies with NCEs, where more than one NCE at more than two dose levels, a control group, a group treated with a commercially available drug with known efficacy are involved. Perhaps the most important thing to remember while designing an animal experiment is the prior knowledge of all the factors that could affect the outcome of the experiment.

References

Aguilar-Nascimento, J.E. (2005): Fundamental steps in experimental design for animal studies. Acta Cir. Bras., 20(1), 1–8.

Bebarta, V., Luyte, D. and Heard, K. (2003): Emergency medicine research: Does use of randomization and blinding affect the results? Acad. Emerg. Med., 10, 684–687.

Beynen, A.C., Festing, M.F.W. and van Montfort, M.A.J. (2001): Design of Animal Experiments. In: Principles of Laboratory Animal Science, Editors, van Zutphen, L.F.M., Baumans, V. and Beynen, A.C. Elsevier Science B.V., The Netherlands.

Bloomsmith, M.A., Schapiro, S.J. and Strobert, E.A. (2006): Preparing chimpanzees for laboratory research. ILAR J., 47, 316–325.

Capitanio, J.P., Kyes, R.C. and Fairbanks, L.A. (2006): Considerations in the selection and conditioning of old world monkeys for laboratory research: Animals from domestic sources. ILAR J., 47, 294–306.

Cox, D.R. (1958): Planning Experiments. John Wiley & Sons, New York, USA.

Dirnagl, U. and Macleod, M.R. (2009): Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. Br. J. Pharmacol., 157(7), 1154–1156.

EPA. (2005): United States Environmental Protection Agency. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001B. http://www.epa.gov/iris/backgr-d.htm.

Festing, M.F.W. (1997): Teaching statistics can save animals, In: Animal Alternatives, Welfare and Ethics. Edited by van Zuphen, L.F.M. and Balls, M. Elsevier Science B.V., Amsterdam, The Netherlands.

Festing, M.F.W. (2003): Principles: the need for better experimental design. Trends Pharmacol. Sci., 24, 341–345.

Festing, M.F.W. and Altman, D.G. (2002): Guidelines for the design and statistical analysis for experiments using laboratory animals. ILAR J., 43, 244–258.

Fisher, R.A. (1935): The Design of Experiments. 8th Edition, 1966. Hafner Press, New York, USA.

Hamada, C. and Ono, H. (2000): The role of biostatistics in pharmacological studies (randomization and statistical evaluation). Nihon Yakurigaku Zasshi, 116(1), 4–11.

Hauschke, D. (1997): Statistical proof of safety in toxicological studies. Drug Inf. J., 31, 357–361.

Hayes, J.P. (1987): The positive approach to negative results in toxicology studies. Ecotox. Environ. Safety, 14(1), 73–77.

Hess, K.R. (2011): Statistical design considerations in animal studies published recently in cancer research. Cancer Res., 71, 625.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Jane Hutton, J. and Altman, D.J. (2009): Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One, 4(11), 1–11.

Kozinetz, C.A. (2011): Application of epidemiologic principles for optimizing preclinical research study design. Int. J. Preclin. Res., 2(1), 63–65.

Lenth, R.V. (2007): Statistical power calculations. J. Anim. Sci. 2007. 85 (E. Suppl.), E24–E29.

Lew, M. (2007): Good statistical practice in pharmacology—Problem 1. Br. J. Pharmacol., 152(3), 295–298.

Lim, H.J. and Hoffmann, R.G. (2007): Study Design: The Basics. In: Topics in Biostatistics. Ambrosius, W.T. (Editor). Humana Press Inc., New Jersey, USA.

Nicholls, A.W., Mortishire-Smith, R.J. and Nicholson, J.K. (2003): NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. Chem. Res. Toxicol., 16, 1395–1404.

Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Mcleod, M., Mignini, L.E., Jayaram, P. and Khan, K.S. (2007): Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ, 334, 197–200.

Savitz, S.I. (2007): A critical appraisal of the NXY-059 neuroprotection studies for acute stroke: a need for more rigorous testing of neuroprotective agents in animal models of stroke. Exper. Neurol., 205, 201–205.

Schapiro, S.J. and Everitt, J.I. (2006): Preparation of animals for use in the laboratory: Issues and challenges for the institutional animal care and use committee (IACUC). ILAR J., 47(1), 370–375.

Vogt, L. and Kloting, I. (1990): Effects of randomization masking diabetes relevant traits in animal experiments. Diabetes Res., 15(3), 131–135.

Workman, P., Aboagye, E.O., Balkwil, F., Balmain, A., Bruder, G., Chaplin, D.J., Double, J.A., Everitt, J., Farningham, D.A.H., Glennie, M.J., Kelland, L.R., Robinson, V., Stratford, I.J., Tozer, G.M., Watson, S., Wedge, S.R. and Eccles, S.A. (2010): Guidelines for the welfare and use of animals in cancer research. British J. Cancer, 102, 1555–1577.

19

How to Select An Appropriate Statistical Tool?

