

Veterinary Biometrics ^{*}

Mark Stevenson [†] Alasdair Noble [‡]

EpiCentre, IVABS

Massey University, Palmerston North, New Zealand

June 18, 2007

Contents

1	Introduction to biometry	5
2	Exploratory data analysis	7
2.1	Quantitative variables	7
2.2	Qualitative variables	8
2.3	Measures of location	9
	Mean	9
	Median	9
	Mode	9
2.4	Measures of variability	10
	Range	10
	Variance and standard deviation	10
	Coefficient of variation	10
	Quartiles, deciles and percentiles	11
	Skewness	11
	Kurtosis	11
	Outliers	11
2.5	Graphs	12
	Bar graphs	12
	Histograms	12

^{*}Notes for Veterinary Biometrics and Epidemiology, taught as course 227.407 within the BVSc program at Massey University.

[†]M.Stevenson@massey.ac.nz

[‡]A.D.Noble@massey.ac.nz

Box and whisker plots	13
Line graphs	13
Pie charts	13
Scatterplots	14
Stem and leaf plots	14
Trellis plots	15
3 Probability distributions	18
3.1 Characteristics of distributions	19
Continuous vs discrete	19
Bounded vs unbounded	19
Parametric vs non-parametric	19
3.2 Ways to express a probability function	19
Probability density	19
Cumulative probability	20
Quantiles of a probability distribution	21
Random values from a specified probability distribution	21
4 The normal distribution	23
5 The binomial distribution	25
6 Confidence intervals	26
6.1 Means and their differences	27
Single sample	27
Two samples: unpaired case	27
Two samples: paired case	28
Non-normal data	28
6.2 Medians and their differences	28
6.3 Proportions and their differences	28
Single sample	28
Two samples: unpaired case	28
Two samples: paired case	29
6.4 Incidence risk	29
6.5 Incidence rate	29
7 Statistical inference	31
7.1 Statistical significance and confidence intervals	32
7.2 Steps involved in testing significance	33
8 Inference for proportions	34

9 Inference for means	36
9.1 Paired versus unpaired tests	37
9.2 Equal or unequal variances	37
10 Correlation coefficients	38
11 Inference for non-parametric distributions	39
12 Analysis of variance	42
12.1 One way ANOVA	42
12.2 Two way ANOVA	46
13 Linear regression	49
13.1 Simple linear regression	49
The model	50
Checking the model	51
13.2 Multiple linear regression	52
The model	52
Model selection	53
Checking the model	53
14 Experimental design	55
14.1 Completely randomised design	55
14.2 Randomised block design	56
14.3 Latin square design	58
14.4 Factorial design	59
14.5 Split plot factorial design	62
14.6 Repeated measures design	63
15 Using a spreadsheet	65
15.1 Formatting cells	65
15.2 Sorting data	66
15.3 Functions	67
Mathematical functions	68
Logical functions	69
Array functions	70
Text functions	71
Statistical functions	71
Probability functions	73
Lookup functions	73
Rounding and truncating functions	75

15.4	Graphs	75
	Error bars	76
	Trend lines	76
	Shewhart charts	76
	Frequency histograms	78
15.5	Shortcuts	79

1 Introduction to biometry

Science is a cultural phenomenon in that different people will require different degrees of evidence for what they believe. This problem is greatly reduced by the notion of selection, by focusing on what can be ruled out with compelling force. Selection involves selection among hypotheses.

The philosopher Karl Popper is credited with the notion of scientific method as selection among competing hypotheses. According to Popper, we can never prove any hypothesis, but we make useful progress by trying to disprove them. If we subject competing hypotheses to testing, the surviving hypotheses will constitute a useful view of the world. In 1893 Löffler and Frosch demonstrated the presence of ultra-microscopic infectious organisms, now known to be viruses. They passed lymph from an animal suffering from foot-and-mouth disease through a filter, which ruled out bacteria (of ordinary size) as the infectious agent. They then infected animals serially, which ruled out any non-replicating poison. Turning to a more modern example, the regulatory regions of genes are frequently dissected by sequentially shortening the sequence under study and testing for expression. In genetics, the gene responsible for a recessive trait in an experimental cross is sometimes identified by a process known as exclusion mapping. Variation in the location of meiotic cross-overs in a backcross allow the investigator to rule out most of the genome, leaving a plausible interval that decreases in size as the data accumulate. The situation is very different if the trait is incompletely penetrant. For a truly recessive trait, the organism that is homozygous for the responsible allele will always possess the trait of interest. For an incompletely penetrant trait, the homozygous genotype may be necessary for the trait, but not sufficient. We cannot then rule out a region simply because it fails to produce the trait in a single homozygous organism. There may be other reasons why the trait will fail to appear. In this case we have to rely on statistical criteria to evaluate the weight of evidence against a region, rather than ruling it out entirely. Progress can still be made, but it will be much slower, more costly, and subject to more doubt. This is a frequent role for statistics, rescuing the strategy of hypothesis rejection, but with a need for replication that makes us pay heavily for the forcefulness of our rejection.

The approach to science described by Popper is often termed *hypothetico-deductive*, as it involves deducing the consequences of hypotheses so that they may be put to the test. Not everything in science, however, will fit clearly into this mould. The demonstrations that amino acids, nucleotides, sugars, and fatty acids can all be generated from presumed pre-biotic conditions is essentially a plausibility argument that is not directed at rejecting any particular hypothesis. Such demonstrations do, however, serve to reject the failure of imagination that seems to cause difficulty when hypotheses involve scales beyond ordinary human experience, as encountered in particle physics, geology, and evolution. It seems more common, however, for too much imagination to be the problem. Humans are so good at recognising patterns that there is a real danger of overinterpreting patterns that are merely due to the play of chance. A major role of statistical inference is to reject chance as an explanation, so that we can have a reasonable assurance that the patterns being interpreted are worthy of interpretation.

As a practicing veterinarian you take on the role (among many others) of clinician, pathologist, diagnostician, economist, psychologist, and therapist. One of the exciting things about practice is that you will be faced with many opportunities where you will need take on all of these roles identify, investigate, and solve problems. Where groups of animals (e.g. herds or flocks) are involved the scientific method outlined above provides the basis for this work. Application of the

scientific method — in particular the testing of hypotheses — is underpinned by a knowledge of biometry, the application of statistics, probability, mathematics, systems analysis, and computer science to studies of the life sciences.

2 Exploratory data analysis

The first step in most statistical analyses is to examine the data using simple graphical and numerical summaries. These descriptions will often be of interest in themselves, but they are also helpful for selecting the most appropriate statistical technique to be used later. Another important use of descriptive statistics is to find unusual values (for example, values that are outside of the feasible range) in the data set, that may either arise from errors in data entry or from subjects whose profiles are very different from those of others (outliers).

Numerical values describing a characteristic of a sample are called statistics. We use statistics to estimate the corresponding parameter of the population from which the sample was taken. Descriptive statistics fall into two categories: (1) measures of location; and (2) measures of variability. Whereas numerical summaries provide an exact description of variables in a data set graphical summaries provide an impression of the overall features.

2.1 Quantitative variables

Quantitative variables are those that can be counted or measured. The main characteristic of this class of data is that the thing being measured could always be measured in a more exact manner: it is plausible to describe something as being partway between two adjacent classes. For instance, on farms the ages of animals are often recorded as their age (in years) at the most recent parturition. In a goat herd, a set of five does could be listed as being 3, 3, 4, 5, and 7 years old. Had a more refined or precise measure been chosen, such as months of age at most recent parturition, those same five does might be listed as having ages of 22, 27, 46, 60, and 89 months. Technically, the animals' age could be measured in days (e.g. the oldest doe turned out to be 2719 days old at her last kidding) yet reported in years, thus giving a value such as 7.45 years or even 7.448219178082 years. It does not matter if the extra precision is useful (knowing a goat's age to 12 decimal places is of doubtful help), but it is plausible. In addition, the individual indices, such as 1, 2, or 3 years, must represent the same interval or distance for data to be quantitative. Therefore a goat that is 3 years old is considered equidistant in age from a 2-year-old and a 4-year-old. This equal interval quality makes continuous data different from ordinal data.

Discrete variables are quantitative variables whose possible values are integers (whole numbers). Examples include: number of offspring born, number of visits to a veterinary practitioner in a year.

Continuous variables are quantitative variables that have an uninterrupted range of values. Examples include: body weight, height, milk yield. Continuous data can be reduced to ordinal or nominal data. Practitioners may measure kilograms of pig sold per sow per year (a continuous variable) yet run an analysis of average piglet weight gains by categorising females into 'low,' 'moderate,' and 'high' producers (creating ordinal data) instead of by the actual kilograms of pig weaned per year. Also, the sow herd can be divided into family or genetic groups, making the classification a nominal category.

2.2 Qualitative variables

Qualitative (also known as categorical) variables are those variables that cannot be characterised by a numerical quantity. There are three types of qualitative variables: (1) ordinal, (2) nominal, and (3) dichotomous.

Ordinal variables are qualitative variables with several categories which have an innate order. Examples include: body condition score (very thin, thin, adequate, fat, very fat). Rapid Mastitis Test findings include trace, 1+, 2+, and 3+. This is the distinction between ordinal data and true quantitative data. With ordinal variables no information exists on the relative differences between ranks (e.g. a cow of body condition score designated as ‘fat’ is not twice as heavy as a cow designated as ‘thin’, Figure 1).

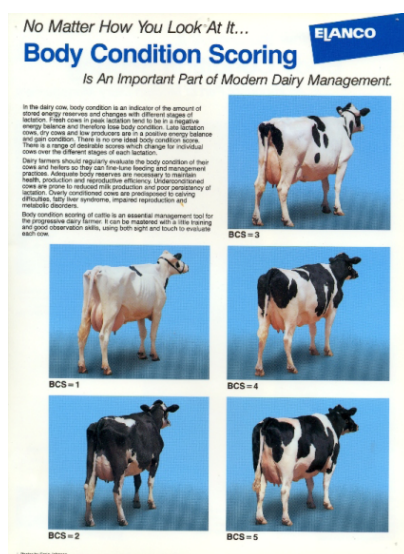


Figure 1: Body condition scores in dairy cattle are an example of ordinal variables.

Nominal variables are qualitative variables with categories that have no natural order. Examples include: disease category might be categorised as alimentary tract, respiratory tract, urinary tract etc. The quick test to check whether nominal data are being used is to change the names to something unrelated and see if information has been lost. For instance, if animal classes of ‘gilt,’ ‘sow,’ and ‘boar’ are changed to ‘apple,’ ‘banana,’ and ‘orange,’ nothing has been lost. The new names still differentiate the groups. They are, therefore, nominal data. Nominal data can be described using frequencies, percentages, or counts. Averages do not apply to nominal data: saying the average animal is three quarters steer and one quarter heifer is meaningless. However, saying a pig herd has 25% sows, 15% gilts, 50% growing pigs, and 10% boars makes sense.

Dichotomous variables are qualitative variables with two classes: yes-no, dead-alive, present-absent. Some dichotomies are based on a natural ordering or preference. For instance, passing an examination is usually preferable to failing it. However, many dichotomies such as gender, offer no natural reason for claiming an inherent ordering. Nonetheless, because there are only two categories, dichotomous data can be treated as ordinal, at least. Although a ranking may not be inherent, ranking female before male or male before female satisfies the mathe-

mathematical requirements for ordinal data. In addition, since only one interval is to be managed, dichotomous data can be mathematically treated as interval data even though the outcome may not make much logical or physical sense (i.e. average gender in a sheep flock may be 0.95 if the dichotomy were male = 0, female = 1, and the flock had 20 ewes for each ram). Hence, treating dichotomous variables depends entirely on the questions asked and the nature of the classification. Some measures can be argued reasonably as representing one data type or the other. For instance, if steer carcass weight is recorded in units of kilograms, it is obviously continuous. However, if carcasses are classified as either ‘big’ or ‘small,’ the measure is obviously ordinal. What about weight classes in 50 kg increments or in 300 kg increments? What if they are just called ‘above average’ and ‘below average’? At some point, even though the basic measure is continuous, categorizations are pushed from continuous to categorical types of data.

Numerical values describing a characteristic of a sample are called statistics. We use statistics to estimate the corresponding parameter of the population from which the sample was taken. Descriptive statistics fall into two categories: (1) measures of location; and (2) measures of variability.

2.3 Measures of location

For a quantitative variable a single figure that indicates the general magnitude of a series of measurements (or samples) is called a measure of location or measure of central tendency. Examples include the mean, median, and mode.

Mean

The mean is defined as the sum of all of the values divided by the number of observations. Where n values from a sample are given by x_1 , the sample mean \bar{x} is an estimate of the population mean, μ .

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

Median

An ordered set of variables is one which is arranged in ascending or descending numerical order. The median is the value that divides an ordered set of variables into two equal parts. For an odd number of values it is the central value, that is the $\frac{1}{2}(n + 1)$ value. For an even number of values, the median is the mean of the $\frac{1}{2}n$ and the $\frac{1}{2}(n + 1)$ values.

Mode

The mode is the value that occurs with the greatest frequency. The mode is not uniquely defined for those distributions where there are two or more values that have the equal highest probability.

2.4 Measures of variability

Range

The range is the difference between the smallest and largest values.

Variance and standard deviation

A natural measure of variability of a distribution is the average of the deviations from the mean. A more convenient measure is the variance which is obtained by taking the average of the squared deviations from the mean. The variance has units that are the square of the units of x (e.g. if we are measuring the number of cows in a group of pens, the variance is expressed in terms of cows²). Where N values from a sample are given by x_i and the population mean is μ , the population variance is:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (2)$$

Often the only data available is a sample of the population of interest, so in this case the sample variance (s^2) is calculated. Where n values from a sample are given by x_i and the sample mean is \bar{x} , the sample variance is:

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{(n - 1)} \quad (3)$$

The standard deviation is the positive square root of the variance. Standard deviation has the same units as the units of the variable under investigation. For normal distributions (only), the areas bounded by 1, 2 and 3 standard deviations either side of the mean contain approximately 68%, 95%, and 99% of the distribution.

If you're asked to estimate the standard deviation of a variable that is normally distributed and you have no idea of what it might be estimate the average and the minimum and maximum value. The standard deviation is the difference between the mean and each bound divided by three.

For example, you estimate average daily milk yield in a herd of dairy cows to be 20 litres with a lower and upper bound of 10 litres and 30 litres, respectively. The standard deviation equals $(30 - 10) \div 3 = 10$ litres.

Coefficient of variation

The coefficient of variation is defined as the standard deviation expressed as a percentage of the mean. The coefficient of variation has no dimensions.

Quartiles, deciles and percentiles

Just as the median divides an ordered set of observations into two equal parts, so quartiles are the three values that divide an ordered set of observations into four equal parts. Thus, one quarter of the values are below the lower quartile, half are below the second quartile (the median), and three quarters are below the upper quartile. Deciles divide the observations into ten equal parts and percentiles into 100 equal parts. The interquartile range is the difference between the upper and lower quartiles (the 75th and 25th percentiles). It defines the interval which contains the middle 50% of the ordered observations. The advantage of using interquartile range as a descriptor of variability is that it is not influenced by outliers and is independent of sample size.

Skewness

Skewness refers to the ‘lopsidedness’ of a distribution. If a distribution has negative skewness (sometimes called left skewed) it has a longer tail to the left than to the right. A positively skewed distribution (right skewed) has a longer tail to the right than to the left. Distributions with zero skewness are usually symmetric. Skewness has no units.

Kurtosis

Kurtosis refers to the ‘peakedness’ of a distribution. The greater the kurtosis value, the more peaked the distribution. A normal distribution has a kurtosis of 3. If a distribution has a kurtosis of less than 3 it will be flatter than a normal distribution.

Table 1: Skewness and kurtosis ranges for various distribution functions.

Distribution	Skewness	Kurtosis
Binomial	$-\infty$ to $+\infty$	1 to $+\infty$
Chi square	0 to 2.83	3 to 15
Exponential	2	9
Lognormal	0 to $+\infty$	3 to $+\infty$
Normal	0	3
Poisson	0 to $+\infty$	3 to $+\infty$
Triangular	-0.56 to +0.56	2.4
Uniform	0	1.8

Outliers

Outliers are sample values that cause ‘surprise’ in relation to the majority of the sample. Outliers may be correct, though they should always be carefully checked for transcription errors. Sample means will be influenced by outliers whereas sample medians will not.

2.5 Graphs

Bar graphs

Bar graphs should be used when you are showing segments of information — typically time series data. The double (or group) vertical bar graph is an effective way to compare groups. One disadvantage of vertical bar graphs is that they lack space for text labelling at the foot of each bar. When category labels are too long, horizontal bar graphs might be a better way of displaying information.

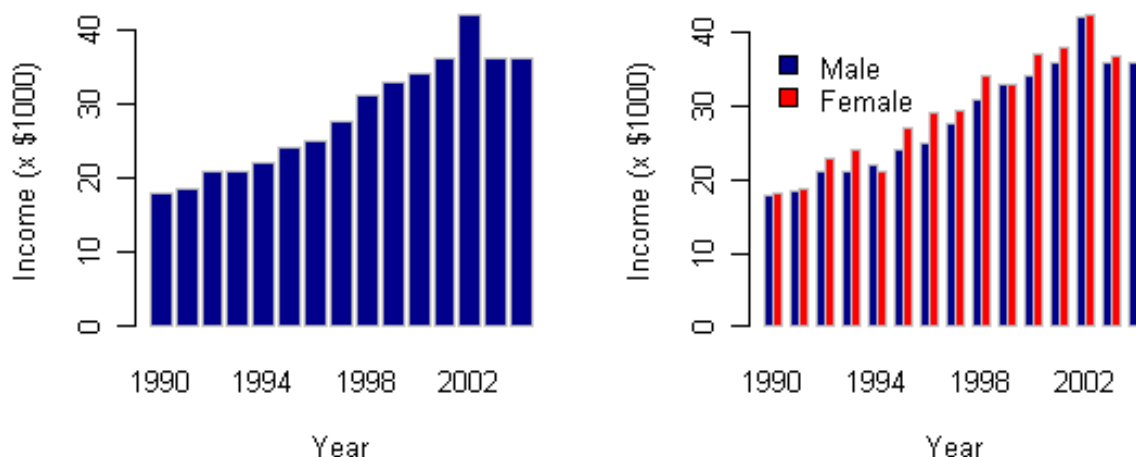


Figure 2: Vertical barplots of annual salaries as a function of time for: (1) male veterinarians (left) and (2) male and female veterinarians.

Histograms

Histograms are used to summarise discrete or continuous data that are measured on an interval scale. A histogram divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles will be drawn of non-uniform height. A histogram has an appearance similar to a vertical bar graph, but when the variables are continuous, there are no gaps between the bars. When the variables are discrete, however, gaps should be left between the bars.

- In a histogram, frequency is measured by the area of the column. In a vertical bar graph, frequency is measured by the height of the bar.
- Histogram are useful to detect any unusual observations (outliers) or any gaps in the data.
- Histograms are often used to illustrate the major features of the distribution of the data.

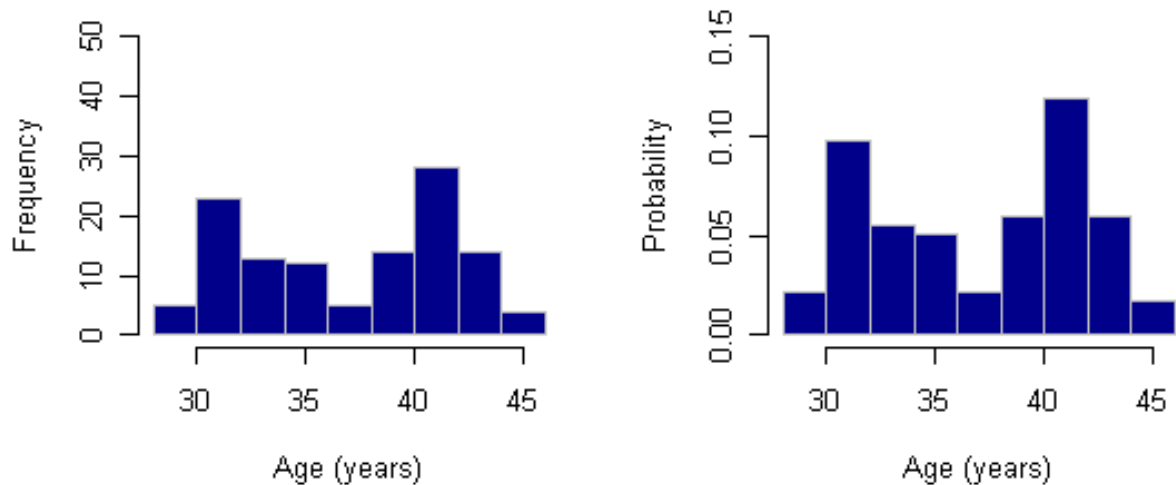


Figure 3: Depression scores in elderly veterinarians: (1) frequency histogram (left) and (2) probability density plot (right).

Box and whisker plots

A box and whisker plot (sometimes called a boxplot) is especially useful for indicating whether a distribution is skewed and whether there are outliers in the data set. Box and whisker plots are also very useful when two or more data sets are being compared.

- The ends of the box are the upper and lower quartiles, so the box spans the interquartile range.
- The median is marked by a line inside the box.
- The whiskers are the two lines outside the box that extend to the highest and lowest observations.

Line graphs

A line graph provides a visual comparison of how two variables are related or vary with each other. The y -axis in a line graph usually indicates quantity (e.g., dollars, litres) or percentage, while the horizontal x -axis often measures units of time.

Although they do not present specific data as well as tables, line graphs are able to show relationships more clearly. Line graphs can also depict multiple series which are usually the best candidate for time series data.

Pie charts

A pie chart is a way of summarising a set of categorical data or displaying the percentage distribution of a given variable. This type of chart is a circle divided into a series of segments.

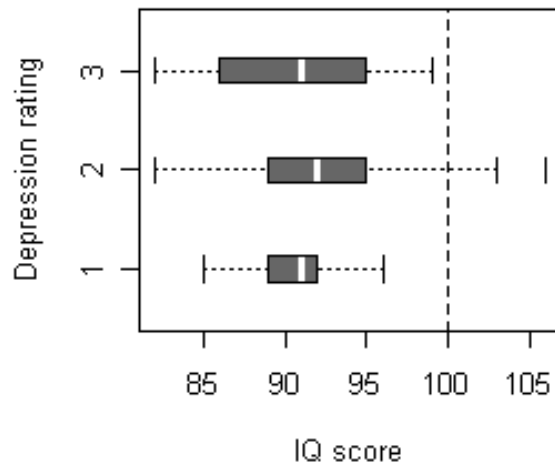


Figure 4: Box and whisker plots showing IQ scores for three levels of depression rating.

