8

Descriptive Statistics

Learning Objectives:

On completing this chapter, you should understand the significance of and how to calculate measures of the central tendency and variability of a dataset:

- mode, median and mean;
- range, variance and standard deviation;
- standard error and confidence intervals.

8.1. Populations and samples

In statistics, the population is the entire group from which data may be collected and conclusions drawn. However, since populations may be very large and inconvenient to work with, statistical analysis is frequently performed on a sample, a smaller group drawn from the population. Assuming the sample is representative of the population, e.g. selected at random and sufficiently large, conclusions made about the sample should be valid for the population as a whole. For example, if we wanted to know whether children born in China in the 1950s were shorter than children born in China in the 1970s, it would be impossible to study the populations, i.e. all children born in China in the 1950s and 1970s, since these are far too large. However, statistical conclusions drawn about samples taken from these two populations should be valid for the whole population, assuming that the samples are unbiased and truly representative of the populations. So far so good. However, depending on the sample size, a statistic calculated for a sample based on the formula for a population may tend to produce a biased result, that is, an overestimate or an underestimate of the true value. For this reason, formulae for calculation statistics from samples often contain a small correction (e.g. $n-1$ in place of n, the number of datapoints) to provide a more accurate answer. For this reason, you always need to be clear whether you are calculating a statistic for a population or a sample, and to use the correct formula (if appropriate).

8.2. The central tendency

Different frequency distributions can be described mathematically by measuring the central tendency and variability of the dataset. The central tendency is a summary measure of the middle of a dataset, which is commonly measured by any of three common descriptive statistics, the '3 Ms': mode, median and mean.

Mode

The mode is the most frequently occurring value in a dataset. It is easy to determine, but is subject to great variation and consequently is of limited value.

Median

The median is the middle value in a dataset, i.e. half the variables have values greater than the median and the other half values which are less. The median is less sensitive to outliers (extreme scores) than the mean and is thus a better measure than the mean for highly skewed distributions, such as family income. Note that the median equals the 50th percentile (P_{50}), i.e. the second quartile (Q_2).

Mean

The mean is the average value of a dataset, i.e. the sum of all the data divided by the number of variables. The arithmetic mean is commonly called the 'average'. When the word 'mean' is used without a modifier, it usually refers to the arithmetic mean. The mean is a good measure of central tendency for symmetrical (e.g. normal) distributions, but can be misleading in skewed distributions since it is influenced by outliers. Therefore, other statistics such as the median may be more informative for distributions which are highly skewed. The mean, median and mode are equal in symmetrical frequency distributions. The mean is higher than the median in positively (right) skewed distributions and lower than the median in negatively (left) skewed distributions.

The formula for the arithmetic mean is:

$$\text{mean} = \frac{\sum X}{N}$$

where $\sum$ means 'sum'; X are the raw datapoints; and N is the number of scores (datapoints).

The geometric mean is the nth root of the product of the scores, for example, the geometric mean of the scores 1, 2, 3 and 4 is the 4th root of $1*2*3*4$, which is the 4th root of $24 = 2.21$. The geometric mean is less affected by extreme values than the arithmetic mean and is useful for some positively skewed distributions. However, the arithmetic mean is far more commonly encountered than the geometric mean.

8.3. Variability

Measurements in biology are frequently quite variable. There are many different sources for this variation, such as biological differences between individuals, resolution of measurement techniques and simple experimental error. It is important to be able to measure and describe the variability in datasets. While the central tendency is a summary measure of the middle of a dataset, variability (or dispersion) measures the amount of scatter in the dataset (e.g. Figure 8.1). Variability is commonly measured by three criteria: range, variance and standard deviation.

Range

Range is the difference between the largest and the smallest value in the dataset. Although it is a crude measure of variability, it is easy to calculate and useful as an outline description of a dataset, for example in box and whisker plots (Section 8.8). However, since the range only takes into account two values from the entire dataset, it may be heavily influenced

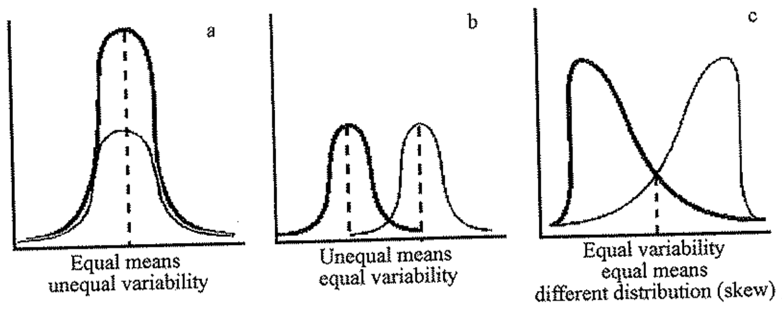

Figure 8.1 Variability

by outliers in the data. Therefore, another criterion is commonly used – the interquartile range, the interval between the 25th and 75th percentiles. In a normally distributed population, the interquartile range contains 50% of the datapoints making up the dataset: $Q_3 - Q_1$. A further measure which is even less subject to extreme scores is the semi-interquartile range, which is half of the interquartile range: $(Q_3 - Q_1)/2$.

Since the semi-interquartile range is little affected by extreme scores, it is a good measure of spread-out or skewed distributions. However, it is more subject to sampling fluctuation (i.e. how much a statistic varies from one sample to another) in normal distributions than the standard deviation (see below), so it not often used for data which are normally distributed.

Variance

The variance of a dataset is more complicated to understand than the range, but is a measure of how 'spread-out' a distribution is. A deviation score is a measure of by how much each point in a frequency distribution lies above or below the mean for the entire dataset:

$$\text{Deviation score} = X - m$$

where X is the raw score and m is the mean for the dataset.

The variance is the mean of the squares of all the deviation scores for a dataset. This represents the amount of deviation of the entire dataset from the mean:

$$\text{Variance of a population} = \sigma^2 = \frac{\sum(X - \mu_x)^2}{N}$$

where $\sum$ is the sum, X is the raw score, μ_x is the mean of the population, and N is the number of datapoints.

$$\text{Variance of a sample} = s^2 = \frac{\sum(X - m)^2}{n - 1}$$

where $\sum$ is the sum, X is the raw score, m is the mean of the sample, and n is the number of datapoints in the sample.

Note that the variance is expressed in squared units, for example, if the raw scores are weight in kg, the variance is kg^2. For this reason, it is more useful to consider the square root of the variance, which is the standard deviation.

Standard deviation

The standard deviation (SD) is the square root of the variance and is the most commonly used measure of how 'spread-out' a distribution is:

$$\text{Standard deviation of a population: } \sigma_x = \sqrt{\frac{\sum(X - \mu_x)^2}{N}}$$

$$\text{Standard deviation of a sample: } S_x = \sqrt{\frac{\sum(X - m)^2}{n - 1}}$$

As with the other measures of data variability, the standard deviation determined from a sample (subset) of a dataset will be biased – since outliers are excluded, it will tend to underestimate the population standard deviation. Hence the formula needs to be modified for samples rather than whole populations. The standard deviation is probably the most useful measure of data spread. As you will see, many formulas in inferential statistics (Chapters 10 and 11) use the standard deviation. Although the standard deviation is less sensitive to extreme scores than the range, it is more sensitive than the semi-interquartile range. For this reason, the standard deviation should at least be supplemented by if not replaced by the semi-interquartile range when the possibility of extreme scores is present or for highly skewed datasets.

8.4. Standard error

Any statistic can have a standard error, which is the standard deviation of the sampling distribution of that statistic. Inferential statistics and significance testing (Chapters 10 and 11), and confidence intervals (below) are all based on standard errors. The standard deviation is an index of how closely individual data points cluster around the mean, thus each standard deviation refers to an individual datapoint. In contrast, standard errors indicate how much sampling fluctuation a summary statistic shows, that is, how good an estimate of the population the statistic is (e.g. the standard error of the mean, σ_m). The standard error of any statistic depends in part on the sample size – in general, the larger the sample size the smaller the standard error.

$$\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{N}}$$

How good an estimate is the mean of a population? One way to determine this is to repeat an experiment many times and to determine the mean of the means. However, this is at best tedious and frequently impossible. Fortunately, the standard error of the mean can be calculated from a single experiment and indicate the variability of the statistic:

$$\sigma_m = \frac{\mathrm{SD}}{\sqrt{N}}$$

We will come back to the use of standard errors again later.

8.5. Confidence intervals

In a normal distribution 68% of datapoints fall within ±1 standard deviations from the mean; 95% of datapoints fall within ±2 standard deviations from the mean (actually ±1.96 standard deviations); and 99.7% of datapoints fall within ±3 standard deviations from the mean (Figure 8.2).

A confidence interval gives an estimated range of values which is likely to include an unknown datapoint. The width of the confidence interval gives us some idea about how uncertain we are about the parameter, for example, a very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter. Confidence intervals are more informative than the results of inferential tests (Chapters 10 and 11), which only help you decide whether to reject a hypothesis, since they provide a probable range of numerical values for statistical parameters.

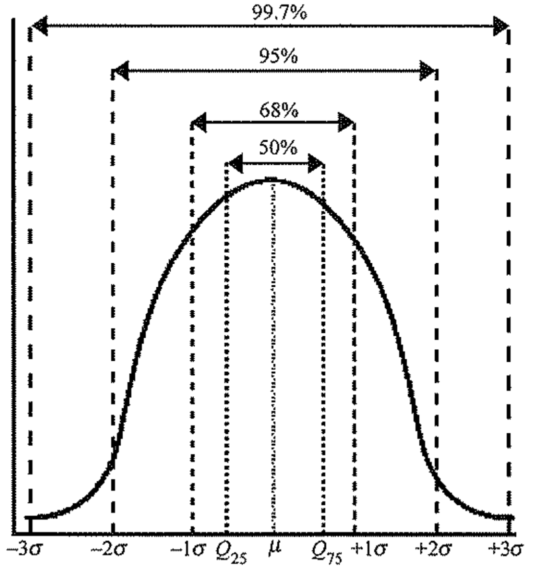

Figure 8.2 Normal distribution

In a normal distribution, since there is less than a 1 in 20 chance of any datapoint falling outside ±2 standard deviations from the mean, we say that this range represents the 95% confidence interval, and the probability (P) of this range containing a particular datapoint is $P = 0.95$ (Chapter 9 contains a more detailed explanation of probability). Similarly, since there is less than a 1 in 99.7 chance of any sample in the population falling outside ±3 standard deviations; this represents a 99% confidence interval for the population, and $P = 0.99$. Confidence intervals can be constructed for any statistical parameter, not just the mean.