Good Statistical Design

Good statistical design is a pivotal factor in animal research. However, replication, randomization and blinding, which are key components of good statistical design, are less often used in animal research (Kilkenny *et al.*, 2009). Hess (2011) reviewed statistical design given in 100 articles on animal experiments published in Cancer Research in 2010. In 14 of the 100 articles, the number of animals used per group was not reported. In none of the 100 articles the method used to determine the number of animals per group was reported. Among the 74 articles in which randomization seemed feasible, only 21 reported that they had randomly allocated animals to treatment groups. None of these articles described how the randomization was carried out. Selection of appropriate statistical tools is very crucial in the analysis of data obtained from toxicological and pharmacological studies. Selection of a non-appropriate statistical tool during the design of a study or using a different statistical tool from that mentioned in the study plan with improper justification may lead to misinterpretation of the data (Kobayashi *et al.*, 2011).

Decision Trees

Several attempts have been made to standardize statistical methodologies for the analysis of data obtained from the toxicological and pharmacological studies. One of the methodologies proposed by several authors is the tree-type algorithms (Gad and Weil, 1986; Healey, 1997; Hamada *et al.*, 1998; Gad, 2006). The tree-type algorithms are called as decision trees, which are graphical representation of decisions involved in the choice of the statistical procedure (Howell, 2008). The decision tree-diagram is an excellent tool for determining the optimum course of action in situations offering several alternatives with uncertain outcomes. The first tree-type algorithm for toxicity studies reported in Japan by Yamazaki *et al.* (1981) is given in Figure 19.1.

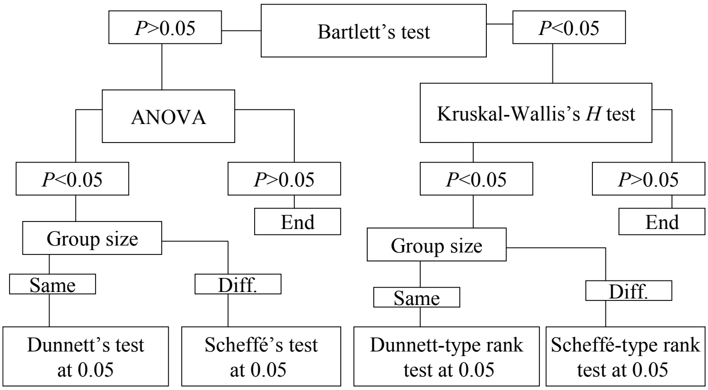


Figure 19.1. The first tree-type algorithm for toxicity studies reported in Japan

This tree-type algorithm was criticized by Kobayashi *et al.* (1995), who identified three major weaknesses which included: selection of a parametric or non-parametric test is based on the highly sensitive Bartlett's homogeneity test; test for normality is not covered in this algorithm; and outliers and dose-dependency are not evaluated.

Hamada *et al.* (1998) proposed a tree-type algorithm for the analysis of quantitative data, which is given in Figure 19.2.

Kobayashi *et al.* (2000) proposed a simple tree-type algorithm for the analysis of quantitative data obtained from toxicological experiments involving more than 2 groups (Figure 19.3).

Sakaki *et al.* (2000) proposed a tree-type algorithm for the analysis of quantitative data, particularly body weight, hematology, and organ weight data, obtained from repeated dose administration studies. This tree-type algorithm does not recommend homogeneity and normality tests; the data are directly analysed by Williams's test (Figure 19.4).

Gad and Weil (1986) proposed a flow chart covering most of the situations that can be encountered in toxicology and pharmacology (Figure 19.5).

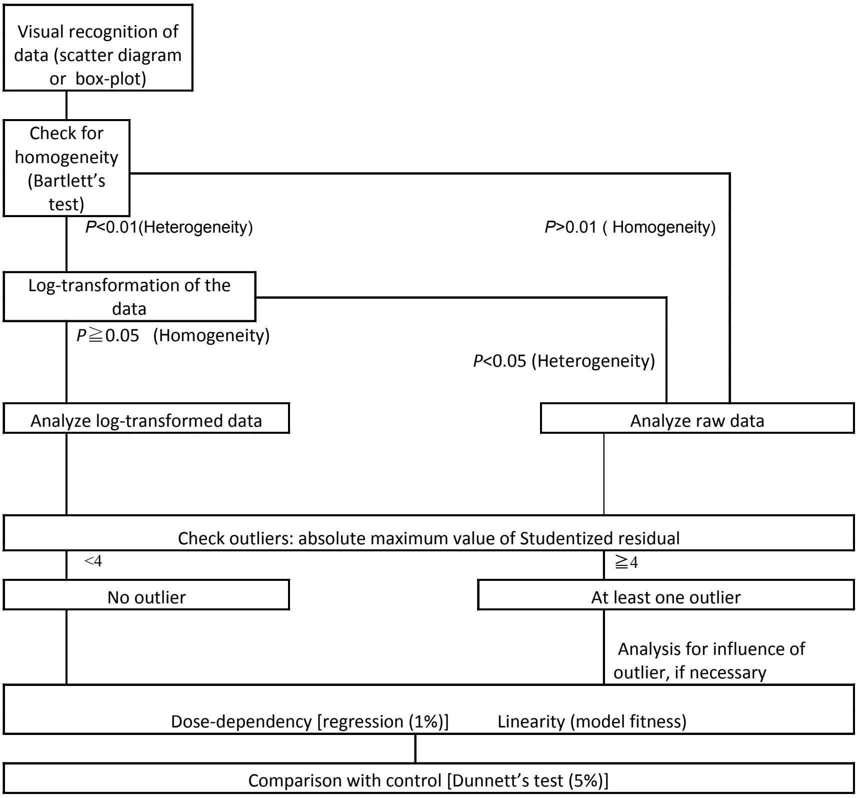

Visual recognition of data (scatter diagram or box-plot)
Check for homogeneity (Bartlett's test)
P<0.01(Heterogeneity)
P>0.01 (Homogeneity)
Log-transformation of the data
P≧0.05 (Homogeneity)
P<0.05 (Heterogeneity)
Analyze log-transformed data
Analyze raw data
Check outliers: absolute maximum value of Studentized residual
<4
≧4
No outlier
At least one outlier
Analysis for influence of outlier, if necessary
Dose-dependency [regression (1%)]
Linearity (model fitness)
Comparison with control [Dunnett's test (5%)]