Each segment represents a particular category. The area of each segment is the same proportion of a circle as the category is of the total data set. The use of the pie chart is quite popular, as the circle provides a visual concept of the whole (100%). Pie charts should be used sparingly for two reasons. Firstly, they are best used for displaying statistical information when there are no more than six categories (otherwise they appear too complex). Secondly, pie charts are not useful when the values of each component are similar because it is difficult to see the differences between slice sizes.

- When drawing a pie chart, ensure that the segments are ordered by size (largest to smallest) and in a clockwise direction.
- Research has shown that many people can make mistakes when trying to compare pie chart values. In general, bar graphs communicate the same message with less chance of misunderstanding.

Scatterplots

Scatterplots are used to indicate the type and strength of relationship between continuous variables.

Stem and leaf plots

A stem and leaf plot looks something like a bar graph. Each number in the data is broken down into a stem and a leaf, thus the name. The stem of the number includes all but the last digit. The leaf of the number will always be a single digit. The main advantages of stem and leaf plots are: (1) the distribution of the data can be readily appreciated, and (2) all of the original data is shown, as part of the plot.

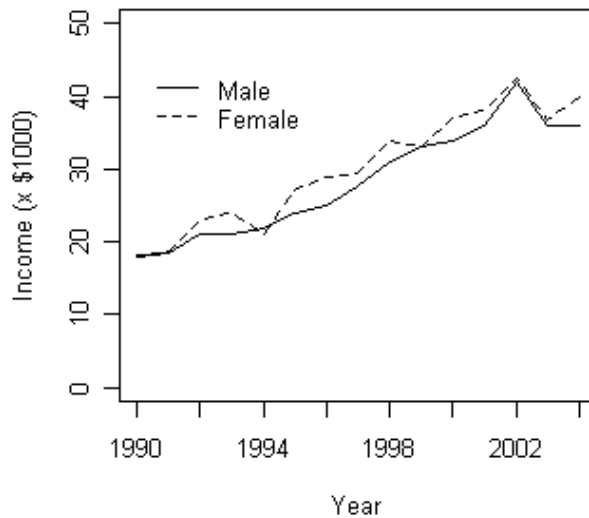


Figure 5: Line plot showing annual salaries of veterinarians as a function of time.



Figure 6: Breakdown of depression scores in elderly veterinarians.

Trellis plots

Trellis (or multipanel) plots are plots are useful for data that are arranged in groups. A plot (line plot, scatter plot, frequency histogram) is produced for each group. Using a common scale for the horizontal and vertical axis allows between group comparisons to be made easily. A trellis plot is shown in Figure 9. Horses were divided into three groups: 2 anthelminic treatment groups (A and B) and a third group which received a placebo (group C). Faecal egg counts were conducted weekly following treatment for a total of 15 weeks. In Figure 9 it is easy to visualise the effect of treatment. Following treatments A and B there is a marked decrease in faecal egg count, followed by a gradual rise in egg count from week 4 on. No such pattern is evident in group C.

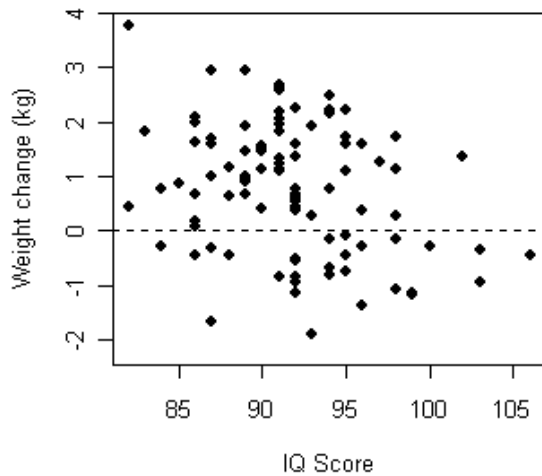


Figure 7: Scatterplot showing weight change (kg) as a function of IQ score.

The decimal point is at the |

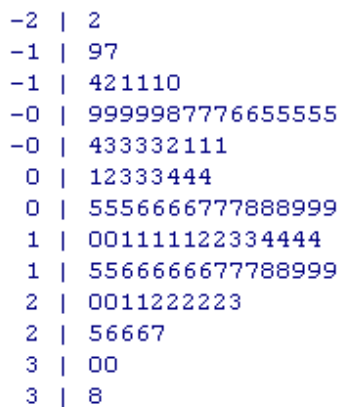


Figure 8: Stem and leafplot showing change in blood pressure (mm Hg) in a group of greyhounds after a period of heavy exercise.

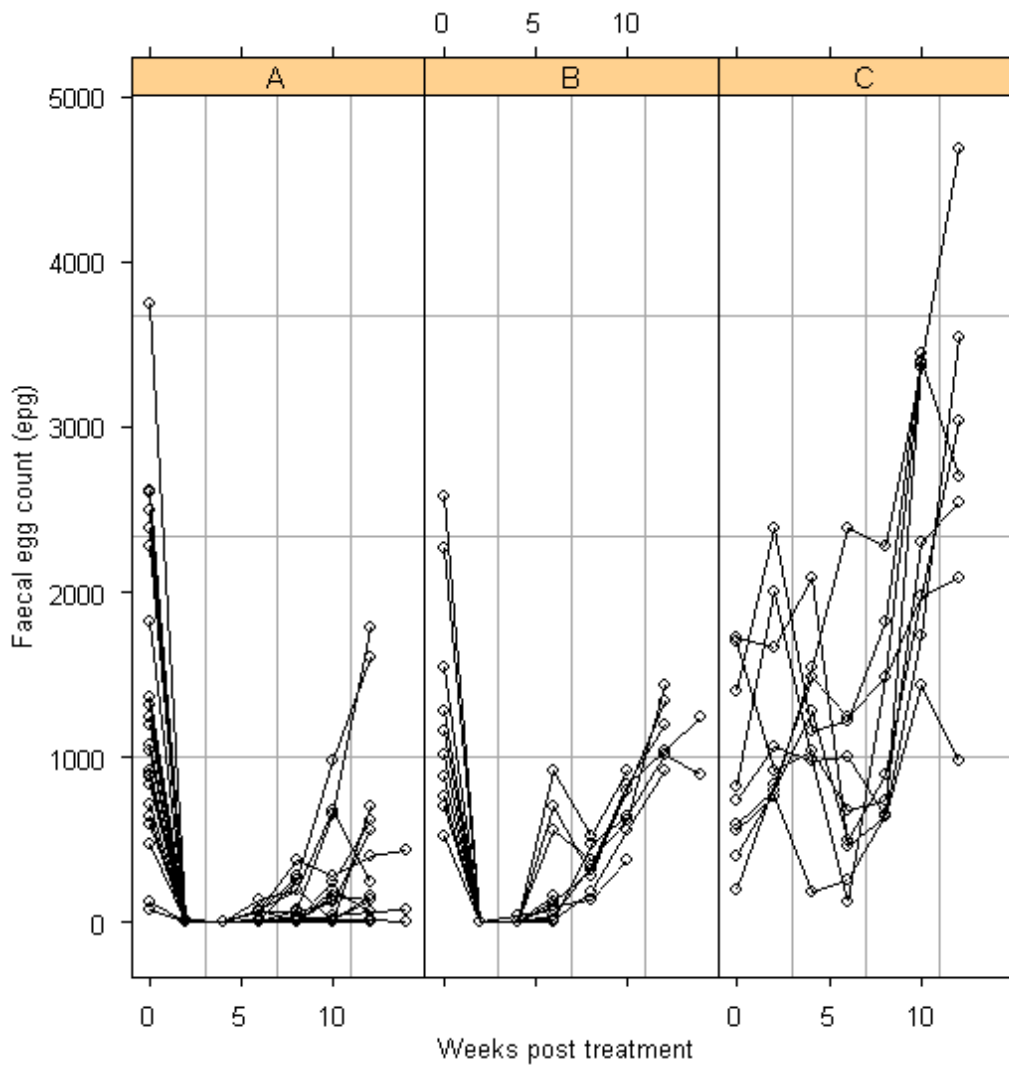


Figure 9: Trellis plot showing faecal egg counts as a function of weeks post anthelmintic treatment for three treatment groups.

3 Probability distributions

In descriptive statistics, we are concerned with learning about the features of a particular data set or describing the observed relationship between variables. In inferential statistics, we would like to be able to draw conclusions about a population given data from a sample drawn from that population.

There are two main modes of statistical inference, usually referred to as frequentist or classical inference, and Bayesian inference. The frequentist school defines the probability of an event as the number of times the event occurs divided by the number of trials, n , as n approaches infinity. For example, the probability that a coin will come up heads is 0.5, since assuming the coin is fair, as the number of trials (flips of the coin) gets larger and larger, the observed proportion will be, on average, closer and closer to 0.5. Similarly, the probability that a surgical technique is successful would be defined as the number of times it is observed to be successful in a large (theoretically infinite) number of trials. While this definition has a certain logic to it, there are some problems. For example, what is the probability it will rain today? Since ‘today’ is a unique event that will not happen an infinite number of times, the above definition cannot be applied. Nevertheless, we often hear statements such as ‘There is a 40% chance of rain today’. Similarly, suppose that a new surgical technique has just been developed, and the surgeon is debating whether or not to apply it to his next patient. Surely the probability of success of the operation compared to the probability of success of the standard procedure for the patients condition will play a large role in the decision, but again, there are as yet no trials (and certainly not an infinite number of trials) upon which to define the probability for this particular patient. While we can conceptualise an infinite number of trials that may occur into the future, this does not help in defining a probability for today’s decision as to which surgery to perform. Clearly, this definition is limited, not only in the case of events that can only happen once, but also because one rarely if ever can observe an infinity of like events.

The second school of thought, often referred to as the Bayesian school, defines the probability of any event occurring as the degree of belief that the event will occur. Therefore, based on the physics of the problem and perhaps a number of test flips, a Bayesian may assert that the probability of a coin flip coming up heads should be close to 0.5. Similarly, based on an assessment that may include both objective data and subjective beliefs about the surgical technique, the surgeon may assert that the probability that the surgery will be successful is 85%. The obvious objection to Bayesian probability statements is that they are ‘subjective’, so different surgeons may state different probabilities of success for the success rate of the surgery, and in general, there is no single ‘correct’ probability statement that may be made about any event, since they reflect personal subjective beliefs. Supporters of the Bayesian viewpoint counter that the frequentist definition of probability is difficult to operationalise in practice, and does not apply to many important situations. Furthermore, the possible lack of agreement on the ‘correct’ probability for any given event can be viewed as an advantage, since it will correctly mirror the range of beliefs that may exist for any event that does not have a large amount of data collected in which to accurately estimate its probability. Hence having a range of probabilities depending on the personal beliefs of a community of surgeons is a useful reflection of reality.

The majority of statistical analyses that appear in medical journals are performed using procedures that arise from the frequentist definition of probability, although over the past 10 years there has been a rapid increase in Bayesian analyses in many applied fields, including medicine.

Many statisticians use the full range of statistical procedures, including Bayesian and frequentist procedures, often switching between the two in analysing the same data set. A key step to learning about frequentist and Bayesian inference is to have a good understanding of the distributional form of the data sets that we work with.

3.1 Characteristics of distributions

Probability distributions can be categorised in terms of being either continuous or discrete, bounded or unbounded, and parametric or non-parametric.

Continuous vs discrete

A discrete distribution may take on a set of identifiable values, each of which has a calculable probability of occurrence (e.g. the number of personnel employed, the number of customers that enter a shop each hour). Discrete distributions include the binomial, hypergeometric, negative binomial and Poisson.

Continuous distributions are used to represent a variable that can take any value within a defined range (e.g. height, weight, milk yield). It is common to use continuous distributions to model variables that are really discrete (e.g. the cost of a project, the number of employees in a large corporation).

Bounded vs unbounded

A distribution that lies within two specified values is said to be bounded. Bounded distributions include uniform, triangular, beta and binomial. Unbounded distributions theoretically extend from minus infinity to plus infinity. Unbounded distributions include the normal and the logistic. Distributions that are contained at one end are partially bounded (e.g. Chi-squared, exponential, Poisson, Weibull).

Parametric vs non-parametric

Parametric distributions are those that require the analyst to have knowledge of the underlying assumptions and associated mathematics. Non-parametric or empirical distributions are defined primarily by the shape of the distribution required (e.g. triangular).

3.2 Ways to express a probability function

Probability density

If an outcome of interest is continuous (e.g. birth weight) then the total range of possible values could be divided into a number of intervals each with a small width (say, 20 g). The set of probabilities corresponding to the birthweight intervals forms a probability distribution. If the intervals are decreased indefinitely then this distribution becomes the probability distribution of a continuous variable. This is known as a probability density function. Figure 10 shows the probability density function for body weight at birth in a group of Friesian calves. In this example the average body weight at birth is 45 kg and the standard deviation 10 kg.

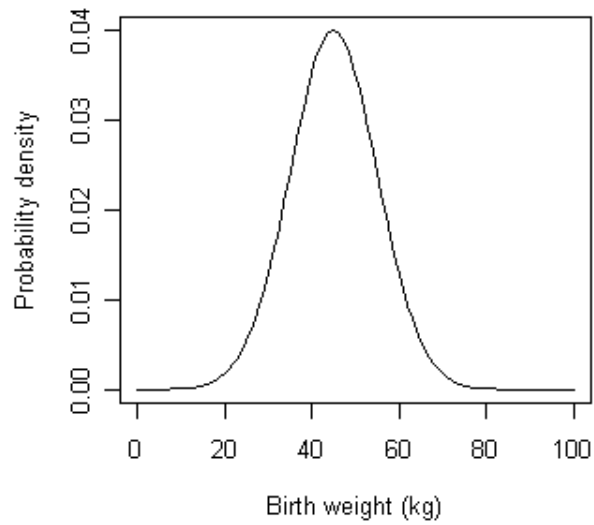


Figure 10: Probability density function for bodyweight at birth in a population of Friesian calves.

Cumulative probability

The data shown in Figure 10 can be re-expressed as a cumulative probability distribution. If the area under the density function in Figure 10 is set to equal one we can plot the cumulative area under the curve as a function of bodyweight at birth (Figure 11). If we plot the distribution in this way we can estimate the probability that a randomly chosen newborn calf will be 30 kg or less (answer: just under 10%).

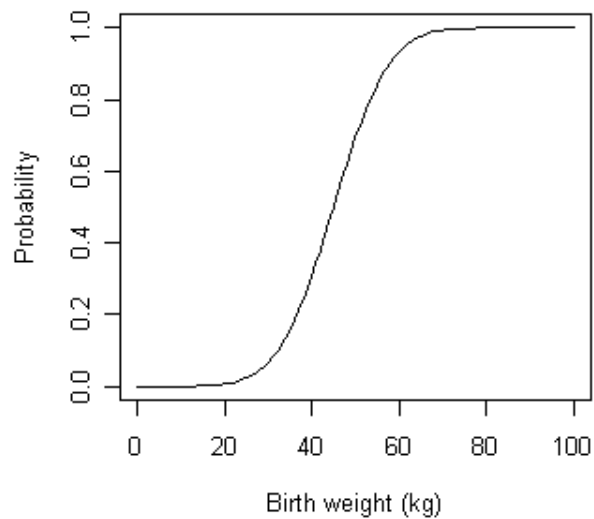


Figure 11: Cumulative probability density function for bodyweight at birth in a population of Friesian calves.

Quantiles of a probability distribution

A quantile plot is similar to the cumulative probability plot with the key difference that the variable of interest (bodyweight, in this example) is plotted on the vertical axis. From Figure 12 we can determine the range of bodyweights at birth for 95% of the population (answer: 28 kg to 61 kg).

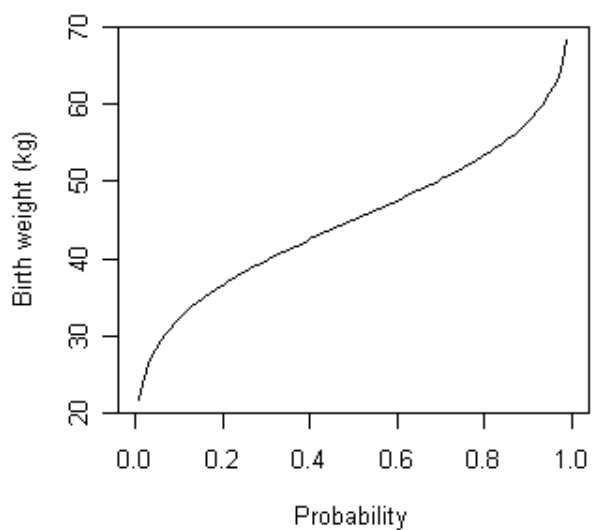


Figure 12: Quantile probability function for bodyweight at birth in a population of Friesian calves.

Random values from a specified probability distribution

The final way to express a probability distribution is to take a series of random draws from the distribution and plot those values as a frequency histogram (Figure 13).

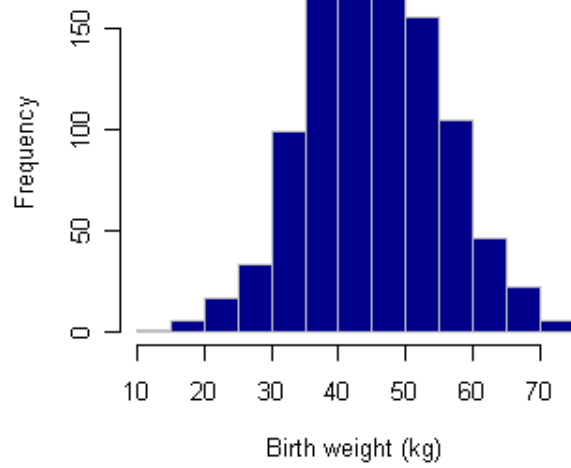


Figure 13: Frequency histogram of 1000 draws from a normal distribution with a mean of 45 and standard deviation of 10.

4 The normal distribution

Perhaps the most commonly used distribution used in statistical practice is the normal distribution. The normal distribution is the familiar ‘bell-shaped’ curve, as shown in Figure 14.

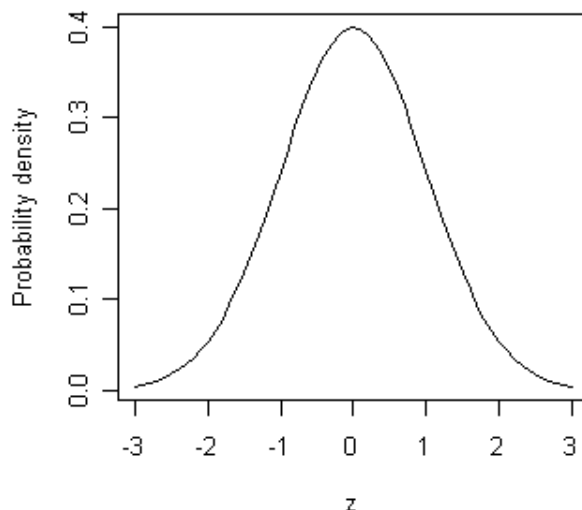


Figure 14: Probability density function for the standardised normal distribution.

Technically, the curve is traced out by the normal density function:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] \quad (4)$$

where \exp denotes the exponential function to the base $e = 2.71828$. The Greek letter μ is the mean of the normal distribution (set to zero in the standard normal curve of Figure 14), and the standard deviation is σ , set to 1 in the standard normal curve. While Figure 14 shows the standard version of the normal curve ($\mu = 0$, $\sigma^2 = 1$), more generally, the mean μ can be any real number and the standard deviation can be any number greater than 0. Changing the mean shifts the curve shown in Figure 14 to the left or right so that it remains centered at the mean, while changing the standard deviation stretches or shrinks the curve around the mean, all while keeping its bell shape. Note that the mean, median and mode (most likely value, i.e., highest point on the curve) of a normal distribution are always the same and equal to μ .

The normal density function has been used to represent the distribution of many measures in biology and medicine. For example, blood pressures, cholesterol levels or bone mineral densities in a population follow a normal distribution with a given mean and standard deviation. It is very unlikely that any of these or other quantities exactly follow a normal distribution. For instance, none of the above mentioned quantities can have negative numbers, while the range of the normal distribution always includes all negative (and all positive) numbers. Nevertheless, for appropriately chosen mean and standard deviation, the probability of out of range numbers will be small, so that this may be of little concern in practice. We may say, for example, that diastolic blood pressure in a given population follows a normal distribution with mean of 80 and a standard deviation of 10, so that the probability of a value less than zero is 6.2×10^{16} ,

and in fact the probability of being less than 50 (three standard deviations below the mean) is only 0.0013. To calculate probabilities associated with the normal distribution, one must find the area under the normal curve. Since this is difficult to calculate by hand statistical tables, spreadsheets, or statistics packages are used.

A study has found that bodyweight of newborn Friesian calves is normally distributed with a mean of 45 kg and SD of 5 kg. What is the probability that a calf will be greater than 55 kg at birth? In Excel:

```
= NORMDIST(55,45,5,TRUE)  
= 0.98
```

The probability that a calf will be up to and including 55 kg is 0.98. The probability that a calf will be greater than 55 kg is $(1 - 0.98) = 0.02$. There is a 2% chance that a calf will be greater than 55 kg at birth.

5 The binomial distribution

One of the most commonly used probability functions is the binomial. The binomial probability function allows one to calculate the probability of obtaining a given number of ‘successes’ in a given number of independent trials. In general, the formula for the binomial probability function is:

$$P_{x,n} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (5)$$

Where where $n!$ is read as ‘ n factorial’, and is shorthand for $n \times (n-1) \times (n-2) \times (n-3) \times \dots \times 3 \times 2 \times 1$. For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, and so on. By convention, $0! = 1$. The probability of a success on each trial is assumed to be p .

Suppose we wish to calculate the probability of $x = 8$ successful surgeries in $n = 10$ operations, where the probability of a successful operation is 70%. From the binomial formula, we can calculate:

$$P(8, 10) = \frac{10!}{8!2!} \times 0.7^8 \times (1-0.7)^2$$

$$P(8, 10) = 0.23$$

So there is a little bit less than a one in four chance of getting 8 successful surgeries in 10 trials. Similarly, the probability of getting 8 or more (that is, 8 or 9 or 10) successful surgeries is found by adding up three probabilities of the above type.

The binomial distribution has a theoretical mean of $n \times p$. For example, if you perform $n = 100$ trials, and on each trial the probability of success is, say, $p = 0.4$ or 40%, then you would intuitively expect $100 \times 0.4 = 40$ successes. The variance of a binomial distribution is $n \times p \times (1-p)$, so that in the above example it would be $100 \times 0.4 \times 0.6 = 24$. Thus the standard deviation is $24^{0.5} = 4.90$, meaning that while on average one expects about 40 successes, one also expects each result to deviate from 40 by an average of about 5 successes.