So when do you use standard deviations, standard errors or confidence intervals?

1. Use standard deviations when you are referring to individual data points. This tells you about the spread of the data.

2. Use standard errors when you are referring to differences between sample statistics, e.g. the mean. This tells you about the accuracy of your estimate.

3. Use confidence intervals when you want to convey the significance of differences between groups.

8.6. Parametric and non-parametric statistics

Statistical methods which depend on estimates of the parameters of populations or probability distributions are referred to as parametric methods, and include: Student's *t*-test; ANOVA (analysis of variance); regression analysis; and correlation analysis. These tests are only meaningful for continuous data which is sampled from a population with an underlying normal distribution, or whose distribution can be rendered normal by mathematical transformation. Non-parametric methods require fewer assumptions about a population or probability distribution and are applicable in a wider range of situations. For example, they can be used with qualitative data, and with quantitative data when no assumption can be made about the population probability distribution.

Non-parametric methods are useful in situations where the assumptions required by parametric methods are questionable. A few of the more commonly used non-parametric methods include: χ^2 test; Wilcoxon signed-rank test; Mann–Whitney–Wilcoxon test; and Spearman rank correlation coefficient. These tests are 'distribution free', i.e. the population from which the sample was drawn does not need to have a normal distribution. Unlike parametric tests which can give erroneous results if applied to the 'wrong sort of data', these methods can be safely used in a wider range of circumstances. Unfortunately, they are less flexible in practice and less powerful than parametric tests.

In cases where both parametric and non-parametric methods are applicable, statisticians usually recommend using parametric methods because they tend to provide better precision.

In statistical jargon, accuracy is a measurement of how close the average of a set of measurements is to the true or target value. Precision is a measure of the closeness of repeated observations to each other without reference to the true or target value, i.e. the reproducibility of the result.

8.7. Choosing an appropriate statistical test

In order to choose an appropriate statistical test, you must answer two questions:

1. What are the features of the dataset being analysed?
2. What is the goal of the analysis?

Table 8.1 summarizes some of the statistical tests which can be used to analyse different datasets. Not all of these tests are described in this book,

Table 8.1 Some of the statistical tests which can be used to analyse different datasets

Goal	Dataset		
	Normal distribution	Non-normal distribution	Binomial distribution
Describe one group	Mean, standard deviation	Median, interquartile range	Proportion
Compare one group to a hypothetical value	One-sample *t*-test	Wilcoxon test	χ^2 or binomial test
Compare two unpaired groups	Unpaired *t*-test	Mann–Whitney test	Fisher's exact test (or χ^2 for large samples)
Compare two paired groups	Paired *t*-test	Wilcoxon test	McNemar's test
Compare three or more unmatched groups	One-way ANOVA	Kruskal–Wallis test	χ^2 test
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane *Q* test
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients
Predict value from another measured variable	Simple regression	Non-parametric regression	Simple logistic regression
Predict value from several measured variables	Multiple regression		Multiple logistic regression

but they have been included in the table for reference purposes. In subsequent chapters we will explore the most frequently employed statistical methods and how they can be used.

8.8. Exploratory data analysis

Hopefully, it will now be clear that one of the most important aspect of statistics is to use the appropriate method rather than a test which may generate a meaningless and misleading answer. Choosing a test largely depends on the nature of the data being analysed, and critically whether this has a normal distribution (so a parametric test can be used) or not (meaning a non-parametric method must be used). This crucial information can be obtained through a process known as ***exploratory data analysis***, which includes many tools designed to reveal possible errors in the data (calculation or experimental errors, typing mistakes, etc.), data outliers, which should be investigated, and the underlying nature of the dataset (e.g. frequency distribution).

Exploratory data analysis comprises many different methods, including descriptive statistics, but the most powerful are graphical methods which literally paint a picture of the dataset. As an example, we will look at some of the most frequently used methods, all of which are easily performed by hand or with commonly available software.

Scatter plots

Consider the three datasets in Table 8.2. At first sight, all three look very similar, with identical means and standard deviations. However, a scatter plot of the data quickly reveals considerable differences between the three datasets (see Figures 8.3 – 8.5).

Frequency distribution histograms

A frequency distribution is a series of rectangles representing the frequencies of the class intervals. Since this was described in the previous chapter (Section 7.5), we will not repeat the description here.

Table 8.2 Datasets 1–3

	Set 1		Set 2		Set 3	
	*x*1	*y*1	*x*2	*y*2	*x*3	*y*3
	10	9.19	10	7.56	8	6.58
	8	8.14	8	6.67	8	6.66
	13	8.74	13	12.74	8	7.71
	9	8.81	9	7.11	8	8.84
	11	9.26	11	6.91	8	8.47
	14	8.10	14	8.84	8	7.04
	6	6.23	6	6.17	8	5.25
	4	3.10	4	6.39	19	12.56
	12	9.13	12	8.15	8	5.56
	7	7.26	7	6.42	8	7.91
	5	4.74	5	5.73	8	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5
Standard deviation	3.3	2.0	3.3	2.0	3.3	2.0

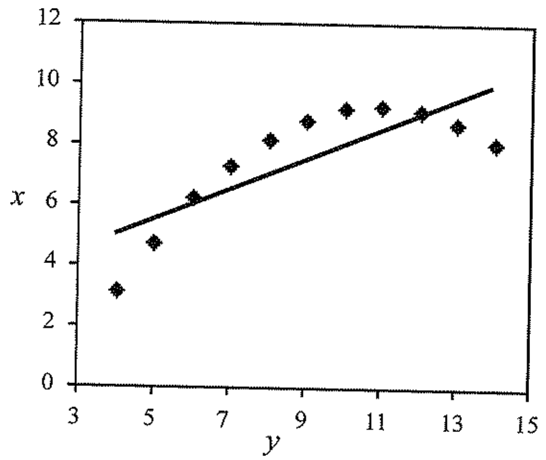

Figure 8.3 Dataset 1: the datapoints all lie on a smooth curve with little scatter – this would appear to be 'good' data

Stem and leaf plots

A stem and leaf plot is like a histogram turned on its side but shows more information – the numerical values of each datapoint in addition to the

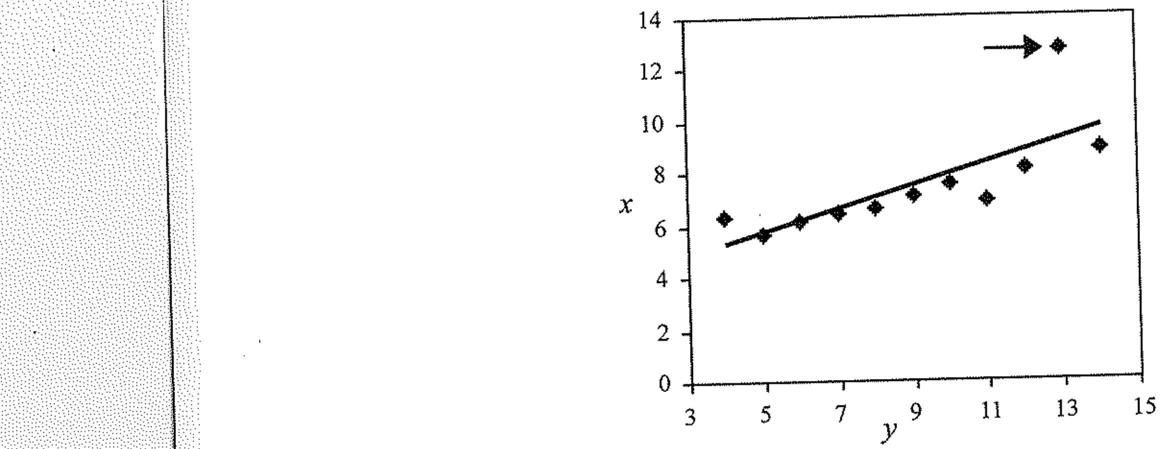

Figure 8.4 Dataset 2: most of the datapoints lie close to a straight line, but one point (arrow) is suspiciously misplaced. This could be the result of either experimental or typographical error, but it is certainly worth investigating the cause before performing further analysis

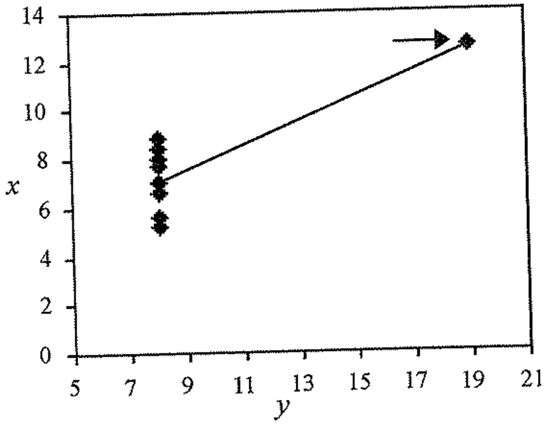

Figure 8.5 Dataset 3: in this example, the single data outlier (arrow) would heavily influence the result of any statistical analysis. It is important to investigate the cause of this isolated datapoint (e.g. experimental error or design) and to consider carefully whether to include this datapoint in any analysis

overall pattern. The dataset – 39, 42, 44, 47, 48, 48, 51, 52, 53, 53, 54, 55, 55, 55, 55, 56, 56, 57, 57, 58, 58, 59, 59, 59, 59, 61, 61, 62, 63, 63, 64, 65, 65, 65, 66, 66, 66, 67, 69, 69, 71, 71, 76, 81, 84, 92 would be represented as:

```
3|9
4|2 4 7 8 8
5|1 2 3 3 3 4 5 5 5 5
5|6 6 7 7 8 8 8 9 9 9 9
6|1 1 2 3 3 4 5 5 5
6|6 6 6 7 9 9
7|1 1 6
8|1 4
9|2
```

The numbers to the left are the 'stem' of the plot – the tens digits in the frequency distribution of the dataset. The numbers to the right are the 'leaves' – the units digits in the frequency distribution, e.g. 5 | 6 represents a score of 56. Scores greater than 99 can be represented as follows: 25 | 6 for 256, 67 | 9 for 679, etc. The effect produced is that of a histogram, but each individual datapoint can be seen. No graphical software is necessary to produce the pattern, which can easily be reproduced in text form. In this example, the data approximates to a normal distribution (with a slight left-skew).