Figure 19.2. Tree-type algorithm for the analysis of quantitative data proposed by Hamada *et al.* (1998)

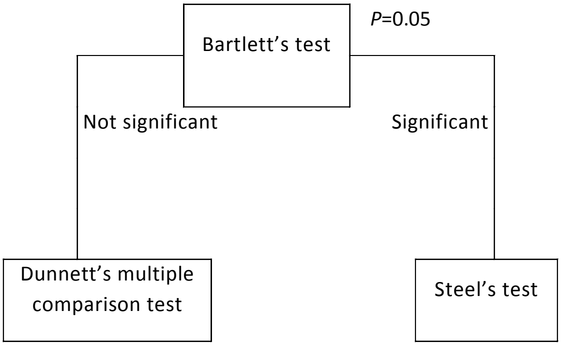


Figure 19.3. The tree-type algorithm for the analysis of toxicological data proposed by Kobayashi *et al.* (2000)

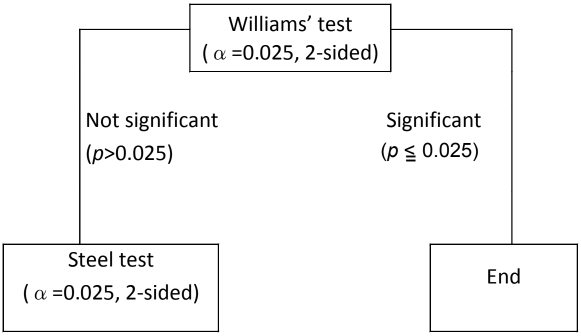


Figure 19.4. The tree-type algorithm for the analysis of quantitative data obtained from repeated dose administration studies proposed by Sakaki *et al.* (2000)

Statistical Procedures Used by National Toxicology Program (NTP), USA

The statistical procedures used in the analysis of data of 2-year toxicity/carcinogenesis studies presented in the Technical Reports of the NTP are given below:

a. Survival Analyses

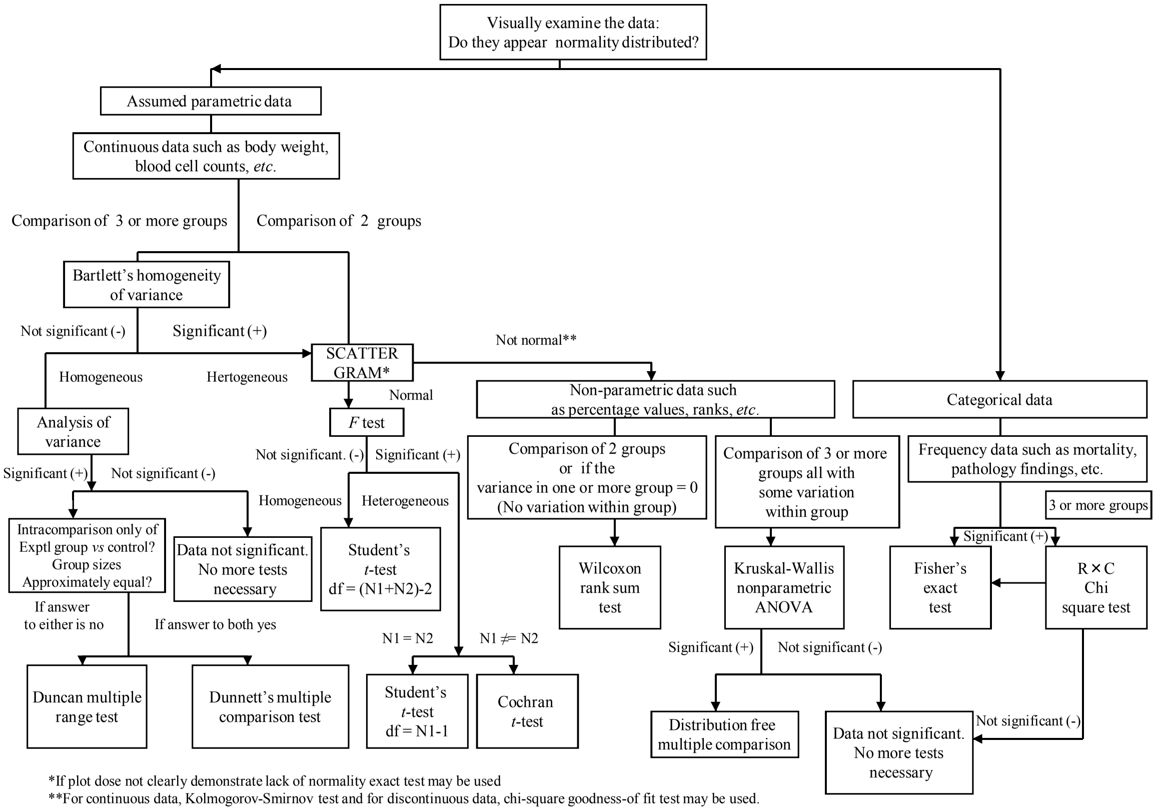
The product-limit procedure of Kaplan and Meier (1958) is used to estimate the probability of survival. Animals found dead due to causes other than natural causes are censored from the survival analyses, while animals dying from natural causes are not censored. Dose-related effects on survival is calculated using Cox's method (Cox, 1972) (for testing two groups for equality) and Tarone's (1975) life table test (to identify dose-related trends). The *P* values are two-sided.