The **normal** distribution may be used as an approximation to the binomial distribution if the sample size n is moderately large and the expected numbers in both classes are not near to zero. The approximation is best when p is at or near 50% or when both np and $n(1-p)$ are at least 5. Since the normal distribution is for a continuous variable an adjustment is necessary, however. This is done by regarding a discrete count as a continuous variable rounded to the nearest integer. Thus a count of 10 is considered as coming from the range 9.5 to 10.5 of a continuous variable. This is known as continuity correction. For the normal approximation to the binomial we substitute the mean of a binomial distribution for μ and the standard deviation of a binomial distribution for σ in the formula for the **standardised normal deviate**.

6 Confidence intervals

While the p-value provides some information concerning the rarity of events as or more extreme than that observed assuming the null hypothesis to true, it provides no information about the true value of what we are measuring. Say, for example we perform the same surgical procedure on 10 patients and find that on 8 of the 10 occasions, we have a successful outcome. In this case the observed probability of success is 80%, but are we confident that if we will have exactly 80 successful operations when we do 100 procedures? Based on our 10 observations, what can we say about where we would expect the true probability of success to be? One way to answer this question is with a confidence interval. Confidence intervals usually have the form estimate \pm ($k \times$ standard error) where the estimate and standard error are calculated from the data, and k is a constant based on the probability distribution that is appropriate for the data.

To continue the example introduced above. If we observe $x = 80$ successful surgical procedures in $n = 100$ operations, the observed probability of success, p' is 0.80. Note that we use the notation p' rather than p to indicate that this is an estimated probability, not necessarily the true probability, which we denote by p . Following the general formula above, a confidence interval for a binomial probability of success is given by:

$$p' - [z_{1-\alpha/2} \times \sqrt{\frac{p' \times (1 - p')}{n}}] \text{ to } p' + [z_{1-\alpha/2} \times \sqrt{\frac{p' \times (1 - p')}{n}}] \quad (6)$$

where $z_{1-\alpha/2} = 1.96$ for the usual case for a 95% confidence interval.

A total of 80 procedures were successful from 100 that were performed. What is the 95% confidence interval for the probability of surgical success?

$p' = 0.80$

$z = \text{NORMSINV}(0.975) = 1.96$

$n = 100$

Lower bound of $p' = p' - [z \times ((p' \times (1 - p')) / n)^{0.5}]$

Lower bound of $p' = 0.72$

Upper bound of $p' = p' + [z \times ((p' \times (1 - p')) / n)^{0.5}]$

Upper bound of $p' = 0.88$

The 95% confidence interval for the probability of surgical success using this procedure is 0.72 to 0.88.

How do we interpret this confidence interval? The confidence interval tells us that when this surgical procedure is used repeatedly we will capture the true value of p on 95% of occasions and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any single application, of course, the interval either does or does not contain the true proportion p . Note that we are careful not to say that our confidence interval has a 95% probability of containing the true parameter value. In other words, we did not say that the true proportion of successful surgeries is in the interval (0.72, 0.88) with 95% probability. This is because the confidence limits and the true rate are both fixed numbers, and it makes no more sense to say that the true rate is in this interval than it does to say that the number 2 is inside the interval (1, 6) with probability 95%. Of course, 2 is inside this interval, just like the number 8 is outside of the interval (1, 6). However, in the procedure that we used to calculate the above confidence interval, we derived upper and lower limits and in repeated uses of this formula across a range of problems, we expect the random

limits to capture the true value 95% of the time and exclude the true value 5% of the time. Despite their somewhat unnatural interpretation, confidence intervals are generally preferred to p-values. This is because they focus attention on the range of values compatible with the data, on a scale that is of direct clinical interest. In the following section we provide formulae that can be used to calculate confidence intervals for a range of outcomes that are common in medical and veterinary practice. These formulae are provided as a reference only.

6.1 Means and their differences

Single sample

The confidence interval for a population mean based on a single sample is given by:

$$\bar{x} - [t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}] \text{ to } \bar{x} + [t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}] \quad (7)$$

Where n equals the number of samples, $t_{1-\alpha/2}$ is the appropriate value from the t distribution with $n - 1$ degrees of freedom associated with a ‘confidence’ of $100(1 - \alpha)\%$ and s is the sample standard deviation.

You are using a new drug to synchronise oestrus in a group of embryo transfer recipients. It is important that an adequate dose of the drug is administered in order for it to be effective. There are 200 recipients in the mob, all of approximately the same age. You decide to weigh 30 of them and to calculate the 95% confidence interval on the estimate of body weight, then to calculate your dose for the upper limit of the confidence interval. The drug is non-toxic, so overdosing will not be a problem. Based on your sample of 30, the average bodyweight is 350 kg with a standard deviation of 50 kg.

$x' = 350$

$s = 50$

$t = \text{TINV}(0.05, 30 - 1) = 2.04$

$n = 30$

Lower bound of $x' = x' - [t \times (s / n^{0.5})]$

Lower bound of $x' = 331$

Upper bound of $x' = x' + [t \times (s / n^{0.5})]$

Upper bound of $x' = 369$

On the basis of our sample of 30 the 95% confidence interval for body weight in this group of recipients is 331 kg to 369 kg.

Two samples: unpaired case

Suppose x_1 and x_2 are the two sample means and s_1 and s_2 the corresponding standard deviations and n_1 and n_2 the sample sizes. Where d equals the difference in the two means the $100(1 - \alpha)\%$ confidence interval for the difference in the two population means is then:

$$d - [t_{1-\alpha/2} \times \sqrt{\frac{s_1^2}{n_1}}] \text{ to } d + [t_{1-\alpha/2} \times \sqrt{\frac{s_1^2}{n_1}}] \quad (8)$$

Where $t_{1-\alpha/2}$ is the appropriate value from the t distribution with $n_1 + n_2 - 2$ degrees of freedom associated with a confidence of $100(1 - \alpha)\%$. If the standard deviations differ considerably then a common pooled estimate is not appropriate unless a suitable transformation of scale can be found.

Two samples: paired case

Paired data arise in studies of repeated measurements. For such data the same formulae as for the single sample case are used to calculate the confidence interval, where x and SD are now the mean and the standard deviation of the individual subject differences.

Non-normal data

Where data are non-normally distributed, confidence intervals can be estimated by transforming the data to achieve approximate normality. Logarithmic transformations are most frequently used.

6.2 Medians and their differences

The median is the value that divides an ordered set of variables into two equal parts. To find the $100(1 - \alpha)\%$ confidence interval for the population median, we calculate the quantities:

$$r = \frac{n}{2} - (z_{1-\alpha/2} \times \frac{\sqrt{n}}{2}) \text{ and } s = 1 + \frac{n}{2} - (z_{1-\alpha/2} \times \frac{\sqrt{n}}{2}) \quad (9)$$

Where $z_{1-\alpha/2}$ is the appropriate value from the standard normal distribution for the $100(1 - \alpha)$ percentile. The values for r and s are rounded to the nearest integers. The r th and s th observations in the ranking are the $100(1 - \alpha)\%$ confidence interval for the population median.

6.3 Proportions and their differences

Single sample

The following is an alternative method for calculating the confidence interval of a proportion. If r is the observed number successes in n trials, then the estimated proportion of success is $p = r/n$. The proportion of failures is $q = 1 - p$. We first calculate the three quantities:

$$A = 2r + z^2 ; B = z\sqrt{z^2 + 4rq} ; C = 2(n + z^2) \quad (10)$$

Where z is $z_{1-\alpha/2}$ from the standard normal distribution. The confidence interval for the population proportion is given by:

$$\frac{(A - B)}{C} \text{ to } \frac{(A + B)}{C} \quad (11)$$

Two samples: unpaired case

Calculate l_1 and u_1 the lower and upper limits that define the $100(1 - \alpha)\%$ confidence interval for the first sample and l_2 and u_2 the lower and upper limits for the second sample using the method described above. The $100(1 - \alpha)\%$ confidence interval for the population difference in proportions is calculated as:

$$D - \sqrt{(p_1 - l_1)^2 + (u_1 - p_1)^2} \text{ to } D + \sqrt{(p_1 - l_1)^2 + (u_1 - p_1)^2} \quad (12)$$

Two samples: paired case

First, as for the unpaired case, calculate separate $100(1 - \alpha)\%$ confidence intervals for p_1 and p_2 , as l_1 and u_1 and l_2 and u_2 . Next, calculate a quantity ϕ which is used to correct for the fact that p_1 and p_2 are not independent (ϕ is a type of correlation coefficient). If any of the quantities $r + s$, $t + u$, $r + t$, or $s + u$ is zero, then $\phi = 0$. Otherwise, calculate:

$$A = (r + s)(t + u)(r + t)(s + u) \quad (13)$$

Then obtain C as follows:

$$\begin{aligned} C &= B - n/2 && \text{if } B \text{ is greater than } n/2 \\ C &= 0 && \text{if } B \text{ is between } 0 \text{ and } n/2 \\ C &= B && \text{if } B \text{ is less than } 0 \end{aligned}$$

Then calculate $\phi = C/\sqrt{A}$. The $100(1 - \alpha)\%$ confidence interval for the population value of the difference between these proportions (D) is then:

$$D - \sqrt{(p_1 - l_1)^2 - 2\phi(p_1 - l_1)(u_2 - p_2) + (u_2 - p_2)^2} \quad (14)$$

$$\text{to} \quad (15)$$

$$D + \sqrt{(p_1 - l_1)^2 - 2\phi(p_1 - l_1)(u_2 - p_2) + (u_2 - p_2)^2} \quad (16)$$

6.4 Incidence risk

A confidence interval for incidence risk can be calculated using the formula for single sample proportions, described above.

6.5 Incidence rate

If rm is the total individual-time at risk then incidence rate, $p = r/n$. The confidence interval for the logarithm of incidence rate is:

$$\ln(p) - z_{1-\alpha/2} \sqrt{\frac{1}{r}} \text{ to } \ln(p) + z_{1-\alpha/2} \sqrt{\frac{1}{r}} \quad (17)$$

Where $z_{1-\alpha/2}$ is from the standard normal distribution.

You record 8 cases of lameness in a herd of dairy cows. The total at risk period was 200 cow-years. What is the 95% confidence interval for the incidence rate of lameness in this herd?

$$r = 8$$

$$n = 200$$

$$z = \text{NORMSINV}(0.975) = 1.96$$

$$\text{Lower bound of } \ln(p) = \ln(r/n) - [z \times (1 / r)^{0.5}]$$

$$\text{Lower bound of } \ln(p) = -3.90$$

$$\text{Lower bound of } p = \exp(-3.90)$$

$$\text{Lower bound of } p = 0.02$$

$$\text{Upper bound of } \ln(p) = \ln(r/n) + [z \times (1 / r)^{0.5}]$$

$$\text{Upper bound of } \ln(p) = -2.53$$

$$\text{Upper bound of } p = \exp(-2.53)$$

$$\text{Upper bound of } p = 0.08$$

The 95% confidence interval for the incidence rate of lameness in this herd was 2 – 8 cases of lameness per 100 cow-years at risk.

7 Statistical inference

Experiments and observational studies are carried out to provide data to answer scientific questions, that is, to test hypotheses.

- Are body weight gains in pigs weaned at 21 days of age greater than those weaned at 28 days?
- Does regular exercise reduce the likelihood of recurrence of FUS in cats?

Data on these two questions may be obtained by carrying out an epidemiological study and a randomised controlled trial respectively. The data then have to be analysed in such a way as to answer the original question. This process is called **hypothesis testing**. The general principles of hypothesis testing are:

- Formulate a null hypothesis. Usually the null hypothesis is that the effect to be tested does not exist.
- Collect data.
- Calculate the probability (p) of these data occurring if the null hypothesis were true.
- If p is large, the data are consistent with the null hypothesis. We conclude that there is no strong evidence that the effect being tested exists (this is not the same as saying that the null hypothesis is true — it may be false but the study was not large enough to detect the departure from the null hypothesis).
- If p is small, we reject the null hypothesis. We conclude that the observed findings were unlikely to have occurred by chance.

The dividing line between ‘large’ and ‘small’ p values is called the significance α (alpha) level. Usually α is chosen as 0.05, 0.01, or 0.001 and a significant result is indicated by ‘ $p < 0.05$ ’ or ‘significant at the α level of 0.05.’ On the other hand, $p > 0.05$ is usually regarded as not statistically significant (NS). Note however that when p is small there are two possibilities:

1. The null hypothesis is true and an event of low probability has occurred by chance.
2. The null hypothesis is not true and can be rejected in favour of the alternative hypothesis.

In the pig weight gain example above, the null hypothesis would be that body weight gains in pigs weaned at 21 days are the same as pigs weaned at 28 days of age. Only if the data appeared inconsistent with this null hypothesis would we feel confident to claim that weight gains were greater in pigs that were weaned early. In the FUS example the null hypothesis would be that the rate of re-occurrence of FUS in cats that were regularly exercised was the same as those who received no exercise. We could conclude that exercise was beneficial only if the data were inconsistent with the null hypothesis.

7.1 Statistical significance and confidence intervals

The use of statistics in biomedical journals over recent decades has increased exponentially. Associated with this increase has been an unfortunate trend away from examining basic results towards an undue concentration on ‘hypothesis testing.’ In this approach, data are examined in relation to a statistical ‘null’ hypothesis and the practice has led to a mistaken belief that studies should aim at attaining ‘statistical significance.’ Contrary to this paradigm is that most research questions in medicine are aimed at determining the magnitude of some factor(s) of interest on an outcome.

The common statements ‘ $p < 0.05$ ’ and ‘ $p = NS$ ’ convey little information about a study’s findings and rely on an arbitrary convention of using the 5% level of statistical significance to define two alternative outcomes: significant or not significant. Furthermore, even precise p values convey nothing about the sizes of the differences between study groups. In addition, there is a tendency to equate statistical significance with medical importance or biological relevance, however small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small.

It is therefore good practice when reporting the results of an analysis involving significance tests to give estimates of the sizes of the effects, both point estimates and confidence intervals. Then readers can make their own interpretation, depending on what they consider to be an important difference (which is not a statistical question). Figure 15) shows the point estimate and 95% confidence intervals for five outcome in relation to the null hypothesis and an arbitrary measure which we call ‘clinical importance.’

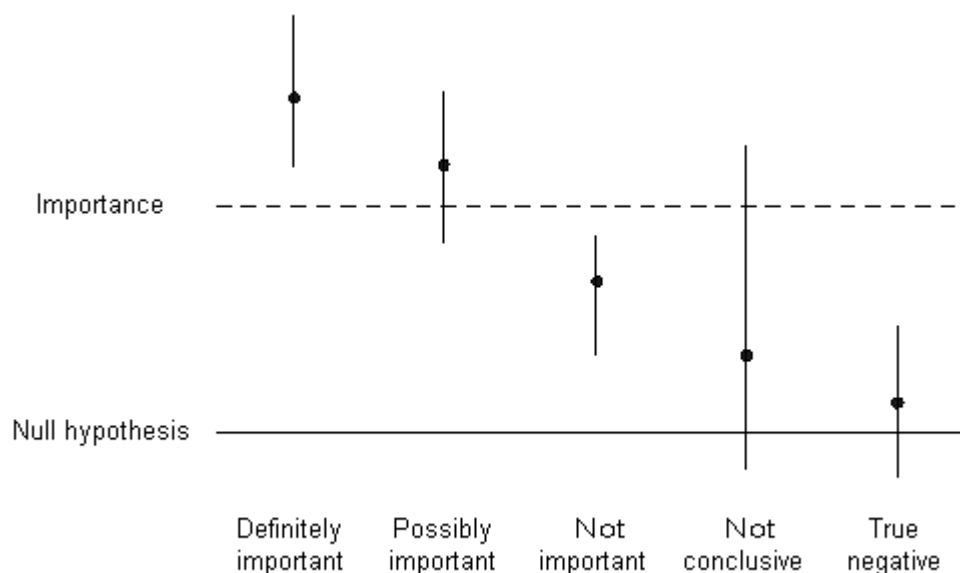


Figure 15: Confidence intervals showing the five possible conclusions in terms of statistical significance and practical importance.

The five possibilities (as shown in Figure 15) are:

1. The difference is significant and certainly large enough to be of practical importance — ‘definitely important.’
2. The difference is significant but it is unclear whether it is large enough to be important — ‘possibly important.’
3. The difference is significant but too small to be of practical importance — ‘not important.’
4. The difference is not significant but may be large enough to be important — ‘not conclusive.’
5. The difference is not significant and also not large enough to be of practical importance — ‘true negative.’

7.2 Steps involved in testing significance

The full answer to any exercise involving a significance test should include:

1. A statement of the null hypothesis.
2. Calculation of test statistic and its associated p value.
3. A statement of conclusion, which should include: (a) the significance or otherwise of the effect being tested, (b) supporting statistics (the test statistic, degrees of freedom, and p value), and (c) an estimate of effect (the point estimate and its confidence interval).

We wish to compare conception rates among cows where oestrus has been induced using a CIDR device and cows where oestrus has occurred naturally. There were 53 services applied to CIDR-induced oestrus events. Of these 53 services, 23 resulted in conception. There were 124 services applied to natural oestrus events. Of these 124 services, 71 resulted in conception. A chi-squared test will be used to compare the two proportions (that is, to test the hypothesis that the proportions $23/53$ and $71/124$ do not differ). The null hypothesis is that conception rates for CIDR-induced oestrus events are equal to conception rates for natural oestrus events.

The chi-squared test statistic, calculated from these data is 2.87. The number of degrees of freedom is 1. The p value corresponding to this test statistic and degrees of freedom is 0.09.

Since our observed p value is greater than 0.05 we accept the null hypothesis and conclude that conception rates for CIDR-induced oestrus events are equal to conception rates for natural oestrus events (chi-squared test statistic = 2.86, $df = 1$, $P = 0.09$). The conception rate for CIDR-induced oestrus events was 43% (95% CI 31% to 57%). The conception rate for natural oestrus events was 57% (95% CI 48% to 66%).

8 Inference for proportions

We wish to compare conception rates among cows where oestrus has been induced using a CIDR device and cows where oestrus has occurred naturally (Table 2).

Table 2: Data from a study comparing conception rates for induced and naturally occurring oestrus events.

	Conceived +	Conceived -	Total
CIDR +	23	30	53
CIDR -	71	53	124
Total	94	83	177

We would like to draw inferences about whether the probability of conception was higher (or lower) for CIDR-induced oestrus events, compared with those oestrus events that occurred naturally. There were 53 services applied to CIDR-induced oestrus events. Of these 53 services, 23 resulted in conception. There were 124 services applied to natural oestrus events. Of these 124 services, 71 resulted in conception. The conception rate for CIDR-induced oestrus events was 43%. The conception rate for natural oestrus events was 57%. As well as reporting the point estimate of conception probability for each group it would also be useful to provide confidence intervals around these estimates. Recall from an earlier chapter the following formula:

$$p' - z \times \sqrt{\frac{p' \times (1 - p')}{n}} \text{ to } p' + z \times \sqrt{\frac{p' \times (1 - p')}{n}} \quad (18)$$

where $z = 1.96$ for the usual case where we calculate a 95% confidence interval. We can now make the statement that the conception rate for CIDR-induced oestrus events was 43% (95% CI 31% to 57%) and the conception rate for natural oestrus events was 57% (95% CI 48% to 66%). Although conception probability for CIDR induced oestrus events was lower than that observed for natural oestrus events we note that the 95% confidence intervals overlap. This raises two possibilities, either: (1) there is no difference between the two groups, or (2) there is a difference between the two groups but we were unable to detect this difference in this particular study because our sample size was too small.

Although confidence intervals are preferred we will also discuss hypothesis testing for proportions, since we often see such tests reported in the literature. Suppose we wish to test the null hypothesis that $p_1 = p_2$, that is, the null hypothesis states that the success rates are identical in the two groups. The expected values for each cell of our 2×2 table are shown in Table 3.

Why do we ‘expect’ to observe this table of data if the null hypothesis is true? We have observed a total number of 94 ‘successes’ (conception positive events) divided among the two groups. If $p_1 = p_2$ and if the sample sizes were equal in the two groups, we would have expected $94/2 = 47$ successes in each group. However, since the sample sizes are not equal, we expect $94 \times 53/177 = 28.15$ to go to the CIDR positive group, and $94 \times 124/177 = 67.45$ to go the natural oestrus (CIDR negative) group. Similarly, expected values for the ‘failures’ can be calculated. Observed discrepancies from these expected values are evidence against the null hypothesis. We calculate the χ^2 (read ‘chi-squared’) test statistic using the following formula:

Table 3: Expected values from a study comparing conception rates for induced and naturally occurring oestrus events when the null hypothesis is true.

	Conceived +	Conceived -	Total
CIDR +	28.15	24.85	53
CIDR -	65.85	58.15	124
Total	94	83	177

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (19)$$

For the CIDR data we have $\chi^2 = 0.94 + 1.07 + 0.40 + 0.46 = 2.87$. We now compare this value with the χ^2 distribution on 1 degree of freedom. The P value corresponding to this test statistic and degrees of freedom is 0.09 (from statistical tables).

In Excel use the CHIDIST function to return a p value given a specified test statistic and degrees of freedom. For this example:
 = CHIDIST(2.87,1)
 = 0.09

We now have evidence (at the alpha level of 0.05) to accept the null hypothesis. More formally, we would make a statement like: ‘we accept the accept the null hypothesis that conception rates for CIDR-induced oestrus events are equal to conception rates for natural oestrus events (χ^2 test statistic = 2.87, df = 1, P = 0.09).’ This finding is consistent with our conclusion from the confidence interval, but note that the confidence interval is more informative than simply looking at the p-value from the χ^2 test, since a range for the difference in proportions is provided.

The χ^2 test can be extended to include tables larger than the so-called 2×2 table in the above example. For example, a 3×2 table could arise if we included (for example) prostaglandin as a means for inducing oestrus events in our comparison. What we would do here is sum over $3 \times 2 = 6$ terms rather than the four terms of a 2×2 table. With a 2×2 table the degrees of freedom is always equal to one, in general, the degrees of freedom for χ^2 tests is given by $(r - 1) \times (c - 1)$, where the number of rows in the table is r , and the number of columns is c . In order for the χ^2 test to be valid, we need to ensure that the expected values for each cell in the table is at least five. Fishers Exact Test should be used if this criterion is not satisfied for a particular table. The Fishers Exact test is valid for tables of any size.

9 Inference for means

We wish to compare total lactation milksolids yields from cows that were fed two different diets. A summary of the data collected is shown in Table 4.

Table 4: Data from a study comparing total lactation milksolids yields in two groups of dairy cows.