Box and whisker plots

This alternative method of examining data makes use of common calculated numerical measures (median, interquartile range), but displays the data in a visual form (Figure 8.6). In the top plot in Figure 8.6, the median value is symmetrically placed in the middle of the box (interquartile range, which by definition covers 50% of the points in the dataset), so this dataset is normally distributed. In the middle plot, the interquartile range (box) is the same, but here the median value is at the right-hand end of the box, meaning that this dataset does not have a normal distribution, but has what is known as a right or positive skew. This dataset would provide inaccurate answers if subjected to parametric tests (unless transformed to a normal distribution first). In the lower plot, the median value is again in

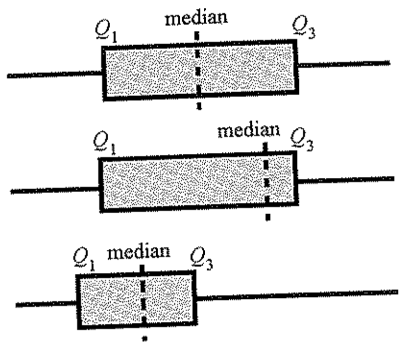

Figure 8.6 Box and whisker plots

the middle of the interquartile range, but this (i.e. the box) covers less of the dataset, meaning that there are more outliers in the data which could reduce the accuracy of any statistical analysis.

Examining the data in question using the above methods *before* applying any statistical tests is vital to any meaningful analysis, ensuring that any numerical summaries of the data or predictions made from the data are valid. Unfortunately, it is frequently overlooked. *Always* carry out some form of exploratory data analysis before proceeding further. Preferably draw at least one picture or graph. Much of statistics is about detecting patterns – something which the human eye and brain are very good at.

Problems (answers in Appendix 1)

Table 8.3 contains a set of data on the microbiological quality of bottled drinking water. In this study, the number of bacterial colony-forming units per millilitre of bottled water was measured for 120 different water samples.

8.1. Construct a grouped frequency distribution table for this dataset.

8.2. Plot a frequency distribution histogram of the data.

8.3. How would you describe this dataset (normal, negative skew or positive skew?)

8.4. Calculate:

(a) the 90th percentile for this dataset;

(b) the 25th percentile for this dataset.

Table 8.3 Microbiological quality of drinking water

Colony forming units mL^{-1}					
9159	6351	9726	8859	5832	6891
3783	7613	9527	9292	7512	6631
848	8799	8259	7645	9166	5864
7478	6758	6038	7952	8166	7078
7999	8492	8712	7718	8352	8659
8652	5791	8392	7698	8185	6951
8952	5184	8005	7912	4664	906
7818	9085	8292	7779	8259	8119
4117	6512	8432	8452	7545	383
8939	8672	6105	8966	8693	7532
9246	7598	6098	9413	8279	8252
4584	8686	7919	4504	6237	9146
6171	4184	8906	5097	7532	8586
6538	8793	6611	7879	6805	8246
7645	9092	8158	8339	8599	9006
7799	8659	7619	9166	8079	5084
2396	8365	8566	7478	8172	7812
5417	7685	8519	1735	8486	6905
8512	8079	7912	8653	7785	8699
6571	7732	8739	7798	7625	7519

8.5. Calculate:

(a) the mean for this dataset;

(b) the median for this dataset;

(c) the mode for this dataset.

8.6. Calculate:

(a) the range for this dataset;

(b) the semi-interquartile range for this dataset;

(c) the variance for this dataset;

(d) the standard deviation for this dataset.

8.7. Exploratory data analysis:

(a) Construct a scatter plot of the following dataset. Are the data normally distributed?

x	10	8	13	9	11	14	6	4	12	7	5
y	11.2	8.1	9.7	9.8	12.8	8	6.2	3	11.8	7.3	5.7

(b) Construct a frequency histogram of the following dataset. Are the data normally distributed?

x	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
y	0	0	1	4	6	9	13	9	8	4

(c) Construct a stem and leaf diagram of the following dataset. Are the data normally distributed?

21, 23, 25, 26, 26, 27, 29, 1, 32, 32, 33, 35, 35, 36, 37, 38, 38, 39, 41, 41, 41, 42, 42, 44, 45, 47, 48, 48, 49, 51, 52, 53, 53, 53, 54, 55, 55, 55, 57, 61, 62, 63, 66, 71, 74, 91.

(d) Sketch a box and whisker plot of the following dataset. Are the data normally distributed? 14, 20, 22, 25, 27, 28, 31, 33, 38, 42, 51, 53, 61, 62, 65, 71, 74, 77, 78, 84, 86, 91. Median = 52, first quartile = 29, third quartile = 73.

9

Probability

LEARNING OBJECTIVES:

On completing this chapter, you should understand the basic principles of probability theory, including:

- how to calculate probability in simple scenarios;
- the difference between selection with and without replacement;
- how to calculate the probability of multiple events.

9.1. Probability theory

Although most people find probability an interesting and enjoyable area of mathematics, why should a biologist need to understand and know how to calculate probabilities? This is because statistical methods depend upon probability theory.

Examples include important activities such as sampling from populations and hypothesis testing (Chapter 10), and probability distributions (Chapter 8).

$$\text{Probability, } P = \frac{\text{number of specific outcomes of a trial}}{\text{total number of possible outcomes of a trial}}$$

The simplest way to understand probabilities is through proportional frequency.

Example

In a group of mice there are 200 white mice and 50 brown mice:

1. Probability, P, is normally written as a decimal, e.g. $P = 0.5$. All probabilities lie between 0 and 1.

2. The proportional frequency of brown mice is $50/250 = 1/5 = 0.2$.

3. If we randomly take one mouse there is a $1/5$ chance of it being brown (0.2).

4. The probability of picking a brown mouse as a single random sample is equivalent to the proportional frequency of brown mice in the group (population).

5. If there were 250 white mice, the probability of selecting a brown mouse would be $0/250 = 0$. The probability of selecting a white mouse would be $250/250 = 1$.

9.2. Replacing or not replacing selections

If we replace the first selection from a population before making a second selection, then the probability of making any given selection is unaltered. Thus, in the above example the probability of picking a brown mouse is still $50/250 = 1/5 = 0.2$. However, if we do not replace our first selection the probability when making the second selection changes.

Example

In a group of mice there are 200 white mice and 50 brown mice:

1. Selection one = a brown mouse.

2. If this is not replaced there are now 249 mice (not 250) and only 49 brown mice (not 50). The probability of picking a brown mouse in the second sample is now $49/249$, not $50/250$. The chance of randomly selecting a brown mouse has decreased (slightly).

3. Similarly, the probability of randomly picking a white mouse in the second sample is now $200/249$ rather than $200/250$ as it would have been in the first selection, i.e. the chance of picking a white mouse in the subsequent selection increases as the chance of picking a brown mouse decreases.

Studying repeated samples (selections) from natural populations is easier if we assume that replacement occurs. This is usually true if the population is large, for example, taking one locust from a swarm of millions will not significantly change the overall population. When the result of the first sample does not affect the probability of the result of subsequent samples, the samples are said to be independent (an important requirement of many of statistical tests).

9.3. Calculating the probability of multiple events

There are two rules of probability:

1. **The SUM or OR rule** – the probability of any one of several distinct events is the sum of their individual probabilities, provided that the events are mutually exclusive (occurrence of one event precludes the others, e.g. selection without replacement).

2. **The PRODUCT or AND rule** – the probability of several distinct events occurring successively or jointly is the product of their individual probabilities, provided that the events are independent (the outcome of one event must have no influence on the others, e.g. tossing a coin).

The number of possible combinations of events is given by the factorial product of the number of events (written as ‘n’) – the product of an integer and all the lower integers, for example, for three events (X, Y, Z), the number of possible combinations $= 3! = 3 * 2 * 1 = 6$:

1	2	3	4	5	6
XYZ	XZY	YXZ	YZX	ZXY	ZYX

Note that these are all different combinations, for example, crossing the road then looking for cars is not the same as looking for cars then crossing the road.

Example

A population of 50 brown mice, 200 white mice, selections with replacement:

1. The probability of three brown mice in three selections:

$$(50/250) * (50/250) * (50/250)$$
$$= (1/5) * (1/5) * (1/5) = 0.008$$

2. The probability of selecting, in order, brown, brown and then white:

$$(50/250) * (50/250) * (200/250)$$
$$= (1/5) * (1/5) * (4/5) = 0.032$$

3. If, however, we are not interested in the order (i.e. brown, brown, white) but just the overall outcome (i.e. two brown, one white), the probability is different. The possible outcome of three selections with replacement is shown in Table 9.1. Thus, the sum of probabilities of a set of mutually exclusive, exhaustive outcomes is 1, but the probability of two brown mice and one white mouse, irrespective of the order of selection is as shown in Table 9.2. Note the difference in outcome between an ordered selection (probability = 0.032) and selection irrespective of order (probability = 0.096) = the sum of all the possible ordered selections.