b. Analysis of neoplasm and non-neoplastic lesion incidences

The Poly-*k* test (Bailer and Portier, 1988; Portier and Bailer, 1989; Piegorsch and Bailer, 1997) is used to assess neoplasm and non-neoplastic lesion prevalence. Tests of significance include pair-wise comparisons of each exposed group with controls and a test for an overall exposure-related trend. Continuity-corrected Poly-3 tests are used in the analysis of lesion incidence. The *P* values are one-sided.

c. Analysis of continuous variables

Organ and body weight data is analyzed with the parametric multiple comparison procedures of Dunnett (1955) and Williams (1971, 1972). Hematology, clinical chemistry, urinalysis, urine concentrating ability,


Visually examine the data:
Do they appear normality distributed?
Assumed parametric data
Continuous data such as body weight, blood cell counts, *etc*.
Comparison of 3 or more groups
Comparison of 2 groups
Bartlett's homogeneity of variance
Not significant (-)
Significant (+)
Homogeneous
Hertogeneous
SCATTER GRAM*
Not normal**
Normal
Analysis of variance
Significant (+)
Not significant (-)
F test
Not significant. (-)
Significant (+)
Homogeneous
Heterogeneous
Intracomparison only of Exptl group *vs* control? Group sizes Approximately equal?
Data not significant. No more tests necessary
Student's *t*-test df = (N1+N2)-2
If answer to either is no
If answer to both yes
N1 = N2
N1 ≠ N2
Duncan multiple range test
Dunnett's multiple comparison test
Student's *t*-test df = N1-1
Cochran *t*-test
Non-parametric data such as percentage values, ranks, *etc*.
Comparison of 2 groups or if the variance in one or more group = 0 (No variation within group)
Comparison of 3 or more groups all with some variation within group
Wilcoxon rank sum test
Kruskal-Wallis nonparametric ANOVA
Significant (+)
Not significant (-)
Distribution free multiple comparison
Data not significant. No more tests necessary
Categorical data
Frequency data such as mortality, pathology findings, etc.
3 or more groups
Significant (+)
Fisher's exact test
R × C Chi square test
Not significant (-)

*If plot dose not clearly demonstrate lack of normality exact test may be used

**For continuous data, Kolmogorov-Smirnov test and for discontinuous data, chi-square goodness-of fit test may be used.

Figure 19.5 Flow chart proposed by Gad and Weil (1986)

cardiopulmonary, cell proliferation, tissue concentrations, spermatid, and epididymal spermatozoal data are analyzed using the non-parametric multiple comparison methods of Shirley (1977), as modified by Williams (1986) and Dunn (1964). Jonckheere's test (Jonckheere, 1954) is used to assess the significance of the dose-related trends and to determine whether a trend-sensitive test (Williams' or Shirley's test) is more appropriate for pair-wise comparisons than a test that does not assume a monotonic dose-related trend (Dunnett's or Dunn's test).

Average severity values are analyzed for significance with the Mann-Whitney *U* test (Hollander and Wolfe, 1973). Vaginal cytology data are transformed to arcsine values and then the treatment effects are investigated by applying a multivariate analysis of variance (Morrison, 1976).

Immunological data is initially tested for homogeneity using Bartlett's test. For data that is determined to be homogeneous, one-way analysis of variance (ANOVA) is conducted. If the ANOVA is significant at $P < 0.05$, Dunnett's multiple range *t*-test is used for multiple treatment-control comparisons. If the data is not homogeneous, the Kruskal-Wallis test or the Wilcoxon rank sum test is used to compare treatment groups with controls groups. The level of statistical significance is set at $P < 0.05$ and $P < 0.01$.

Values are routinely presented as mean ± standard error.

Decision Tree Produced by OECD

OECD produced a decision tree for analyzing data in long-term toxicology studies by summarizing common statistical procedures (OECD, 2010). This decision tree, more or less similar to an approach used by the US National Toxicology Program, is given in Figure 19.6.

A detailed description on this decision tree is given in the guideline (OECD, 2010) by providing explanation on each circled number given in the Figure.

Decision tree has also been used *in vitro* assays and pharmacological experiments. Decision-tree approaches were proposed for the analysis of the chromosome aberration assay (Kim *et al.*, 2000; Hothorn, 2002) and for evaluating drug-specific effects of quantitative pharmaco-EEG (Dago *et al.,* 1994).