Diet	n	Mean (SD)	Median (Q1, Q3)	Range
A	140	502 (80)	590 (510, 600)	306, 676
B	26	526 (54)	524 (480, 610)	430, 680

At a first glance it appears that cows on diet B produced more than cows on diet A. Cows on diet A produced an average of 502 kg milksolids. Cows on diet B produced an average of 526 kg milksolids. As well as reporting the point estimates total lactation yields for the two groups we can provide confidence intervals around these estimates. Recall that the confidence interval for a population mean is derived using the mean of a sample and its standard error from a sample of size n . The confidence interval is given by:

$$\bar{x} - [t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}] \text{ to } \bar{x} + [t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}] \quad (20)$$

Where n equals the number of samples, $t_{1-\alpha/2}$ is the appropriate value from the t distribution with $n - 1$ degrees of freedom associated with a ‘confidence’ of $100(1 - \alpha)\%$ and s is the sample standard deviation. Using this equation the 95% confidence interval for total lactation milksolids yield for cows on diet A was 489 kg to 515 kg compared with 504 kg to 548 kg for cows on diet B. While the observed mean lactation milksolids yield for cows on diet A was less than that of cows on diet B the two confidence intervals are wide and largely overlap so that we do not seem to have strong evidence from this data for an effect of diet on total lactation yield.

As with proportions, hypothesis tests are available to supplement the confidence intervals, although we would again advise that once confidence intervals are calculated, hypothesis tests add little (if any) additional clinically useful information. Nevertheless, for the sake of completeness, we provide the formulae for one sample tests for a single mean and the difference between two means. To test the null hypothesis that a single mean μ has value μ_0 , we calculate the test statistic:

$$s = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (21)$$

and use this test statistic to determine the appropriate p-value from a t distribution with $n - 1$ degrees of freedom. Say, for example we would like to test the hypothesis that total lactation milksolids yield from cows on diet A was equal to 510 kg. The test statistic in this case equals -1.18. The p-value corresponding to a test statistic of -1.18 in this case is 0.12. We accept the null hypothesis and conclude that the mean of the lactation yields for cows on diet A is 510 kg milksolids. The next thing we might want to do is test the hypothesis that the total lactation milksolids yields from cows on the two diets are equal. The null hypothesis is that total lactation

yields for cows on each diet are the same. The alternative hypothesis is that they differ. We calculate the test statistic for a difference in two means as follows:

$$s = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (22)$$

Applying the above formula our test statistic is -1.91. The corresponding p value is 0.03. Is this ‘significant’? To answer this question we must choose between a one- and two-sided test. A one-sided test is used when the direction of the difference between two groups is known in advance or when differences observed in the opposite direction are not of interest or are not possible. For example, a drug may increase the length of long bones but will not decrease it. A study of changes in bone length is thus not concerned with the probability that bones will be shorter at the end of the study, only the probability that they will be longer. A one-sided test is appropriate in this circumstance. If the bones could plausibly be either shorted or longer, a two-sided test would be appropriate. For a one-sided test, the critical probability value for declaring a test significant equals α . For a two-sided test, the critical probability value for declaring a test significant equals $\alpha/2$. Thus in the usual case, where α is set to 0.05 the critical probability value to declare significance for a two-sided test is 0.025. Returning to the example, the p value for our lactation yield data is 0.03. Because a two-sided test is appropriate in this case (there might be a positive or negative difference in milksolids yield) the critical probability value is 0.025. Since 0.03 is greater than 0.025, we accept the null hypothesis and conclude that total lactation yields for cows on the two diets are the same.

9.1 Paired versus unpaired tests

The total lactation milksolids data is an example of where we are comparing two independent samples. In some experiments, for example, if one wishes to compare quality of life before and after surgery is performed, a paired design is appropriate. Here one would subtract the value measured on an appropriate quality of life scale before the surgery to that measured on the same scale after the surgery to create a single set of before to after differences. Once this subtraction has been done for each patient, one in fact has reduced the two sets of before and after values on each patient to a single set of numbers representing the differences. Therefore, paired data can be analysed using the same formulae as used for single sample analyses. Paired designs are often more efficient than unpaired designs.

9.2 Equal or unequal variances

The tests and confidence intervals given above assume that the variances in the two groups are unequal. Slightly more efficient formulae can be derived if the variances are the same, as a single pooled estimate of the variance can be derived from combining the information in both samples together. We do not discuss pooled variances further here, firstly because in practice the difference in analyses done with pooled or unpooled variances is usually quite small, and secondly because it is rarely appropriate to pool the variances, since the variability is usually not exactly the same in both groups.

10 Correlation coefficients

In a study of factors influencing reproductive performance of dairy cows, calving to conception intervals were plotted as a function of age, as shown in Figure 16.

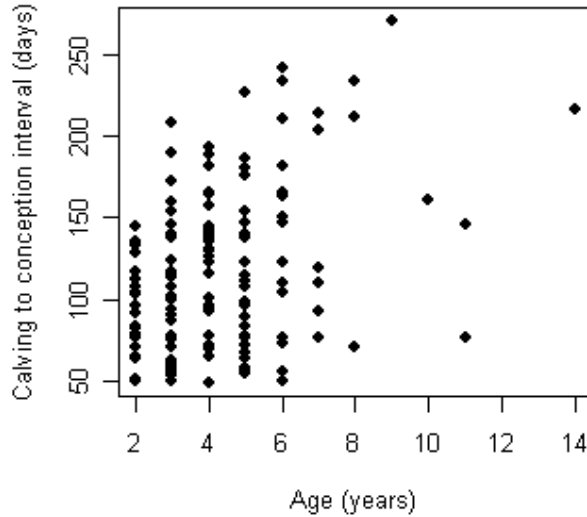


Figure 16: Scatter plot showing calving to conception interval as a function of age (in years).

Looking at Figure 16 there seems to be a moderate relationship between age and calving to conception interval: as cows get older calving to conception intervals increase. We can use Pearson's correlation coefficient to measure the association between age and conception interval.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (23)$$

Pearson's correlation coefficient ranges from -1 (perfect negative correlation: when one variable increases, the other decreases) to 1 (perfect positive correlation: when one variable increases, the other also increases) with 0 indicating no relationship. Applied to the data from Figure 16 we find that the correlation $r = 0.32$, which is, as expected, moderate. Pearson's correlation measures the strength of the linear (straight line) relationship between two variables, but may not work well if the relationship is highly nonlinear. For example, if the relationship between two variables follows a parabolic curve, the correlation may appear to be near 0, even though a strong (but nonlinear) relationship exists. One must also be aware that even very high correlations do not necessarily imply that a causal relationship exists between two variables.

Spearman's correlation coefficient is a nonparametric version of Pearson's correlation coefficient, wherein one first ranks the data (separately for each variable), leading to pairs of data of ranks, rather than 'raw' values. Pearson's formula is then applied to the ranked data to obtain the Spearman correlation.

11 Inference for non-parametric distributions

So far, statistical inferences on populations have been made by assuming that the data we are working with has properties consistent with a defined distribution (e.g. the Normal distribution). Parameters for that distribution can then be estimated, based on the observed data. Once the parameters have been estimated (for example, the mean and/or variance for a Normal distribution) we say that the distribution is fully specified. This is known as parametric inference. Sometimes we may be unwilling to specify in advance the general shape of the distribution, and prefer to base the inference only on the data, without a parametric model. In this case, we use distribution free, or nonparametric methods. Consider the following data which relates to the number of days 24 dogs were hospitalised following conventional cruciate ligament repair and after application of a new method, which we call a ‘non-invasive’ procedure:

Table 5: Number of days hospitalised for 24 dogs undergoing conventional and non-invasive cruciate ligament repair surgery.

Surgery method	Days hospitalised
Conventional	21, 12, 11, 28, 3, 10, 9, 5, 7, 10, 6
Non-invasive	4, 3, 4, 5, 20, 22, 5, 12, 15, 5, 1, 14, 3

We plot the hospitalisation times as frequency histograms to visualisation the distributional properties of the data, Figure 17.

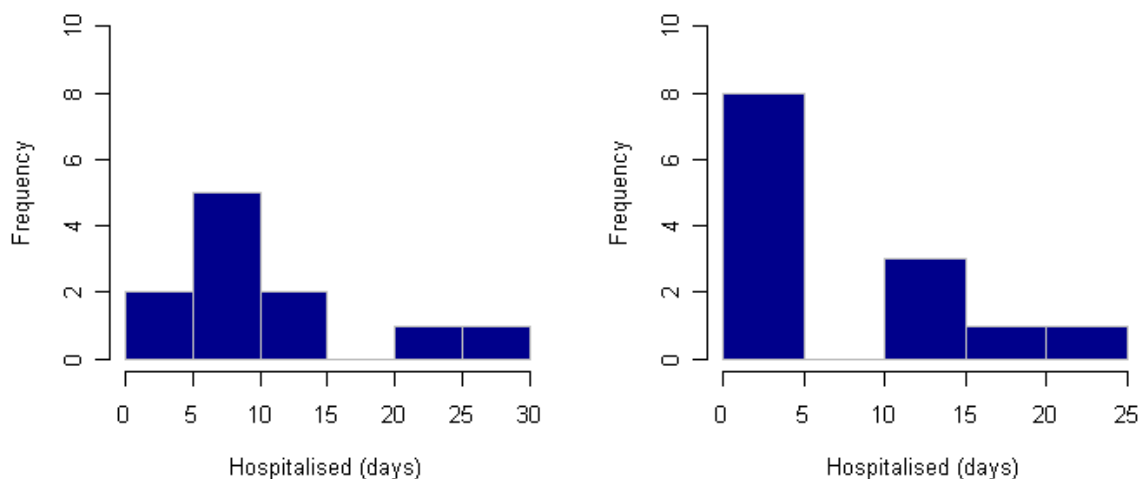


Figure 17: Frequency histograms showing the number of days of hospitalisation following: (1) conventional cruciate ligament repair (left); and (2) non-invasive cruciate ligament repair (right).

From Figure 17 it is clear that hospitalisation times are non-normally distributed. This being the case it is no longer appropriate to use tests of similarity of group means. Rather, the null and alternative hypotheses here are: H_0 : There is no treatment effect, the non-invasive method of cruciate ligament repair gives rise to convalescence days similar to those from the conventional surgery group. H_1 : Non-invasive surgery gives rise to different values for convalescence days compared with those from the conventional surgery group.

In order to test these hypotheses, the first step is to order and rank the data from lowest to highest values, keeping track of which data points belong to each treatment group.

Table 6: Number of days hospitalised for 24 dogs undergoing conventional and non-invasive cruciate ligament repair surgery, ranked data.

Group	Days	Rank	Ties
Non-invasive	1	1	1
Non-invasive	3	2	2.5
Conventional	3	3	2.5
Non-invasive	4	4	4.5
Non-invasive	4	5	4.5
Non-invasive	5	6	7.5
Non-invasive	5	7	7.5
Non-invasive	5	8	7.5
Conventional	5	9	7.5
Conventional	6	10	10
Conventional	7	11	11
Conventional	7	12	12
Conventional	10	13	13.5
Conventional	10	14	13.5
Conventional	11	15	15
Conventional	12	16	16.5
Non-invasive	12	17	16.5
Non-invasive	13	18	18
Non-invasive	14	19	19
Non-invasive	15	20	20
Non-invasive	20	21	21
Conventional	21	22	22
Non-invasive	22	23	23
Conventional	28	24	24

By ranking the data, we simply sort the data from the smallest to the largest value regardless of group membership, and assign a rank to each data point depending on where its value lies in relation to other values in the data set. The lowest value receives a rank of 1, the second lowest a rank of 2, and so on. Since there are many ‘ties’ in this data set, we need to rank the data accounting for the ties, which we do by grouping all tied values together, and distributing the sum of the available ranks evenly among the tied values. For example, the second and third lowest values in this data set are both equal to 3, and there is a total of $2 + 3 = 5$ ranks to be divided among them. Hence each of these values receives a rank of $5/2 = 2.5$. Similarly, the 6th through 9th values are all tied at 5. There are $6 + 7 + 8 + 9 = 30$ total ranks to divide up amongst 4 tied values, so each receives a value of $30/4 = 7.5$, and so on. The next step is to sum the ranks for the values belonging to the conventional surgery group, which gives 147.5. We now reason as follows: There is a total of $1 + 2 + 3 + \dots + 23 + 24 = 300$ ranks

that can be distributed among the conventional and non-invasive surgery groups. If the sample sizes were equal, and if the null hypothesis were true, we would expect that these ranks should divide equally among the two groups, so that each would have a sum of ranks of 150. Now the sample sizes are not quite equal, so that here we expect $300 \times (11/24) = 137.5$ of the ranks to go to the conventional group, and $300 \times (13/24) = 162.5$ of the ranks to go to the non-invasive group (note that $137.5 + 162.5 = 300$ which is the total sum of ranks available). We have in fact observed a sum of ranks of 147.5 in the conventional group, which is higher than expected. Is it high enough that we can reject the null hypothesis?

For this we must refer to computer programs that will calculate the probability of obtaining a sum of ranks of 147.5 or greater given that the null hypothesis of no treatment difference is true. Most statistical computer packages will carry out this calculation, which in this case gives 0.58. We accept the null hypothesis and conclude that the non-invasive method of cruciate ligament repair gives rise to convalescence days similar to those from the conventional surgery group.

This nonparametric test is called the Wilcoxon rank sum test. The equivalent unpaired t-test for the same data also give a p-value of $p = 0.58$, so that the same conclusion is reached. Since the two tests do not always provide the same conclusions, which of these tests is to be preferred? The answer is situation specific. Remember that the t-test assumes either that the data are from a normal distribution (so here, it would imply that the days to convalescence are approximately normally distributed), or that the sample size is large. Figure 17 shows that the data are skewed towards the right, so that normality is unlikely, and the sample sizes of 11 and 13 are not large. Hence in this example the nonparametric test is preferred, since the assumptions behind the t-test do not hold. In general, if the assumptions required by a parametric test may not hold, a nonparametric test is to be preferred, while if the distributional assumptions do hold, a parametric test provides better power compared to a nonparametric test. The Wilcoxon rank sum test is appropriate for unpaired designs. A similar test exists for paired designs, called the Wilcoxon signed rank test.

12 Analysis of variance

Whereas the t test is a hypothesis test used to compare two groups (or categories) of a continuous variable, analysis of variance (ANOVA) is used to compare three or more groups. ANOVA is closely related to regression analysis: regression analysis is used to quantify the effect of continuous and categorical variables on an outcome, ANOVA quantifies the effect of categorical variables.

One way ANOVA (also known as univariate ANOVA, simple ANOVA, single classification ANOVA, or one-factor ANOVA) assesses the effect of a single categorical explanatory variable (a factor) on a single continuous outcome variable. It tests whether the groups formed by the explanatory variable seem similar (specifically that they have the same pattern of dispersion as measured by comparing estimates of group variances). If the groups are not similar, then it is concluded that the explanatory variable has an effect on the outcome variable.

Two way ANOVA assesses the effect of two categorical explanatory variables on a single continuous outcome variable. Multiway ANOVA assesses the effect of three or more categorical explanatory variables on a single continuous outcome variable. It should be noted that as the number of explanatory variables increases, the number of potential interactions increases. Two explanatory variables (A and B) have a single first-order interaction (AB). Three explanatory variables (A, B, and C) have three first order interactions (AB, AC, and BC) and one second-order interaction (ABC). Four explanatory variables (A, B, C, and D) have six first-order (AB, AC, AD, BC, BC, and CD), three second-order (ABC, ACD, and BCD), and one third-order (ABCD) interaction.

Analysis of covariance (ANCOVA) assesses the effect of one or more categorical variables while controlling for the effect of other explanatory variables (called covariates) on a single outcome variable.

12.1 One way ANOVA

Many experiments involve exposing experimental units to a set of categorical variables known as factors. A factor might be drug treatment for a particular illness with levels corresponding to a placebo and (say) two or three alternative treatments. Alternatively, a factor might be mineral fertiliser, where four levels represent four different mixtures of nitrogen, phosphorus and potassium. Factors are often used in experimental designs to represent statistical blocks; these are internally homogeneous units in which each of the experimental treatments is repeated. Blocks may be different fields in an agricultural trial, different genotypes in a plant physiology experiment, or different growth chambers in a study of insect photoperiodism. It is important to understand that regression and ANOVA are identical analytical techniques except for the nature of the explanatory variables. In ANOVA the explanatory variables are all categorical. In regression the explanatory variables can be a mixture of categorical or continuous variables. Having said this, it should be noted that some experiments combine regression and analysis of variance by fitting a series of regression lines, one in each of several levels of a given factor (this is called analysis of covariance, ANCOVA).

The emphasis in ANOVA has traditionally been on hypothesis testing. The aim of ANOVA is to estimate means and standard errors of differences between means. Comparing two means using

a t -test involves calculating the difference between the two means, dividing by the standard error of the difference, and then comparing the resulting statistic with the value of Student's t from tables. The means are said to be significantly different when the calculated value of t is greater than the critical value. For large samples ($n > 30$) a useful rule of thumb is that a t -value greater than 2 is significant. In ANOVA we are concerned with cases where we want to compare three or more means. For the two sample case, the t -test and the ANOVA are identical, and the t -test is to be preferred because it is simpler.

Suppose we have just two levels of a single factor. We plot the data in the order in which they were measured: first for the first level of the factor, y_A (observations 1 to 5) and then for the second level, y_B (observations 6 to 10) as shown in Figure 18.

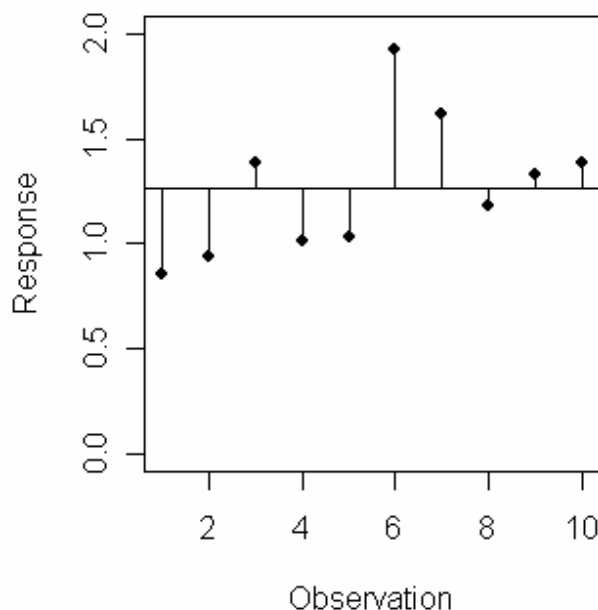


Figure 18: Observation number as a function of response. Observations 1 to 5: factor 1; observations 6 to 10: factor 2.

Figure 18 shows the variation in the data set as a whole. We can calculate the difference between each data point and the overall mean. The total sum of squares (SST) is the sum of the squares of these differences:

$$SST = \sum y - \bar{y}^2 \quad (24)$$

Next we can look at the means for each group (y_A and y_B) and consider the sum of squares of the differences between each y value and its own group mean. We call this SSE, the error sum of squares:

$$SSE = \sum y_A - \bar{y}_A^2 + \sum y_B - \bar{y}_B^2 \quad (25)$$

Conceptually, what we are doing is shown in Figure 19. If the two means were identical then the SSE should be equal to the SST, because the two horizontal lines in Figure 19 would be in the same position as the single line in Figure 18. If the group means were different, then SSE will be less than SST. In the limit, SSE would be zero if the replicates from each group fell exactly on their respective means. This is how analysis of variance works: you can make inferences about

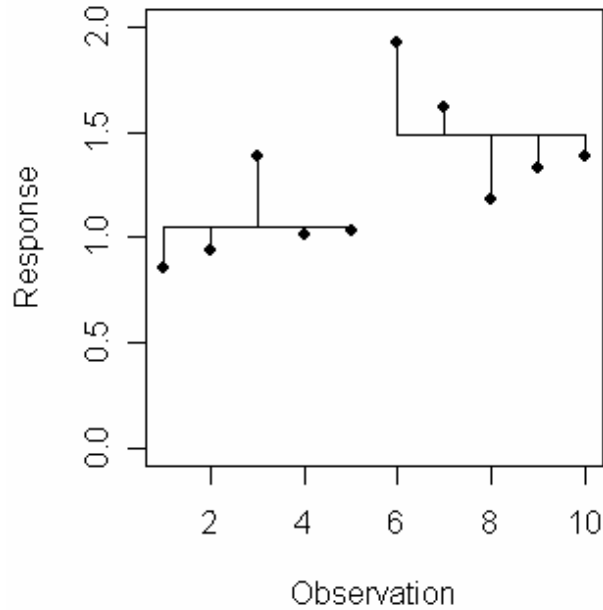


Figure 19: Observation number as a function of response. Observations 1 to 5: factor 1; observations 6 to 10: factor 2.

differences between means by looking at the sums of squares. The group sums of squares, SSA , equals the difference between SST and SSE :

$$SSA = SST - SSE \quad (26)$$

We convert the sums of squares into variances by dividing by their degrees of freedom. In our example, there are two levels of the factor and so there is $(2 - 1) = 1$ degree of freedom for SSA . In general, we might have a levels of any factor and hence $a - 1$ degrees of freedom for the factor effects. If each factor level were replicated n times, then there would be $n - 1$ degrees of freedom for error within each level of the factor (we lose one degree of freedom for each individual group mean estimated from the data). Since there are a levels, there would be $a \times (n - 1)$ degrees of freedom for error in the whole experiment. The total number of subjects in the whole experiment is $a \times n$, so the total degrees of freedom equals $(an - 1)$. The single degree is lost for estimation of the overall mean). As a check in more complicated designs, it is useful to make sure that the individual component degrees of freedom add up to the correct total.

$$kn - 1 = k - 1 + k(n - 1) \quad (27)$$

$$kn - 1 = k - 1 + kn - k \quad (28)$$

The calculations for turning the sums of squares into variances are carried out in an ANOVA table, as shown in Table 7.

Each row in the sums of squares column is divided by appropriate degrees of freedom column to give the variances in the mean square column. The significance of the difference between the means is then assessed using an F test (a variance ratio test). The group variance MSA is divided by the error variance, s^2 , and the value of this test statistic is compared with the critical

Table 7: Calculations for a one way analysis of variance.

Source of variation	SS	df	MS	F statistic	P
Factor A	SSA	$a - 1$	$MSA = SSA/(a - 1)$	MSA/s^2	
Error	SSE	$a(n - 1)$	$s^2 = [SSE/a(n - 1)]$		
Total	SST	$an - 1$			

value of F using the quantiles of the F distribution, with $P = 0.95$, $a - 1$ degrees of freedom in the numerator, and $a(n - 1)$ degrees of freedom in the denominator. If the test statistic is larger than the critical value we reject the null hypothesis that all means are the same and accept the alternative hypothesis, that at least one of the means is significantly different from the others. If the test statistic is less than the critical value, then it could have arisen due to chance alone, and so we accept the null hypothesis.