Table 9.1 Possible outcome of three selections with replacement

Selection outcome			Probability of selection			Probability of outcome	
1	2	3	1	2	3	Sum	Total
B	B	B	1/5	1/5	1/5	(1/5) * (1/5) * (1/5)	0.008
B	W	B	1/5	4/5	1/5	(1/5) * (4/5) * (1/5)	0.032
B	B	W	1/5	1/5	4/5	(1/5) * (1/5) * (4/5)	0.032
B	W	W	1/5	4/5	4/5	(1/5) * (4/5) * (4/5)	0.128
W	B	B	4/5	1/5	1/5	(4/5) * (1/5) * (1/5)	0.032
W	W	B	4/5	4/5	1/5	(4/5) * (4/5) * (1/5)	0.128
W	B	W	4/5	1/5	4/5	(4/5) * (1/5) * (4/5)	0.128
W	W	W	4/5	4/5	4/5	(4/5) * (4/5) * (4/5)	0.512
						Total	1.0

Table 9.2 Probability of two brown mice and one white mouse irrespective of order of selection

Selection outcome			Probability of selection			Probability of outcome	
1	2	3	1	2	3	Sum	Total
B	W	B	1/5	4/5	1/5	(1/5) * (4/5) * (1/5)	0.032
B	B	W	1/5	1/5	4/5	(1/5) * (1/5) * (4/5)	0.032
W	B	B	4/5	1/5	1/5	(4/5) * (1/5) * (1/5)	0.032
						Total	0.096

9.4. The binomial distribution

The binomial probability distribution describes what will happen when there are only two possible outcomes of an event, e.g. tossing a coin (heads or tails) or selections from a population consisting of two types of member (e.g. brown and white mice).

Such binary variables turn out to occur quite frequently in biology. In its simplest form, the binomial expansion summarizes the possible outcomes for any number of samples when there are only two possible outcomes (e.g. brown and white mice). For independent events, the binomial distribution is given by:

$$(P + Q)^n$$

where P is the probability of one of the possible events, Q is the probability of the second event ($Q = 1 - P$), and n is the number of trials in the series.

For samples of 1 ($n = 1$): $(P + Q)^1 = (P + Q)$,

For samples of 2 ($n = 2$): $(P + Q)^2 = P^2 + 2PQ + Q^2$,

For samples of 3 ($n = 3$): $(P + Q)^3 = P^3 + 3P^2Q + 3PQ^2 + Q^3$, etc.

To return to the mice, these expansions of the binomial equation describe all the possible outcomes from the experiment above. If P = brown mice and Q = white mice, for three samples from the population ($n = 3$) there is: one way of obtaining three brown mice (BBB) = P^3; three ways of obtaining two brown mice and one white mouse (BBW:BWB:WBB) = $3P^2Q$; three ways of obtaining one brown mouse and 2 white mice (BWW:WBW:WWB) = $3PQ^2$; and one way of obtaining three white mice (WWW) = Q^3.

These are all the possible outcomes. In the population from which the samples were drawn:

$$\begin{aligned}&50 \text{ brown mice, } P = 50/250 = 0.2\\&200 \text{ white mice, } Q = 200/250 = 0.8\end{aligned}$$

and we can therefore calculate the distribution of outcomes from the binomial equation. In this example we can calculate the probability of two brown mice and one white mouse being selected as:

$$3P^2Q = 3(0.2)^2(0.8) = 0.096$$

This method is acceptable when there is a small number of samples and a small number of outcomes, but gets progressively more difficult as the sample size increases. For example, try using this method to calculate how many different ways there are to select seven brown mice and six white mice in 13 selections. To perform such calculations as this, we can use the following equation:

$$\text{Number of outcomes} = \frac{n!}{r!(n-r)!}$$

where n is the number of selections and r is the number of one of the outcomes (remember '!' = factorial).

Example

For two brown mice and one white mouse (i.e. BBW, BWB, WBB), the number of outcomes is:

$$\frac{3!}{2!(3-2)!} = \frac{3*2*1}{2*1(1)} = \frac{6}{2} = 3$$

So for seven brown mice and six white mice, the number of possible outcomes is given by:

$$\frac{13!}{7!(13-7)!} = \frac{13!}{7!*6!} = \frac{13*12*11*10*9*8}{6*5*4*3*2*1}$$
$$= \frac{1\,235\,520}{720} = 1716$$

If we know the probability of the outcome for a single selection (e.g. probability of selecting a brown or a white mouse), we can calculate the total probability for the outcome using:

$$P(r) = \frac{n!}{r!(n-r)!} * p^r(1-p)^{n-r}$$

where P is the total probability of the outcome (e.g. two brown mice and one white mouse), p is the probability of the event that occurs r times, and $(1-p)$ is the probability of the event that occurs $n-r$ times.

In our example of two brown mice and one white mouse:

$$\frac{3!}{2!(3-2)!} * \left(\frac{50}{250}\right)^2 * \left(\frac{200}{250}\right)^1$$
$$= 3*(0.2)^2*(0.8)^1 = 3*0.04*0.8 = 0.096$$

In practice, rather than actually performing such calculations, it is more usual to look up the probability of an event from a pre-calculated table of binomial probabilities (Appendix 3).

A particular importance of probability theory in statistics is that it controls sampling of populations and can be used to determine how large a sample needs to be taken from a population in order for an experiment to be successful, i.e. to have a statistically meaningful outcome.

Example

Suppose that 4% of students carry an inherited defect in the (mythical) *statz* gene which restricts the ability of carriers to understand statistics. The only way to determine if someone is a carrier is to select individuals from the population at random and test them. If the number of students tested is too small there is a risk of not finding any carriers, but if it is too large, it will not be possible to mark all the tests. What sample size is required to give a good likelihood of sampling affected individuals? The binomial distribution can be used in a case such as this because the variable is binary, that is, each individual will or will not carry the defective gene. If 4% of students are carriers of the gene, then $P = 0.04$ (*statz*$^-$) and $Q = 0.96$ (*statz*$^+$). To find the probability of finding some (i.e. one or more) carriers of the gene, the most common method is to calculate is the probability of no cases [i.e. $P(0)$] for a given sample size, e.g. 10. Using the binomial equation, if the number of carriers, r, is 0, and the number of trials, n, is 10, we can calculate the probability of testing 10 individuals and finding no carriers:

$$P(r) = \frac{n!}{r!(n-r)!} * p^r(1-p)^{n-r}$$
$$P(0) = \frac{10!}{0!(10-0)!} * 0.04^0(1-0.04)^{10-0}$$

(NB. Any number raised to the power 0 is 1 and any number raised to the power 1 is itself, e.g. $20^0 = 1$ and $20^1 = 20$, so $1! = 0! = 1$.)

$$P(0) = 1 * 1 * 0.96^{10}$$
$$P(0) = 0.67$$

Thus, if 4% of students are carriers, there is a 67% chance that a sample of 10 students will fail to find any carriers. This tells us a sample size of 10 is too small to give a reasonable chance of finding at least one carrier, so we need to test a larger sample of the population:

1. If the number of students tested is 20, $P(0) = 0.96^{20} = 0.44$, i.e. there is now a 56% chance of finding an affected carrier $(1 - 0.44 = 0.56)$.

2. If the number of students tested is 40, $P(0) = 0.96^{40} = 0.20$, i.e. a 80% chance of finding an affected carrier $(1 - 0.2 = 0.8)$, etc.

Of course, the lower the frequency of any characteristic in a population, the higher the probability of not finding any positive results in a small sample. For example, if only 1% of students are *statz*$^+$ there is only a 10% chance of finding a carrier in a sample of size 10, i.e. $P(0) = 0.99^{10} = 0.9$. This method is useful to determine the minimum sample number needed to obtain at least one positive result from a sample for any binary variable, for example to find at least one affected carrier in a random sample. For other types of variable which may be continuous and normally distributed, the usual method of determining sample sizes to use the standard deviation (Chapter 8).

9.5. Coincidences

When working with large numbers (populations), probability theory has some unexpected results. Many apparently unexpected coincidences are

merely the result of probability theory operating on very large populations, for example:

1. The chances of winning the UK National Lottery jackpot with a single ticket are about 14 million to one. What are the chances of someone who buys one Lotto ticket every week winning the jackpot twice within a year? Astronomical? Not necessarily. The chance of any single person (e.g. you) winning the jackpot twice within a year are approximately 10^{14} to one, but if 25 million people each buy one ticket every week, the chance of *anyone* winning the jackpot twice within a year are much greater – less than 100 to one.

2. What are the chances that someone else in a group of people has the same birthday as you?

$$P = 1 - (364/365 * 363/365 * 362/365 \ldots)$$

In a group of 22 people, there is a 50% chance that two people have the same birthday ($P = 0.5$). In a group of 120 people it is likely that someone else has the same birthday as you (work it out yourself).

3. In a large grassy field, the chances of putting your finger on a particular blade of grass are millions to one, but if you reach down and touch the ground, the chance of touching any blade of grass is nearly 100% ($P = 1$).

Why do 'coincidences' matter? They matter because, when you are trying to determine if an event is statistically significant or not, the seemingly logical 'expected' answer can be very misleading – events which might seem very unlikely to occur by chance can do precisely that if enough cases are involved. Consider a statistical analysis of whether banging your head against a hard surface can cure the common cold. Many studies of this problem are conducted, each with 95% confidence limit ($P = 0.05$). As soon as 20 studies have been performed, there will be, on average, at least one scientific paper published which proves that banging your head against a wall cures colds. Yet if you had a cold, what would you do?