Though the decision trees are used in the statistical analysis of data of various toxicological studies (Krores *et al.*, 2004), critics point out that, 'although there are efficiency gains in the application of flow charts, there

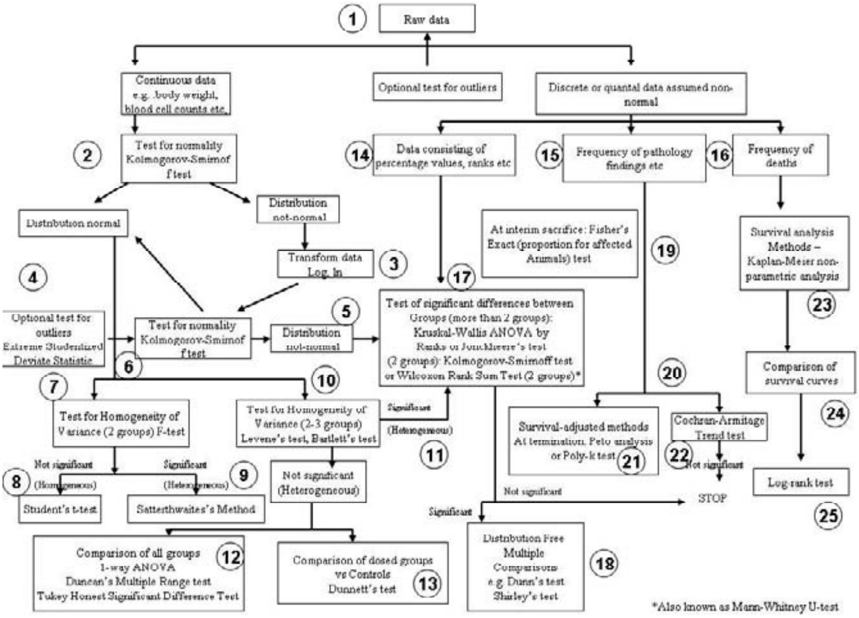


Figure 19.6. Decision tree produced by OECD for the analysis of data in long-term toxicology studies (OECD, 2010)

is a 'deskilling' of the task, an over-emphasis on significance testing for decision making, and vulnerability to artefactual results'. There is also the methodological problem with a multiple testing procedure where one hypothesis test is used to select another test which can complicate quantifying the true probability values associated with various comparisons (OECD, 2010).

Incongruence in Selection of a Statistical Tool

Nomura (1994) compared the tree-type algorithms used at the contract research laboratories in Japan and other countries. He observed that the countries developed their own tree-type algorithms. Kobayashi *et al.* (2011) compared the statistical tools used for analysing the data of repeated dose toxicity studies with rodents conducted in 45 countries, with that of Japan. The study revealed there was no congruence among the countries in the use of statistical tools for analysing the data obtained from the above studies. For example, to analyse the data obtained from repeated dose toxicity studies with rodents, Scheffé's multiple range and Dunnett type (joint type Dunnett) tests are commonly used in Japan, but in other countries use of these statistical tools is not so common. In most of the countries, the data are generally not tested for normality. The authors observed that out of 127 studies examined, data of only 6 studies were analysed for both homogeneity of variance and normal distribution.

The decision trees mentioned above are developed based on the classical statistical principles sidelining biological principles. For example a sensitive Bartlett's test for examining homogeneity of variance may not be suitable in most of the animal studies. The below mentioned decision trees or flow charts are developed providing due consideration to biological principles:

Selection of a Statistical Tool—Suggested Decision Tress or Flow Charts

1. Selection of a statistical tool when the data show a normal or non-normal distribution (Kobayashi *et al.*, 2008).

Situation 1 (Number of Groups, 2)

When the data of each group show a normal distribution by Shapiro-Wilk's *W* test, then the *F*-test is applied. If the *F*-test is insignificant, the data are analysed using Student's *t*-test and if it is significant, Aspin-Welch's *t*-test is used to analyse the data.

When the data of any group show a non-normal distribution by Shapiro-Wilk's *W* test, they are subjected to Mann-Whitney's *U* test (Rank sum test).

Flow chart of situation 1 is given in Figure 19.7.

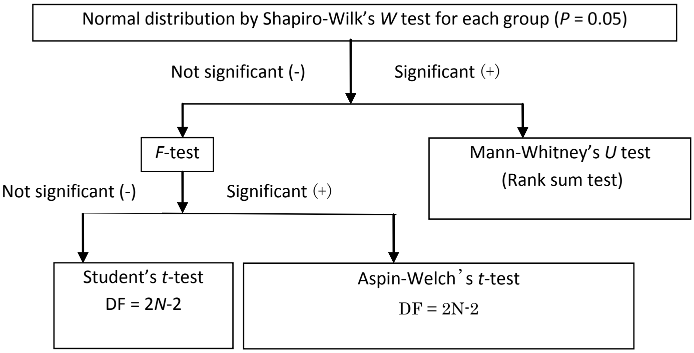


Figure 19.7. Flow chart for selecting the statistical tool when the data show a normal or non-normal distribution (Situation 1, Number of group = 2)

Situation 2 (Number of Groups, ≥3)

When each group shows a normal distribution by Shapiro-Wilk's *W* test, the Dunnett's multiple comparison test is used. When control group or all groups do not show a normal distribution, non-parametric Steel's test (Dunnett's separate type test) is used. When normal distribution is not observed by one or two treatment groups, they are excluded from the analysis and the remaining groups are analyzed by Dunnett's multiple comparison test. The clinical relevance of the excluded groups is assessed in the light of other observations.

