Study Guide 12: How to Choose a Statistical Test

Dr David Chinn,
Research & Development Office,
Queen Margaret Hospital, Dunfermline, Fife.
davidchinn@nhs.net 01383 623623 (ext 20943)
Alternative contact: Dr Amanda Wood amanda.wood3@nhs.net 01383 623623 (ext 20941)

Contents

Disclaimer

I am an epidemiologist, not a statistician. These notes are written from my experience of working in the field of medical research for over 40 years. I have sought to give what I hope is a clear and simple overview when choosing a statistical test. I do not profess to be an expert in statistics and a 'proper' statistician reading this guide may take issue with some of my explanations. Accordingly, I would encourage the reader to refer to one of the many excellent introductory books available on statistics for further guidance; some titles are given in the references and further reading.

(1) Overview and learning outcomes

This guide is directed at those with some basic knowledge of statistics who require a better understanding of which statistical test to use under different circumstances. This is a very basic account and a detailed description of each test is beyond the remit of this guide. We suggest you use this guide to identify the test you require and consult one of the many introductory texts on the subject. Alternatively, you could discuss your requirements with a statistician, preferably at the design stage. After reading this guide you should be able to:

- Cite the considerations needed when choosing a statistical test
- Describe different data types
- Cite the four common statistical inferences made in studies
- Be familiar with the properties and uses of common statistical distributions
- Know which statistical tests to use in which circumstances
- Be familiar with the limitations of the tests

Note: Examples on the use and interpretation of some commonly used tests are provided in the NHS Fife study guide ‘How to make sense of numbers’.

Associated NHS Fife study guides:
10 Introduction to medical statistics
11 How to calculate sample size and statistical power
13 How to make sense of numbers

(2) The purpose of statistics

Statistics is concerned with estimation and describing ‘uncertainty’ by measuring the variability within- and between-persons and the source and size of this variability. We use *descriptive statistics* to represent visually and numerically summaries of statistical information. Data can be summarised visually in bar charts, pie charts and histograms, and numerically as measures of central tendency (mean, median, mode) and of dispersion (Variance, Standard Deviation (SD), Standard Error (SE), Inter-Quartile Range (IQR)). Further details of these terms are covered in the NHS Fife study guide ‘An introduction to medical statistics’.

In comparison, we use *inferential statistics* to make generalisations, or inferences, between two or more target populations from which representative samples are drawn. Some characteristics are measured in the samples and probability estimates made to test hypotheses of, for example, equivalent values (a null hypothesis) using a test of statistical significance. The choice of which test to use depends mainly on (1) the type of data, (2) the inferences you wish to make and (3) the distribution of the data in the target population and in the sample drawn from it.

(3) Types of data

There are essentially two types of data and measurement scales.

(3.1) Categorical, also called Qualitative data (all or none)

Categorical data (data in categories) is mutually exclusive (you can only be in one category or another) and may be *nominal* or *ordinal* in character. Data are described as nominal if they cannot be ordered because there is no natural order. Examples include marital status, smoking habit, religion, eye colour, nationality and vital status (dead or alive).

Data are described as ordinal where there is a ranked order but the magnitude of the difference between adjacent categories is not identical. Examples include results from a road race. Runners are categorised as first, second, third etc. The winner may have run the race in 60 minutes, the person coming second may have taken 62 minutes (2 minutes behind), and the person in third place may have taken 75 minutes (13 minutes behind the person coming second). Another example is where patients are asked to report pain on a categorical scale of 'no pain', 'a little pain', 'a lot of pain', 'the worst imaginable pain'. We cannot assume the difference between being in 'no pain' and 'a little pain' is the same magnitude as the difference between being in 'a little pain' and 'a lot of pain'. Other examples of ordinal scales are the Borg scale (breathless scores) and those often used in patient satisfaction surveys.

(3.2) Interval, also called Quantitative data

Data are derived from a count, or a standard measurement, and have a frequency distribution. The numbers can be *discrete* or *continuous*. Discrete data are integers (whole numbers) where the magnitude of the difference between adjacent categories is identical (unlike ordinal data above). Examples include the number of children in a family (0, 1, 2, 3, 4, 5 etc), length of stay (in days), number of asthma attacks in a year, number of GP visits in a year, or the number of beds on a hospital ward. The difference between 3 and 4 beds (i.e. 1 bed between adjacent categories) is the same as the difference between 6 and 7 beds.

Continuous data include measures such as height, body mass (weight), haemoglobin level, and age which can take any value within a range. The data are measured in standard units, with clear meaning attached to the difference between measures whatever the magnitude of the measurement (for example, the difference between a body mass of 21-26 kg (i.e. 5kg) is the same size as that for a difference between 65-70 kg).

There are differences in interpretation, however, between discrete and continuous data. We can calculate the average number of children in a family (for example 2.50) and the average body mass (for example, 68.9 kg) but the numbers have different interpretations. It is possible for a person to have a body mass of exactly 68.9 kg but not possible for a family to have 2.5 children!

Another form of data is the *ratio* which is also measured in standard units but the scale has a true zero which represents a total absence of the variable (for example, time, length, volume, mass). Age, height, blood volume and body weight are examples of ratio variables where it is possible to state, for example, that one person is twice as old as another. This is not the case with non-ratio variables.

Continuous variables can be treated to generate ordinal or nominal data. For example, age can be reported as a continuous measure, or in age groups (ordinal data) or as a dichotomous variable such as 'young' versus 'old' (nominal data).

(4) Inferences to be made

The inferences depend on the research question. There are four main types of inference:

(1) The difference in a measurement between two groups. For example, the difference between healthy men and women in haemoglobin level (continuous data), or the prevalence of depression (categorical data).
(2) The evaluation of two or more interventions or treatments in a group of patients. For example, in menopausal women the evaluation of a drug on

diastolic blood pressure (continuous data), or of two methods of diagnosing osteoporosis (categorical data).

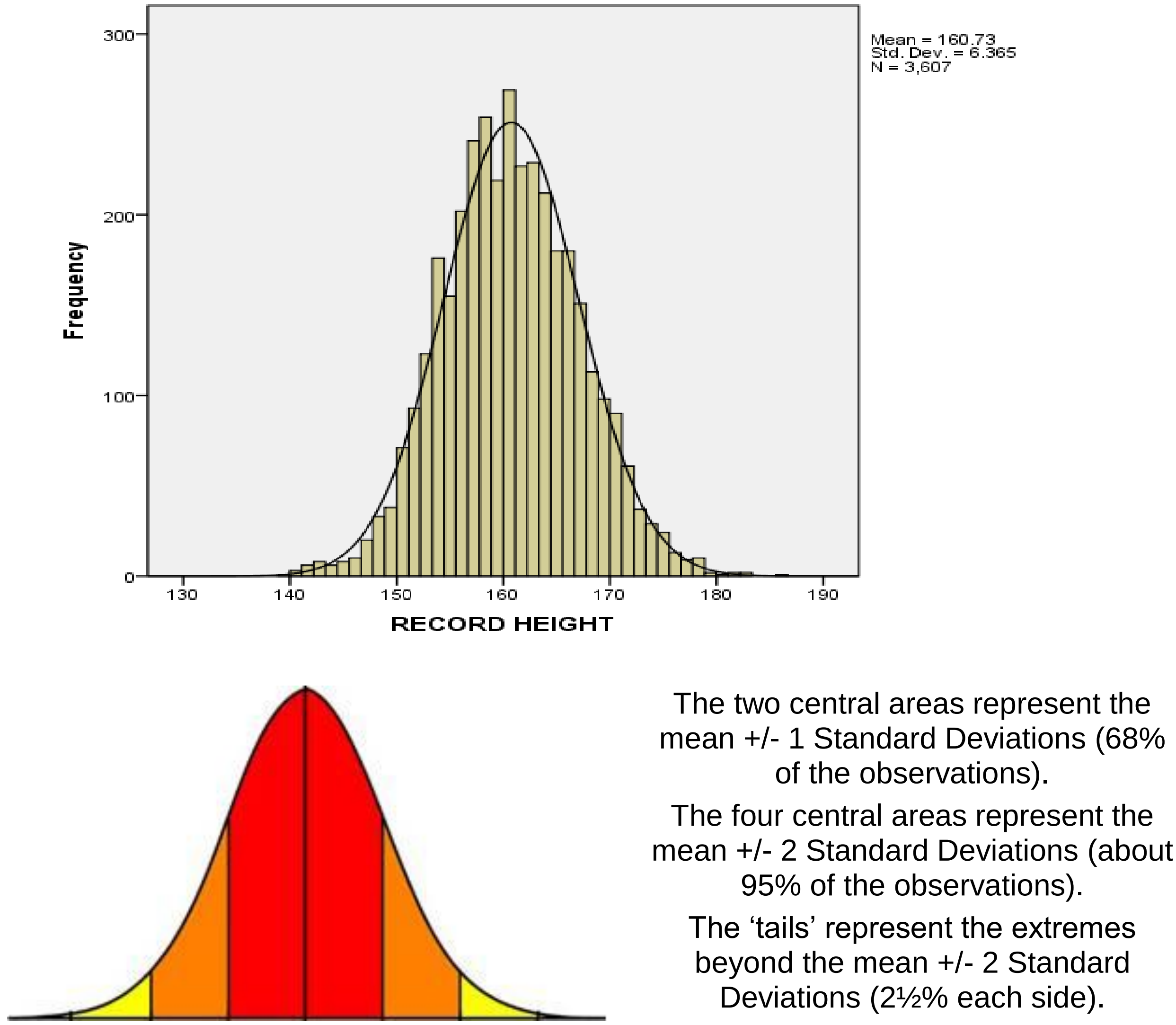
(3) The relationship between two variables. For example, in school age boys the relationship between lung capacity and body size (continuous data), or between school attended and achievement of an academic standard (categorical data).

(4) The trend in a variable of interest. For example, in patients with type I diabetes the survival pattern following diagnosis (continuous data), or the change in annual incidence of developing blindness (categorical data).

(5) Common distributions of data

(5.1) Normal or Gaussian distribution

Continuous data can be grouped into categories and the frequency of observations within each category plotted in a histogram which provides a useful visual summary, particularly of a large amount of data (Figure 1).



The two central areas represent the mean +/- 1 Standard Deviations (68% of the observations).

The four central areas represent the mean +/- 2 Standard Deviations (about 95% of the observations).

The 'tails' represent the extremes beyond the mean +/- 2 Standard Deviations (2½% each side).

Figure 1. Upper plot: the height of 3,607 adult women recorded in the Scottish Health Survey, 1998 with the 'Normal' curve superimposed.
Lower plot: the 'Normal' distribution and its properties

In Figure 1, the distribution is bell-shaped and is referred to as the Normal, or Gaussian distribution. The properties of the Normal distribution are:

(1) Distribution is bell-shaped,
(2) Distribution is symmetrical about the mean,
(3) The mean = median = mode,
(4) The distribution is characterised completely by two parameters (the mean and standard deviation):
   a. The mean +/- 1 standard deviations encompasses 68% of the observations,
   b. The mean +/- 2 standard deviations encompasses about 95% of the observations,
   c. 2½% of observations have a value = mean – 2 standard deviations,
   d. 2½% of observations have a value = mean + 2 standard deviations

(5.2) Binomial distribution

The binomial distribution is a theoretical distribution relevant for a trial when the outcome can take only one of two values (e.g. heads or tail with a coin toss, or success or failure with a treatment). The properties of the binomial distribution are used when making inferences about proportions. Consider a study to determine the effect on female fertility following treatment with a new drug. Women either conceive (a success) or not (a failure). In a study where each woman has the same probability of success the Binomial random variable is the observed number of conceptions (successes). The two parameters that describe the distribution are the number of women in the study (n) and the true probability of success for each woman (*p*). Then the number of expected women with a successful conception is n x *p* (also called the mean) and the variance is [n x *p* (1 – *p*)]. These values are used to calculate the confidence intervals associated with the chance of a successful conception. The calculations can be complex when the number of participating women is small. However, the calculations are simpler when the Binomial distribution approximates the Normal distribution which is the case when n x *p* and n(1-*p*) are greater than 5.

Example: Six women successfully conceived in a study of 30 women undergoing fertility treatment with a new drug. The proportion of successes was 6/30 = 0.2 and the 95% confidence interval (that is the range of proportions in which we are 95% confident that the *true* rate of success lies) is 0.05 to 0.34. For *p*=0.2 and n=30, n x *p* =6 and n(1-*p*)=24 so this is just within limits where it is safe to assume the Normal distribution applies. If either n x *p* or n(1-*p*) are less than 5 the calculations are much more complicated! In general, it is best practice to use a sample size large enough to assume the Normal distribution applies. This requires careful planning at the design stage by guessing what you think will be the likely proportion of successes achieved. For example, in this fertility study if you thought the likely success rate would be 0.1 (10%) then for n x *p* and n(1-*p*) to equal 5 or more you would need to recruit at least 50 women (50 x 0.1 = 5, and 50 x 0.9 = 45). In general, when planning such studies use of a small sample size will be associated with a very large confidence interval and you have to ask yourself whether the effort is worthwhile.

(5.3) Poisson distribution

The Poisson distribution arises from a simple probability distribution, just like the Binomial distribution but relates to the number of counts, or events that occur randomly and independently in a time interval. Examples include the number of admissions to hospital in a day. Unlike the Normal and Binomial distributions, the

Poisson distribution is characterised by only one parameter, the average *rate* observed from which the number of events per unit time can be calculated (referred to as the mean). Each value of a mean relates to a separate distribution so the Poisson is a family of distributions. We can calculate the probability of a certain number of admissions on any particular day by applying the distribution relevant to that mean. The calculations are complex but, for practical purposes the distribution obtained fits more closely with the Normal distribution when the mean is greater than 10. Assuming the Poisson distribution fits the Normal distribution simplifies the calculations when deriving the confidence intervals.

The Poisson distribution is used, for example, in mortality studies in a population where deaths occur randomly and independently of one another over the course of a year. The distribution obtained is based on the number of deaths observed (the mean) and the Poisson distribution can be used to compare deaths between different subsets of the same population, or between different time periods.

Example: A study was undertaken of deaths from cirrhosis of the liver in male qualified medical practitioners in England and Wales. There were 14 deaths when only 4.49 would have been expected from the age-specific rates estimated from the general male population. The standardised mortality ratio (that is the number of deaths observed divided by the number expected multiplied by 100) is 311, so about 3 times more deaths than expected. We can calculate the 95% confidence interval of this estimate by using the Poisson distribution and assuming the deaths occur randomly and are independent of one another. Because the number of deaths observed is 14 (i.e. greater than 10) we can assume the Poisson distribution associated with this number approximates the Normal distribution. The *approximate* 95% confidence interval is then calculated using estimates from the Normal distribution which yields values for the standardised mortality ratio of 148 to 474. This interval does not include 100 (the expected value if the null hypothesis were true) so the difference is statistically significantly different from 100 at the 5% level. Hence, the high mortality amongst doctors cannot be ascribed to chance. If the number of observed deaths had been less than 10 the calculations would have been more complex and beyond the description here.
Source: An Introduction to Medical Statistics. 2nd ed. Martin Bland, 1995, Oxford Medical Publications.

(6) Parametric and non-parametric distributions (interval data)

Many interval measurements in medicine conform to a bell-shaped (Normal) distribution. Examples of variables which have a normal (or approximately normal) distribution are heights in adulthood (see Figure 1), blood pressure, haemoglobin concentration and lung capacity in *healthy* people. These are referred to as *parametric* distributions. However, some variables do not fit with a Normal distribution. These include length of stay (Figure 2) which is skewed to the right (positive skew) and gestational age at birth (Figure 3) which is skewed to the left (negative skew). These are referred to as *non-parametric* distributions.

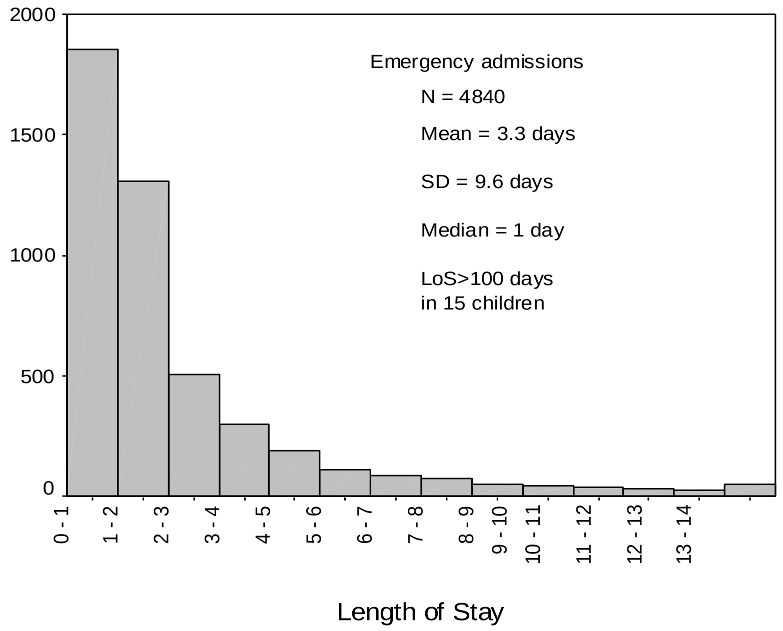


Figure 2. Length of stay (days) in 4840 children admitted to hospitals in one Hospital Trust over 3 years. An example of a positively skewed distribution.

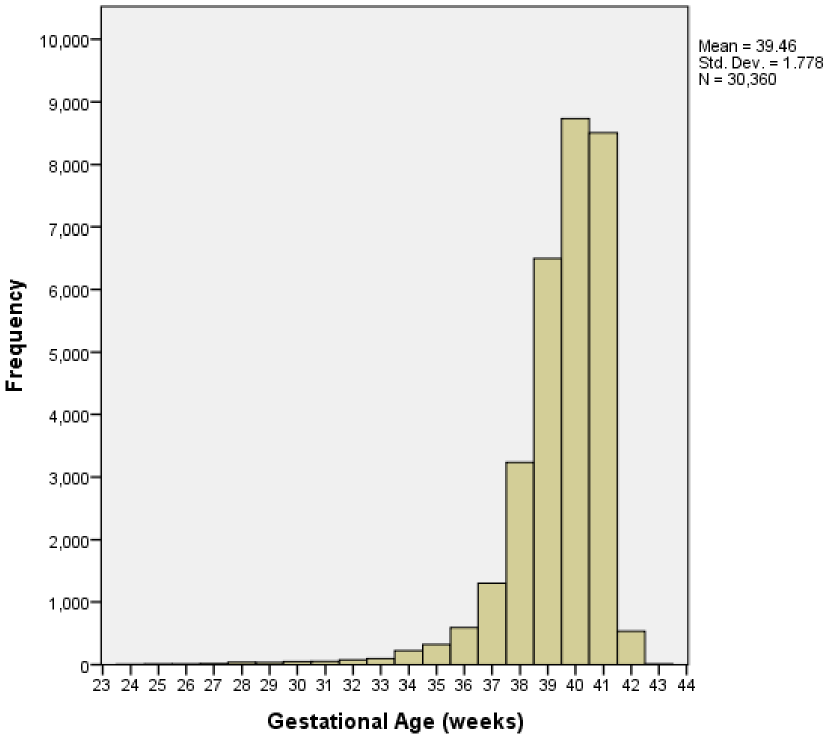


Figure 3. Gestational age in 30,360 births in Fife, 2003 - 2012. An example of a negatively skewed distribution.

The shape of the distribution (parametric or non-parametric) influences the choice of statistical test used for analysing interval data. The t-test (a parametric test) is commonly used to compare the means of two samples. The assumptions underlying the use of the t-test are:

1) the data come from a Normal distribution
2) the samples are not too small
3) the samples do not contain outliers (particularly a problem for small samples)

4) For comparison of 2 samples:
   a) the samples are of equal or nearly equal size
   b) the variances are equal or approximately so (but not critical as the calculations can allow for any marked difference in the spread of data)

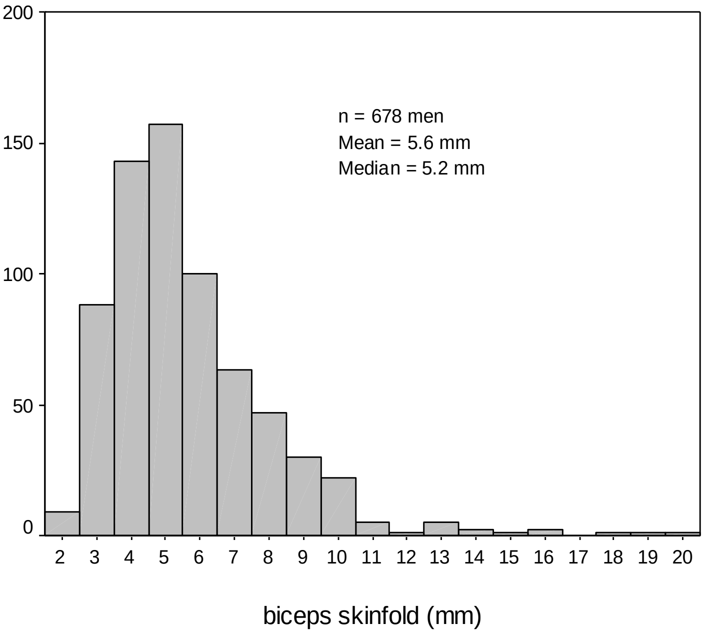
If these conditions are not met you will need to either (i) transform the data so that it does conform to the Normal distribution or (ii) use non-parametric tests which make no assumptions about the distribution of the data.

(6.1) Transforming data

To use a t-test on data that are not normally distributed (skewed either positively or negatively) you must *transform* the data by subjecting it to a mathematical function so that it does fit a Normal, or approximately Normal distribution. Data can be transformed by taking logs or calculating the reciprocal (1/x) for positively skewed, and square (x^2) or cube (x^3) functions for negatively skewed data (but watch out for zero and negative values where the transformation used can lead to errors in outcomes). Using a t-test on data that are not normally distributed can lead to the wrong conclusions (see section 14, page 16).

The decision whether to transform the data and use a parametric test depends on the question you are asking. For example, if you want to know by how much the means of two groups differ then you will need to transform the data and use a parametric test such as the t-test which allows you to calculate a confidence interval for the difference in means. However, if you merely want to know if the distribution of the data from the two groups differ significantly (yes or no) then you can use a *non-parametric* test which makes no assumptions about the distribution of the data. It is possible to calculate a confidence interval but the calculations are very complex and not routinely made by statistical packages. The non-parametric equivalents to the t-test are the Mann-Whitney U Test and the Wilcoxon rank sum test.

An example of a transformation is the biceps skin fold data in which the data are log transformed (= Log to base 10 of the biceps skinfold) to change the shape of the distribution to fit more closely with that for a Normal distribution (Figure 4).



Biceps skinfold (mm)

n	678	
Mean	5.6	
Median	5.2	
SD	2.3	
Minimum	2.2	
Maximum	20.4	
Percentiles:		
	25%	4.1
	50%	5.2
	75%	6.8
IQR 4.1 to 6.8		

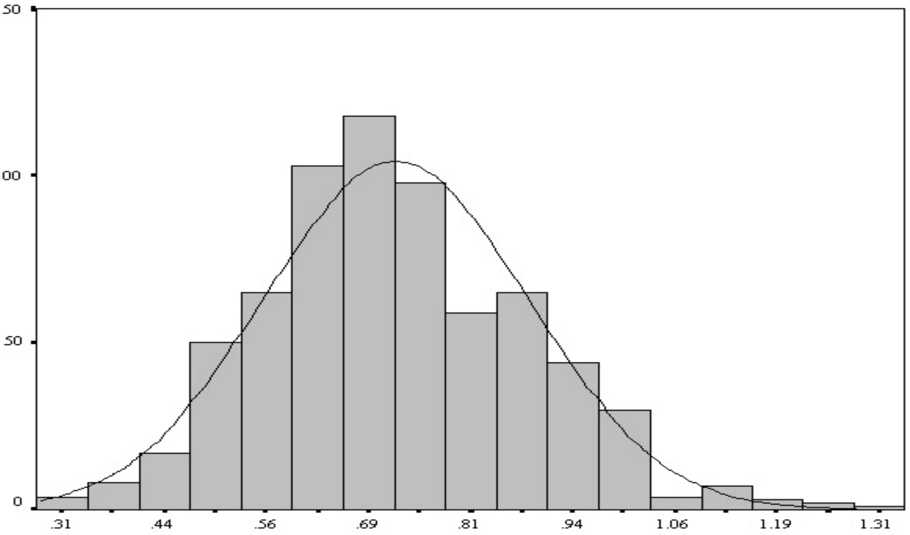


Figure 4. An example of a skewed distribution: the biceps skin fold (mm) and its log transformation (lower plot)

If you are unable to plot the data as a histogram a simple measure of skewness can be made from the summary data using the *Pearson's skewness coefficient*.

Pearson's skewness coefficient = (mean – median) / standard deviation …….(eqn 1)

For a normal distribution the mean equals the median so a value of 0 indicates the variable is normally distributed. A coefficient greater than 0.2 or less than -0.2 indicates the distribution is skewed. The coefficient calculated for the data in Figure 2 is 0.24 and that for Figure 4 (untransformed data) is 0.17. The data in Figure 2 would have to be analysed using non-parametric tests as it is unlikely that a suitable transformation would convert it to a normal distribution. The untransformed data in Figure 4 could be analysed using parametric techniques but it would be best to transform it to ensure the data fully met the assumptions for using these techniques.

(7) Assessing the difference between two independent groups

The choice of test depends on the type of data and on its distribution, if interval (Table 1).

Table 1. The choice of statistical test when comparing the difference between two groups

Type of data	Distribution Normal or approximately so	Size of sample	Statistical test
Interval, continuous	yes	≥ 30 each sample	*z*-test (Normal distribution for means) or t-test (unmatched) *
			F test or Levene's test for comparing variances **
	yes	< 30	t-test (unmatched) *
			F test or Levene's test for comparing variances
	no	any	Mann-Whitney U test or Wilcoxon rank sum test
Interval, discrete	-	any	Kolmogorov-Smirnov 2-sample test
Ordinal	-	any	Mann-Whitney U test or Wilcoxon rank sum test
Nominal	-	Large, all with expected frequencies >5	Chi-square test, odds ratio
		Small, at least one with an expected frequency of <5	Chi-square test with Yates' correction or Fischer's Exact test

* If other parametric assumptions met, otherwise use non-parametric equivalent.
** Levene's test is more robust than the F-test to departures of normality.

(8) Assessing the differences between more than two independent groups

A one-way Analysis of Variance (ANOVA) is used for assessing differences between the means of three or more groups of continuous data that are normally distributed or approximately so (Table 2). The null hypothesis is that there is no difference in the means of the different groups. The analysis will identify if the means do differ but will not identify which group or groups are significantly different from the others. For this you need to run post-hoc tests such as Duncan's Multiple Range tests. Further discussion on this application is beyond the remit of this simple guide. A one-way ANOVA applied to two groups is the equivalent of the unmatched (independent groups) t-test.

Table 2. The choice of statistical test when comparing the difference between more than two independent groups

Type of data	Distribution Normal or approximately so?	Statistical test
Interval, continuous	yes	ANOVA *
		F test or Levene's test for comparing variances **
	no	Kruskal-Wallis test
Interval, discrete	-	Kruskal-Wallis test
Ordinal	-	Kruskal-Wallis test
Nominal	-	Chi-square test, ***
		Chi-square test for trend

* If other parametric assumptions met, otherwise use non-parametric equivalent

** Levene's test is more robust than the F-test to departures of normality.

*** Valid if the sample size is large and >80% of cells have an expected frequency of >5. If the sample size is small and >20% of cells have an expected frequency of <5 then use the Chi-square test but reduce the number of categories by collapsing cells or excluding categories.

(9) Paired samples

In some studies interval data are collected as pairs of observations, for example, a measurement of lung capacity before and after administration of an inhaled drug. In this case the null hypothesis states the mean difference in *change* in lung capacity is zero. The difference between the pre- and post-drug results is calculated for each participant and the mean change and its standard error compared with zero. When the *differences* are normally distributed we use the paired t-test and when they are not we use the equivalent non-parametric test, the Wilcoxon matched pairs signed rank test. For categorical data the appropriate non-parametric tests are the Sign test and McNemar's test (Table 3).

Table 3. The choice of statistical test for paired samples

Type of data	Distribution Normal or approximately so?	Size of sample	Statistical test
Interval, continuous	yes	≥ 30 paired observations	*z*-test (Normal distribution for means) or t-test (paired)
	yes	< 30	t-test (paired)
	no	any	Sign test or Wilcoxon matched pairs signed rank test
Ordinal	-	any	Sign test
Nominal	-	any	McNemar's test

(10) Assessing the relationship between two variables

The relationship between two variables can be described via a correlation analysis which will measure the strength of association (correlation) between the two variables and by regression analysis which is more informative in describing the actual relationship in numerical terms. The Pearson correlation coefficient describes linear relationships and is best used when both variables are normally distributed. The non-parametric equivalent of Pearson's correlation is Spearman's Rank correlation which is best used when both variables are not normally distributed, or at least one variable is measured on an ordinal scale, or the sample size is small.

In a simple regression analysis one continuous variable (called the independent variable, the explanatory variable or a covariate) is used to predict another (called the dependent variable). Two or more independent variables (continuous, discrete, ordinal or nominal) can be used to predict the dependent variable when the process is known as multiple regression or analysis of covariance.

When the dependent variable is a binary variable (takes one of two values, for example, dead or alive) the appropriate regression technique is logistic regression.

(11) Assessing the trend in a variable of interest.

Trends over time can be assessed using, for example, survival analyses. Survival curves of patients diagnosed with a particular condition can be compared using the Kaplan-Meier curve. A formal comparison of survival curves from two or more groups can be made using the Log-rank test (Figure 5).

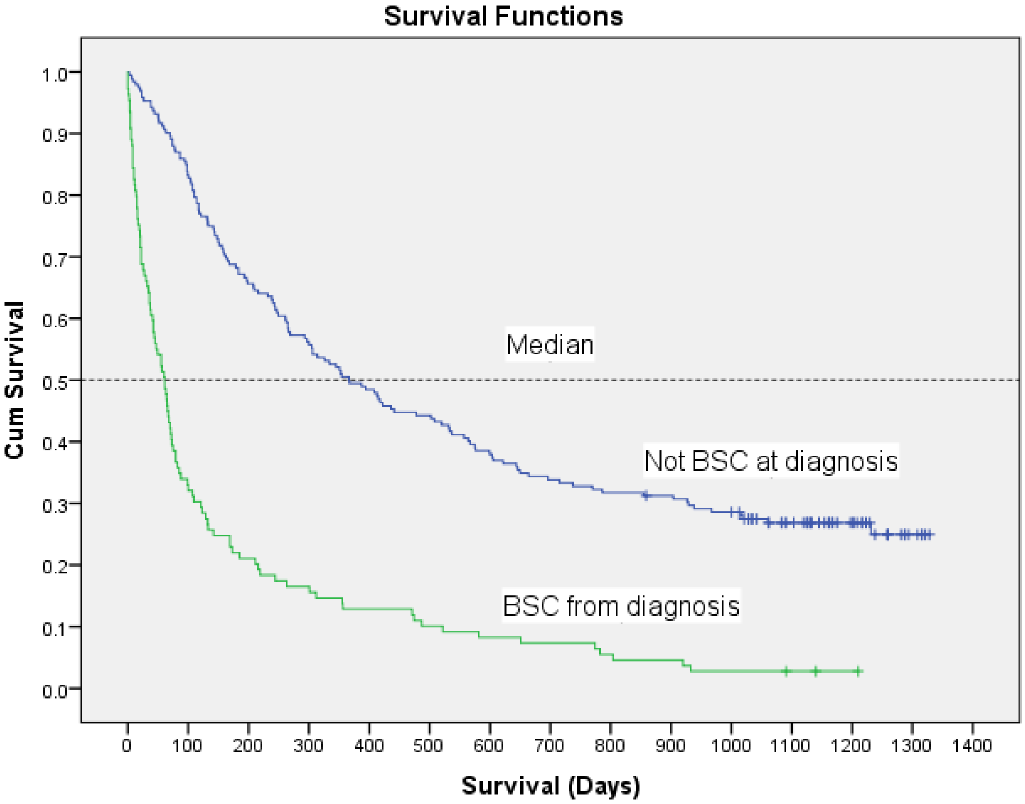


Figure 5. Survival curves in 301 patients with lung cancer.
BSC = Best Supportive Care

At diagnosis 192 patients had cancers that were treatable and 109 patients had advanced disease (untreatable) and were referred for ‘Best Supportive Care’ (BSC). Median survival in those with treatable cancers was 366 days (95%CI 274 – 458) and in those considered BSC was 61 days (95% CI 43 – 79). The Log-rank test was highly significant (P<0.00001).

The Cox proportional hazards regression model should be used to adjust survival estimates for factors known, or suspected of influencing survival including prognostic factors such as age, disease stage, treatments etc.

(12) Assessing agreement

The agreement between two instruments, or two observers measuring the same thing, or the repeatability of a single observer making duplicate measurements under identical conditions can be assessed using Cohen’s Kappa, if the data are categorical, or by the paired t-test, if the data are interval and the *differences* between repeated measures are normal distributed. If the differences are not normally distributed you should consider using the non-parametric equivalent of the paired t-test, namely the Sign test or the Wilcoxon matched pairs signed rank test. An additional procedure to compare the repeatability of interval data is to calculate the Intraclass Correlation Coefficient (ICC), which is similar to the Pearson correlation coefficient. However, you need to be cautious in interpreting a Pearson correlation coefficient obtained from repeated measures and a better approach is to plot the difference between repeated measures against the mean of the two measures (known as a ‘Bland-Altman plot’). A practical example showing the correct (and incorrect) method of assessing repeatability is given in the NHS Fife Study Guide ‘How to make sense of numbers’.

(13) Summary of when to use the tests and their limitations

General rules:

- Use parametric techniques on interval data taken from a population in which it is known that the data are normally distributed.
- Use non-parametric techniques on interval data when the distribution from which the data came is unknown or if there is reason to suspect that the data are not normally distributed.
- Use non-parametric techniques when the data are nominal or ordinal.

(13.1) Parametric tests

The z-test

Use to compare the means of two groups when the data are interval, parametric assumptions are met and the samples are large in size (more than 30).

The unmatched t-test (for unrelated or independent data)

Use to compare the means of two independent groups when data are interval and parametric assumptions are met (*see* Section 6 above). In reality the t-test is robust to small departures from a normal distribution and to small differences in variance between the two samples. The variances can be compared with the F-test or

Levene's test when the data are only approximately normally distributed. The F-test calculates the ratio of the largest to the smallest variance (F = Variance1 / Variance 2, where variance 1 is greater than variance 2). When both variances are equal the ratio is 1 which indicates the spread (width) of the two distributions is equal. As the width of one distribution increases compared with the other the ratio of variances will increase. It is safe to use the t-test even when the ratio of variances is up to about 4. When the widths of the distributions do differ materially the t-test can still be used though special adjustments are made to the calculation of results. Statistical packages such as SPSS will report the F-test (or Levene's test for comparing variances) and both results for the t-test, one assuming the variances do not differ significantly and the other assuming they do.

The Paired t-test (for related data)

Use to compare the mean difference in paired, continuous measurements when parametric assumptions are met.

ANOVA for unrelated data (one-way)

Use to compare the means of three or more groups when data are continuous and parametric assumptions are met (both that for normality of the distributions and equality of variance).

Limitations: ANOVA is fairly robust to moderate departures from normality but less so to unequal variances. Levene's test can be used to compare the variances of the different groups. The analysis will identify if the means do differ but will not identify which group or groups are significantly different from the others.

Pearson correlation coefficient

Use to determine the <u>linear</u> association between two continuous variables that are normally distributed (but not critical).

Limitations: The Pearson correlation coefficient should not be used if the relationship is non-linear, in the presence of outliers, when the variables are measured over more than one distinct group, when one of the variables is fixed in advance and <u>never</u> for assessing agreement between observers (or techniques).

(13.2) Non-parametric tests

The Chi-square test (χ^2)

Use to compare two or more proportions for nominal data using actual counts.

Limitations: In a 2 x 2 table (comparing two groups) the <u>expected</u> frequencies in all four cells should be at least 5. Otherwise, use Fischer's Exact test. When there are more than two categories for one or both groups being compared at least 80% of the expected frequencies should be at least 5. If the condition is not met some categories should be combined until the condition is met. No cell should be 'empty' (i.e. contain zero observations).

The Chi-square test (χ^2) *test for trend*

Use to analyse categorical data when one of the variables has multiple categories that are ordered (e.g. by age groups, or time periods).

The Mann-Whitney U test

Use to compare the distributions of two unmatched (independent) groups when data are interval and parametric assumptions are not met, or when data are ordinal (rank ordered).

Limitations: the test assumes the distribution of one or both samples is skewed. The test should not be used to compare the median of a positively skewed distribution with that from a negatively skewed distribution. It may not always be possible to check this if the sample sizes are small but if it is known that the distributions of the variable in the two populations from which the samples are drawn are essentially different then the test should not be used. In these circumstances the data from one sample may be transformed and reshaped to ensure the two distributions do match, at least approximately so.

The Wilcoxon rank sum test

Use as for the Mann-Whitney U test which is an equivalent.

The Wilcoxon matched pairs signed rank test

Use to compare differences in the medians of paired, continuous measurements when parametric assumptions are not met, or to ordinal data. It is also called the Wilcoxon signed ranks test in some books.

The Kruskal-Wallis test

Use to compare the distributions of three or more independent groups when data are continuous and parametric assumptions are not met, or when data are ordinal (rank ordered).

Limitations: if only three samples there must be at least 5 observations in each sample. The test will identify if the distributions differ significantly between the groups but will not identify which groups differ from each other.

The Sign test

Use to compare the median of paired observations (interval or ordinal).

Limitations: the sign test is a simple test based only on the sign of differences between paired observations. A more powerful test is the Wilcoxon matched pairs signed rank test (also called the Wilcoxon signed ranks test) which takes into account not just the sign of the difference but also its rank.

McNemar's test

Use for nominal data when analysing pairs of observations that are dichotomous (present or absent).

Limitations: the test is based on the number of discordant pairs of which there should be at least 10. If there are less than 10 the test statistic should be calculated using exact binomial probabilities taken from the binomial distribution.

Spearman's rank correlation coefficient (Spearman's Rho)

Use to determine the association between two continuous variables that are not normally distributed, or with ordinal data.

Limitations: there should be at least 7 pairs of observations.

The Kolmogorov-Smirnov 2-sample test

Use to compare two frequency distributions of discrete data.

(14) An example of the consequences of using the wrong test

Blood loss during surgery was compared using two different surgical procedures for the same operation. The researchers collected data on 319 operations. Initially they compared the mean blood loss using an independent samples unmatched t-test and concluded that blood loss was not significantly different between the two surgical techniques (P=0.108). However, the data did not conform to a normal (bell-shaped) distribution so the t-test (a parametric test which assumes the data are normally distributed) was inappropriate. A Mann-Whitney U test, a non-parametric test which makes no assumptions about the distribution of the data suggested that the distribution of blood loss <u>was</u> significantly different between the two techniques (P=0.002). Use of the wrong statistical test had resulted in an incorrect inference (interpretation). The summary statistics are shown in Table 4.

Table 4 Blood loss during surgery comparing two surgical techniques

Blood loss (ml)	Technique 1	Technique 2	Statistical test	
n	274	45		
Mean	640	789		
Median	500	700	t-test	Mann-
SD	589	444		Whitney U
Minimum	100	250	P = 0.108	
Maximum	7500	2500		P = 0.002
25th percentile	400	500		
75th percentile	750	1000		

In a normal distribution the mean and median are identical, or at least very similar. This was not the case from the summary statistics which suggested the distribution of blood loss was skewed. The mean was greater than the median so the distribution was likely to be positively skewed as in Figure 4 above. In addition, the standard deviation (SD) was large in relation to the mean. Another 'rule' is that for data that can only take positive values (such as blood loss) the data are likely to be skewed when the mean minus 2 x SD is less than zero, which was the case in this example. In addition, the Pearson's skewness coefficient (Page 9) was 0.24 for technique 1 and 0.20 for technique 2. A histogram of the data confirmed the skewness of the distribution (Figure 6).

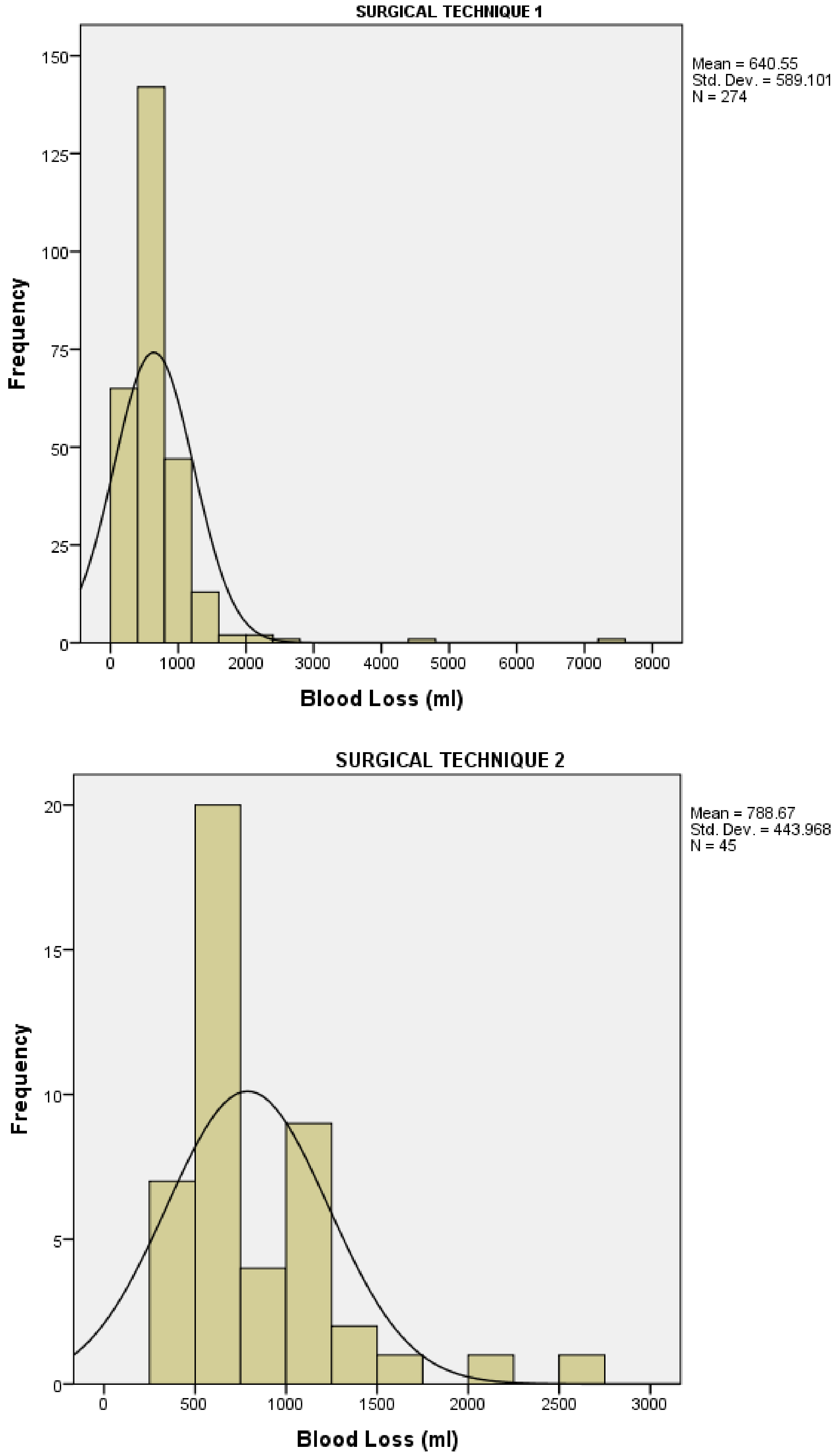


Figure 6. The distribution of blood loss in two surgical techniques

(15) Exercises

When planning a study it is important to also plan the data analysis. The choice of a statistical test must be made based on the type of data (categorical (*nominal, ordinal*), continuous (*discrete, interval*)), the inferences you wish to make (that is the comparison of interest such as comparing groups, testing relationships) and, for continuous data its distribution (Normal or non-Normal).

(15.1) A study was being planned to compare the prevalence of stress incontinence in 80 men and 80 women with chronic obstructive pulmonary disease. Which of the following statistical tests would be appropriate to test the null hypothesis that the prevalence was the same in both groups?
(a) The paired t-test?
(b) The independent samples (unmatched) t-test?
(c) The Mann-Whitney U test?
(d) The Chi-square test?
(e) Fisher's exact test?

(15.2) GP data are to be used to compare the age of onset of asthma between adult males and females. Which of the following statistical tests would be appropriate to test the null hypothesis that the age of onset was the same in both groups?
(a) McNemar's test?
(b) The independent samples (unmatched) t-test?
(c) The Mann-Whitney U test?
(d) The Chi-square test?
(e) The Wilcoxon rank sum test?
(f) One-way analysis of variance?

(15.3) A study is being planned to investigate the mean number of days lost from school in children with asthma over the course of a year. For each child with asthma the researchers will select an age- and gender-matched control child free of asthma. It is envisaged that the raw data on days lost will not fit a normal distribution. Which of the following tests would be appropriate to test the null hypothesis that the difference in mean days lost is zero?
(a) The paired t-test?
(b) The independent samples t-test?
(c) The Mann-Whitney U test?
(d) The Wilcoxon matched pairs signed rank test?
(e) The sign test?
(f) McNemar's test?

(15.4) In the planned study in 15.3 the researchers also wish to investigate if the children with asthma and their matched controls are exposed to secondary cigarette smoke in their homes. Which of the following tests would be appropriate to test the null hypothesis of no difference in exposure to second-hand smoke between children with asthma and their matched controls?
(a) The paired t-test?
(b) The independent samples t-test?
(c) The Mann-Whitney U test?
(d) The Wilcoxon matched pairs signed rank test?
(e) McNemar's test?

(15.5) Researchers wish to investigate if there is an association between lung capacity and height in women. Lung capacity will be measured in a random sample of 1000 women aged between 20 and 70 years who are free of respiratory symptoms. Which is the appropriate approach to analysing these data and which statistical measure could be used to assess the level of association? What additional factors may the researchers have to take into account?

(16) References and further reading

A-Z of Medical Statistics. Pereira Maxwell F. 1998, Arnold.

An Introduction to Medical Statistics. 3rd ed. Martin Bland, 2000, Oxford Medical Publications (see Chapter 14)

Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999:8; 135-160. doi: 10.1177/096228029900800204. http://smm.sagepub.com/cgi/content/abstract/8/2/135

Essential Medical Statistics. 2nd ed. Betty Kirkwood & Jonathan Sterne, 2003, Blackwell Scientific Publications.

Interpreting statistical findings. A guide for health professional and students. Walker J, Almond P. 2010. Open University Press.

Medical Statistics at a Glance. 3rd ed. Aviva Petrie & Caroline Sabin, 2009, Blackwell Publishing.

Practical Statistics for Medical Research. Douglas G Altman, 1991, Chapman and Hall. (new ed due 2011)

Statistical Questions in Evidence-Based Medicine. Martin Bland & Janet Peacock, 2000, Oxford Medical Publications (see Chapter 14).

Appendix 1: Answers to the exercises

(15.1) The data are categorical (stress incontinence present or absent) and the analysis will involve a comparison of proportions in a large sample. Hence, the appropriate test is the chi-square test which would be a 2 x 2 contingency table (columns: Male / Female, rows: stress incontinence present / absent). Alternatively, the Fisher's exact test could be used which is valid for any 2 x 2 contingency table and would have been appropriate particularly if the sample size was going to be small.
See Table 1

(15.2) Age of onset is a continuous variable and the analysis will involve a comparison of means (or medians if the data are not normally distributed) between two independent groups (adult males and females). The appropriate tests are the independent samples t-test (if the data are normally distributed or approximately so) and the Mann-Whitney U test or the Wilcoxon rank sum test if the data are not normally distributed. *See Table 1*

(15.3) Number of days lost from school is a continuous variable. The analysis involves a comparison of days lost in children with asthma compared with their matched control. Hence, this will involve a paired analysis. It is irrelevant if the raw data on days lost are not normally distributed as the hypothesis is comparing the mean of the *difference*s between each child with asthma and their matched control. The paired t-test would be appropriate assuming the differences are normally distributed. However, if the distribution of the differences in days lost was skewed you could use a Wilcoxon matched pairs signed rank test or a sign test. Use of the paired t-test will allow you to calculate a 95% confidence interval on the difference in days lost whereas use of the Wilcoxon matched pairs signed rank test or sign test will not.
See Table 2

(15.4) Parental smoking habit is a categorical (nominal) variable to be analysed as a dichotomous measure (smoker or smokers in the household versus no smokers in the household). The analysis will involve a paired comparison between children with asthma and their matched control. The McNemar's test is the appropriate statistical test as it will allow for the matched design.
See Table 2

(15.5) The researchers will need to plot the lung capacity (vertical axis) against height (horizontal axis) and calculate the correlation coefficient. If the data are normally distributed they could use Pearson's correlation coefficient (which measures the strength of the linear relationship between the variables) but if one or both variables are skewed in distribution they should use the non-parametric equivalent which is the Spearman's rank correlation coefficient. The researchers may also need to consider the women's smoking habits and their ethnicity as these are factors known to influence lung capacity. The relationship between lung capacity and height could be investigated further taking into account differences in smoking habits and ethnicity using multiple regression, also known as analysis of covariance. *See section 10, page 12*

Appendix 2: Glossary Sources: [adapted from A-Z of Medical Statistics (Pereira Maxwell F) and Medical Statistics at a Glance, 3rd ed. (Aviva Petrie & Caroline Sabin)]

Analysis of covariance	A special form of analysis of variance that compares values for a dependent variable between groups of individuals after adjusting for the effects of one or more explanatory variables.
Analysis of variance	An analysis comparing the means of two or more groups of observations by splitting the variance of a variable into its component parts, each attributed to a particular factor.
Bar chart	A chart illustrating the distribution of a categorical or discrete variable by showing a separate bar for each category where its length is proportional to the relative frequency in that category. Bar charts can be represented either horizontally or vertically.
Binomial distribution	A discrete probability distribution of a binary (dichotomous) variable used to draw inferences about proportions.
Categorical data	Data in which an individual value of a variable can be ascribed to one of a number of distinct categories.
Chi-squared test	A significance test for comparing two or more proportions from independent groups. The observed proportion in each group is compared with the expected proportion based on a null hypothesis.
Cohen’s Kappa	A measure of agreement for categorical data.
Confidence interval, CI	A range of values in which the true mean for a population is likely to lie. It usually has a proportion assigned to it (for example 95%) to give it an element of precision.
Continuous variable	A numerical variable which can theoretically take any value within a given range (for example, height, weight, blood pressure).
Correlation coefficient	A measure of the linear association (a straight line in a scatter plot) between quantitative or ordinal variables.
Covariate	*see* Independent variable
Cox proportional hazards regression	A regression method for modelling the risk (or hazard) of an event (usually death) occurring at a given time. The model can contain multiple characteristics, for example, age, gender, smoking habit, presence of coexisting diseases etc).
Dependent variable	A variable (usually denoted by *y*) the value of which is predicted in a regression equation.
Dichotomous	Division into two classes or groups.

Discrete data	A numerical variable that can only take integer values (whole numbers).
Explanatory variable	*see* Independent variable
Frequency distribution	A display of data values from the lowest to the highest, along with a count of the number of times each value occurred.
Histogram	A graphic display of data frequency using rectangular bars with heights equal to the frequency count.
Hypothesis	A statement of the relationship between 2 or more study variables.
Independent variable	A variable (usually denoted by *x*) used to predict the dependent variable in a regression equation. Also called the predictor variable, the explanatory variable or a covariate.
Interval data	See continuous variable
Intraclass correlation coefficient, ICC	A measure of reliability or agreement for interval data. Its calculation is similar to the Pearson's correlation coefficient but the ICC is more appropriate for assessing agreement.
Inter-quartile range, IQR	A measure of the variability of a set of measurements. The difference between the 25th and 75th percentiles containing the middle 50% of observations from a distribution of a continuous variable.
Kaplan-Meier plot	A survival curve in which survival probabilities are plotted against time from baseline (*see* Log-rank test).
Logistic regression	A statistical procedure to derive an equation to model a binary categorical variable (two categories) from one or more other variables that can be interval or categorical in nature. It is used to predict a probability (from 0 to 1) of occurrence of an event from a set of conditions.
Log-rank test	A non-parametric statistical test to compare two survival curves (*see* Kaplan-Meier plot)
Mean	The average value or measure of central tendency. The mean is obtained by dividing the sum of values by the total number of values.
Median	Middle value when data are ordered. The value that splits the sample in two equal sized parts.
Mode	The value that occurs most frequently.
Multiple regression	*See* Regression

Nominal data	A categorical variable for which the categories have no natural order (e.g. gender, eye colour, religion)
Non-parametric	Refers to data and tests of significance which makes no assumptions about the distribution of the data. Data that are skewed in distribution (to the right or left) are described as non-parametric.
Normal (Gaussian) distribution	A continuous probability distribution that is bell-shaped and symmetrical; its parameters are the mean and variance.
Ordinal data	A categorical variable for which the categories are ordered (e.g. severity of pain scores)
Outlier	Values in a set of observations which are much higher, or lower, than the ‘average’.
Parameter	A measurable characteristic of a population (e.g. average and standard deviation of blood pressure for a group of individuals).
Parametric	Refers to data in which the distribution is bell-shaped (Normal or Gaussian). Statistical tests that rely on data being distributed this way are called parametric tests.
Pie chart	A circular diagram in which the separate categories of a variable are represented as a fraction of 360 degrees. Each section of the ‘pie’ is proportional to the frequency in that category.
Poisson distribution	A discrete probability distribution of a variable representing the number of events that occur randomly and independently at a fixed average rate.
Predictor variable	*see* Independent variable
Qualitative data	See categorical data
Quantitative data	Data that can take either discrete or continuous values.
Regression	A statistical procedure to derive an equation to predict an outcome, interval variable (also called the dependent variable) from one or more other variables (also called the independent variable, the explanatory variable, the predictor variable). The independent variable can be interval or categorical in nature. When there is only one independent variable the procedure is referred to as ’simple regression’. When there is more than one independent variable the procedure is referred to as ‘multiple regression’.

Regression coefficient	The slope of the line of best fit in a plot between two variables. It represents the increase in an outcome variable from a unit increase in the predictor variable. For example, in a plot of total lung capacity against height in women the regression coefficient is 6.60 litres/metre which means that for every increase in one metre in height the lung capacity increases by 6.60 litres.
Significance level (P-Value)	In the context of significance tests, the P-value represents the probability that a given difference (or a difference more extreme) is observed in a study sample (between means, proportions etc) when in reality such a difference does not exist in the population from which the sample was drawn. In effect it's the probability of getting a wrong answer by deciding that two populations differ in some way when in fact they do not. In statistical parlance, it is the probability of rejecting a null hypothesis of no difference between two populations when in fact the null hypothesis is true.
Simple regression	*See* Regression
Standard deviation, SD	A measure of variability of data. The standard deviation is the average of the deviation of individual values from the mean measured in the same units as the mean.
Standard error (of the mean), SE	A measure of precision of the sample mean. Estimates of a population *mean* value will vary from sample to sample. The distribution of these values is called the sampling distribution. The SE is the 'standard deviation' of this distribution.
Standard score (also, z-score)	Refers to how many standard deviations away from the mean a particular score is located.
T-test	A statistical test used to determine if the means of 2 groups are significantly different.
Variable	Any quantity that varies (e.g. blood pressure).
Variance	A measure of variability of data equal to the square of the standard deviation.
Z-score	A standard score, expressed in terms of standard deviations from the mean.
Z-test	A significance test used for comparing a mean or a proportion between two groups.