Another way of thinking about ANOVA is to consider the relative amounts of sampling variation between replicates in the same group (i.e. between individual samples in the same group), and between different groups (i.e. between-group). When the variation between replicates within a group is large compared with the variation between groups we are likely to conclude that the difference between the treatment means is not significant. Only if the variation between replicates within groups is relatively small compared to the differences between groups will we be justified in concluding that the group means are significantly different.

The definitions of the various sums of squares can now be formalised, and ways found of calculating their values from samples. The total sum of squares, SST, is defined as:

$$SST = \sum y^2 - \frac{(\sum y)^2}{kn} \tag{29}$$

Note that we divide by the total number of numbers we added together to get $\sum y$ (the grand total of all the y 's) which is $(a \times n)$. It turns out that the formula that we used to define SSE is rather difficult to calculate, so we calculate the group sums of squares SSA, and obtain SSE by difference. The group sum of squares, SSA, is defined as:

$$SSA = \frac{\sum C^2}{n} - \frac{(\sum y)^2}{kn} \tag{30}$$

where the new term is C, the group total. This is the sum of all the n replicates within a given level. Each of the a different treatment totals is squared, added up, and then divided by n (the number of numbers added together to get the treatment total). The formula is slightly different if there is unequal replication in different groups. Notice the symmetry of the equation. The second term on the right hand side is also divided by the number of numbers that were added together ($a \times n$) to get the total ($\sum y$) which is squared in the numerator. Finally:

$$SSE = SST - SSA \tag{31}$$

to give all the elements required for completion of the ANOVA table.

We now work through an example. The data come from a plant growth experiment in which the response variable is growth (mm) and the categorical explanatory variable is a factor called

photoperiod with four levels: very short, short, long and very long daily exposure to light. There were six replicates of each treatment. A one way ANOVA was carried out to test for differences in plant growth among the different levels of photoperiod.

Table 8: Data from a plant growth experiment.

Replicate	Photoperiod				Total
	Very short	Short	Long	Very long	
1	2	3	3	4	12
2	3	4	5	6	18
3	1	2	1	2	6
4	1	1	2	2	6
5	2	2	2	2	8
6	1	1	2	3	7
Total	10	13	15	19	57

Table 9: Plant growth experiment. One way ANOVA to test for differences in plant growth among different levels of photoperiod.

Source of variation	SS	df	MS	F statistic	P
Group (photoperiod)	7.12	3	2.37	1.46	0.26
Error	32.5	20	1.62		
Total	39.6	23			

We conclude from Table 9 that there is no significant difference in growth among the four levels of photoperiod. An F statistic of 1.46 will arise due to chance when the means are all the same with a probability greater than 1 in 4 (for significance, we want this probability to be less than 0.05). The differences we noted at the beginning are not significant because the variance is so large ($s^2 = 1.625$) and the error degrees of freedom ($df = 20$) rather small.

12.2 Two way ANOVA

ANOVA can be extended to include two factors (instead of one). The key feature of two way ANOVA is that the interaction between the two factors can be assessed. Formulae for a two way ANOVA table are shown in Table 10.

Continuing the plant growth analysis example presented earlier, we fitted a 4-level factor for photoperiod (the group effect), but the unexplained variation (SSE) was very large (32.5). So large in fact, that the differences between mean growth in the different photoperiods was not significant. Fortunately, the experiment was well designed. Material from six plant genotypes (G1, .. G6) was cloned and photoperiod treatments were allocated at random to each of four clones of the same genotype. The data are shown in Table 11.

The interpretation of the experiment is completely different. We now conclude that photoperiod has a highly significant effect on mean growth. The moral is that good experimental design,

Table 10: Calculations for a two way analysis of variance.

Source of variation	SS	df	MS	F statistic	P
Factor A	SSA	$a - 1$	$MSA = SSA/(a - 1)$	MSA/s^2	
Factor B	SSB	$b - 1$	$MSB = SSB/(b - 1)$	MSB/s^2	
Error	SSE	$ab(n - 1)$	$s^2 = [SSE/ab(n - 1)]$		
Total	SST	$abn - 1$			

Table 11: Data from a plant growth experiment.

Genotype	Photoperiod				Total
	Very short	Short	Long	Very long	
G1	2	3	3	4	12
G2	3	4	5	6	18
G3	1	2	1	2	6
G4	1	1	2	2	6
G5	2	2	2	2	8
G6	1	1	2	3	7
Total	10	13	15	19	57

aimed at reducing variation between individuals (blocking by genotype in this case) can pay dividends in terms of the likelihood of detecting biologically important differences.

Table 12: Plant growth experiment. Two way ANOVA to test for differences in plant growth among different levels of photoperiod and genotype.

Source of variation	SS	df	MS	F statistic	P
Photoperiod (group)	7.12	3	2.37	7.70	< 0.01
Genotype (group)	27.9	5	5.57	18.1	< 0.01
Error	4.62	15	0.31		
Total	39.6	23			

13 Linear regression

Univariate statistical techniques, which describe or draw inferences about the characteristics of a single variable or measurement, are of limited use for clinical decision making. The simple t-test and the confidence interval for a difference in means compare the measurements for an outcome of interest in two different groups. Often, however, researchers are interested in comparing results between several groups, or determining the associations between an outcome and other continuous variables. The correlation coefficient measures the linear relationship between two variables, ranging from 1 (a perfect positive linear relationship) through zero (no linear relationship) to -1 (a perfect negative linear relationship). However, while it provides the strength of the association, it cannot be used for prediction and is restricted to linear relationships between two variables. Regression models generalise both the correlation coefficient and the t-test. Simple linear regression models can be used for inferring the relationship between two variables, and prediction of the outcome variable based on the explanatory is possible. Simple regression can be further generalised to multiple regression, where we simultaneously consider more than one possible predictor variable, and logistic regression, where we allow response variables to be dichotomous (e.g. disease present, disease absent) rather than continuous.

A common thread runs through all of the procedures discussed here. The primary interest is to explain the variability in the response by taking into account the structure of its relationship with the possible predictors. If division into groups or the use of an independent continuous regression variable can explain a substantial part of the total variability in the response variable, we can develop a much better understanding of the response variable and more precise predictions of the responses for future individuals. We will first describe simple linear regression and how it is used, along with some of the diagnostic checks to evaluate the fit of the model to the data. Multiple regression, the generalisation of simple linear regression when more than one independent variable is used to predict the outcome variable, as well as model selection techniques will follow.

13.1 Simple linear regression

A study was conducted to investigate the influence of a set of predictors on calving to conception intervals in a group of dairy cows. In this study the researchers were interested in determining whether the following variables influenced calving to conception interval: (1) calving to first oestrus interval, (2) calving to first service interval, (3) age, and (4) the presence of health events such as endometritis, mastitis, cystic ovarian disease, pyometra, milk fever, and retained foetal membranes (RFM). The data set is comprised of records for 162 cows.

Given this data, we might want to compare the mean calving to conception interval for RFM positive and RFM negative cows. The mean calving to conception interval in the RFM positive group is 156 days with a standard deviation of 39 days. The mean calving to conception interval in the RFM negative group is 112 days with a standard deviation of 47 days. Compared with RFM negative cows, RFM positive cows took, on average, 44 days (95% CI 26 – 62 days) longer to conceive. The confidence interval excludes zero, so there is reasonable evidence to indicate that mean calving to conception interval in the two groups is different. Comparing two groups is easy enough, but what if we want to now look at the average calving to conception interval using information about another variable, such as age? We can't simply calculate mean values,

because there are no natural groups into which we can divide age. We could divide age into young cows and old cows, but this would waste information about the differences that exist between subjects within the young and old age groups. An alternative approach would be to calculate the correlation coefficient between age and calving to conception interval. This gives a correlation of 0.32, with a 95% confidence interval of 0.17 – 0.46 indicating that there is a statistically significant (but moderate in value) association between the two variables (the confidence interval does not include zero). However, we may want to know more about the relationship, for example by how much the response (calving to conception interval) varies as the predictor (age) changes. The common approach to this problem is known as simple linear regression. Simple linear regression is a method for deriving and describing the straight line that best describes the relationship (if any) between two continuous variables. Consider Figure 20 which is a scatterplot with values of the response along the vertical axis and values of the predictor along the horizontal axis. It is clear that there is great variability in calving to conception interval in our study subjects, but that there may be a trend towards an increased conception interval with increases in age. We would like to fit the best straight line to this set of points, which would represent the trend, if any.

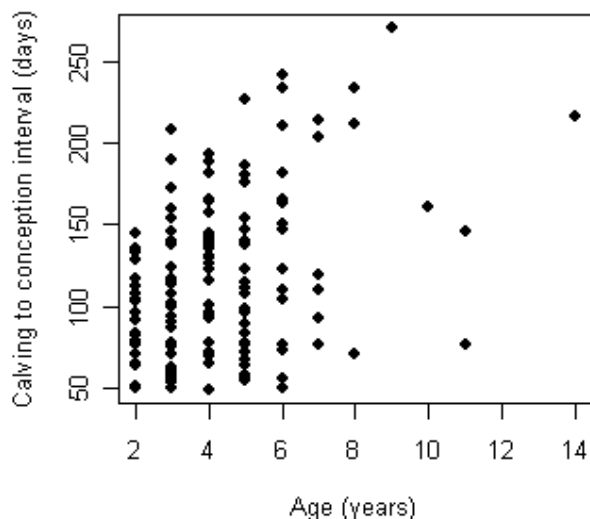


Figure 20: Scatter plot showing calving to conception interval as a function of age (in years).

The model

The simple linear regression model for this problem is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{32}$$

where y_i and x_i equals the calving to conception interval and the age of the i th cow, respectively. The parameters β_0 and β_1 represent the coefficients that relate y to x . In particular, β_1 is the slope of the regression line, so that a one unit change in x leads to a change of β_1 in y , and β_0 is the intercept, which is the predicted value of y when $x = 0$. The basic assumption made

is that the model is linear (that is, the relationship between calving to conception interval and age is a straight line). Of course, it is extremely unlikely that this model will be exactly correct for all data points, in that we do not expect the values of x and y for all cows to lie exactly on a straight line, and we must account for this. We do so by adding a small amount of error, ϵ_i , to the model. The ϵ_i are often referred to as the residuals, or random errors of the model. While we usually do not expect these errors to follow any predetermined pattern, we do make assumptions about their behaviour. The standard set of assumptions is that they:

1. Independent from each other;
2. Follow a Normal distribution with zero mean; and
3. Have a constant standard deviation σ throughout the range of the data.

Estimation of the coefficients β_0 and β_1 is done using the method of least squares. Briefly, this method chooses the estimated coefficients so as to minimise the sum of the squared differences between the estimated values and the actual values. In other words, we minimise the sum of the ϵ_i s. In our example, the value of β_1 represents the increase in the number of days between calving to conception interval corresponding to an increase of one year of age. The value of β_0 is the mean calving to conception interval for an individual of age zero. Applying the least squares technique to our data we find that the ‘best’ straight line regression model is $y = 84.9 + 7.82x$. This means that calving to conception interval increases (on average) by around 8 days for each additional year of life. Note that β_0 may not be clinically meaningful. It requires that the value $x = 0$ is meaningful, which is not the case for this example. This illustrates an important general principle: regression equations are only safely used over the range of the data they were calculated from. For example, a regression equation showing the decline of bone mineral density with age calculated on female subjects over age 50 would not necessarily be applicable for predicting bone density in young adult females or in males. It is also useful to look at σ_ϵ , the standard deviation of the residuals. For this model $\sigma_\epsilon = 46$ days indicating the size of the residual error standard deviation (measuring the unexplained variability in the data).

Checking the model

When using regression models, it is important to be aware of and understand the underlying assumptions of the model. First, the model is linear: a unit increase in age corresponds to an increase of 8 days in calving to conception interval. It is conceivable that this assumption is not correct, that the rate of increase in calving to conception interval is higher for older cows compared to younger cows, for example. The scatterplot shown in Figure 20 provides a useful start for investigating this assumption. If a curved or other (non-linear) pattern is observed, steps can be taken to address it, usually by transforming the x or y variable. For example, rather than using x , we might try x^2 , or rather than y we might try $\log(y)$. Another essential part of the model checking process is to look at the residuals. Recall that the model virtually never fits perfectly and we assume that the true model fits with a small amount of error added. We cannot observe the actual values of these errors, but we can get an idea of what they look like by plotting the residuals as a function of the fitted values and checking that residuals follow a normal distribution, as shown in Figure 21.

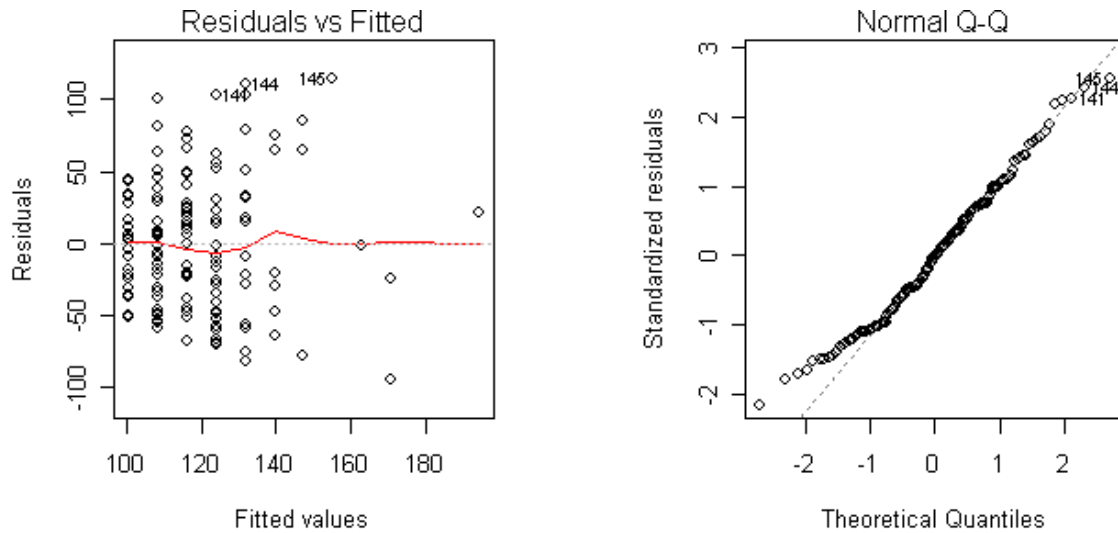


Figure 21: Diagnostic plots from a simple linear regression model where age (in years) is a predictor of calving to conception interval in a population of dairy cows.

If this is the case, what can be done? Several possibilities exist, but the most common approach is to transform either the response or the predictors to achieve a model with better residual properties. A thorough discussion of the options for transformations can be found in Neter et al.³

13.2 Multiple linear regression

We now consider the situation where we have multiple predictor variables, each of which has an independent influence on the outcome. Perhaps we would like to examine the influence of age and presence of retained foetal membranes on calving to conception interval. One approach would be to try several simple linear regression models, one for each of the possible predictor variables. This will give information about their individual effects on calving to conception interval but not on the joint effect of all variables together, or on the effect of one of the variables adjusted for the presence of another. We can use multiple regression, which simultaneously associates many predictors to a continuous response, to accomplish both these tasks.

The model

A multiple regression model with two predictor variables is written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad (33)$$

The model works in essentially the same way as the simple linear regression model but interpretation of the coefficients changes slightly because of the presence of other variables in the model. Now, β_1 refers to the change in y associated with a unit change in x_1 after adjusting for the effect of x_2 . In other words, for a fixed level of x_2 , the mean change in y when x_1 increases by 1 is β_1 . The analogous interpretation holds for β_2 when x_1 is held constant.

In our dairy cow data set, suppose we want to look at the effects of the presence of a RFM diagnosis and age on calving to conception interval. Let x_1 denote age and x_2 presence or absence of an RFM diagnosis. The fitted model is $y = 86.26 + 6.36x_1 + 35.27x_2$ which means: (1) unit increases in age, after adjusting for the presence of a RFM diagnosis, increases calving to conception intervals increases by 6 days, and (2) the presence of RFM at calving, after adjusting for the effect of age, increases calving to conception intervals by 35 days. For this model $\sigma_\epsilon = 44$ days (an improvement on the 46 days from the simple linear model).

Model selection

To model calving to conception intervals using this data set we have several options. There are nine variables in the data set: (1) calving to first oestrus interval, (2) calving to first service interval, (3) age, and (4) the presence of health events such as endometritis, mastitis, cystic ovarian disease, pyometra, milk fever, and retained foetal membranes (RFM) which gives 2^9 different combinations of two variables. How should we decide how many and which independent variables we should include in the model? Having too few independent variables will result in a model with sub-optimal predictions, while having too many will add unwanted variability to the estimation process. When faced with multiple possible predictors, how do we decide: (1) which to consider as part of a model and (2) which model provides the best fit, given the data and our knowledge of the problem? This is a complicated process, and there are many factors that influence the way in which it is done. Many researchers use model selection techniques, such as backwards or forwards stepwise regression which systematically builds a model based on a variety of criteria. However, it is often possible that models that have good statistical properties are lacking in real clinical properties, in that variables that are in the model may be of less interest to research than variables that are out of the model. Further, stepwise model selection with a large number of predictors can often give results that are statistically significant purely by chance. Stepwise procedures decide between two models using one of several possible criteria. Some of the most common are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). These (and others) evaluate models based on criteria such as fit (better fitting models are better) and complexity (in general, if two models fit the data equivalently well, the less complex model is better). So while stepwise regression can be useful, it has to be used with caution to be sure that the results make sense from the clinical and research perspective. It is better to have a good understanding of the variables in the dataset. Ideally, descriptive statistics and clinical knowledge will indicate an order in which to consider variables in the model. For example, in non randomised studies it is customary to use age and sex of patients in the model even when they are weak predictors of the response, just because it is appropriate to discuss the models by comparing the results for patients of the same sex and age.

Checking the model

Diagnostic tests to check model assumptions for multiple regression models operate in a way completely analogous to simple linear regression. Checking linearity using scatterplots of the response against the various predictors and plots of the residuals are the two main methods for determining whether the model is appropriate. For certain problems, there are simple solutions to the violation of assumptions. If the pattern of the residuals appears curved, a polynomial

model (using higher powers of the predictors) may be appropriate. If the variance of the residuals appears unequal across the range, a weighted regression approach may help.

14 Experimental design

14.1 Completely randomised design

A completely randomised design is concerned with the comparison of t population (treatment) means m_1, m_2, \dots, m_t . We assume that there are t different populations from which we are to draw independent random samples of sizes n_1, n_2, \dots, n_t respectively, resulting in (n_1, n_2, \dots, n_t) experimental units (e.g. people, animals) on which a measurement is made. The treatments are randomly allocated to the experimental units such that n_1 receive treatment 1, n_2 receive treatment 2 and so on. The objective of the experiment is to make inferences about the corresponding treatment means. The model for a completely randomised design can be written in the form:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (34)$$

Where:

- y_{ij} the response on treatment i in subject j
- μ an overall mean
- α_i the effect due to treatment i
- ϵ_{ij} random error

A study was conducted where 12 lactating dairy cows were randomly assigned to one of three treatment groups. Change in milk yield over a defined follow up period are shown in Table 13. The results of a one way ANOVA are shown in Table 14.

Table 13: Data from a completely randomised design.

Treat 1	Treat 2	Treat 3
4.3	4.3	5.0
3.7	4.0	4.3
3.0	2.7	4.3
2.3	2.3	3.6

Table 14: ANOVA from a completely randomised design.

Source of variation	SS	df	MS	F statistic	P
Group (treat)	2.53	2	1.27	1.8	0.21
Error	6.07	9	0.67		
Total	8.60	11			

We conclude from Table 14 that there is no significant difference in milk yield change among the three levels of treatment ($F_{2,9} = 1.8$; $p = 0.21$).

Advantages of completely randomised designs:

- Easy to construct the design.
- Easy to analyse, even though sample sizes may be different.
- Design can be used for any number of treatments.

Disadvantages of completely randomised designs:

- Best suited for relatively few treatments.
- The experimental units to which the treatments are applied must be as homogenous as possible. Extraneous sources of variation tend to inflate the error term.

14.2 Randomised block design

A randomised block design is where we group the experimental units into blocks that are as homogenous as possible, and then randomly allocate each treatment to each experimental unit in each of the blocks. Randomised blocking increases statistical power by isolating the blocking variable and separating or partitioning out the variance due to this factor. The variance due to this factor does not contribute to the residual term as a result. The model for a randomised block design can be written in the form:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad (35)$$

Where:

- y_{ijk} the response on treatment i in block j in subject k
- μ an overall mean
- α_i the effect due to treatment i
- β_j the effect due to block j
- ϵ_{ijk} random error

A study was conducted where 12 lactating dairy cows were randomly assigned to one of three treatment groups. Change in milk yield over a defined follow up period was the outcome. If we were able to match the subjects on initial bodyweight (into groups of 3) and assign one member of each block to each treatment then we would have a randomised block design. Initial bodyweight then becomes the blocking factor and three subjects of similar weight are placed into each block and then one subject within each block is randomly assigned to each of the three treatments:

Table 15: Data from a randomised block design.

Weight	Treat 1	Treat 2	Treat 3
1	4.3	4.3	5.0
2	3.7	4.0	4.3
3	3.0	2.7	4.3
4	2.3	2.3	3.6

Table 16: ANOVA from a randomised block design.

Source of variation	SS	df	MS	F statistic	P
Group (treat)	2.53	2	1.27	13.95	< 0.01
Group (weight)	5.53	3	1.84	20.29	< 0.01
Error	0.54	6	0.09		
Total	8.60	11			

By adding the blocking factor into a 2 way ANOVA, we are able to partition the variance and remove the impact of subject initial weight from the estimated error variance. This has resulted in treatment being detected as a significant effect. For animal experiments, blocks might be litters, weight of animals prior to experiment or on previous history of the animals. In humans, classifications used for blocking include: age, sex, height, weight, social class, medical history, race, and so on.

The usual approach to randomised blocking is to have the number of subjects in each block equal to the number of treatments. One should not force subjects into inappropriate blocks simply to get blocks of the same size as the number of treatments. An incomplete block design results when the number of subjects in each block is unequal.

It is of interest to determine whether blocking has been beneficial in terms of increasing our precision for comparing treatment means in a given experiment. Some texts say that the blocking term in the ANOVA model should be statistically significant since that clearly demonstrates that blocking was justified. In a similar vein, if the blocking term is non-significant, it should perhaps be removed from the model since it has not increased the precision and may be taking degrees of freedom away from the error term. We can estimate the benefit of blocking with a term called the ‘relative efficiency’ of the randomised block design. This is an attempt to produce a ratio of the estimate of variance of the i th treatment mean if the same data were examined using a completely randomised design versus a randomised block design. Normally, we don’t do the two analyses, but we can estimate relative efficiency from the randomised block design using the following formula:

$$\text{Relative efficiency} = \frac{(b - 1)MS_{block} + (b - 1)MS_{error}}{(bt - 1)MS_{error}} \quad (36)$$

Where:

The relative efficiency tells you the number of observations (or replications) of each treatment required to obtain the same precision in a completely randomised design. If, for example,

b	the number of blocks
t	the number of treatments within blocks

the relative efficiency was 9.68 then we would conclude that approximately 10 times as many observations of each treatment would be required in a completely randomised design to obtain the same precision for treatment comparisons as was obtained with the randomised block design that was used.

The relative efficiency for the data shown in Table 15 using this blocking design was 6.3. This means that 6.3 times as many subjects would be required to obtain the same result if we used a completely randomised design as opposed to the randomised block design.

Advantages of randomised block designs:

- Useful for comparing t treatments in the presence of a single extraneous source of variability.
- Easy to analyse, even though sample sizes may be different.
- Missing values easily handled.
- Design easy to construct.
- Design can be used for any number of treatments in any number of blocks.

Disadvantages of randomised block designs:

- Best suited for relatively few treatments.
- Only controls for one source of variability (due to blocks). Extraneous sources of variation tend to inflate the error term.
- The effect of each treatment on the response must be approximately the same from block to block.

14.3 Latin square design

A Latin square design can be thought of as a randomised block design with more than one block. The model for a Latin square design can be written in the form:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl} \quad (37)$$

Where:

y_{ijkl}	the response on treatment i in block j in block j in subject l
μ	an overall mean
α_i	the effect due to treatment i
β_i	the effect due to block j
γ_k	the effect due to block k
ϵ_{ijkl}	random error

Latin square designs employ the blocking principle for two confounding (nuisance) factors. The levels of the confounding factors are assigned to the rows and the columns of a square; then the cells of the square identify treatment levels. A study was conducted where 12 lactating dairy cows were randomly assigned to one of three treatment groups. In an experiment change in milk yield over a defined follow up period was the outcome. Initial bodyweight and pen were thought to be important confounders. The data for this Latin square design are shown in Table 17.

Table 17: Data from a Latin square design.

Weight	Treat 1	Treat 2	Treat 3
1	Pen A	Pen B	Pen C
2	Pen B	Pen C	Pen A
3	Pen C	Pen A	Pen B
4	Pen A	Pen B	Pen C

Advantages of Latin square designs:

- Useful for comparing t treatments in the presence of two extraneous sources of variability.
- Easy to analyse, even though sample sizes may be different.
- Missing values easily handled.
- Design easy to construct.
- Design can be used for any number of treatments in any number of blocks.

14.4 Factorial design

Rather than comparing the effects (or means) of the levels of one experimental factor, we might wish to examine the effects of two or more experimental factors. For example, we may wish to examine the effect of fertilisers on the yield of various varieties of a crop. The model for an $a \times b$ factorial design can be written in the form:

Table 18: Data from a Latin square design in long format.

Weight	Treat	Pen	Yield
1	1	A	4.3
1	2	B	4.3
1	3	C	5.0
2	1	B	3.7
2	2	C	4.0
2	3	A	4.3
3	1	C	3.0
3	2	A	2.7
3	3	B	4.3
4	1	A	3.0
4	2	B	2.7
4	3	C	4.3

Table 19: ANOVA from a Latin square design.

Source of variation	SS	df	MS	F statistic	P
Group (weight)	2.17	2	1.08	7.87	0.11
Group (treat)	1.50	2	0.75	5.45	0.15
Group (pen)	0.17	2	0.09	0.63	0.61
Error	4.38	2	0.14		
Total	8.60	9	2.06		

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \epsilon_{ijk} \quad (38)$$

Where:

A situation where there are important implications of interaction effects is in the use of drugs. Drugs ingested in combination may have effects that are unpredictable from the effects of the individual drugs acting alone. The example we will use is the suppression of basal metabolism by alcohol and barbiturates. The dependent variable is some measure of basal metabolism. The two independent variables are alcohol consumption (present or absent) and barbiturate use (present or absent). Suppose that there is no interaction between alcohol and barbiturate use. Figure 22a shows how this might work. The effect of alcohol is exactly the same whether barbiturates are present or not. If the two drugs interacted we get the situation shown in Figure 22b. The effect of alcohol without barbiturates (the solid line) results in a moderate decrease of metabolism. The effect of alcohol in combination with barbiturate use (the dashed line) produces a more dramatic decrease in metabolism. An interaction between two factors exists when:

1. The effects of factor A vary at the different levels of factor B;
2. The values of one or more contrasts in factor A vary at different levels of factor B;

y_{ijk}	the response on the observation taken at the i th level of factor A and the j th level of factor B in subject k
μ	an overall mean
α_i	the effect due to treatment i
β_j	an effect due to the j th level of factor B
$\alpha_i\beta_j$	an effect due to interaction between the i th level of factor A and the j th level of factor B
ϵ_{ijk}	random error

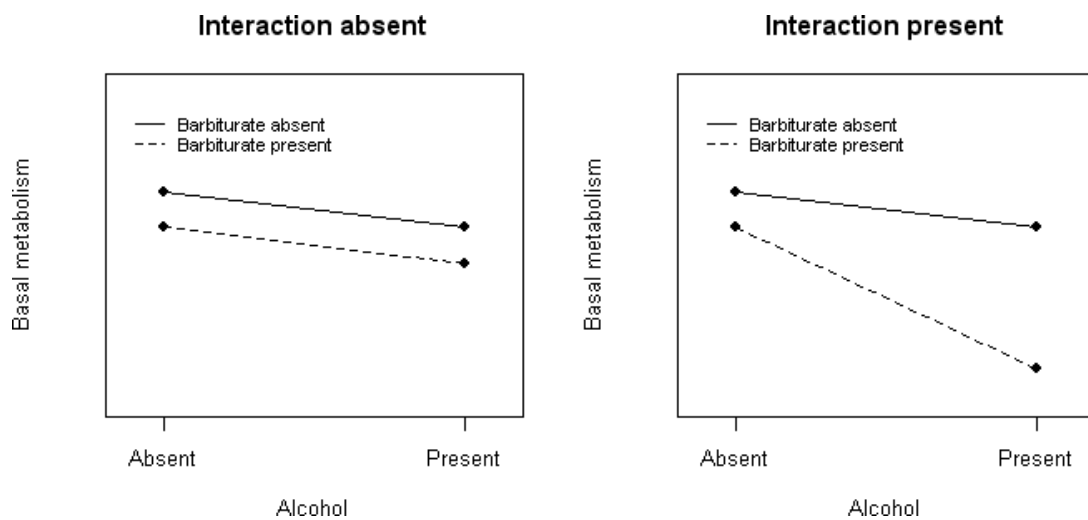


Figure 22: Interaction diagrams showing the possible effect of alcohol consumption and barbiturate use on basal metabolism: (a) no interaction, and (b) interaction.

3. The simple effects of factor A are not the same at all levels of factor B; and
4. The differences among the cell means representing the effect of factor A at one level of factor B don't equal to the corresponding difference at another level of factor B.

A complete factorial describes the situation where responses are measured at every possible combination of the factors involved. If there are a levels of factor A and b levels of factor B, then the experiment would be called an $a \times b$ factorial design. In this case the total number of treatments is $t = a \times b$. Incomplete or fractional factorial designs are often used in agriculture and industrial applications. As the number of interactions increases experiments become more complex to design and analyse. An example follows: an experiment using a 2×2 factorial design examined the effect of water temperature and water flow on growth rates in a particular species of fish, Table 20.

To better understand the nature of any interaction that might be present, plot the responses against the levels for a given factor. The levels the factor are marked on the horizontal axis and the population means for the treatment combinations are plotted on the vertical axis. Points corresponding to the same level of the other factor are joined by straight lines.

Table 20: Data from a factorial design.

Temperature	Still flow	Rapid flow
Warm	0,1,2,2,3,4	6,7,8,8,9,10
Cold	2,3,4,4,5,6	1,2,3,3,4,5

Table 21: ANOVA from a factorial design

Source of variation	SS	df	MS	F statistic	P
Group (temp)	13.5	1	13.5	6.75	0.02
Group (flow)	37.5	1	37.5	18.75	< 0.01
Temp \times Flow	73.5	1	73.5	36.75	< 0.01
Error	40	20	2.00		
Total	8.60	11			

If an interaction between two factors is significant, the F-statistic may be either smaller, similar in size, or larger than the corresponding F-statistics for the two main effects. If the interaction F-statistic is smaller than those of the main effect terms then it suggests the main effects are most important (i.e. there is a real main effect) but that this difference is modified by the interaction. One would therefore present and discuss both the main effect means and the interaction results. If the interaction F-statistic is similar in size or larger than the F-statistics for the main effect terms, then interpretation of the main effects means is meaningless and only the means of the interaction terms should be presented and discussed.

14.5 Split plot factorial design

The split plot factorial design is one of the most widely used designs in behavioural and agricultural sciences and is often combined with repeated measurements on subjects. It contains features of two block designs: completely randomised and randomised block design. Conceptually it involves larger plots and smaller sub-plots. For example, in agricultural a common design is to apply different methods of preparing the land (disc plough vs tiller vs chemical to kill plants) to a series of plots then divide each plot into a series of sub-plots and within each apply several different sowing rates and fertiliser applications. A typical animal design is shown in Table 22.

In Table 22 a total of 32 animals are used. Animals are first assigned to blocks and preferably this should be done to ensure that blocks are similar (e.g. weight, age etc). Each block is then randomly assigned to one of two levels of Treatment 1 (Drug A, Drug). Within each block, animals are randomly assigned to one of the four levels of Treatment 2 (Drug C, Drug D, Drug E, or Drug F).

The split plot approach generates two error terms: one to test the main effect of the main plot factor and the other to test the effect of the subplot factor and the interaction (time and time \times treatment). The error term for the time and time \times treatment interaction is usually smaller and so these effects have greater power.

Table 22: Data from a split plot factorial design.

Block	Treat 1	Treat 2			
		Drug C	Drug D	Drug E	Drug F
1	Drug A	3	4	7	7
2	Drug A	6	5	8	8
3	Drug A	3	4	7	9
4	Drug A	3	3	6	8
5	Drug B	2	1	5	10
6	Drug B	2	3	6	10
7	Drug B	2	4	5	9
8	Drug B	2	3	6	11

Table 23: ANOVA from a split plot factorial design.

Source of variation	SS	df	MS	F statistic	P
Group (Treat 1)	3.12	1	3.12	5.05	0.04
Group (Treat 2)	193.2	3	64.42	104.22	< 0.01
Treat 1 \times Treat 2	18.62	3	6.21	10.05	< 0.01
Treat 1 \times block	9.37	6	1.56	2.53	0.06
Error	11.12	18	0.62		

14.6 Repeated measures design

When several measurements are made on the same experimental unit, they tend to be correlated and this must be taken into account when the analysis is performed. In a repeated measures ANOVA design, the same response variable is measured at different times (e.g. individual cow somatic cell counts measured at monthly intervals throughout a lactation). The model for a repeated measures design can be written in the form:

$$y_{ijk} = \mu + \alpha_i + \gamma_{ik} + \tau_{ij} + \epsilon_{ijk} \quad (39)$$

y_{ijk}	the response on treatment i in subject j and measurement k
μ	an overall mean
α_i	the effect due to treatment i
γ_{ik}	the effect of treatment i on each of k measurements
τ_{ij}	the effect of treatment i in subject j
ϵ_{ijk}	random error

The design is powerful because it controls for individual variation. Subjects serve as their own controls, so the variability owing to individual differences is eliminated from the error (residual) term. The classic repeated measures split plot factorial design is shown in Table 24.

Table 24: Data from a repeated measures design.

Treatment	Animal	Time		
		1	2	3
Drug A	1			
Drug A	2			
Drug A	3			
Drug B	4			
Drug B	5			
Drug B	6			
Drug C	7			
Drug C	8			
Drug C	9			

15 Using a spreadsheet

Microsoft Excel is a spreadsheet software package that allows the user to store, manipulate, analyses and graph data. Each Excel file consists of a workbook that can hold a maximum of 256 worksheets. In each worksheet there can be up to 256 columns and 65,536 rows (a total of 16,777,216 cells). Within a worksheet columns are identified by letters (A,B,C, ... AA,AB, ...) rows are identified by numbers (1,2,3, ...). Individual cells are referenced by their column and row identifiers: cell A1 is the cell in the top left corner of the worksheet. Numbers, text and dates can be entered into these cells.

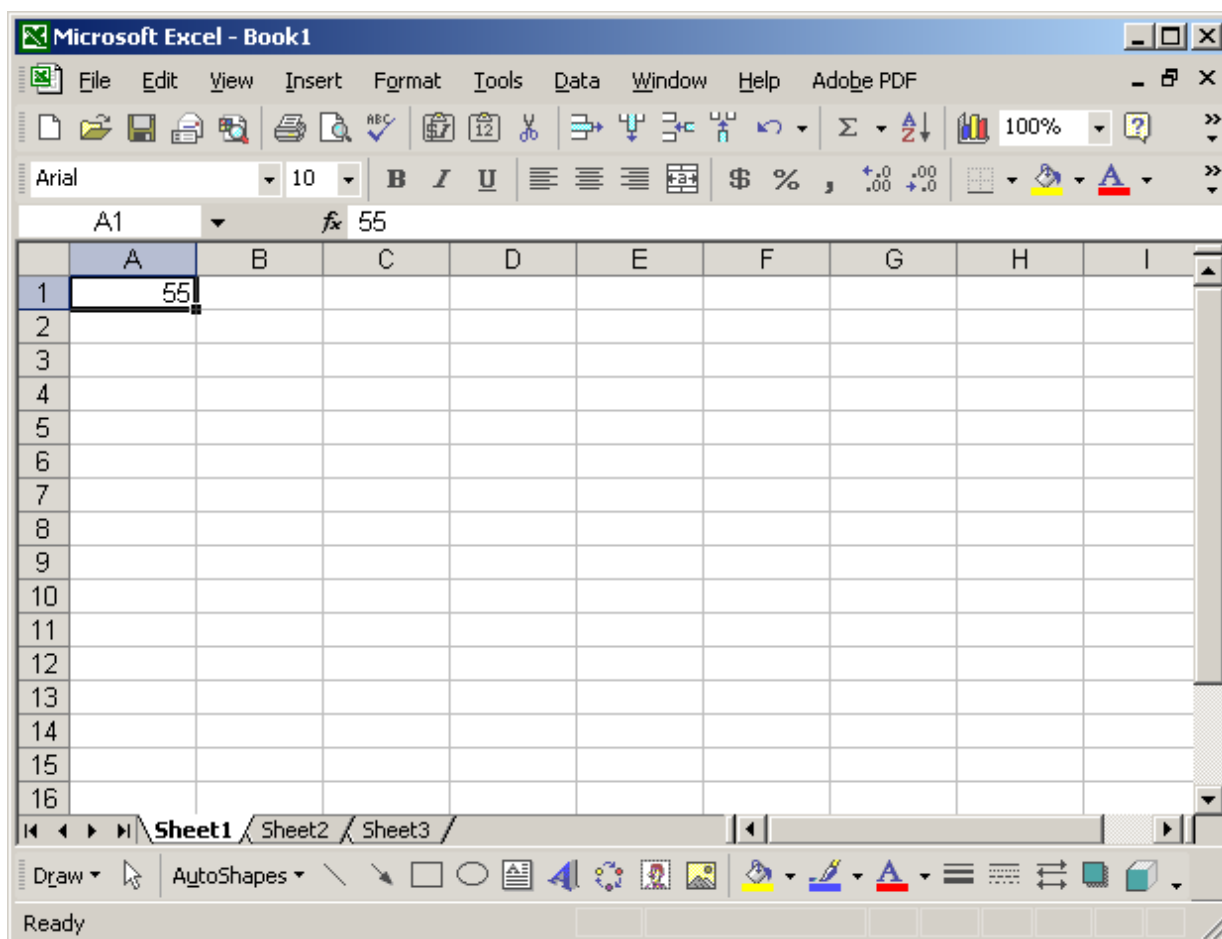


Figure 23: The Excel worksheet. The number '55' has been entered into cell A1.

15.1 Formatting cells

A cell can be formatted to display either text or numbers. To format a cell open the format dialogue box (-FORMAT-CELLS from the menu bar) and select one of the options on the tab titled 'Number.' Another way to access the format dialogue box is to right click on a selected cell and select 'Format Cells' from the drop down list. Using the format dialogue box it is also

possible to change the font of displayed numbers (or text), change the colour of displayed text, add borders around cells and so on.

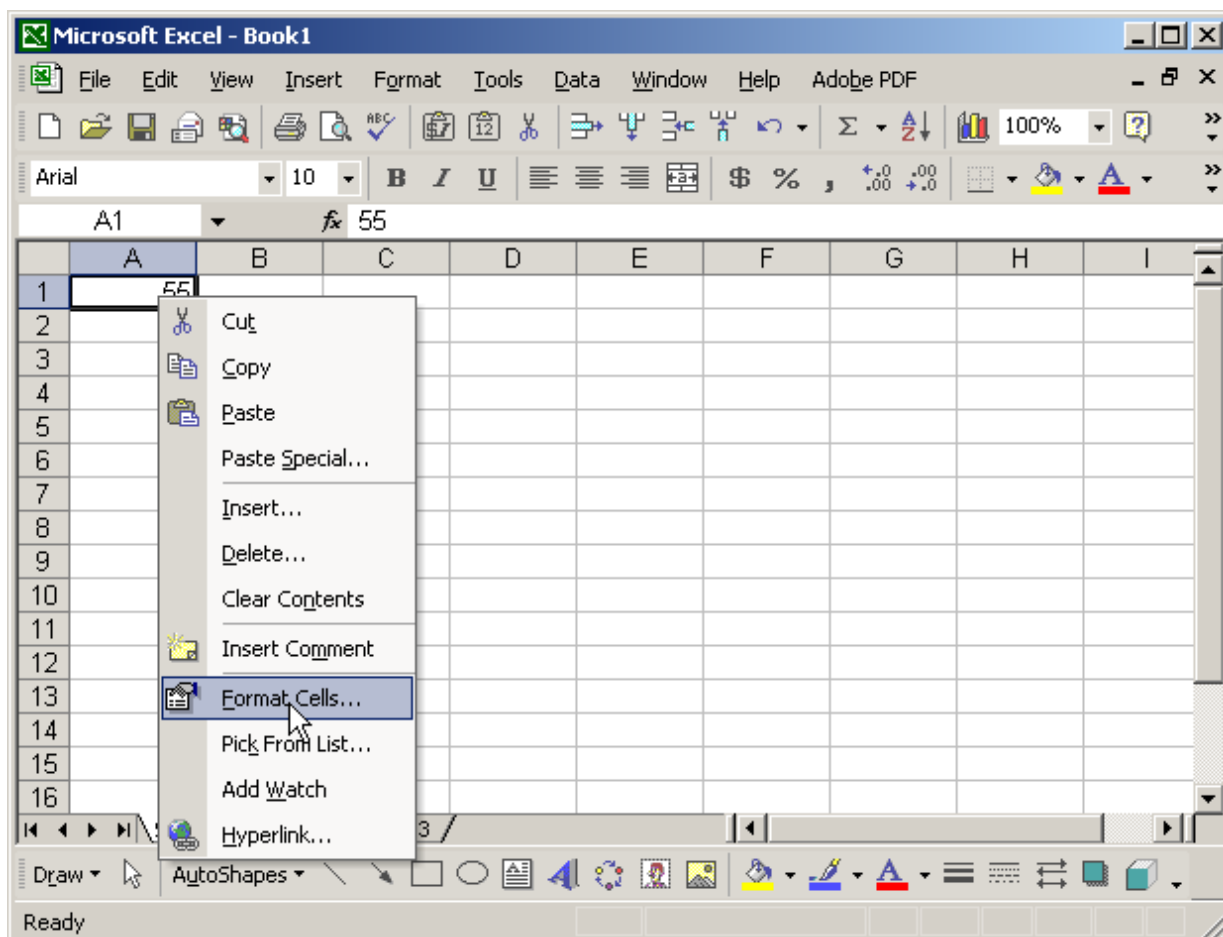


Figure 24: Formatting cells in MS Excel. A cell is selected and the right mouse clicked. Select the Format Cells ... option.

15.2 Sorting data

Once data has been entered into the cells of a worksheet it is possible to sort it (e.g. from lowest to highest, or alphabetical order) using the `-DATA-SORT` function. When data is sorted text and number formats are sorted separately. If you have numbers and a combination of numbers and text in your list (e.g. 1, 230, 23a, 34b, ...) and you want these values to be sorted together you will need to change the numbers to text before you perform the sort. Once you have done your sort, you might need to change your numbers back to numeric format. Use the paste special (`-EDIT-PASTE SPECIAL`) function to multiply all numbers by 1.

If you have values in multiple columns and highlight a single column then Excel will sort that single column independent of the others. Be careful! If each row of your worksheet has several items of information related to particular subject (patient, farm etc) sorting in this fashion will result in your data becoming scrambled. To avoid this problem, highlight all columns before selecting `-DATA-SORT`.

Table 25: Numeric formats in MS Excel.

Format	Description
General	Cells have no specific number format.
Number	Used to display numbers, allows user to specify the displayed number of decimal points.
Dates	Used to format dates and times.
Currency	Used for entering currency values (aligns decimal points and insert a dollar sign).
Percentage	Multiplies the selected number by 100 and displays it with a percent sign.
Scientific	Displays the number in scientific notation format.

15.3 Functions

Excel allows the user to manipulate and analyse data by entering formula into cells. The different types of formula can be grouped into a number of categories including:

- Mathematical
- Logical
- Array
- Statistical
- Look-up
- Rounding and truncating

Formulae in Excel start with an equal sign (=) followed by the formula syntax. Formulae can contain either numbers e.g. =2+3 or a reference to another cell in the spreadsheet e.g. =A1+A2. In some cases you will need to stipulate a range of numbers. You can enter either the cell references themselves (e.g. A1,A2,A3) or the first and last cell in the range are identified for example A1 to A10 is entered as A1:A10.

If you type a single quote (') in a cell before anything else Excel will interpret the contents of the cell as text and display the cell contents exactly as it has been entered (without the '). For example, if you type '=3+4, 7 will be displayed. However, if you type '=3+4 then =3+4 will be displayed. This feature is useful for making your calculations clear to others using your spreadsheet.

When referring to a cell on its own or as part of an array the reference can be *relative* or *absolute*. An absolute reference is differentiated from a relative reference by the inclusion of \$ in front of the column and/or row reference. For example, the relative reference of the top left cell in the worksheet is A1 and the absolute reference is \$A\$1. The difference between the two is apparent only when the formula is pasted to another location. An absolute reference always refers to the particular cell no matter where in the worksheet it is pasted. In contrast, when a relative reference is pasted to another cell the cell it refers to is changed based on where the original cell was in relation to the new cell. For example if the reference in cell B1 was to A1 and it was copied to D1 then the reference will now refer to cell C1.

To make a relative reference absolute highlight the cell reference and press the F4 key. \$A\$1 will be entered, rendering the cell absolute. Click the F4 key again will cycle through \$A1, A\$1, and A1 making the cell column absolute, row absolute, and relative.

Mathematical functions

Mathematical functions can be entered by directly typing them into a cell or by using the function wizard. To open the function wizard click -INSERT-FUNCTION.

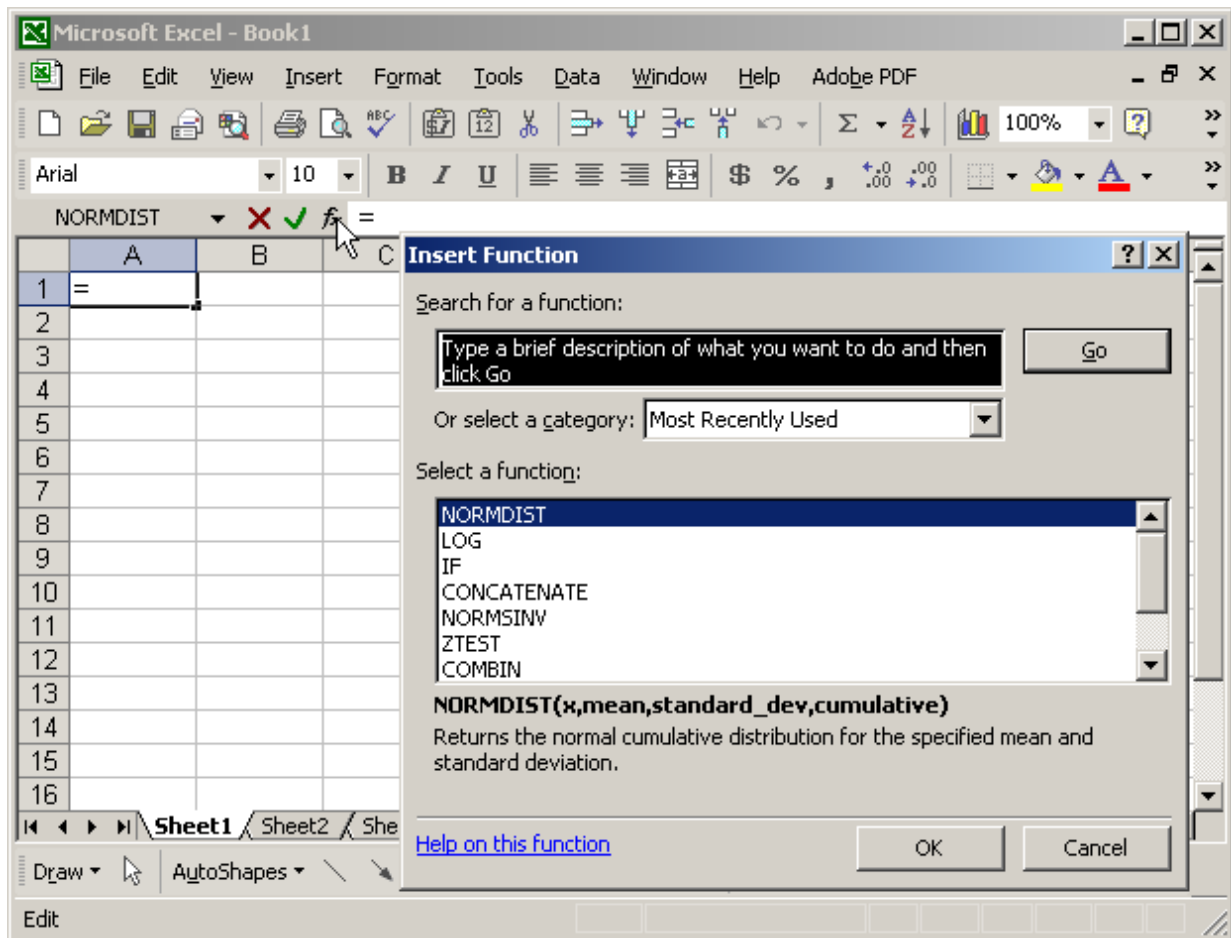


Figure 25: The Excel function wizard.

The keyboard operators for some of the more simple mathematical functions are:

- addition (+)
- subtraction (-)
- multiplication (*)
- division (/)

Table 26: Syntax for commonly used mathematical functions.

Function	Description	Example
SUM	Sums a list of numbers or cell references.	= SUM(A1, B1:B10)
MAX	Returns the maximum number in a list of numbers.	= MAX(A1, A20)
SQRT	Returns the square root.	= SQRT(9)
EXP	Returns e raised to given number.	= EXP(A1)
LN	Returns natural of log of a given number.	= LN(A1)
LOG10	Returns base-10 log of a number.	= LOG(10)
LOG	Returns base-n log of a number.	= LOG(2,5)
POWER	Returns the result of one number raised to another.	= POWER(A1,A2)

After entering the name of a function into a cell Excel can be made to prompt you for the required values by entering CTL-A or CTL-SHIFT-A.

If several operators (+, -, *, /) occur in the same formula the program will calculate the higher order precedence first. That is when entering a formula Excel will calculate parentheses, any to the power of another number, multiplication/division and then addition/subtraction. For example, when calculating =3+4*(3-1)² Excel will calculate 3 - 1 to give 2, it will then square 2 to get 4 and multiply 4 by 4 to give 16 and finally add 3 to give 19. If events are of the same precedence (e.g. multiplication and division or addition and subtraction) then Excel will work from left to right. If you have any doubt as to how Excel will deal with a formula, enter brackets to ensure the formula behaves as you intend it to.

Absolute and relative cell references can be used to complete running totals. For example, Figure 28 shows a spreadsheet of month and sales and the running total. The running total (column C) was calculated by entering =SUM(\$B\$2:B2) and copying the formula into the other cells in column C. When the formula is pasted the start of the array (cell \$B\$2) stays fixed because it is an absolute reference but the end of the array is free to move depending on what row the formula is pasted into.

When using functions the formula not the cell value is recorded. To copy the actual results to a new location select -EDIT-PASTE SPECIAL and then highlight value only.

Logical functions

Logical functions are used to compare two numerical values or strings and return a TRUE or FALSE value. Logic operators can also be used to return one value if certain criteria are met and another if not.

The IF function can be used to specify the value to be returned if the condition is TRUE and FALSE. The general form of the IF function is:

= IF(logical condition, value if true, value if false)

For example =IF(weight<990,NA(),1) returns an empty cell if weight is greater than 990 and a 1 if it is less than 900. Up to seven IF statements can be nested within a single formula:

	A	B	C	D	E
1	Month	Sales	Running total		
2	1	100	100	=SUM(\$B\$2:B2)	
3	2	300	400	=SUM(\$B\$2:B3)	
4	3	150	550	=SUM(\$B\$2:B4)	
5	4	200	750	=SUM(\$B\$2:B5)	
6	5	500	1250	=SUM(\$B\$2:B6)	
7	6	400	1650	=SUM(\$B\$2:B7)	
8					
9					
10					
11					
12					
13					
14					
15					

Figure 26: Use of sum and absolute and relative cell references to calculate a running total of sales by month.

Table 27: Syntax for commonly used logical functions.

Function	Description	Example
AND	TRUE when both conditions are met and FALSE otherwise.	= AND(A2>=1,A2<=10)
OR	TRUE if either condition met and FALSE otherwise.	= OR(A2<1,A2>10)
=	TRUE if values are the same and FALSE otherwise.	= (A1=A2)
NOT	TRUE if values are different and FALSE otherwise.	= NOT(A1=A2)

=IF(weight<300,5,IF(weight>800,15,10)) returns 5 if weight is less than 300, 10 if weight is between 300 and 800, and 15 if weight is greater than 800.

It is possible to count and sum values that meet certain criteria, for example sales for each agent, costs from a supplier. This is done using the COUNTIF and SUMIF functions. The arguments for these functions are:

= COUNTIF(range, criteria)
 = SUMIF(range, criteria, sum range)

Array functions

Array formulae operate on more than one cell at a time and return more than one value. Array formulae can be recognised by different enclosing brackets {} that are entered by Excel after the formula is entered. To ensure results of an array formula are written to an array (rather than a single cell) you must press SHIFT + CTRL + ENTER when entering the formula. If you just press ENTER then the value will be returned to a single cell.

Figure 28 contains data arranged in three columns (columns A, B, and C). We wish to create a frequency histogram of these data. First of all we create a list of ‘bin’ values (cells E2:E11).

	A	B	C	D	E	F	G	H	I
1	Caller	Jan Cost		Caller	Count	Cost			
2	Fred	1.50		Agatha	3	4.47			
3	Charlie	2.00		Carol					
4	Agatha	0.76		Charlie					
5	Carol	3.65		Fred					
6	Charlie	0.85		Susan					
7	Fred	3.21		Total					
8	Susan	1.92							
9	Fred	0.85		E2 = COUNTIF(\$A\$2:\$A\$16, D2)					
10	Susan	0.85		F2 = SUMIF(\$A\$2:\$A\$16, D2, \$B\$2:\$B\$16)					
11	Agatha	2.96							
12	Agatha	0.75							
13	Carol	4.56							
14	Charlie	0.65							
15	Fred	3.00							
16	Susan	2.45							

Figure 27: Using the COUNTIF and SUMIF functions in MS Excel.

We now want to determine the frequency of each bin value - to do this we use the FREQUENCY function. Select the cells F2:F11, type =FREQUENCY(A2:C21,E2:E11), click CTRL + SHIFT + ENTER.

Text functions

Text functions change numbers to a text or string. You can specify what format the text will be displayed in by using some of text functions.

To convert a number stored as text to a number put the number 1 in an empty cell and copy the cell. Then highlight the cells you wish to change and select -EDIT-PASTE SPECIAL from the menu bar. In the past special dialog box and tick multiple in the operator section.

Statistical functions

It is often necessary to summaries data using statistics such as the average, standard error, minimum and maximum. In Excel it is possible to calculate a number of values for a data set by typing in the statistical function you require.

Excel also has an Analysis ToolPak add-in which contains a number of statistical functions including:

- ANOVA
- Correlation
- Descriptive statistics
- t-tests

Table 28: Syntax for commonly used text functions.

Syntax	Description
DOLLAR()	Returns a number as text in currency format.
FIXED()	Returns as a text rounded to the specified number of digits.
T()	Return the value as text.
VALUE()	Returns the number value of a string.
CONCENTENTATE	Combines two or more strings together.
FIND()	Locates a substring (case sensitive).
LEFT()	Extracts characters from the left.
LEN()	Returns the number of characters in a string.
LOWER()	Converts a string to lower case.
MID()	Extracts a substring.
PROPER()	Capitalizes the first letter of each word in the string.
REPLACE()	Replaces a substring.
RIGHT()	Extracts characters from the right of a string.
SEARCH()	Locates a substring (not case sensitive).
TRIM()	Removes leading and trailing spaces.
UPPER()	Converts a string to upper case.

Table 29: Syntax for commonly used statistical functions.

Function	Description	Example
AVERAGE	Returns the mean of a series of numbers.	= AVERAGE(B1:B10)
CORREL	Returns the correlation between two values.	= CORRELMAX(A1:A20, B1:B2)
COUNTBLANK	Counts the number of cells that have no value.	= COUNTBLANK(A1:E20)
LARGE	Returns the kth largest value in the data set.	= LARGE(A1:E20, 3)
MEDIAN	Returns the median.	= MEDIAN(A1:A20)
MODE	Returns the mode.	= MODE(A1:A20)
STD	Returns the standard deviation of a range of numbers.	= STD(A1:A20)

	A	B	C	D	E	F	G	H	I
1	COL A	COL B	COL C		BIN	FREQ	%		
2	55	316	223		50	3	5%		
3	124	93	163		100	6	10%		
4	211	41	231		150	9	15%		
5	118	113	400		200	13	22%		
6	262	1	201		250	7	12%		
7	167	479	205		300	6	10%		
8	489	15	89		350	6	10%		
9	179	248	125		400	4	7%		
10	456	153	289		450	2	3%		
11	289	500	198		500	4	7%		
12	126	114	303						
13	151	279	347		F2 = {FREQUENCY(A2:C21,E2:E11)}				
14	250	175	93		G2 = {FREQUENCY(A2:C21,E2:E11)/COUNTA(A2:C21)}				
15	166	113	356						
16	152	384	157						

Figure 28: Using ARRAY functions in MS Excel.

- Moving averages
- Histograms

To access the ToolPak see -TOOLS-DATA ANALYSIS from the menu bar. If the feature is not available then click -TOOLS-ADD INS and select 'Analysis ToolPak.' If the option to select Analysis ToolPak is not visible you will need the Microsoft Office installation media to install the Analysis module into your version of Microsoft Office. The descriptive statistics option in the Analysis ToolPak is useful for calculating a range of summary statistics.

Probability functions

Syntax for the various functions related to probability distributions are shown in Table 30.

To return the z value for a given probability value:

= NORMSINV(0.975)
= 1.96

To return the probability of a given value of z:

= NORMSDIST(1.96)
= 0.975

Lookup functions

Lookup functions return a value either from a one-row or one-column range or from an array. There are several lookup functions available within Excel, as listed below. These notes will outline the use of the VLOOKUP function. The arguments for the VLOOKUP function are:

Table 30: Syntax for commonly used probability distributions.

Function	Description	Example
BETADIST	Cumulative beta pdf.	= BETADIST(x, alpha, beta, A, B)
BETAINV	Inverse of the cumulative beta pdf.	= BETAINV(x, alpha, beta, A, B)
BINOMDIST	Individual term binomial distribution probability.	= BINOMDIST(no_s, trials, prob_s, cumul)
CHIDIST	One-tailed probability of the chi-squared distribution.	= CHIDIST(x, dof)
CHIINV	Inverse of the one-tailed probability of the chi-squared distribution.	= CHIINV(x, dof)
CHITEST	Value from the chi-squared distribution for the statistic and <i>df</i> .	= CHITEST(act_range, exp_range)
EXPONDIST	Exponential distribution.	= EXPONDIST(x, lambda, cumul)
FDIST	F probability distribution.	= FDIST(x, dof1, dof2)
FINV	Inverse of the F probability distribution.	= FINV(prob, dof1, dof2)
GAMMADIST	Gamma distribution.	= GAMMADIST(x, alpha, beta, cumul)
GAMMAINV	Inverse of the gamma cumulative distribution.	= GAMMAINV(prob, alpha, beta)
LOGNORMDIST	Cumulative lognormal distribution.	= LOGNORMDIST(x, mean, sd)
NEGBINOMDIST	Negative binomial distribution.	= NEGBINOMDIST(no_f, no_s, prob_s)
NORMDIST	Normal distribution for the specified mean and SD.	= NORMDIST(x, mean, sd, cumul)
NORMINV	Inverse of the normal distribution for the specified mean and SD.	= NORMINV(prob, mean, sd)
NORMSDIST	Standard normal cumulative distribution function.	= NORMSDIST(z)
NORMSINV	Inverse of the standard normal cumulative distribution.	= NORMSINV(prob)
POISSON	Poisson distribution.	= POISSON(x, mean, cumul)
TDIST	Percentage points for the Student's <i>t</i> -distribution.	= TDIST(x, dof, tails)
TINV	<i>t</i> -value of the <i>t</i> distribution as a function of the probability <i>df</i> .	= TINV(prob, dof)

Table 31: Lookup functions in MS Excel.

Function name	Purpose
VLOOKUP	Returns value in a cell based on the value in another and looking to and array.
CHOOSE	Returns a specific value from a list of values (up to 29) supplied as arguments.
HLOOKUP	Same as VLOOKUP but in the horizontal direction.
INDEX	Returns a value or a reference to a value from within a table or range.
MATCH	Locates the position within a vector.

= VLOOKUP(lookup_value, table_array, col_index_num, range_lookup)

Figure 29 shows a spreadsheet where columns A, B, C and D list details of patients recorded at a hospital clinic. Patient identifiers have been entered into column F and in column G the VLOOKUP function has been used to: (1) read the ID value recorded in column F, (2) find the row with a matching ID variable in the array A1:D10, and (3) return the appropriate value for SEX.

	A	B	C	D	E	F	G	H	I
1	ID	SEX	AGE	STATUS		ID	SEX		
2	1	Male	45	ILL		1	Male		
3	2	Female	22	HEALTHY		6	Female		
4	3	Female	21	HEALTHY		8	Female		
5	4	Male	16	ILL		9	Male		
6	5	Male	18	HEALTHY		2	Female		
7	6	Female	29	HEALTHY		1	Male		
8	7	Male	62	ILL		1	Male		
9	8	Female	81	HEALTHY		4	Male		
10	9	Male	40	ILL		7	Male		
11						12	#N/A		
12									
13									
14									
15									
16									

Figure 29: Using the VLOOKUP function in MS Excel.

In the above example, note the use of 0 for the range_lookup argument. range_lookup is a logical value that specifies whether you want VLOOKUP to find an exact match or an approximate match. If TRUE or omitted, an approximate match is returned. In other words, if an exact match is not found, the next largest value that is less than lookup_value is returned. If FALSE, VLOOKUP will find an exact match. If one is not found, the error value #N/A is returned.

Rounding and truncating functions

Excel can change the way in which a number is displayed (leaving the entered number unchanged). It is possible to change the actual stored number by using rounding and truncating functions.

15.4 Graphs

To enter a graph into an Excel worksheet click -INSERT-CHART from the menu bar to open the chart wizard. Select the type of graph you require and follow the prompts to produce a chart. Once a chart is completed you can right click on a feature within the chart to alter its features.

Table 32: Functions that change the way numbers are stored in a MS Excel spreadsheet.

Function	Description	Example
ABS	Returns the absolute value.	= ABS(-12.55)
FLOOR	Rounds the number towards the nearest multiple.	= FLOOR(1.255, 0.5)
INT	Returns the nearest integer.	= INT(2.4)
ROUND	Rounds a number to the required digits.	= ROUND(1.378, 2)
ROUNDDOWN	Rounds the number down the required number of digits.	= ROUNDDOWN(1.378, 2)
ROUNDUP	Rounds the number up the required number of digits.	= ROUNDUP(1.378, 2)

Error bars

To add error bars right click on the data points, select Format Data Series from the drop down menu, then select the Y Error Bars tab. Error bars can be either a fixed value, a percentage, a number of standard deviations or custom defined. The custom feature is extremely useful if each point has a different standard error, you can enter the standard errors for each point in a column beside the value and then enter the cell range into custom + and/or -. It is also possible to enter different values for the + and - error bars.

Trend lines

Trend lines can also be added to data points by right clicking on the data series and selecting the Add Trendline option. This will open the Trendline dialog box shown in Figure 32. Select the type of trend line you require and use the options to display the formula on the graph.

Shewhart charts

Often we need to monitor variables and how they change over time. A typical example in manufacturing would involve monitoring the number of defective units per batch of goods processed. If an increase in the number of defective units is detected then steps can be taken to investigate (and rectify) the problem. Graphs are a useful way to visualise this process. Here, date (and/or time) is plotted on the horizontal axis and a value for the monitored parameter is plotted on the vertical axis.

A problem that often arises is: two ‘abnormal’ values have been recorded over a short period of time. Is this sufficient to justify an investigation, or are these values just part of ‘random fluctuation’? Shewhart charts (also known as control charts) can be useful in this situation by displaying the level of normal, random variation in a data and by revealing the observations that indicate real change. The steps to create a Shewhart chart are as follows:

- Create a time series chart.
- Add a centre line (usually the mean) for central reference.

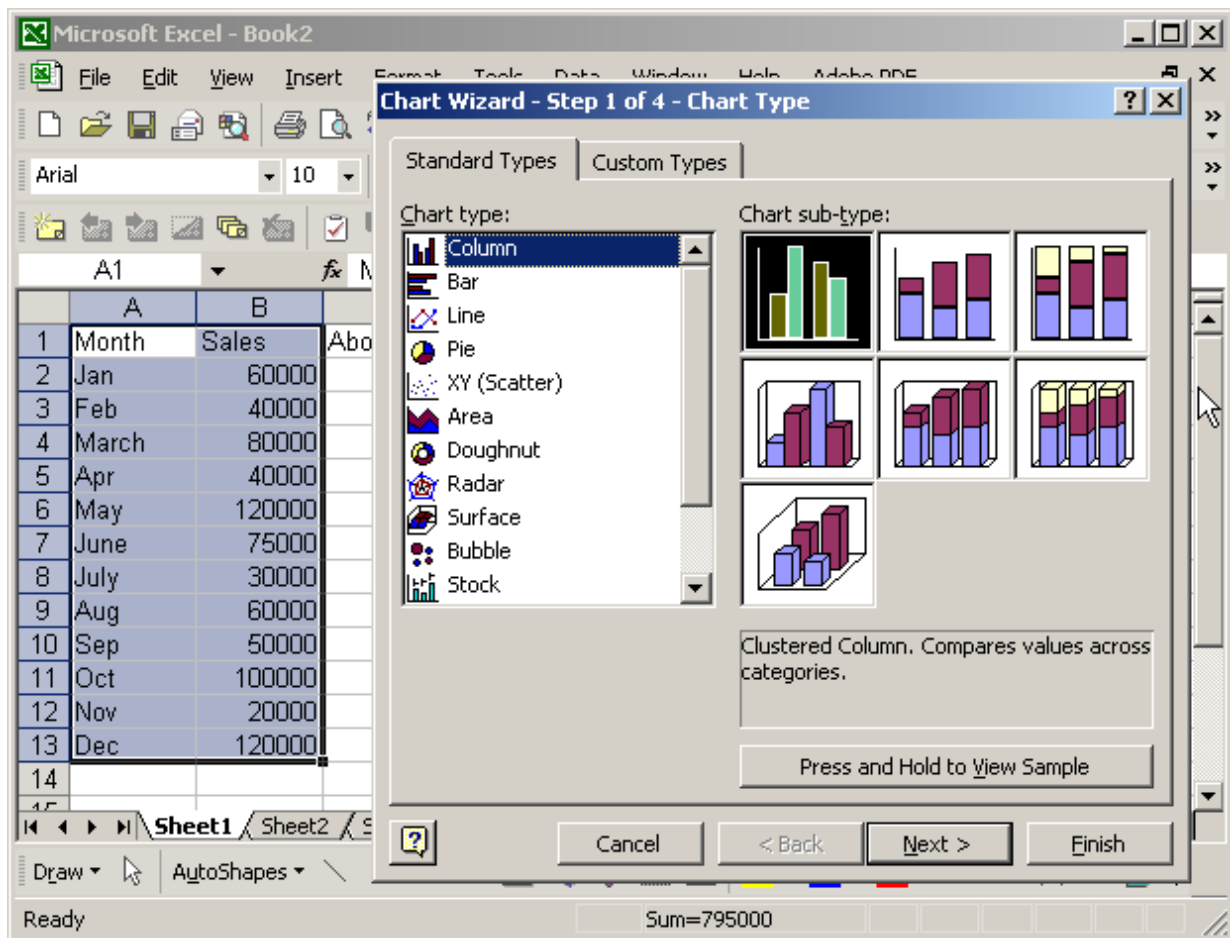


Figure 30: The MS Excel chart wizard.

- Add control limits, computed from the data and based on the common cause variation only, equidistant on either side of the centre line.
- Apply tests to distinguish between data points resulting from special causes and data points resulting from common causes only.

When a single data point is observed outside of one of the control limits, the probability is that this point is not a real change is around 3 out of 1000. Additional tests can be applied to indicate, with a high probability, that a real change has occurred. Using several tests simultaneously increases the sensitivity of inferences but also increases the possibility of a false positive. Wheeler (1993) uses three simple tests:

- Test 1: a single point outside of control limits.
- Test 2: three out of four consecutive points closer to the control limit than to the centre line.
- Test 3: eight or more successive points on one side of the centre line.

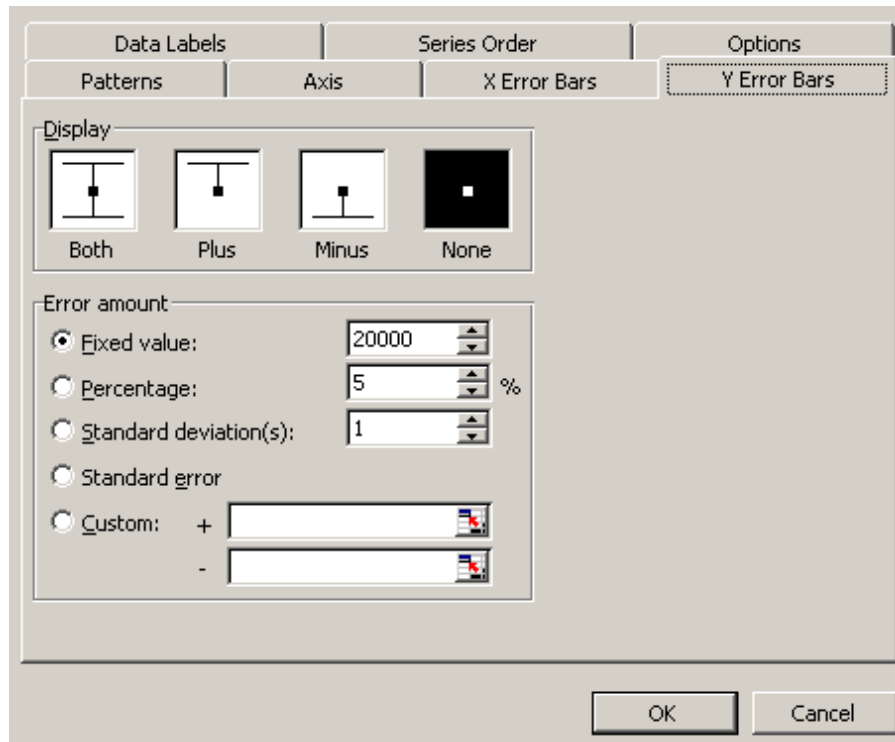


Figure 31: Options for error bar charts.

Figure 33 shows weekly corrected feed conversion ratios (cFCR) measured for batches of broilers on a poultry farm. cFCR values range between 1.5 and 2.0 kg/kg. Date and cFCR have been recorded in columns A and B. In column C the IF function has been used to return the value recorded in column B if the average of the previous three weeks cFCR is greater than 1.85 kg/kg and #N/A otherwise.

Frequency histograms

Histograms are used to summarise discrete or continuous data that are measured on an interval scale. A histogram divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles will be drawn of non-uniform height. A histogram has an appearance similar to a vertical bar graph, but when the variables are continuous, there are no gaps between the bars. When the variables are discrete, however, gaps should be left between the bars. The data shown in column A in Figure 34 is to be plotted as a frequency histogram. The data bins to be used are shown in column C.

With the Analysis ToolPak loaded, click -TOOLS-DATA ANALYSIS... and select the Histogram option. Define the input range (cells A2:A100 in this example), the bin range (cells C2:C21E1) and click OK to produce a graph similar to Figure 35. Note that this graph has been manipulated to eliminate spaces between the bars of the histogram.

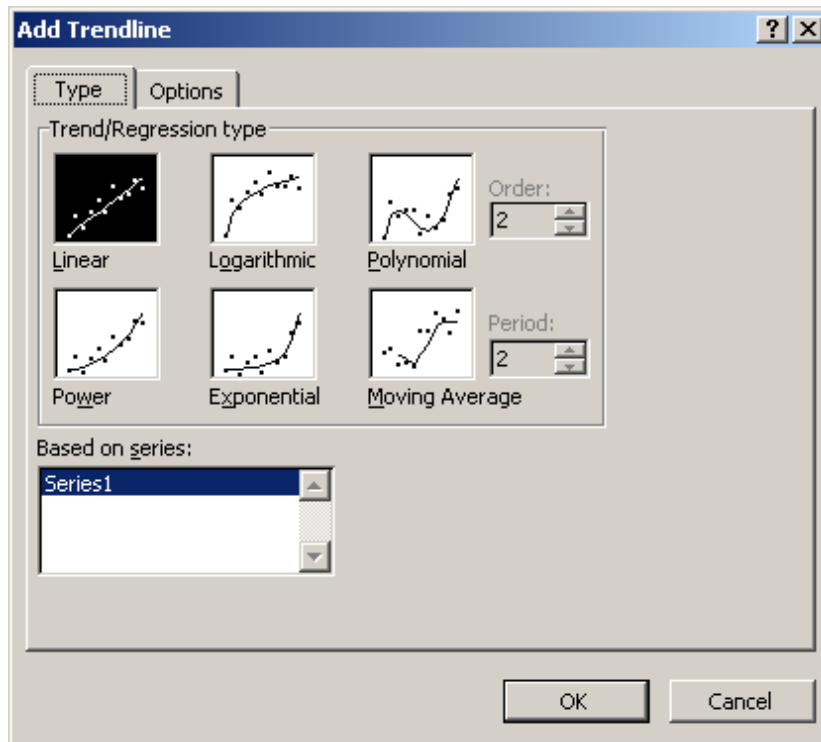


Figure 32: Options for adding trendlines to scatterplot graphs.

The histogram function in Excel works with data formatted as dates, as well as numbers. This facility is useful for plotting epidemic curves.

15.5 Shortcuts

Table 33: Keyboard shortcuts for selecting data.

Key combination	Description
CTRL+SHIFT+*	Selects the current range around the active cell, the largest rectangle of data surrounded by white space.
SHIFT+↑↓	Select cells around and active cell.
END+SHIFT+↑ ↓	Select cells around and active cell to the first non blank cell in the direction selected.
CTRL+A	Selects all the cells in the spreadsheet.
CTRL+SPACEBAR	Selects an entire column.
CTRL+SHIFT	Selects an entire row.
CTRL+SHIFT+END	Extend the selection to the last cell in the worksheet.
CTRL+SHIFT+HOME	Extend the selection to the last non blank cell in the same column or row as the active cell.

Table 34: Keyboard shortcuts for moving around a workbook.

Key combination	Description
PAGE UP	Move up on screen.
PAGE DOWN	Move down on screen.
ALT+PAGE UP	Move one screen to the left.
ALT+PAGE DOWN	Move one screen to the right.
CTRL+HOME	Move to the beginning of the worksheet.
CTRL+END	Move to the end of the worksheet.
CTRL+PAGE DOWN	Move to the next sheet in the workbook.
CTRL+PAGE UP	Move to the previous sheet in the workbook.
CTRL+F6	Move to the next workbook or window.
CTRL+TAB	Move to the next workbook or window.

Table 35: Miscellaneous shortcuts.

Key combination	Description
F4	Toggles between absolute and relative references.
SHIFT+F9	Recalculates the active worksheet.
CTRL+ALT+F9	Recalculates all open worksheets.
CTRL+C	Copies the highlighted section.
CTRL+X	Cuts the highlighted section.
CTRL+V	Pastes the selection on the Windows clipboard.
CTRL+;	Enter the current date.
CTRL+SHIFT+:	Enter the current time.
CTRL+ENTER	Fill the selected cells with the current entry.
SHIFT+F2	Create a cell comment.
ALT+back space	Undo the last command.
ALT+ENTER	Repeat the last command.
CTRL+F	Choose a font.

Table 36: Error messages.

Error	Cause
#DIV/0	Attempting to divide by zero or by an empty cell.
#NAME	Formula contains an undefined variable or function name.
#N/A	No value is available; this can be entered as =NA().
#NULL	The result has no value.
#NUM	Numeric overflow for example when asking for the square root of a negative value.
#REF	Invalid cell reference.
VALUE	Invalid argument (for example asking for the sum when the cell contains text).
#####	The column is too narrow to view the result or the formula returns an invalid date e.g. prior to 1900 or negative time

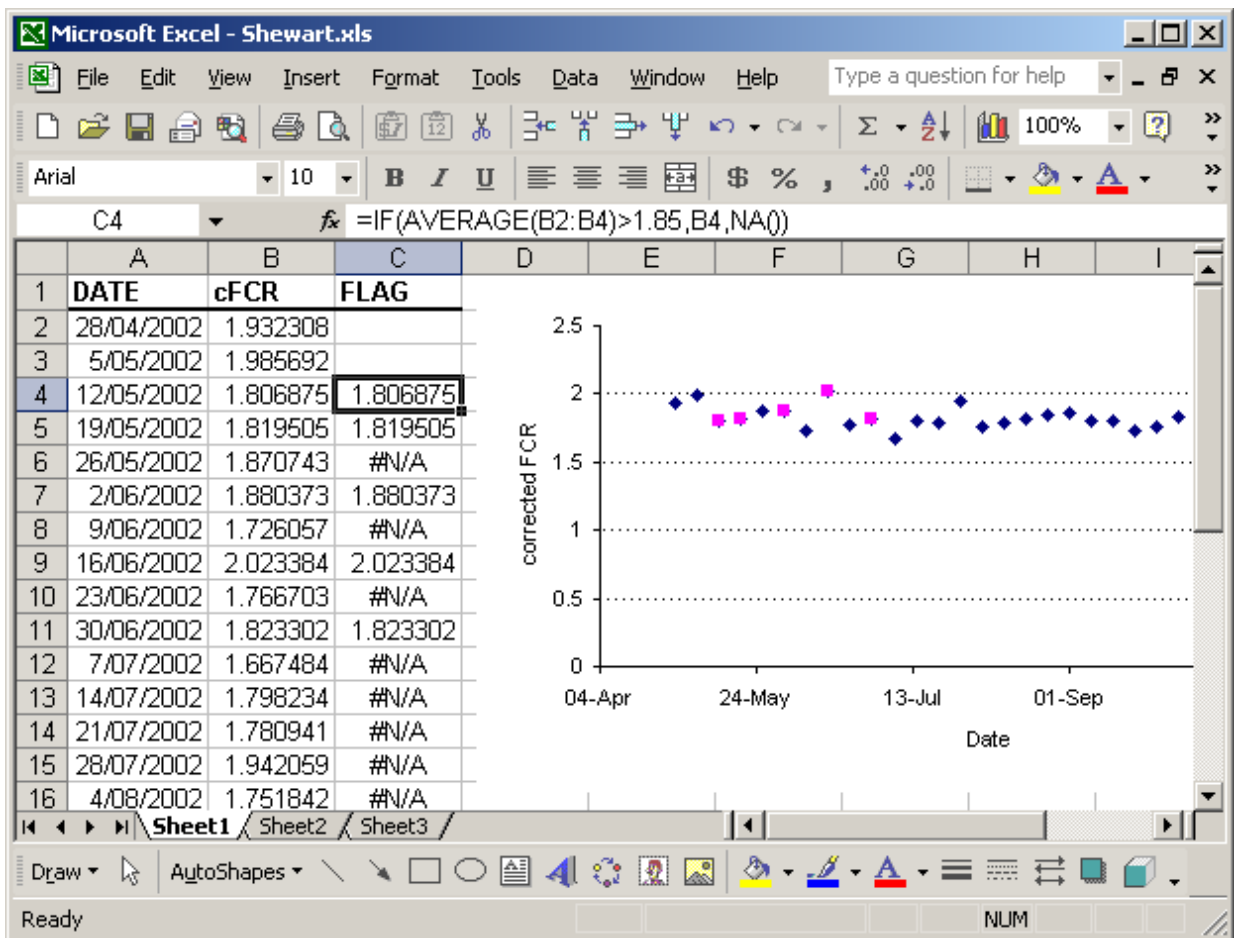


Figure 33: A time series chart. Those data points where the average of the readings taken over the previous three days are shown in pink.

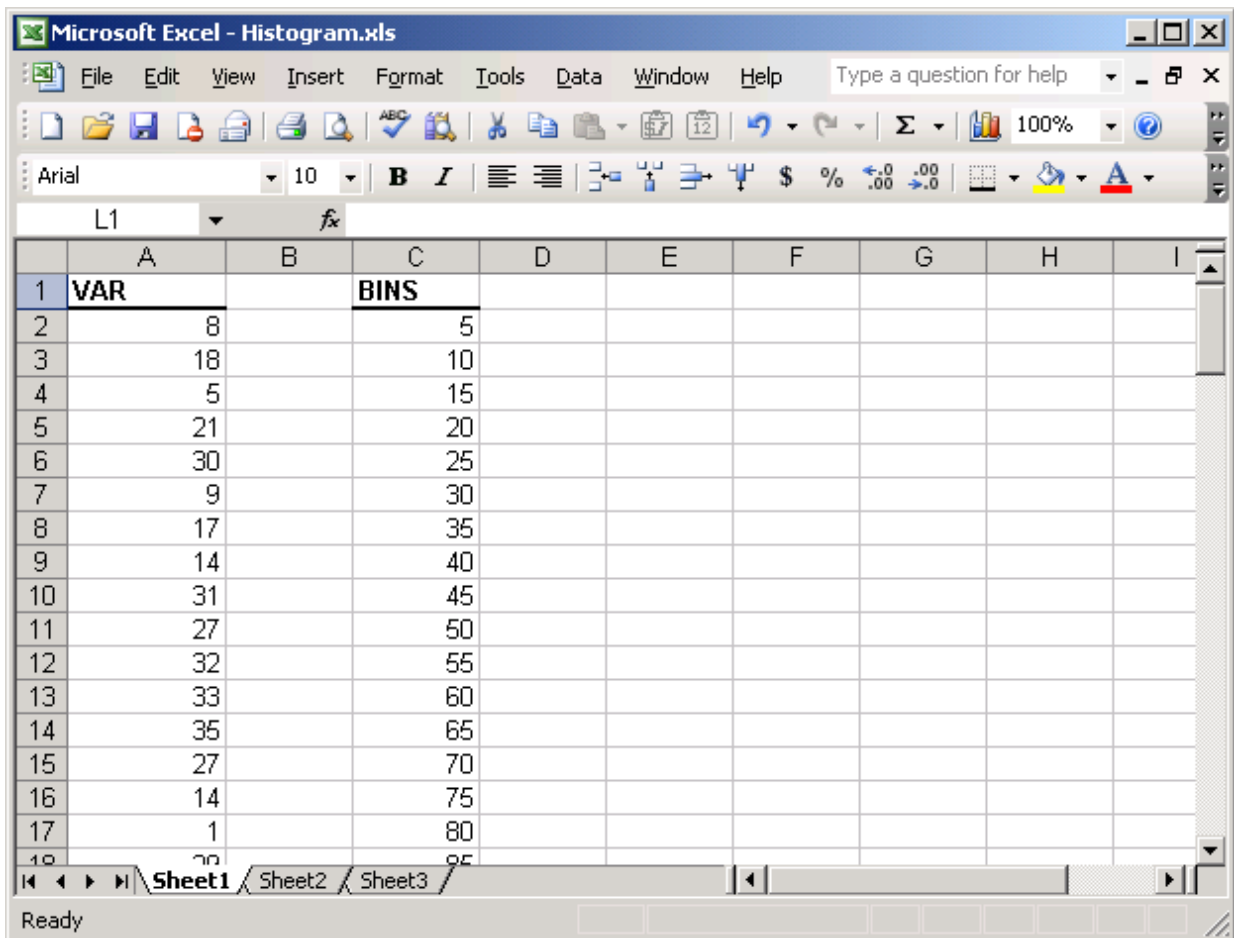


Figure 34: Frequency histograms in MS Excel. The data to be plotted are shown in column A. The bins to be used for the histogram are shown in column C.

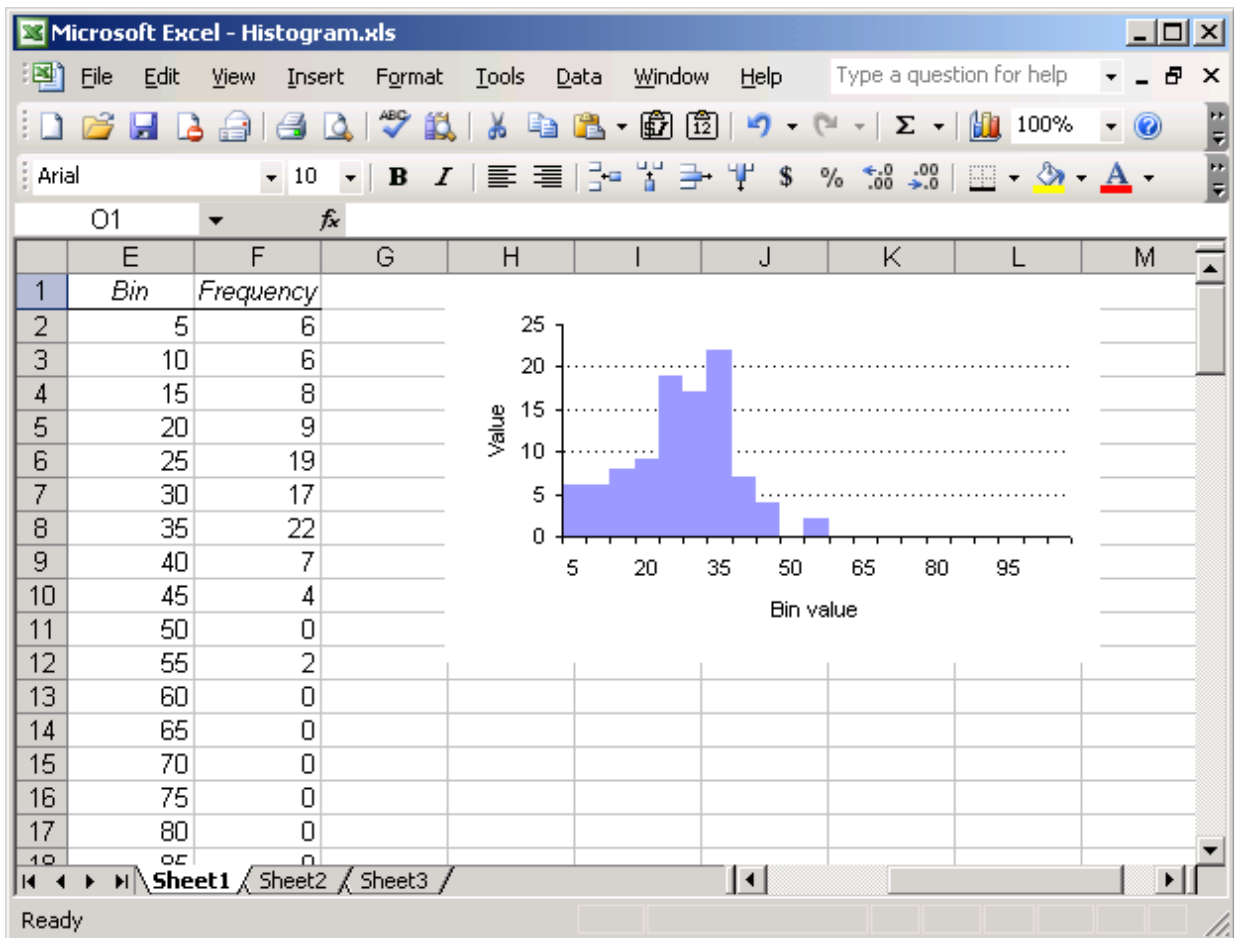


Figure 35: Frequency histograms in MS Excel. Bin and frequency values are shown in cells E1 to F22. A frequency histogram, based on these data, is shown in cells H1:L12.

References

- Altman, D., Machin, D., Bryant, T., & Gardner, M. (2001). *Statistics with Confidence*. London: British Medical Journal.
- Armitage, P., Berry, G., & Mathews, J. (2002). *Statistical Methods in Medical Research*. London: Blackwell Publications.
- Bland, M. (2000). *An Introduction to Medical Statistics*. New York: Oxford University Press.
- Campbell, M. (2006). *Statistics at Square Two*. London: British Medical Journal Publishing Group.
- Campbell, M., & Swinscow, T. (2002). *Statistics at Square One*. London: British Medical Journal Publishing Group.
- Dawson-Saunders, B., & Trapp, R. (1994). *Basic and Clinical Biostatistics* (2nd ed.). New York: Prentice Hall International Inc.
- Devore, J., & Peck, R. (1996). *Statistics: The Exploration and Analysis of Data*. London: Brooks/Cole Publishing Company.
- Everitt, B. (2002). *Modern Medical Statistics: A Practical Guide*. London: Edward Arnold Publishers Ltd.
- Hill, A., & Hill, I. (1991). *Bradford Hill's Principles of Medical Statistics*. London: Edward Arnold.
- Lang, T., & Secic, M. (1997). *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. Philadelphia, Pennsylvania: American College of Physicians.
- Murray, N. (2002). *Import Risk Analysis: Animals and Animal Products*. Wellington New Zealand: New Zealand Ministry of Agriculture and Forestry.
- Petrie, A., & Watson, P. (2005). *Statistics for Veterinary and Animal Science*. London: Blackwell Science.
- Popper, K. (1972). *The Logic of Scientific Discovery*. London: Hutchinson.
- Selvin, S. (1996). *Statistical Analysis of Epidemiological Data*. London: Oxford University Press.