Flow chart of situation 2 is given in Figure 19.8.

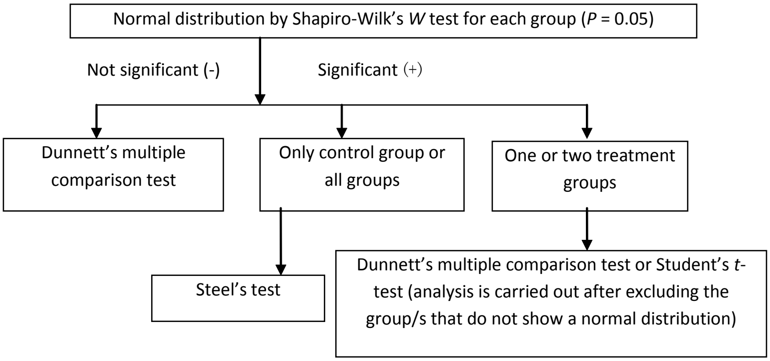

Normal distribution by Shapiro-Wilk's *W* test for each group (*P* = 0.05)
Not significant (-)
Significant (+)
Dunnett's multiple comparison test
Only control group or all groups
One or two treatment groups
Steel's test
Dunnett's multiple comparison test or Student's *t*-test (analysis is carried out after excluding the group/s that do not show a normal distribution)

Figure 19.8. Flow chart for selecting the statistical tool when the data show a normal or non-normal distribution (Situation 2, Number of group ≥ 3)

2. Analysis of qualitative data of urinalyses and pathological findings.

Analysis of qualitative data of urinalysis and pathological findings presented in 2×2 and 4×4 Tables is given in Table 19.1.

Statistical Tools Suggested for the Analysis of Toxicology Data

The suggested statistical tools for the analysis of parametric and non-parametric data are given in Table 19.2 and for the comparison of two and multi-groups are given in Table 19.3.

Use of Statistics in Toxicology-Limitations

There are limitations in the use of statistics in toxicology. According to Gad and Weil (1986), the limitations are: 1. statistics cannot make poor data better; 2. statistical significance may not imply biological significance; 3. an effect that may have biological significance may not be statistically significant; 4. the lack of statistical significance does not prove safety. Statistical analysis cannot rescue poor data resulting from a flawed design or

Table 19.1. Analysis of qualitative data of urinalyses and pathological findings (Kobayashi, 2010)

Incidence

2×2 Table	
Control: Observed (+)	Control: None (–)
Treatment: Observed (+)	Treatment: None (–)
(1) Chi square test	
(2) Fisher's test	
Note: Small numerical values (0–5) are not suitable for Chi square analysis in the four-values data set (Control: +, – and Treatment: +, –). Fisher's test (one-sided) is suitable for the data with small numerical values.	

4×4 Table, Grades and number of findings with the grades in Groups				
Group	No finding (–)	Slight (+)	Moderate (++)	Marked (+++)
Control	10	1	0	0
Low	4	3	2	1
Mid	1	4	3	2
High	0	3	4	3
Note 1: If Chi square analysis by 4×4 Table shows a significant difference, Control Group *vs* Low dose Group, Control Group *vs* Mid dose Group and Control Group *vs* High dose Group are analysed by 2×4 Table by division. Note 2: If the number of animals in a group is ≥5, use of Mann-Whitney's *U* test is preferred. Note 3: Cochran-Armitage trend test is the preferred tool for examining dose-related pattern.				

Table 19.2. Parametric and non-parametric statistical tools for the analysis of data obtained from toxicology studies

Group settings	Parametric test	Non-parametric test
Only two groups	Student, ◎Aspin-Welch, Cochran-Cox *t*-tests	Mann-Whitney *U* test, Wilcoxon test
Three or more group	ANOVA	Kruskal-Wallis rank sum test
	◎Dunnett's multiple comparison test, General, multiple comparison test	Nonparametric type Dunnett's rank sum test
		◎Steel's test
	Tukey's multiple range test (the size of the group is the same)	Nonparametric type Tukey's rank sum test
	Tukey-Kramer's multiple range test (the size of the group is different)	◎Steel-Dwass' test
	◎Duncan's multiple range test	Nonparametric type Duncan's rank sum test
	Scheffé's multiple comparison test	Nonparamteric type Scheffé's rank sum test
	◎Williams's *t*-test (analyzes the difference of the mean values between each treated group and control, when the mean value of the treated groups changes in one direction.)	Shirley-Williams's test
	—	Jonckheere's trend test

◎Tests recommended.