Problems (answers in Appendix 1)

9.1. Cystic fibrosis (CF) is the most common recessive genetic disorder in Caucasians – approximately one person in 2500 carries one copy of the CF gene, which occurs with equal frequency in males and females. If a couple are both carriers of the CF gene and have a child, the following probabilities apply: normal child, non-carrier, $P = 0.25$; normal child, carrier, $P = 0.50$; child with cystic fibrosis, $P = 0.25$. What is the probability that the couple will have:

(a) Two children (either sex) who do not carry the CF gene?

(b) One son who is a carrier?

(c) Two daughters, one who is a carrier and one who has cystic fibrosis?

(d) Two daughters with cystic fibrosis?

9.2. In order to study great crested newt (*Triturus cristatus*) populations, 150 newts are harmlessly marked with a temporary non-toxic dye. Fifteen newts are then returned to each of 10 ponds known to contain this species. One week later, the ponds are fished again and, of 351 newts caught, 54 are marked.

(a) Estimate the total population of great crested newts in these 10 ponds.

(b) If one pond has a population of 107 newts (15 marked), what is the probability of catching marked (M) and unmarked (U) newts in this order: UUMUUUMU?

9.3. In a health survey, 19 of 60 men and 12 of 40 women are found to smoke cigarettes.

(a) What is the probability of a randomly selected individual being a male who smokes?

(b) What is the probability of a randomly selected individual smoking?

(c) What is the probability of a randomly selected male smoking?

(d) What is the probability that a randomly selected smoker is male?

9.4. The probability of being infected with HIV from each single exposure to one of the following events is approximately: unprotected sexual intercourse with an HIV carrier, 0.005; sharing an infected needle for intravenous drug use, 0.007; needlestick injuries in healthcare workers, 0.003. The cumulative probability of being infected $P(i)$ after n occurrences is given by the formula:

$$P(i) = 1 - (1 - k)^n$$

where k is the probability of being infected with HIV from each single exposure and $n =$ the number of occurrences. What is the probability of

being infected with HIV after:

(a) five occurrences of unprotected sexual intercourse with an HIV carrier;

(b) nine occurrences of sharing an infected needle for intravenous drug use;

(c) one needlestick injury in a healthcare worker who subsequently has unprotected sexual intercourse with an HIV carrier three times.

9.5. In a practical class, you are given three tubes of an enzyme (A B C) needed to perform an experiment you only have time to do once. A kind demonstrator has told you that only one of the tubes contains active enzyme – the other two are inactive. You choose tube A. To help you further, the demonstrator tells you that tube B contains inactive enzyme. Should you stick with tube A or switch to tube C for the experiment? Explain why.

10

Inferential Statistics

Learning Objectives:

On completing this chapter, you should understand:

- how to draw reliable conclusions about samples taken from larger populations;
- how to compare different populations;
- when to use various inferential statistical methods;
- when *not* to use particular inferential statistical methods.

10.1. Statistical inference

To infer means to conclude from evidence. Statistical inference allows the formation of conclusions about almost any parameter of a sample taken from a larger population, for example, are conclusions based on a sample valid for the whole population? It also allows the formation of conclusions about the difference between populations with regard to any given parameter. There are two methods of reaching these sorts of statistical inference, estimation and hypothesis testing.

Estimation

In estimation, a sample from a population is studied and an inference is made about the population based on the sample. The key to estimation is the probability with which particular values will occur during sampling; this allows the inference about the population to be made. The values which occur are inevitably based on the sampling distribution of the population. The key to making an accurate inference about a population therefore depends on random sampling, i.e. where each possible sample of the same size has the same probability of being selected from the population. In real life, it is often surprisingly difficult to take truly random samples from a population. Shortcuts are frequently taken, e.g. every third item on a list, 'expert' opinion, or simply taking the first *n* results obtained. Estimation is a relatively crude method of making population inferences. A much better method and the one which is normally used in statistical analysis is hypothesis testing.

Hypothesis testing

To answer a statistical question, the question is translated into a hypothesis – a statement which can be subjected to test. Depending on the result of the test, the hypothesis is accepted or rejected. The hypothesis tested is known as the null hypothesis (H_0). This must be in the form of a true/false statement. For every null hypothesis, there is an alternative hypothesis (H_A). Constructing and testing hypotheses is an important skill, but the best way to construct a hypothesis is not necessarily obvious:

1. If one of the two hypotheses is 'simpler' it is given priority so that a more 'complicated' theory is not adopted unless there is sufficient evidence against the simpler one (Occam's Razor: 'If there are two possible explanations always accept the simplest').

2. In general, it is 'simpler' to propose that there is no difference between two sets of results than to say that there is a difference.

3. The null hypothesis has priority and is not rejected unless there is strong statistical evidence against it.

The outcome of hypothesis testing is to 'reject H_0' or 'do not reject H_0'. If we conclude 'do not reject H_0', this does not necessarily mean that the null hypothesis is true, only that there is insufficient evidence against H_0 and in favour of H_A. Hypothesis testing never proves that the null hypothesis is true, just as rejecting the null hypothesis suggests but does not prove that the alternative hypothesis may be true.

In order to decide whether to accept or reject the null hypothesis, the level of significance (α) required of the result must be decided. In general terms:

$\alpha = 0.05$ – significant (confidence interval 95%, $P = 1 - 0.95 = 0.05$), most commonly used;

$\alpha = 0.01$ – highly significant (confidence interval 99%, $P = 1 - 0.99 = 0.01$), strong statistical evidence;

$\alpha = 0.001$ – very highly significant (confidence interval 99.9%, $P = 1 - 0.999 = 0.001$), rarely used.

The level of significance allows us to state whether or not there is a 'significant difference' (note that this is a technical term which should only be used in the correct context) between populations, that is, whether any difference between populations is a matter of chance, due to experimental error, or so small as to be unimportant.

10.2. Procedure for hypothesis testing

1. Define H_0 and H_A, based on the guidelines given above.

2. Choose a value for α. Note that this should be done before performing the test, not when looking at the result.

3. Calculate the value of the test statistic.

4. Compare the calculated value with a table of the critical values of the test statistic.

5. If the calculated value of the test statistic is *less than* the critical value from the table, accept the null hypothesis (H_0). Note that this does not mean that the null hypothesis has been conclusively proved, only that it has not been rejected.

6. If the calculated value of the test statistic is *greater than or equal to* the critical value from the table, reject the null hypothesis (H_0) and accept the alternative hypothesis (H_A).

Note that very small *P*-values (e.g. 0.001) do not signify large statistical differences, only that the observed differences are highly improbable given the null hypothesis tested. *P*-values indicate how sure you can be that there is a real difference, not the size of the difference. For example, a very small *P*-value can arise when any difference is tiny but the sample sizes very large. Conversely, a large *P*-value can arise when the effect is large but the sample size is small.

10.3. Standard scores (*z*-scores)

z-scores define the position of a score in relation to the mean using the standard deviation as a unit of measurement. They are therefore useful for comparing datapoints in different distributions.

$$z = (\text{score} - \text{mean})/\text{standard deviation}$$

The *z*-score is the number of standard deviations by which the score departs from the sample mean. Since this technique normalizes distributions, *z*-scores can be used to compare data from different sets, e.g. a student's performance on two different exams (e.g. did Joe Blogg's performance on test 1 and test 2 improve or decline?):

1. Joe B scored 71.2% on exam 1 (mean = 65.4%, SD = 3.55) $z = (71.2 - 65.4)/3.55 = 1.63$.

2. Joe B scored 66.8% on exam 2 (mean = 61.1%, SD = 2.54) $z = (66.8 - 61.1)/2.54 = 2.24$.

3. Conclusion – Joe B did better, compared with the rest of his classmates, on exam 2 than on exam 1, even though his mark was lower in the second exam.

Note that the *z*-score is a parametric statistic (Chapter 8), and is only meaningful when it refers to a normal distribution – calculating a *z*-score from a skewed dataset may not produce a meaningful number. Comparing *z*-scores for different distributions is also meaningless unless: the datasets being compared are as similar as possible (e.g. response to different doses of a drug under the same physiological conditions); and the shapes of the distributions being compared are as similar as possible.

10.4. Student's *t*-test (*t*-test)

Biological systems are complex, with many different interacting factors. To compensate for this, the most common experimental design in biology involves comparing experimental results with those obtained under control conditions. To interpret this type of experiment, we must be able to make objective decisions about the nature of any differences between the experimental and control results – is there a statistically significant difference or are the results due to experimental error or random chance (e.g. sampling error)? A frequently used test of statistical significance is Student's *t*-test (or simply *t*-test), devised by William Gosset ('Student') in 1908. The *t*-test is used to compare two groups and has two variants:

1. Paired *t*-test – used when each data point in one group corresponds to a matching data point in the other group.

2. Unpaired *t*-test – used whether or not the groups contain matching datapoints.

The *t*-test is a parametric test which assumes that the data analysed:

- Is continuous, interval data comprising a whole population or is sampled randomly from a larger population.

- Has a normal distribution (Chapter 8).

- If the sample size (n) is < 30, the variances (Chapter 8) of the two groups should be similar (*t*-tests can be used to compare groups with different variances if $n > 30$).

- The sample size should not differ hugely between the groups (e.g. < 50%).

If you use the *t*-test under other circumstances, the results may be misleading. In other situations, non-parametric tests should be used to compare the groups, for example, the Wilcoxon signed rank test for paired data and the Wilcoxon rank sum test or Mann–Whitney test for unpaired data (not covered in this book). The *t*-test can only be used to compare two groups. To compare three or more groups, other tests must be used, for example, analysis of variance between groups (ANOVA; see Section

10.5). In general though, the *t*-test is quite robust and produces approximately correct results in many circumstances.

The paired *t*-test is used to investigate the relationship between two groups where there is a meaningful one-to-one correspondence between the data points in one group and those in the other, for example a variable measured at the same time points under experimental and control conditions. It is not sufficient that the two groups simply have the same number of datapoints. The advantage of the paired *t*-test is that the formula procedure involved is fairly simple.

Procedure

1. Start with the hypothesis (H_0) that the mean of each group is equal, that is, there is no significant difference between the means of the two groups, e.g. control and experimental data. The alternative hypothesis (H_A) is therefore that the means of the groups are not equal. We test this by considering the variance (standard deviation) of each group.

2. Set a value for α (significance level, e.g. 0.05).

3. Calculate the difference for each pair (i.e. the variable measured at the same time point under experimental and controlled conditions).

4. Plot a histogram of the differences between data pairs to confirm that they are normally distributed – if not, stop.

5. Calculate the mean of all the differences between pairs (d_{av}) and the standard deviation of the differences (SD).

6. The value of *t* can then be calculated from the following formula:

$$t = \frac{d_{\mathrm{av}}}{\mathrm{SD}/\sqrt{N}}$$

where d_{av} is the mean difference, i.e. the sum of the differences of all the datapoints (set 1 point 1 – set 2 point 1, etc.) divided by the number of pairs; SD is the standard deviation of the differences between all the pairs; and *N* is the number of pairs. NB. The sign of *t* (+/–) does not matter; assume that *t* is positive.

7. The calculated value of *t* can then be looked up in a table of the *t* distribution (Appendix 3, or obtained from appropriate software). To do this, you need to know the 'degrees of freedom' (df) for the test. The result of any statistical test is influenced by the population size, for example it is more accurate to make 200 measurements than 20 measurements. Since the number of observations (population size) affects the value of statistics such as *t*, when we calculate or look up *t*, we need to take the population size into account – this is what degrees of freedom does. For a paired *t*-test:

$$\mathrm{df} = n - 1 \text{ (number of pairs} - 1)$$

To look up *t*, you also need to determine whether you are performing a one-tailed or two-tailed test. In any statistical test we can never be 100% sure that we have to reject (or accept) the null hypothesis. There is therefore the possibility of making an error as shown in Table 10.1.

Table 10.1 The possibility of making an error

		Null hypothesis	
		True	False
Decision	Reject	Type I error	Correct
	Accept	Correct	Type II error

Falsely rejecting a true null hypothesis is called a type I error. The probability of committing a type I error is always equal to the significance level of the test, α. Failure to reject a false null hypothesis is called a type II error. The 'power' of a statistical test refers to the probability of correctly claiming a significant result. As scientists are generally cautious, it is considered 'worse' to make a type I error than a type II error; we thus reduce the possibility of making a type I error by having a stringent rejection limit, 5% ($\alpha = 0.05$). However, as we reduce the possibility of making one type of error, we increase the possibility of making the other type. Whether you use a one- or two-tailed test depends on your testing hypothesis.

1. One-tailed test – used where there is some basis (e.g. previous experimental observation) to predict the direction of the difference, such as expectation of a significant difference between the groups. In some circumstances, one-tailed tests can be valuable, for example if it is proposed that a new drug is more effective in the treatment of a disease than an existing drug. The new drug should only be adopted if there is a significant improvement in treatment outcome.

2. Two-tailed test – used where there is no basis to assume that there may be a significant difference between the groups; this is the test most frequently used. The result of a two-tailed test does not tell you if any difference between groups is 'greater than' or 'less than', only that there is a significant difference.

They are called 'tails' because of the region of retention and regions of rejection on a graph of the distribution of the test statistic (Figure 10.1). Note that the alternative hypothesis states 'there is a difference'; it does not state why there is a difference or whether the difference between the two groups is 'greater than' or 'less than'. If the alternative hypothesis had specified the nature of the difference, this would have been a one-tailed hypothesis. However, if the alternative hypothesis does not specify the nature of the difference, we can accept either a reduction or an increase and it is therefore a two-tailed hypothesis. For a variety of reasons two-tailed hypotheses are safer than one-tailed. Statistical tables are sometimes tabulated only for one-tailed hypotheses. To convert them to two-tailed, double the value of α. A table of critical values of t for Student's t distribution is given in Appendix 3.

If the calculated value of t is greater than or equal to the critical value of the test statistic, the null hypothesis is rejected, that is, there is evidence of a statistically significant difference between the groups. If the calculated value of t is less than the critical value, the null hypothesis is accepted – there is no evidence of a statistically significant difference between the two groups.

The unpaired t-test does not require that the two groups be paired in any way, or even of equal sizes. A typical example might be comparing a variable in two experimental groups of patients, one treated with drug A and one treated with drug B. Such situations are common in medicine where an accepted treatment already exists and it would not be ethical to withhold this from a control group. Here, we wish to know if the differences between the groups are 'real' (statistically significant) or could have arisen by chance. The calculations involved in an unpaired t-test are slightly more complicated than for the paired test. Note that the unpaired t-test is equivalent to one-way ANOVA (Section 10.5), used to test for a difference in means between two groups.

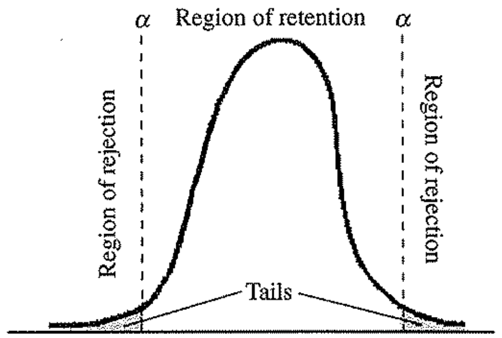

Figure 10.1 'Tails'

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{(SE_A)^2 + (SE_B)^2}}$$

where $\bar{X}$ is the mean of groups A and B, respectively, and $SE = SD/\sqrt{N}$.

For an unpaired t-test:

$$df = (nA + nB) - 2$$

where n is the number of values in the two groups being compared. Note that this is different from the calculation of the number of degrees of freedom for a paired t-test. Compare the calculated value of t with the critical value in a table of the t distribution (Appendix 3). Remember that the sign of t (+ / −) does not matter, and assume that t is positive. If the calculated value of t is greater than or equal to the critical value, the null hypothesis is rejected – there is evidence of a statistically significant difference between the groups. If the calculated value of t is less than the critical value, the null hypothesis is accepted – there is no evidence of a statistically significant difference between the groups.

Example

Consider the data from the following experiment. A total of 12 readings were taken, six under control and six under experimental conditions (Table 10.2). Before starting to do a t-test, several questions must be answered:

1. Are the datapoints for the control and experimental groups paired?

   No, they are just replicate observations, so we need to perform an unpaired t-test.

2. Are the data normally distributed?

Table 10.2 Experimental data

	Experimental, group A	Control, group B
	11.2	10.3
	13.1	12.6
	9.3	8.4
	10.2	9.3
	9.6	10.8
	9.8	8.9
Variance	1.68	1.96
SD	1.30	1.40
$SE = SD/\sqrt{N}$	0.53	0.57

Yes, approximately:

	Experimental	Control
Mean	10.53	10.05
Median	10.00	9.80

3. H_0: 'There is no difference between the populations of measurements from which samples have been drawn' (H_A: there is a difference).

4. Set the value of $\alpha = 0.05$ (i.e. a 95% confidence interval).

5. Are the variances of the two groups similar?

   Yes, approximately (1.68 vs 1.96).

6. Since all the requirements for a t-test have been met, we can proceed:

$$t = \frac{10.53 - 10.05}{\sqrt{(0.53)^2 + (0.57)^2}} = 0.62$$

7. Is this a one-tailed or a two-tailed test?

   Two-tailed, since we have no firm basis to assume the nature of any difference between the groups.

8. How many degrees of freedom are there?

$$df = (n_A - 1) + (n_B - 1) = 10$$

9. From the table of critical values of t (Appendix 3), we can see that for a two-tailed test with $df = 10$ and $\alpha = 0.05$, t would have to be 2.228 or greater for $> 5\%$ (0.05) of pairs of samples to differ by the observed amount.

10. Since $t_{calc} = 0.62$ and $t_{crit} = 2.228$, the null hypothesis is accepted. The conclusion is that there is no evidence of a statistically significant difference (at the 95% confidence level) between the experimental and the control groups in this experiment.

10.5. Analysis of variance (ANOVA)

Student's t-test can only be used for comparison of two groups. Although it is possible to perform many pair-wise comparisons to analyse all the possible combinations involving more than two groups, this is undesirable because it is tedious, but more importantly because it increases the possibility of type I errors (Section 10.4). However, ANOVA can compare two or more groups. ANOVA is a parametric test which assumes that:

1. The data analysed is continuous, interval data comprising a whole population or sampled randomly from a population.

2. The data has a normal distribution. Moderate departure from the normal distribution does not unduly disturb the outcome of ANOVA, especially as sample sizes increase, but highly skewed datasets result in inaccurate conclusions.

3. The groups are independent of each other.

4. The variances in the groups should be similar. For ANOVA, this is more important to accuracy that normal distribution of the data.

5. For two-way ANOVA, the sample size the groups is equal (for one-way ANOVA, sample sizes need not be equal, but should not differ hugely between the groups). This is because the results of ANOVA

tests can be upset by different variances in the groups, but this effect is minimized if the groups are of the same or similar sizes.

ANOVA tests come in various forms:

1. One-way (or one-factor) ANOVA – tests the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). Student's *t*-test is actually a particular application of one-way ANOVA (two groups compared) and results in the same conclusions.

2. Two-way (or two-factor) ANOVA – simultaneously tests the hypothesis that the means of two variables ('factors') from two or more groups are equal (drawn from populations with the same mean), for example the difference between a control and an experimental variable, or whether there is a difference between alcohol consumption and liver disease in several different countries. It does not include more than one sampling per group. This test allows comments to be made about the interaction between factors as well as between groups.

3. Repeated measures ANOVA – used when members of a random sample are measured under different conditions. As the sample is exposed to each condition, the measurement of the dependent variable is repeated. Using standard ANOVA is not appropriate because it fails to take into account correlation between the repeated measures, violating the assumption of independence. This approach can be used for several reasons, such as where research requires repeated measures, for example, longitudinal research which measures each sample member at each of several ages – age is a repeated factor.

The *F*-ratio ('Fisher ratio') compares the variance within sample groups ('inherent variance') with the variance between groups ('treatment effect') and is the basis for ANOVA:

$$F = \text{variance between groups/variance within sample groups}$$

ANOVA works by comparing the relationship between the variability within groups, across groups and the total. The actual ANOVA calculation itself is quite laborious and best performed using statistical software (Appendix 2). If you insist on knowing the equations involved, they can be looked up in a statistics textbook or software manual. This chapter will concentrate on how to use ANOVA. The basic procedure is similar to that

for performing a *t*-test:

1. Formulate the null hypothesis, i.e. that the means of the groups are equal.

2. Choose a confidence interval and set the significance level accordingly, e.g. CI = 95%, $\alpha = 0.05$.

3. Calculate the test statistic (F) (best done using software).

4. Compare the calculated value of F with a table of critical values of F.

5. If the calculated value of the F is less than the critical value from the table, accept the null hypothesis (H_0). If the calculated value of F is greater than or equal to the critical value from the table, reject the null hypothesis (H_0) and accept the alternative hypothesis (H_A).

Examples

Tables 10.3–10.5 show an example of one-way ANOVA. The null hypothesis is that there is no difference between the four groups being

Table 10.3 Experimental data

Pain Score for three analgesics			
Aspirin	Paracetemol (Acetaminophen)	Ibuprophen	Control (no drug)
5	4	4	5
4	4	4	5
5	3	5	5
3	4	3	4
5	5	3	5
5	3	5	5
4	4	3	5

Table 10.4 Summary

Groups	Count	Sum	Average	Variance
Aspirin	7	31	4.43	0.62
Paracetemol	7	27	3.86	0.48
Ibuprophen	7	27	3.86	0.81
Control (no drug)	7	34	4.86	0.14

Table 10.5 ANOVA

Source of variation	SS	df	F	F_{crit}
Between groups	4.96	3	3.23	3.01
Within groups	12.29	24		
Total	17.25	27		

compared. In this example, with a significance level of 95% ($\alpha = 0.05$), since the calculated value of F (3.23) is greater than F_{crit} (3.01), we reject the null hypothesis that the three drugs perform equally. The null hypothesis would have been rejected if even one of the groups differed significantly from the other three. A *post-hoc* comparison or series of individual pair-wise comparisons would have to be performed to determine which pair or pairs of means caused rejection of the null hypothesis, but since this was not part of the original question, we cannot address this directly here. If ANOVA is performed on three or more groups and it finds a significant difference, then a *post-hoc* test (also called pair-wise comparisons, multiple comparison tests, and multiple range tests) needs to be performed in order to make multiple comparisons between the groups. By comparing pairs of groups in every possible combination, the differences among them are revealed. There are various *post-hoc* tests which can be used, such as the 'Bonferonni', 'Scheffe', 'Tukey' and 'LSD' (least significant difference) tests. It is beyond the scope of this chapter to go into *post-hoc* tests, so you will need to consult other sources or software manuals if you are ever in a position to need such tests.

Table 10.6 Experimental results

Apple codling moth (Cydia pomonella) caught in pheromone traps		
	Bait 1	Bait 2
Orchard 1	19	20
	22	22
	19	18
	18	19
	20	19
	21	20
Orchard 2	22	21
	19	19
	19	18
	18	18
	20	20
	21	22

Table 10.7 Anova: two-factor without replication

Summary	Count	Sum	Average	Variance
Orchard 1	2	39	19.5	0.5
	2	44	22	0
	2	37	18.5	0.5
	2	37	18.5	0.5
	2	39	19.5	0.5
	2	41	20.5	0.5
Orchard 2	2	43	21.5	0.5
	2	38	19	0
	2	37	18.5	0.5
	2	36	18	0
	2	40	20	0
	2	43	21.5	0.5
Bait 1	12	238	19.83	1.97
Bait 2	12	236	19.67	2.06

Table 10.8 ANOVA

Source of variation	SS	df	F	F_{crit}
Rows	40.5	11	10.57	2.82
Columns	0.17	1	0.48	4.84
Error	3.83	11		
Total	44.5	23		

Tables 10.6–10.8 show an example of two-way ANOVA. As always, the null hypothesis is that there is no difference between the groups being compared. In this example, with a significance level of 95% ($\alpha = 0.05$), the calculated value of F (10.57) for the table rows (orchard 1 vs orchard 2) is greater than F_{crit} (2.82), so the hypothesis that there is no difference between the orchards is rejected. However, the calculated value of F (0.48) for the table columns (bait 1 vs bait 2) is less than F_{crit} (4.84), so the hypothesis that there is no difference between the pheromone baits is accepted. This example only compares two groups, so it is relatively easy to interpret the outcome.

10.6. χ^2-test

This is an example of a non-parametric test which, unlike Student's *t*-test and ANOVA, makes no assumptions about the distribution of the data. χ^2 (pronounced 'kye-squared') is used when data consists of nominal or ordinal variables rather than quantitative variables, when we are interested in how many members fall into given descriptive categories (not for quantitative measurements, such as weight, etc.).

The χ^2-test of independence asks 'Are two variables of interest independent (not related) or related (dependent)?' and deals with nominal and ordinal variable expressed as integers, that is, variables which fall into different, mutually exclusive categories. This is distinct from the *t*-test, which deals with interval variables, although the ANOVA test can also be performed on nominal data (Chapter 7). The χ^2-test investigates whether the proportions of certain categories are different in different groups. When the variables are independent, knowledge of one variable gives no information about the other variable. When they are dependent, knowledge of one variable is predictive of the value of the other variable, for example:

1. Is level of education related to level of income?
2. Is membership of a political party related to a person's preferred television station?
3. Is there a relationship between gender and examination performance?

The χ^2-test has two main uses: comparing the distribution of one category variable (nominal or ordinal) with another; and comparing an observed distribution with a theoretically expected one.

The expectation might be that the data would be normally distributed, or that particular attributes (e.g. treatment and disease) are independent, meaning there is no closer association than might be expected by chance. In the first case, a table of values for a normal distribution would be the source of the expected values. In the second, the expected values would be calculated assuming independence (random distribution). The χ^2-test is a non-parametric test which assumes that the data analysed:

1. Consist of nominal or ordinal variables.
2. Consist of entire populations or are randomly sampled from the population.

3. No single data point should be zero (if so, use Fisher's exact test; Section 10.7).
4. All the objects counted should be independent of one another.
5. Eighty per cent of the expected frequencies should be 5 or more (if not, try aggregating groups or use Fisher's exact test for small sample sizes; Section 10.7).

If you use the χ^2-test under other circumstances, the results may be misleading. The χ^2-test is by default one-tailed and can only be carried out on raw data (not percentages, proportions or other derived data). The basis of the χ^2-test is:

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

Note that acceptance or rejection of the null hypothesis can only be interpreted strictly in terms of the question asked, for example 'There is a difference between the observed and expected frequencies' or 'There is no difference between the groups' and not extrapolated to 'There is a difference between the observed and expected frequencies because . . .'.

Example A: comparing the distribution of one category variable with another

Of 120 male and 100 female applicants to a university, 90 male and 40 female had work experience. Does the gender of an applicant to university correspond to whether or not they have prior work experience?

The starting point for most χ^2 analyses of this type is to construct a contingency table, a table showing how the values of one variable are related to ('contingent on') the values of one or more other variables:

		Work experience		
		Yes	No	Total
Gender of applicant	Male	90	30	120
	Female	40	60	100
	Total	130	90	220

Next, formulate the null hypothesis (H_0): male and female applicants have equivalent work experience (H_A, male and female applicants have different work experience). Set a confidence interval, e.g. CI = 95%, so $\alpha = 0.05$. Calculate χ^2:

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

		Work experience		
		Yes	No	Total
Gender of applicant	Male	*a*	*b*	*a* + *b*
	Female	*c*	*d*	*c* + *d*
	Total	*a* + *c*	*b* + *d*	*n*

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{220(90 \times 60 - 30 \times 40)^2}{(90+30)(40+60)(90+40)(30+60)}$$

$$\chi^2 = \frac{220(5400 - 1200)^2}{120 \times 100 \times 130 \times 90}$$

$$\chi^2 = \frac{3\,880\,800\,000}{140\,400\,000}$$

$$\chi^2 = 27.64$$

As explained earlier, the distribution of χ^2 depends upon the number of degrees of freedom (df) in the test:

$$\text{df} = (\text{number of columns} - 1) * (\text{number of rows} - 1)$$

For the above test, df = (2 − 1) ∗ (2 − 1) = 1. Look up the calculated value of χ^2 in a table of critical values of the χ^2 distribution (Appendix 3). If the calculated value of χ^2 is greater than the critical value of χ^2 (from the table), reject H_0. If the calculated value of χ^2 is less than the critical value of χ^2 (from the table), accept H_0.

In this example, $\chi^2 = 27.64$, greater than the critical value for 1 df, so equivalent work experience. Note that, from the test result alone, we cannot say whether males or females have greater work experience, only that the two groups are not equal. In this example, it is fairly easy to work out which group has greater work experience by simply scrutinizing the table. The χ^2-test has simply proved that the difference between the two groups is statistically significant (at a 95% confidence level). Of course, the differences between groups are not always as clear-cut as in this example.

Alternative method: χ^2 calculation using observed and expected values

An alternative method of calculating χ^2 for the above example is to calculate the expected distributions assuming the null hypothesis to be true: 130 students out of a total of 220 had work experience. If the proportion of males and females with work experience were equivalent we would expect: males with experience = (130/200) ∗ 120 = 71. Table 10.9 shows the contingency table.

Table 10.9 Contingency table

	Observed (O)		Expected (E)		O − E		$(O - E)^2/E$	
	Yes	No	Yes	No	Yes	No	Yes	No
Male	90	30	71	49	19	− 19	5.1	7.4
Female	40	60	59	41	− 19	19	6.1	8.8
Total	130	90	130	90	0	0	11.2	16.2

$\chi^2 = 11.2 + 16.2 = 27.4$. From the table of critical values of χ^2, the calculated value is greater than the critical value, so the null hypothesis is rejected. The advantage of this method is that it can be applied to problems where there are more than two groups, for example:

1. Each of a group of 1350 students were immunized with one of five influenza vaccines under test. Is there any evidence that any one influenza vaccine is better than the others based on the numbers of students who developed influenza and those who did not?

2. We can produce a table with observed and expected values (not shown here). The overall χ^2 value will inform us whether there are differences

3. The sums of the $(O - E)^2/E$ for each vaccine will provide information about the contribution of each vaccine to the overall χ^2 – the vaccine contributing the most to the overall difference will have the largest $(O - E)^2/E$.

Example B: comparing an observed distribution with a theoretically expected one

Using the method of observed and expected values we can use the χ^2-test to compare an observed distribution with a theoretically expected one. For example, in a population of mice:

Colour	Observed	Expected from genetic theory
White	380	51%
Brown	330	40.8%
Black	74	8.2%

Do the proportions observed differ from those expected? Formulate the null hypothesis (H_0): the observed distribution does not differ from the expected distribution (H_A, the observed distribution differs from the expected distribution). Set a confidence interval, e.g. CI = 95%, so $\alpha = 0.05$. Table 10.10 shows the contingency table.

Table 10.10 Contingency table

Colour	Observed	Theoretical proportion	Expected	O – E	$(O - E)^2/E$
White	380	0.510	400 (0.510 * 784)	– 20	1.0
Brown	330	0.408	320 (0.408 * 784)	10	0.3125
Black	74	0.082	64 (0.082 * 784)	10	1.5625
Total	784	1.0	784	0	2.8750

Calculate $\chi^2 = 2.875$. Calculate df:

$$\text{df} = (\text{number of columns} - 1) * (\text{number of rows} - 1)$$

$$(\text{columns} = \text{observed, expected} = 2;\ \text{rows} = \text{white, brown, black} = 3)$$
$$= (2 - 1) * (3 - 1) = 1 * 2 = 2$$

From the table of critical values of χ^2 (Appendix 3), the calculated value of χ^2 is less than the critical value, so the null hypothesis is accepted.

Although the χ^2-test is, strictly speaking, non-parametric, it still has limitations. All the objects counted should be independent of one another, so the outcome of counting one should not influence the outcome of counting any of the others. Eighty per cent of the expected frequencies should be 5 or more. If this is not the case, it is sometimes possible to get around this difficulty by aggregating (combining) groups. Also, no single data point should be zero. This can present an insuperable problem. For datasets where many of the values are less than 5 or any are equal to 0, it is necessary to substitute Fisher's exact test for the χ^2-test (Section 10.7).

10.7. Fisher's exact test

Sir Ronald Aylmer Fisher (1890–1962) 'the father of modern statistics', developed the concept of likelihood:

> The likelihood of a parameter is proportional to the probability of the data and it gives a function which usually has a single maximum value, called the maximum likelihood.

He also contributed to the development of methods suitable for small samples and studied hypothesis testing. Fisher's exact test is an alternative to χ^2 for testing the hypothesis that there is a statistically significant difference between two groups. It has the advantage that it does not make any approximations (Fisher's exact test), and so is suitable for small sample sizes. Fisher's exact test is a non-parametric test which assumes that:

1. The data analysed consist of nominal or ordinal variables.

2. The data consist of entire populations or be randomly sampled from the population, as in all significance tests.

3. The value of the first unit sampled has no effect on the value of the second unit – independent observations. Pooling data from

before–after tests or matched samples would violate this assumption.

4. A given case may fall in only one class – mutual exclusivity.

The formula for calculating Fisher's exact test is not complex, but can be tedious. Where:

$$\begin{array}{ccc} a & b & r_1 \\ c & d & r_2 \\ c_1 & c_2 & n \end{array}$$

$$P = (r_1!\ r_2!\ c_1!\ c_2!)/n!\ a!\ b!\ c!\ d!$$

As long as the criteria for test have been met, you can perform Fisher's test using statistics software or one of the many online calculators (search the internet for 'Fisher's' 'exact' and 'calculator').

Problems (answers in Appendix 1)

10.1. The heights of a group of girls and a group of boys was measured. The frequency of measurements in both groups was found to have a normal distribution:

	Girls	Boys
Mean	1.25 m	1.29 m
Standard deviation	6 cm	5 cm

(a) Susan's height is 1.31 m. What is her *z*-score?

(b) Michael's height is 1.31 m. What is his *z*-score?

(c) Sally's *z*-score is − 1.2. Is she taller or shorter than the average for her group?

(d) True or false: the boys' *z*-scores are higher than the girls' *z*-scores (explain your answer).

(e) What percentage of boys are taller than 1.39 m?

10.2. A group of 12 patients with high blood pressure is treated with drug A for 3 months. At the end of the treatment period, their blood pressure is measured and treatment with drug B started. After a further 3 months, their blood pressure is measured again. Analyse the data from this trial using Student's *t*-test:

	Drug A	Drug B
Patient 1	189	186
Patient 2	181	181
Patient 3	175	179
Patient 4	186	189
Patient 5	179	175
Patient 6	191	189
Patient 7	180	183
Patient 8	183	181
Patient 9	183	186
Patient 10	189	190
Patient 11	176	176
Patient 12	186	183

(a) What sort of *t*-test should you perform to analyse these data?

(b) Should you use a one tailed or two-tailed test?

(c) How many degrees of freedom are there in this test?

(d) Is there a statistically significant difference at the 95% confidence level in the blood pressure of the patients after treatment with the two drugs?

10.3. In a study of the acidification of lakes, pH measurements were made of a series of lakes draining into two different rivers, A and B. Analyse the data from this trial using Student's *t*-test:

A		B	
6.97	7.20	5.93	6.70
5.88	7.81	4.88	6.81
6.41	6.98	5.71	6.18
6.85	7.42	5.85	6.42
6.24	5.59	5.24	4.59
6.26	6.77	7.86	6.77
5.01	5.84	4.01	5.24
7.64	8.41	6.64	7.31
6.40	6.59	7.20	6.29
6.72	7.10	6.32	6.10

(a) What sort of t-test should you perform to analyse these data?

(b) Should you use a one-tailed or two-tailed test?

(c) How many degrees of freedom are there in this test?

(d) Is there a statistically significant difference at the 95% confidence level in the pH readings of the lakes draining into the two rivers?

10.4. The number of eggs in robins' nests in three different areas of woodland were counted and found to be:

A: 2, 0, 1, 1, 1, 3, 1, 3, 2, 1, 1, 2, 2, 2, 1, 3, 3, 1, 2, 0, 1, 1, 1, 1, 0
B: 2, 1, 2, 0, 1, 5, 1, 2, 3, 2, 1, 2, 2, 2, 0, 3, 2, 0, 1, 1, 0, 1, 0, 0, 1
C: 2, 0, 2, 0, 2, 5, 1, 2, 2, 1, 0, 1, 3, 2, 3, 2, 1, 1, 0, 1, 2, 1, 1, 4, 2

Can you perform an ANOVA test to demonstrate whether or not there a statistically significant difference at the 95% confidence level between the three woodlands?

10.5. A biologist measures the preference of three-spined sticklebacks (*Gasterosteus aculeatus*) for various food items. In a 3 h period, fish of length less than 4 cm consumed 14 *Daphnia galeata*, 14 *Daphnia magna* and 36 *Daphnia pulex*, while fish longer than 4 cm consumed 6 *Daphnia galeata*, 24 *Daphnia magna* and 31 *Daphnia pulex*. Use the χ^2-test to compare the distribution of these variables and decide whether there is a statistically significant difference at the 95% confidence level between the feeding behaviour of the larger and the smaller sticklebacks.

(a) Construct a contingency table for the data.

(b) Formulate the null hypothesis for this experiment.

(c) How many degrees of freedom are there in this case?

(d) Calculate χ^2.

(e) Is there a statistically significant difference at the 95% confidence level between the feeding behaviour of the larger and the smaller sticklebacks?

10.6. A group of 353 cancer patients are treated with a new drug. Of the patients who receive this treatment, 229 survive for more than 5 years after the commencement of treatment. Compare this result with a control group of 529 similar patients treated with the previously accepted drug therapy, 310 of whom survive for more than 5 years after the commencement of treatment. Is there a statistically significant difference at the 95% confidence level between the survival rates of the patients who received the new drug

11

Correlation and Regression

Learning Objectives:

On completing this chapter, you should understand:

- the differences between correlation and regression;
- when to use each;
- the limitations of these tests.

11.1. Regression or correlation?

The correlation between two or more variables demonstrates the degree to which the variables are related. Linear regression demonstrates the relationship between selected values of X and observed values of Y, from which the most probable value of Y can be predicted for any value of X. Both correlation and regression are based on geometry and graphs and plots. Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations. Calculate linear correlation if:

- You measured both X and Y in each subject and wish to quantify how well they are associated.

- Do not calculate a correlation coefficient if you manipulated both variables, for example salt intake (in diet) and blood pressure (by drug