Table 19.3. Statistical tools suggested for the comparison of two and multi-groups

Group setting	Comparison	Analysis
Only two groups	Only one time	Aspin-Welch's *t*-test
Control(x_0), Low dose(x_1), Mid-dose(x_2), High dose(x_3)	Analysis of difference of the chisel between control group and each dose group (the analysis frequency is three times)	Dunnett's multiple comparison test; Williams's *t*-test (assumption: data possess a dose-dependency)
Control, Drug A, Drug B, Drug C or Group A, Group B, Group C, Group D	Analysis of difference between control group and each drug or group (total number of comparisons made is three)	Dunnett's multiple comparison test
	Comparison of all pairs (total number of comparisons made is six)	Tukey's multiple range test; Duncan's multiple range test
Control(x_0), Low dose(x_1), Mid dose(x_2), High dose(x_3), Reference drug (R_1)	Analysis of difference between control group and Reference drug followed by comparison of control group with each dose group.	Dunnett's test or Williams's *t*-test. Examine if there is a significant difference between x_0 and R_1 by *t*-test; if there is a significance, then compare the control with x_1, x_2 and x_3, excluding R_1 using the tests of Dunnett or Williams.

a poorly conducted study. An appropriate data analysis will follow directly from a correct experimental design (including the selection of statistical methods to be applied) and implementation (OECD, 2010). According to Altman and Bland (1994), 'failing to reject the hypothesis often leads to the conclusion of evidence in favour of safety, simply because absence of evidence is not evidence of absence'.

References

Altman, D. and Bland, M. (1994): Regression towards the mean. British Med. J., 308, 1499.

Bailer, A.J., and Portier, C.J. (1988): Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics, 44, 417–431.

Cox, D.R. (1972): Regression models and life-tables. J.R. Stat. Soc., B34, 187–220.

Dago, K.T., Luthringer, R., Lengellé, R., Rinaudo, G. and Macher, J.P. (1994): Statistical Decision Tree: A Tool for Studying Pharmaco-EEG Effects of CNS-Active Drugs. Neuropsychobiol., 29, 91–96.

Dunn, O.J. (1964): Multiple comparisons using rank sums. Technometrics, 6, 241–252.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc., 50, 1096–1121.

Gad, S. (2006): Statistics and Experimental Design for Toxicologists and Pharmacologists. 4th Edition, Taylor and Francis, Boca Raton, FL, USA.

Gad, S. and Weil, C.W. (1986): Statistics and Experimental Design for Toxicologists. The Telford Press Inc., New Jersey, U.S.A.

Hamada, C., Yoshino, K., Matsumoto, K., Nomura, M. and Yoshimura, I. (1998): Tree-type algorithm for statistical analysis in chronic toxicity studies. J. Toxicol. Sci., 23(3), 173–181.

Healey, G.F. (1997): How to achieve standardization of statistical methods in toxicology. Drug Inf. J., 31–32, 327–334.

Hess, K.R. (2011): Statistical design considerations in animal studies published recently in cancer research. Cancer Res., 15, 71(2), 625.

Hollander, M. and Wolfe, D.A. (1973): Nonparametric Statistical Methods, John Wiley and Sons, New York, USA.

Hothorn, L.A. (2002): Selected biostatistical aspects of the validation of *in vitro* toxicological assays. ATLA, 30 (Supp. 2), 93–98.

Howell, D.C. (2008): Fundamental Statistics for the Behavioral Sciences. 6th Edition, Thomson Wadsworth, Belmont, USA.

Jonckheere, A.R. (1954): A distribution-free *k*-sample test against ordered alternatives. Biometricka, 41, 133–145.

Kaplan, E.L. and Meier, P. (1958): Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc., 53, 457–481.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D. Hutton, J. and Altman, D.J. (2009): Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoSOne, 4(11), 1–11.

Kim, B.S., Zhao, B., Kim, H.J. and Cho, M. (2000): The statistical analysis of the *in vitro* chromosome aberration assay using Chinese hamster ovary cells. Mutation Res., 469, 243–252.

Kobayashi, K. (2010): Trend of statistics used for toxicity studies. Yakuji-niposha, Tokyo, Japan.

Kobayashi, K., Kanamori, M., Ohori, K. and Takeuchi, H. (2000): A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies in rodents. San Ei Shi., 42(4), 125–129.

Kobayashi, K., Pillai, K.S., Guhatakurta, S., Cherian, K.M. and Ohnishi, M. (2011): Statistical tools for analysing the data obtained from repeated dose toxicity studies with rodents: A comparison of the statistical tools used in Japan with that of used in other countries. J. Environ. Biol., 32(1), 11–16.

Kobayashi, K., Pillai, K.S., Suzuki, M. and Wang, J. (2008): Do we need to examine the quantitative data obtained from toxicity studies for both normality and homogeneity of variance? J. Environ. Biol., 29(1), 47–52.

Kobayashi, K., Watanabe, K. and Inoue, H. (1995): Questioning the usefulness of the non-parametric analysis of quantitative data by transformation into ranked data in toxicity studies. J. Toxicol. Sci., 20(1), 47–53.

Krores, R., Renwick, A.G., Cheeseman, M., Kleiner, J., Piersma, A., Schilter, B., Schlatter, J., van Schothorst, F., Vos, J.G. and Wurtzen, G. (2004): Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. Food Chem. Toxicol., 42(1), 65–83.

Morrison, D.F. (1976): Multivariate Statistical Methods, 2nd Edition, McGraw-Hill Book Co., New York, USA.

Nomura, M. (1994): International comparison of statistical analysis methods for toxicological study. Jap. Soc. Biopharm. Statistics, 40, 1–36.

NTP. National Toxicology Program, USA. http://ntp.niehs.nih.gov/go/10007

OECD (2010): Organisation for Economic Cooperation and Development. OECD Draft Guidance Document N° 116 on the Design and Conduct of Chronic Toxicity and carcinogenicity Studies, Supporting TG 451, 452, 453. OECD, Paris, France.

Piegorsch, W.W. and Bailer, A.J. (1997): Statistics for Environmental Biology and Toxicology, Section 6.3.2., Chapman and Hall, London.

Portier, C.J. and Bailer, A.J. (1989) : Testing for increased carcinogenicity using a survival-adjusted quantal response test. Fund. Appl. Toxicol., 12, 731–737.

Sakaki, H., Igarashi, T., Ikeda, Y., Mizoguchi, K., Omichi, T., Kadota, M., Kawada, T., Takizawa, O., Tsukamoto, Y., Terai, K., Tozuka, A., Hirata, J., Handa, H., Mizuma, Z., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 71–81.

Shirley, E. (1977): A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, 33, 386–389.

Tarone, R.E. (1975): Tests for trend in life table analysis. Biometrika, 62, 679–682.

Williams, D.A. (1971): A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, 27, 103–117.

Williams, D.A. (1972): The comparison of several dose levels with a zero dose control. Biometrics, 28, 519–531.

Williams, D.A. (1986): A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. Biometrics, 42, 182–186.

Yamazaki, M., Noguchi, Y., Tanda, M. and Shintani, S. (1981): Statistical method appropriate for general toxicological studies in rats. J. Takeda Res. Lab., 40 (3), 163–187.

Appendices

Appendix 1. Coefficient for Shapiro-Wilk *W* Test (Conover, 1999)

i\n	2	3	4	5	6	7	8	9	**10**
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.0588	**0.5739**
2	-	0.0000	0.1667	0.2413	0.2806	0.3031	0.3164	0.3244	**0.3291**
3	-	-	-	0.0000	0.0875	0.1401	0.1743	0.1976	**0.2141**
4	-	-	-	-	-	0.0000	0.0561	0.0947	**0.1224**
5	-	-	-	-	-	-	-	0.0000	**0.0399**

i\n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1449	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	-	-	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	-	-	-	-	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	-	-	-	-	-	-	0.0000	0.0163	0.0303	0.0422
10	-	-	-	-	-	-	-	-	0.0000	0.0140

i\n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	00822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	-	-	0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13	-	-	-	-	0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14	-	-	-	-	-	-	0.0000	0.0084	0.0159	0.0227
15	-	-	-	-	-	-	-	-	0.0000	0.0076

Appendix 1. contd....

Appendix 1. contd....

i\n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737
3	0.2475	0.2462	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368
4	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098
5	0.1874	0.1878	0.1880	0.1882	0.1883	0.1833	0.1883	0.1881	0.1880	0.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513	0.1520	0.1526
8	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376
9	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237
10	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	0.0585	0.0629	0.0699	0.0706	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870
13	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759
14	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651
15	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17	-	-	0.0000	0.0062	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343
18	-	-	-	-	0.0000	0.0057	0.0110	0.0158	0.0203	0.0244
19	-	-	-	-	-	-	0.0000	0.0053	0.0101	0.0146
20	-	-	-	-	-	-	-	-	0.0000	0.0049

i\n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.6917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0379	0.0411	0.0422	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0355	0.0361	0.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22		-	0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23	-	-	-	-	0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
24	-	-	-	-	-	-	0.0000	0.0037	0.0071	0.0104
25	-	-	-	-	-	-	-	-	0.0000	0.0035

Conover, W.J. (1999): Practical Nonparametric Statistics, 3rd Edition, John Wiley & Sons, Inc. New York, USA.

Appendix 2. Quantiles of the Shapiro-Wilk Test Statistic (Tsubaki and Tsubaki, 2001)

n	0.01	0.02	**0.05**	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.387	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.386	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	**0.978**	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.852	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989

n	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.919	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	0.928	0.936	0.946	0.954	0.974	0.985	0.988	0.990	0.991
48	0.929	0.937	0.947	0.954	0.974	0.985	0.988	0.990	0.991
49	0.929	0.937	0.947	0.955	0.974	0.985	0.988	0.990	0.991
50	0.930	0.938	0.947	0.955	0.974	0.985	0.988	0.990	0.991

Tsubaki, M. and Tsubaki, H. (2001). Statistical Method in Medical Research, 3rd Edition. Scientist, Tokyo, Japan.

Appendix 3. Z Score for Normal Distribution (Gad and Weil, 1988)

Proportional Parts										
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2746	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0584	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0121	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0018	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0013	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0009	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0006	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0005	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
3.4	0.0003	0.0003	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Gad, S. and Weil, C.S. (1988). Statistics and Experimental Design for Toxicologists. Telford Press, New Jersey, USA.

Statistics plays an important role in pharmacology and related subjects like, toxicology, and drug discovery and development. Improper statistical tool selection to analyze the data obtained from studies conducted in these subjects may result in wrongful interpretation of the performance-or safety of drugs. Thus it is imperative to have some understanding of the subject.

This book has been written with an objective to communicate statistical tools in simple language. Utmost care has been taken to avoid complicated mathematical equations, which readers may find difficult to assimilate. The examples used in the book are similar to those that the scientists encounter regularly in their research. The authors have provided cognitive clues for selection of appropriate statistical tools to analyse the data obtained from the studies and also to interpret the result of the statistical analysis.

K16056

9 781578 087525

CRC Press
Taylor & Francis Group
an **informa** business
www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK