An introduction to Medical Statistics

Beatson Cancer Charity
Determined to beat cancer

The beatson
West of Scotland Cancer Centre

University of Glasgow
Via Veritas Vita

What is Statistics

> *"**Statistics** is the science of collecting, summarizing, presenting and interpreting data, and of using them to estimate the magnitude of associations and test hypotheses."*
>
> *Kirkwood and Sterne, Essential Statistics*

Descriptive statistics is concerned with summarising and presenting data. It is an essential step before further analysis is conducted.

It helps form subjective impressions of answers to research questions.

To produce appropriate descriptive statistics requires knowledge of different data types. Broadly, data are either *numerical* or *categorical.*

Inferential statistics provides a framework to make subjective impressions more objective by quantifying uncertainty. Through estimation and testing it quantifies the magnitude of associations and indicates the strength of evidence that these are "real".

Statistics is the core science of evidence based medicine.

Terminology: Probability

In layman terms probability is often regarded as the degree of belief that an event will happen.

For the most part when statisticians refer to probability (denoted as **P** or **p**), they are referring to the long term frequency of an event occurring. This provides a measure of the chance that an event will happen, usually under a set of specific assumptions. The probability of the event not occurring is 1-p.

By definition probability can take any value between 0 (the event will definitely not occur) and 1 (the event will definitely occur).

Probability is sometimes expressed as a percentage 0 -100%.

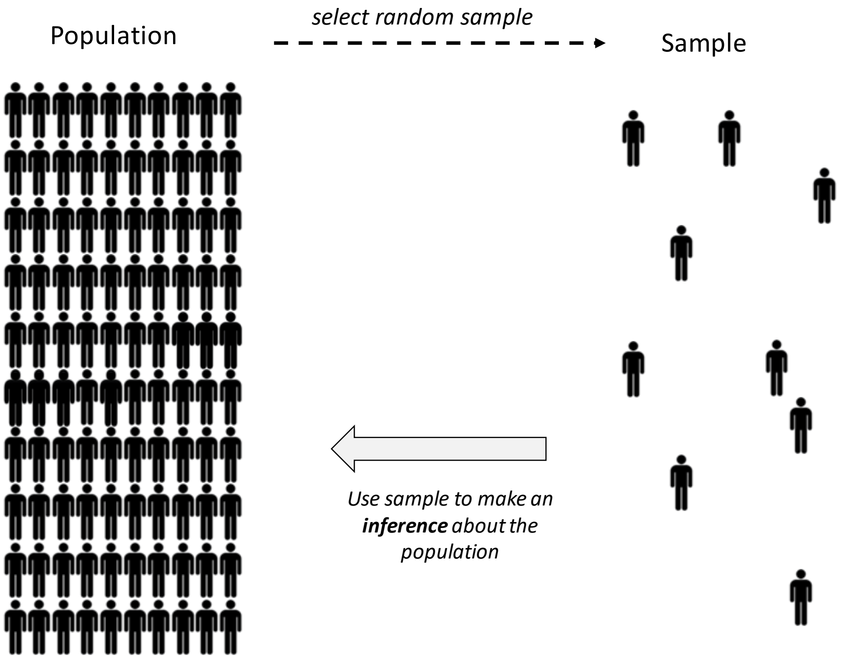
Populations and Samples

Statistics describes and makes inferences about '***populations***'. A **population** is the entire group of individuals whose characteristics we are interested in.

This is achieved by using a '***sample***' of individuals from the entire population.

Using the **sample,** we measure a characteristic whose value can vary from individual to individual (a '***variable***') – age, tumour stage, blood pressure, performance status, weight, sex, occupation, and length of survival following radiotherapy are all examples of variables.

The values observed among individuals within the sample are called ***observations*** or ***data***.



We *summarise* the values of the characteristic in the sample (the summary is called the ***sample statistic –*** for example 'average' age). We then make an inference about the corresponding summary value (called the ***population parameter***) in the entire population

An important assumption is that the sample is **representative** of the entire population. To achieve this, individuals in the sample must be selected **randomly** from the entire population. Random selection means that *every person in the population has the same probability of being selected to the sample*.

The above strategy is called **simple random sampling**. There are other selection strategies, they include ***stratified random sampling***, and ***cluster sampling***. At their core is random selection, and the notion that we cannot predict which individual will be included in the sample.

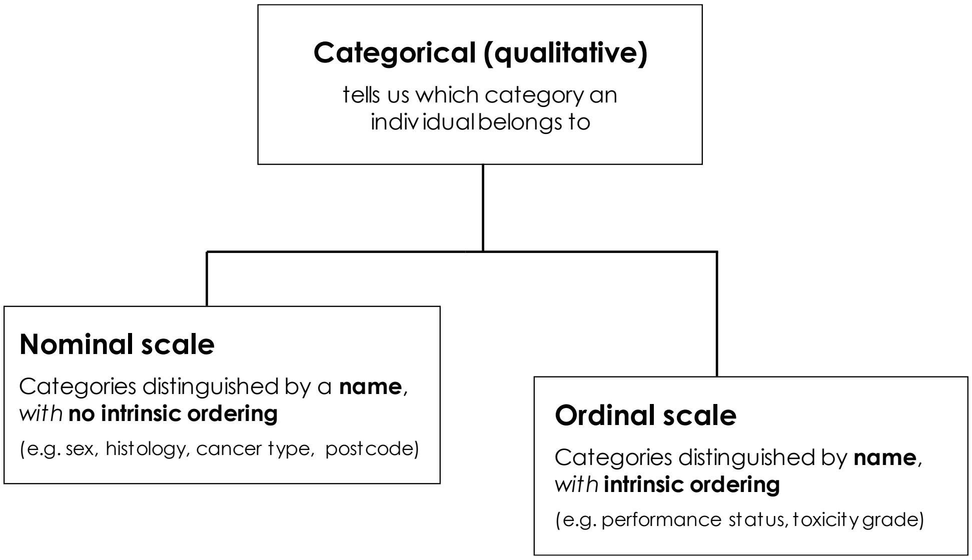
Selection bias

Bias is a type of error that systematically skews results in a certain direction. Selection bias is a common type of error. The decision about who to include in a study can throw findings into doubt. Selection bias can occur when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random (i.e. with observational studies such as cohort, case-control and cross-sectional studies). Random selection helps reduce selection bias by ensuring each individual has an equal chance of being selected.

Even then selection bias occurs when people agree or decline to participate in a study. Those who choose to join (i.e. who self-select into the study) may share a characteristic that makes them different from non-participants. Often, selection bias is unavoidable. That's why it's important for researchers to examine their study design for this type of bias and find ways to adjust for it, and to acknowledge it in their study report.

Types of Data

Broadly, data are either numerical or categorical. Other terms to describe this being quantitative or qualitative.

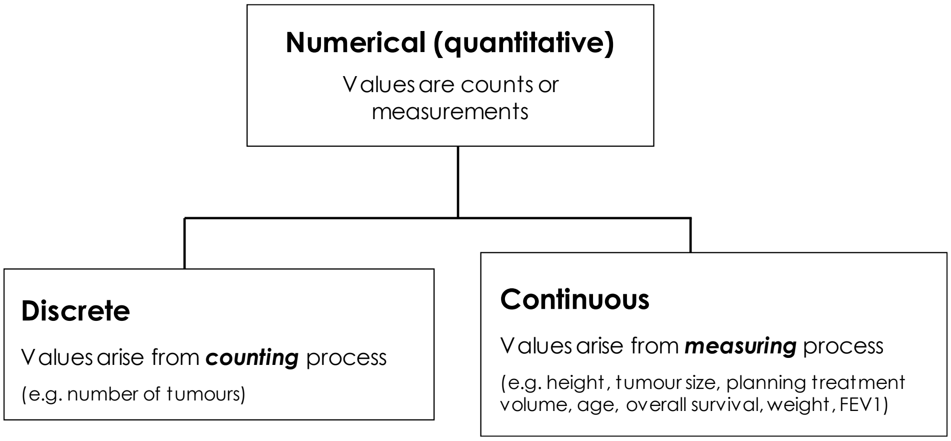


Categorical (qualitative)
tells us which category an individual belongs to

Nominal scale
Categories distinguished by a **name**, *with* **no intrinsic ordering**
(e.g. sex, histology, cancer type, postcode)

Ordinal scale
Categories distinguished by **name**, *with* **intrinsic ordering**
(e.g. performance status, toxicity grade)

A categorical measure with only two categories (for example alive or dead) is called **dichotomous**. Sometimes the categories of a dichotomous variable are labeled 0 and 1, and are called a **binary variable**.



Numerical (quantitative)
Values are counts or measurements

Discrete
Values arise from ***counting*** process
(e.g. number of tumours)

Continuous
Values arise from ***measuring*** process
(e.g. height, tumour size, planning treatment volume, age, overall survival, weight, FEV1)

Paired data

The majority of this course will be concerned with comparing characteristics measured in two separate groups of individuals. In some circumstances, however, data may consist of pairs of outcome measurements. We might wish, for example, to carry out a study where the assessment of tumour response to radiotherapy is based on comparing tumour size measurements in a group of lung cancer patients, before and after they received treatment. For each person, we therefore have a pair of measures, tumour size after treatment and tumour size before treatment. It is important to take this pairing in the data into account when assessing how much on average the treatment has affected tumour size.

Presenting Data

The use of graphics and descriptive statistics is an important step to identifying the main features of data, for detecting **outliers**, and identifying data which has been recorded incorrectly. **Outliers** are extreme observations which are not consistent with the rest of the data. The presence of outliers can distort statistical techniques.

Frequency Distribution Tables

A count of the number of times something occurs is called *a* ***frequency count***.
Within a data set we may list the data values and count how many times each value occurs. A table with the set of data values and frequency of each value is called a ***frequency distribution table.***

Number of brain metastases	Number of patients (n)	percent of patients (%)
1	19	59.4
2	4	12.5
3 or more	9	28.1
Total	32	100

Table 1 Number of brain metastases among renal cancer patients treated with cranial radiotherapy

The ***most*** *frequent* category is called the **mode**. In the above table the mode would be 1 metastasis, which occurred in a total of 19 real cancer patients.

The percentages in a frequency distribution table are sometimes expressed as relative frequencies or proportions. In the above table, the proportions would 0.594, 0.125, 0.281.

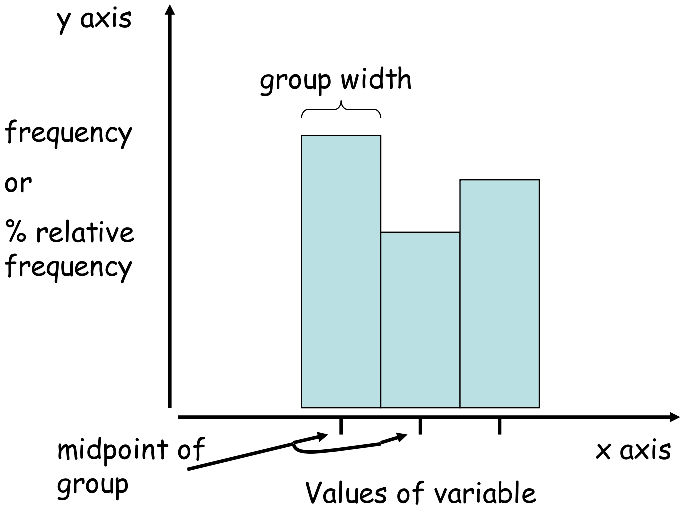
Grouped Frequency Tables

When the set of potential data values is large the data are organized in groups (examples – age groups 20-29, 30-39, 40-49, 50-59, 60-69 etc). A count of the number of data values in each group (**group frequency**) is made. The frequency distribution table will then include the data groupings and the frequency of each group.

Graphs for Numeric Data

Histograms

Consists of a set of rectangles which present the frequency distribution of a variable. It allows a visualization of the shape of the data.



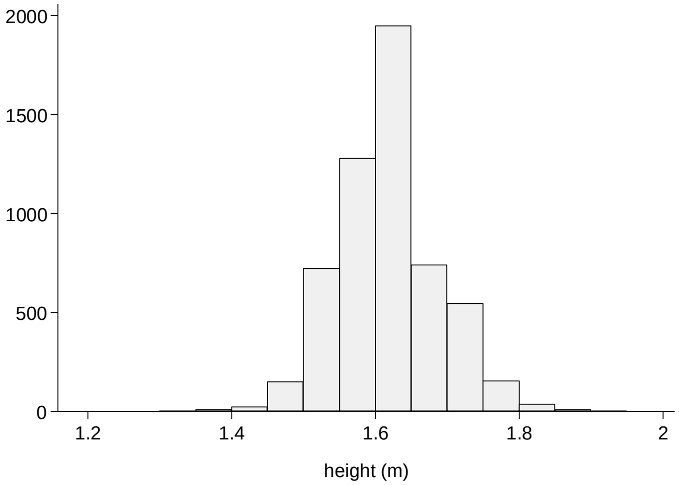
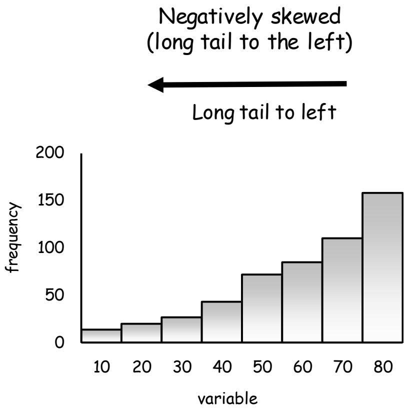
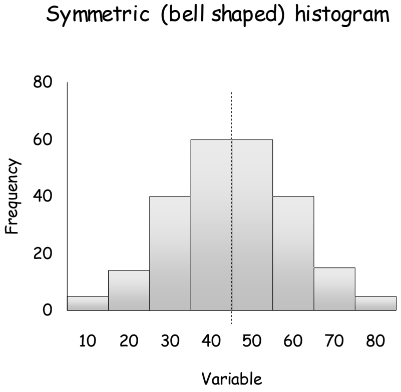
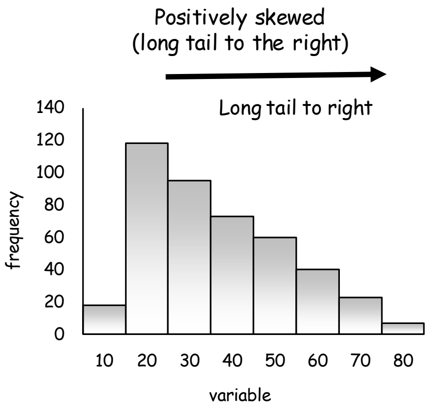


Figure 1 Histogram of the heights of 5,682 women aged 25-64

The histogram above shows that the *most frequent* grouping of heights occurs at 1.6-1.65 cm. There is a tendency for the bars of the histogram to cluster around this central **modal group** in a symmetric fashion. This is a classic example of a **Normal distribution** which displays a symmetric bell shape. Not all histograms are symmetric; skewness can exist in the distribution of values.







Dot-plot

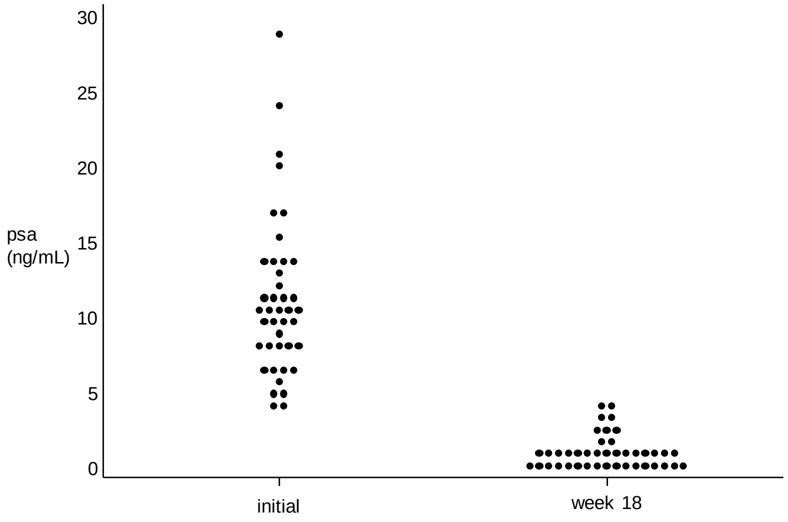
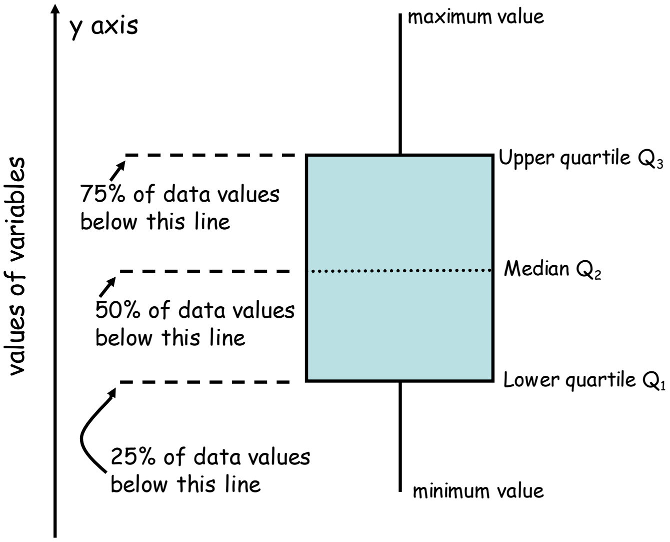


Figure 2 PSA levels before treatment and at week 18 among 41 men who received SABR for prostate cancer

Dot plots are a simple method of conveying as much information as possible by showing all of the data. It retains individual subject values and clearly demonstrates differences between groups. An additional advantage is that outliers will be detected.

Box plot

A boxplot is used for discreet and continuous data. It indicates 5 important statistics which describe the distribution of observations. The first two statistics are the **minimum** and **maximum** values. The other statistics are called the **lower quartile** (**Q1**), **median** (**Q2**), **upper quartile** (**Q3**). These indicate the values of the variable of interest which divide the sample into four equal groups – each group containing 25% (quartiles) of the sample.



The **median** is the *middle observation* in a set of data. 50% of the sample lie below, and 50% lie above the median. It can be seen from the figure that 50% of the data sample lie between the lower quartile (Q_1) and the upper quartile (Q_3)

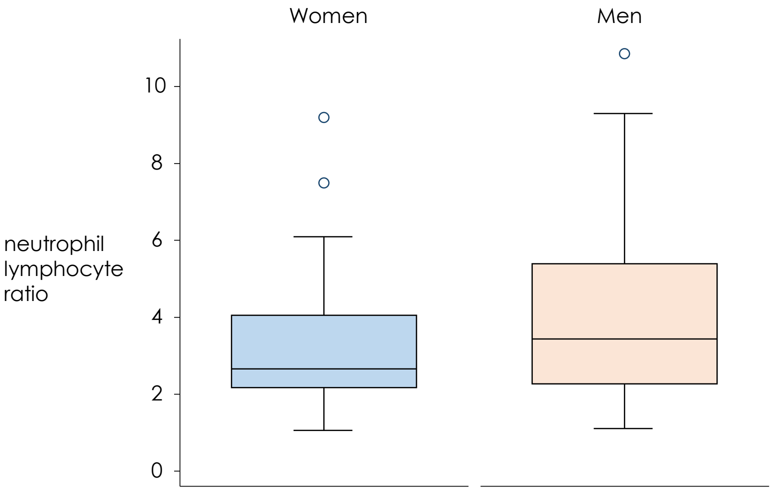
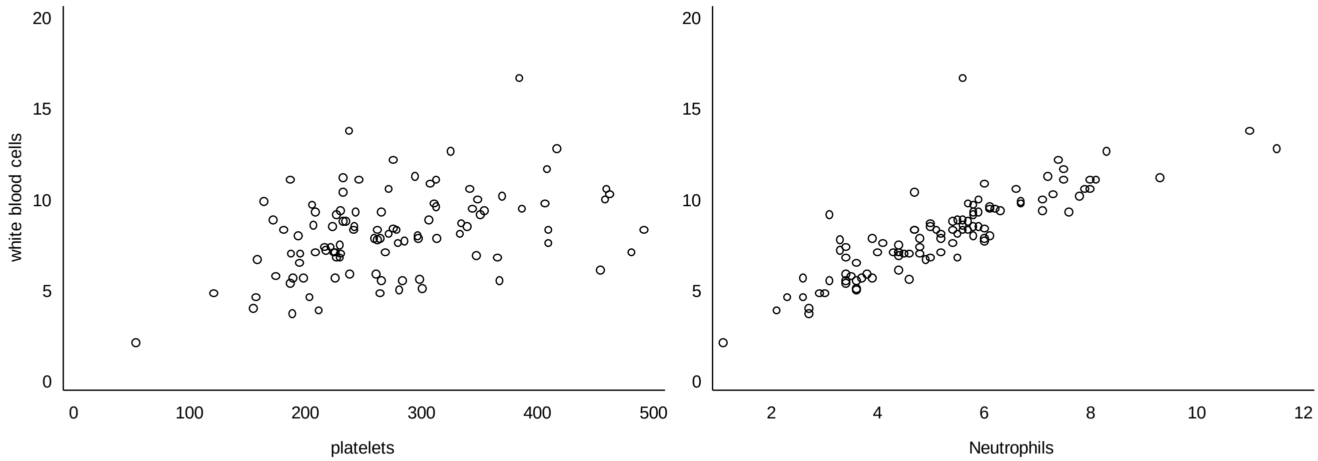


Figure 3 Pre-treatment neutrophil lymphocyte ratio in men and women who received SABR for lung cancer

Graphs for bivariate continuous data

Scatterplots

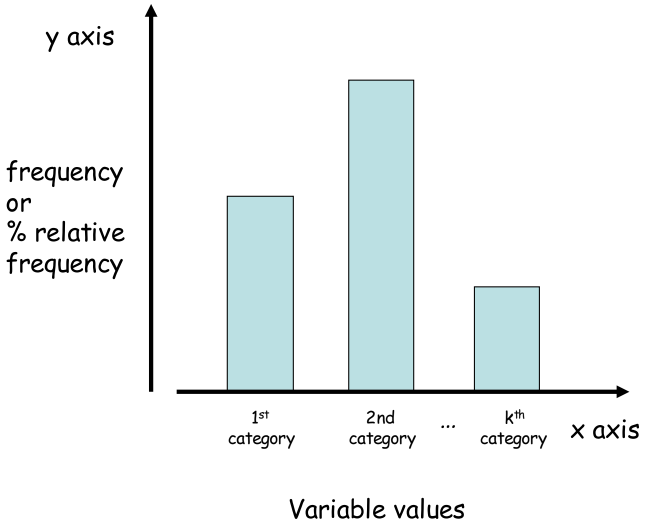


Scatterplots shows the relationship between two numeric variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot. Two variables are positively associated when higher values of one variable tend to accompany higher values of the other, and lower values tend to occur together. The figures above suggest a weak positive association between platelets and white blood cell counts, and a stronger positive association between neutrophil counts and white blood cells. This is not surprising as the white blood cell count includes neutrophil counts. The scatter, however, indicates one individual with a higher than expected white blood cell count.

Graphs for Categorical Data

Bar Charts

Bar Charts are typically used for categorical data (but can be employed for discrete numerical data). Data can be nominal or ordinal.



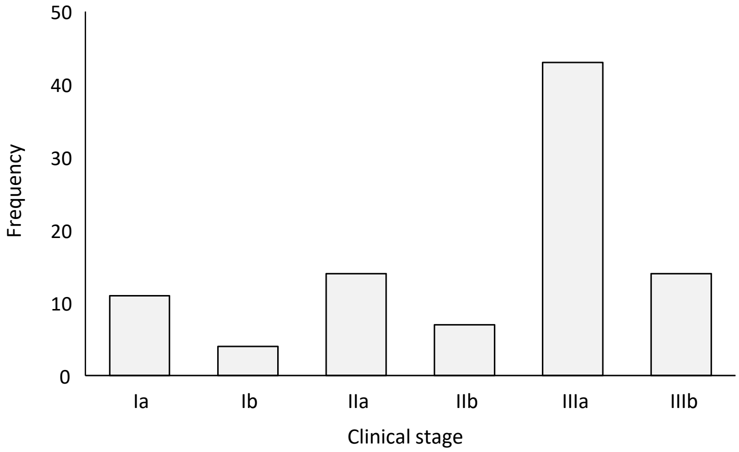


Figure 4 Stage among 93 radical radiotherapy NSC lung cancer patients

Pie chart

Pie charts illustrate nominal data. Each slice of the pie represents a category. The size of the slice is proportional to the relative frequency of that category. It is best utilized with 3-5 categories.

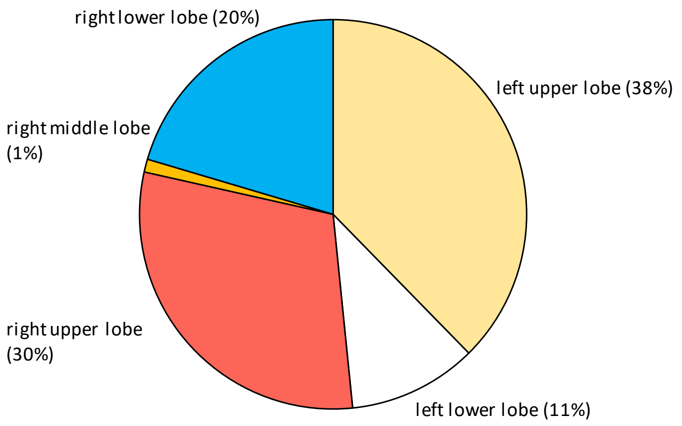


Figure 5 Tumour location among 93 patients who received radical radiotherapy for lung cancer

Tables and Graphs for Bivariate Categorical data

Contingency Tables

When the interest is in the relationship between two qualitative variables a contingency table (a cross-tabulation) is created. This has the categories for one variable as rows and the categories of the other variable as columns. Each cell of the table has a count of the number of individuals in both categories.

The following example concerns the two-year follow-up of 42 men and 60 women following stereotactic radiotherapy for lung cancer. In total 25 patients had died and 77 were still alive at the two year follow-up. The table below shows a cross-tabulation of status at two-years by sex.

Status at two years	Women (n)	Men (n)	Total (n)
Died	48	29	77
Alive	12	13	25
Total	60	42	102

A total of 12 women and 13 men had died. There are different total numbers of men and women in the sample, so simply comparing 12 with 13 is not informative. The numbers need to be expressed relative to the total number of men and women (relative frequency or percentage).

Status at two years	Women (n)	Men (n)	Total (n)
Died	48 (80%)	29 (69%)	77 (75%)
Alive	12 (20%)	13 (31%)	25 (25%)
Total	60 (100%)	42 (100%)	102 (100%)

31% of men had died but only 20% of women. There is a suggestion of a possible association between sex and outcome following SABR for lung cancer.

Clustered (grouped), stacked (component), percentage component bar charts

We can create subsets of the data based on the categories of one variable. We can then examine the distribution of another variable within each subset.

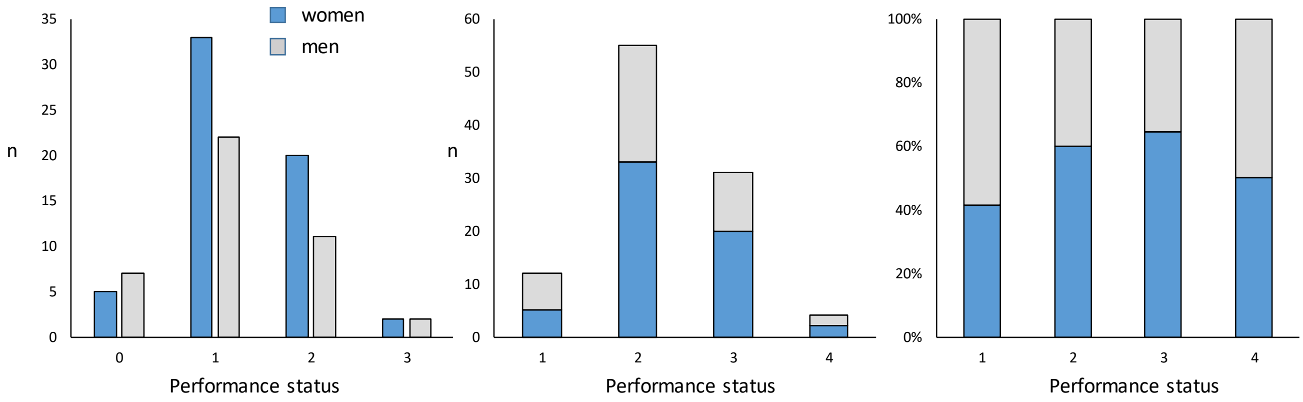


Figure 6 Distribution of performance status by sex among 102 NSCLC patients presented in different ways

The figures above show the breakdown of men and women with NSCLC by performance score. Overall it suggests that while there are more women than men in the sample, there is a greater number of men with performance status 0 than women.

Descriptive measures

A descriptive measure is a numerical value, which summarises a set of data. We represent the number of patients in a sample by the letter **n**. We often represent the individual values of a variable with a small letter **x**. The value for patient 1 would be x_1, the value for patient 2 would be x_2. In a sample of *n* patients, the value for patient n would be x_n.

Measures of Location or Central Tendency

These give the location of the center of the data. Representative measures are ***mean, median,*** and ***mode***. They are referred to as 'average' values – i.e. 'something in the middle'.

Sample mean ($\bar{x}$)is calculated as

$$mean = \frac{\sum x}{n}$$

$\Sigma x = x_1+x_2+x_3+\ldots+x_n$ is the sum of all values across the subjects, and n is the total number of subjects.

Sample median is the *middle value* when the observations are ranked from lowest to highest. If n is even, it is the mean of the middle two values. Its interpretation is that 50% of data values are above the median; 50% are below the median.

Quartiles (Q1, Q2, and Q3) - when the observations are ranked from lowest to highest the quartiles divide a set of data into four parts of equal frequency. 25% of the data values are smaller than Q_1(**lower quartile**). 50% of the data values are smaller than Q_2 (**median**). 25% of the data values are larger than Q_3 (**upper quartile**).

Sample mode is the *most frequently* occurring value. This term is seldom used.

Measures of Dispersion

Knowing the "average" value of data is not very informative by itself. We also need to know how "concentrated" or "spread out" the data are. That is, we need to know something about the "variability" of the data.

Measures of dispersion are ways of quantifying this numerically. They describe the degree to which the data vary about their average value, their scatter or spread. Representative measures: ***range, standard deviation,*** and ***interquartile range***.

Range is simply the difference between the smallest (**minimum**) and largest (**maximum**) values in the sample.

Standard deviation (**SD** or **sd**) is a measure of how far away observations are from the **sample mean**.

It is calculated as

$$SD = \sqrt{\frac{\sum(x - mean)^2}{n-1}}$$

Which is the square root of the sum of squared differences from the **sample mean** divided by **(n-1)**

Interquartile range (IQR) is simply the lower quartile and upper quartile (Q1, Q3). Sometimes it is expressed as the value of Q3-Q1.

Example

Mean and standard deviation

10 patients with lung cancer had a pretreatment PET scan and SUV_{max} was measured.

Original Data: 1.8, 8.9, 2.7, 9.4, 5.4, 16.0, 5.8, 17.9, 13.1, 6.6

The **mean** $SUV_{max} = \frac{1.8+8.9+2.7+9.4+5.4+16.0+5.8+17.9+13.1+6.6}{10} = 8.8$

The calculation of the **standard deviation** involves subtracting the mean SUVmax from each observation

-7, 0.1, -6.1, 0.6, -3.4, 7.2, -3.0, 9.1, 4.3, -2.2

squaring these values

49.0, 0.01, 37.21, 0.36, 11.56, 51.84, 9.0, 82.81, 18.49, 4.84

summing these values together

$$265.12$$

dividing by the number of observations minus 1

$$\frac{265.12}{(10-1)} = 29.46$$

taking the square root

standard deviation = √29.46 = 5.43

Median and interquartile range

Ordering the data from lowest to highest allows identification of the median and quartiles

Ordered data: 1.8, 2.7, 5.4, 5.8, 6.6, 8.9, 9.4, 13.1, 16.0, 17.9

The **median** SUV_{max} = (6.6+8.9)/2 =7.75; **Q1**=5.4 and **Q3**=13.1

The **minimum** value is 1.8, and the **maximum** value is 17.9. The **range** = 17.9-1.8 =16.1.
The **interquartile range (IQR)** is (5.4, 13.1) *or* 13.1-5.4 =7.7.

Remember to report means with standard deviations (8.76 (SD 5.43)), and medians with interquartile ranges (7.75 (IQR 5.4-13.1)). Do not mix means with the interquartile range or medians with standard deviations.

Which measures to choose?

The mode should be used when calculating a measure of center for nominal categorical variables. When the variable is numeric with a symmetric distribution, then the mean is proper measure of center. In a case of numeric variables with skewed distribution, the median is a good choice for the measure of center. The median is less influenced by outlier (extreme) values.

The sample mode, the sample median and the sample mean have corresponding population measures. That is, we assume that the variable in question has a population mode, population median, population mean, which are all unknown. The sample mode, the sample median and the sample mean are used to estimate the values of these corresponding unknown population parameters.

Normal distribution

The “normal distribution” is referred to frequently in statistics. It’s a symmetrical, bell-shaped distribution of data. The Normal distribution is a cornerstone of statistics as many statistical methods are built around it. If it did not exist statisticians would have had to invent it.

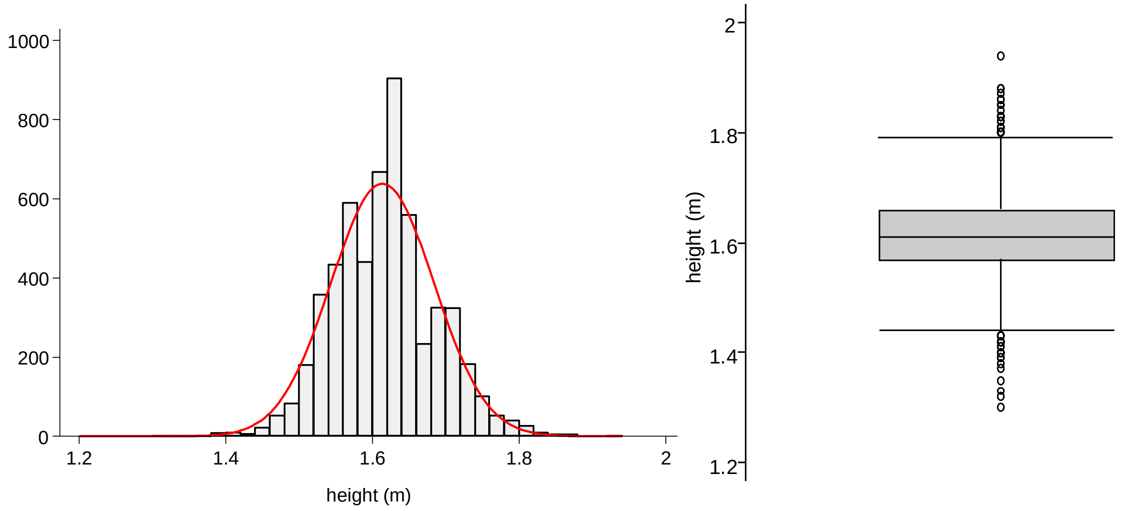


Figure 7 Height of 5,628 women aged 25-64

It is often the case that the histogram of a continuous variable will display the characteristic bell-shaped distribution of the Normal distribution. The height of women above shows a Normal distribution; the box plot shows a symmetric distribution of values above and below the median line.

The Normal distribution is completely described by two population parameters μ and σ, where **μ** represents the **population mean** (the center of the distribution) and **σ** the **population standard deviation**. One property of the Normal distribution is that exactly 95% of the distribution lies between -

$$\mu -1.96\sigma \text{ and } \mu +1.96\sigma$$

This is called a **reference range,** in this example, it is the range in which 95% of the population lie.

In practice the two parameters of the Normal distribution are estimated from the sample data. The **sample mean** (**$\bar{x}$**) and the **sample deviation** (**SD**). If a sample is taken from a Normal distribution, and provided that the sample is not too small, then approximately 95% of the *sample* will be covered by

$$\bar{x}\text{-1.96SD and } \bar{x} \text{ +1.96SD}$$

In the example above the mean height ($\bar{x}$) of women was 1.61 m and the standard deviation (SD) was 0.07 m (i.e. 7cm). Approximately 95% of the sample will be cover in the range 1.61-1.96x0.07 to 1.61+1.96x0.07 which is 1.47m to1.75m.

The standard error

Imagine another random sample of women was selected to investigate their height. The values in this second sample will vary from woman to women and we would expect the mean value for this new group to be different but **not too different** from that obtained in the first sample. The precision with which the sample mean is estimated can be measured by the standard deviation of the mean, this is called the **standard error, SE**.

$$\text{SE=SD/}\surd\text{n}$$

In the above example of 5,867 women the standard error of the mean would be 0.07/√5628 = 0.07/75 = 0.0009 m. Because the sample size is very large the standard error of the mean is very small in this example. If the sample size was smaller, say 20 women, the standard error of the mean would have been larger 0.0156 m. Remember that sample size does not influence the size of the sample standard deviation, it influences the precision of the estimated parameters.

Confidence intervals

Confidence intervals define the range of values within which the population mean μ is likely to lie. A 95% confidence interval for the population mean is defined by

$$\bar{x}\text{-1.96SE and } \bar{x} \text{ +1.96SE}$$

In the case of height among the sample of 5,867 women the 95% confidence interval (95% CI) would be (1.608m to 1.611m). The interpretation of the confidence interval is that 95% of intervals will contain the true population mean. This is interpreted as a 95% chance that the population mean will be contained in the interval. The interval in this example is very narrow due to the large sample size. If the sample size had been 20, the interval would have been wider at (1.58m to 1.64m).

Skewed distributions

Many statistical tests require data to be Normally distributed. In practice distributions are sometimes not symmetric and can be skewed. Often they display a long right hand tail (positive skew) or long left hand skew (negative skew). Skewed distributions can be made approximately Normal by transforming the original data. In the case of positively skewed data taking the log of the data can transform the data into an approximate Normal distribution. In the case of negatively skewed data (long left tail) squaring the data may help.

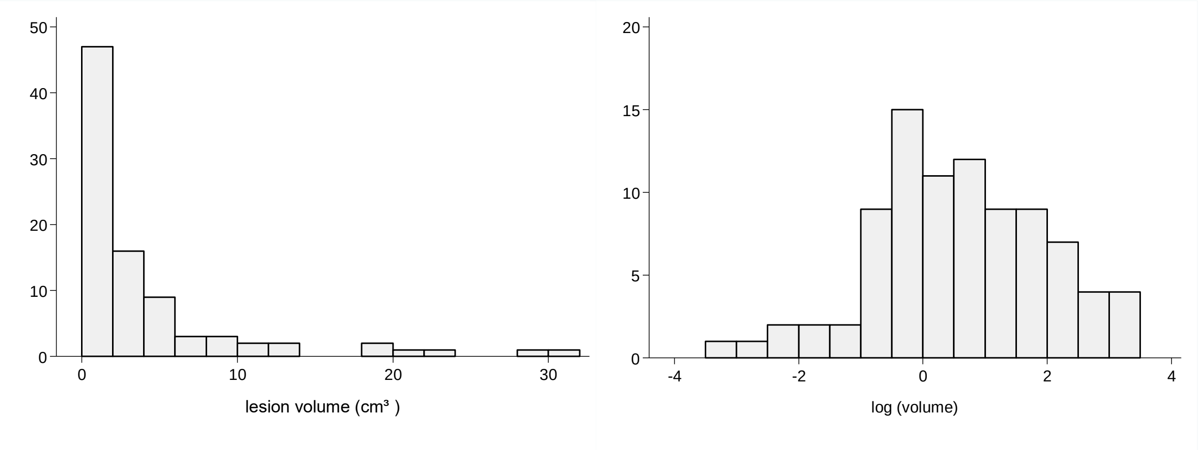


Figure 8 Tumour volume measured from the CT of 88 NSCL patients. Original and log transformed data.

If data are not Normally distributed, then **non-parametric methods** which do not make assumptions that the data come from a Normal distribution may need to be used.

The central limit theorem

The distribution of sample means will be nearly Normally distributed (whatever the distribution of measurements among individuals). It will get closer to a Normal distribution as the sample size increases. This feature of mean values comes from the Central Limit Theorem and is very useful in the analysis of proportions.

Concepts in Statistical inference

Until the end of the 17th Century Europeans assumed that all swans were white. The hypothesis that "All swans are white" was assumed to be true but was rejected by the sighting of a black swan by Willem de Vlamingh in 1697. The black swan resulted in a rejection of the original **null hypothesis (H_0):** *"All swans are white"* in favour of the **alternative hypothesis (H_1):** *"All swans are not white"*.

Statistical inference follows a similar logical process. Having come up with a research question the procedure is to collect data from a sample of individuals, formulate an appropriate null hypothesis, assume it to be true, and seek evidence to refute it.

Null Hypothesis

The **null hypothesis, H_0** is a statement of *'no difference'* or *'no effect'* which is *assumed* to be true.

For example, in a clinical trial of a new drug for hypertension, the null hypothesis might be that the new drug has a similar average effect on blood pressure as another drug in current use – i.e. that there is no difference between the drugs.

***H_0**: there is no difference in the effect on blood pressure between the two drugs*

Alternative Hypothesis

The **alternative hypothesis** (**H_1** or **H_A**) is the negation of the null hypothesis. It holds if the null hypothesis is not true. The alternative hypothesis relates more directly to the theory we are interested in. In the anti-hypertensive example, we might have:

***H_1:** the effects of the two anti-hypertensive drugs are not equal*

The test-statistic and p-values

Having set up the null hypothesis, the probability that the observed data (or more extreme data) would be obtained if the null hypothesis were true is evaluated. This is done by calculating a numerical summary called a **test statistic** (calculated from the sample data) which is known to have a specific probability distribution. The test statistics is used to test the null hypothesis. The value of the test statistic is related to the specific probability distribution to obtain a **P-value**. The smaller the P-value, the greater the evidence against the null hypothesis.

Using the P-value

A P-value less than 0.05 (**p<0.05**), is conventionally considered enough evidence to reject the null hypothesis. P<0.05 suggests only a small chance that the observed results (or more extreme results) would have occurred if the null hypothesis were true. The null hypothesis is then rejected in favour of the alternative hypothesis and the results described as statistically significant at the 5% level.

In contrast, a P-value equal to or greater than 0.05, suggests insufficient evidence to reject the null hypothesis. The null hypothesis is not rejected, and the results are described as not statistically significant at the 5% level. This does not mean that the null hypothesis is true - just that it cannot be rejected.

One or two-tailed test?

In the above example the alternative hypothesis did not specify the direction for the difference in the effects of the two anti-hypertensive medications, i.e. it did not state whether the new drug provides better blood pressure control than the current drug or vice versa. This is known as a **two-tailed test** because it allows for either eventuality. In some circumstances, a **one-tailed test** in which the direction of the difference is specified in $\mathbf{H_1}$ may be carried out. In general, one-tailed tests are discouraged as it is unlikely that we can know beforehand which direction will occur.

Making a decision - Type I and type II errors

A **type I** error leads to the conclusion that an effect or relationship exists when in fact it does not. Type I errors results from the incorrect rejection of a true null hypothesis (a "false positive"),

A **type II** error is a failure to detect an effect that is present. It results from incorrectly retaining a false null hypothesis (a "false negative").

In the previous example of a clinical trial of a new drug for hypertension, the null hypothesis and alternative hypotheses were

***H_0*: there is no difference in the effect on blood pressure between the two drugs**

***H_1*: the effects of the two drugs on blood pressure are not equal**

A **type I** error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them. A **type II** error would occur if we failed to reject the null hypothesis when the there was a real difference in effect of the two drugs.

The probability of making a Type I error (denoted by a (alpha)) is simply the chosen significance level (conventionally 5%).

Probability (type I error) = probability (reject null when true) = alpha = a

The chance of making a Type II error is denoted by β (beta);

Probability (type II error) = probability (fail to reject null hypothesis when false) = β

The complement of β is (1-β). This is the Probability of not making a type II error.

Probability (not type II error) = probability (reject null hypothesis when false) = 1-β

(1 – β) is called the **power** of the test. The **power**, therefore, **is the probability of rejecting the null hypothesis when it is false**; i.e. it is the chance (usually expressed as a percentage) of detecting, as statistically significant, a real treatment effect.

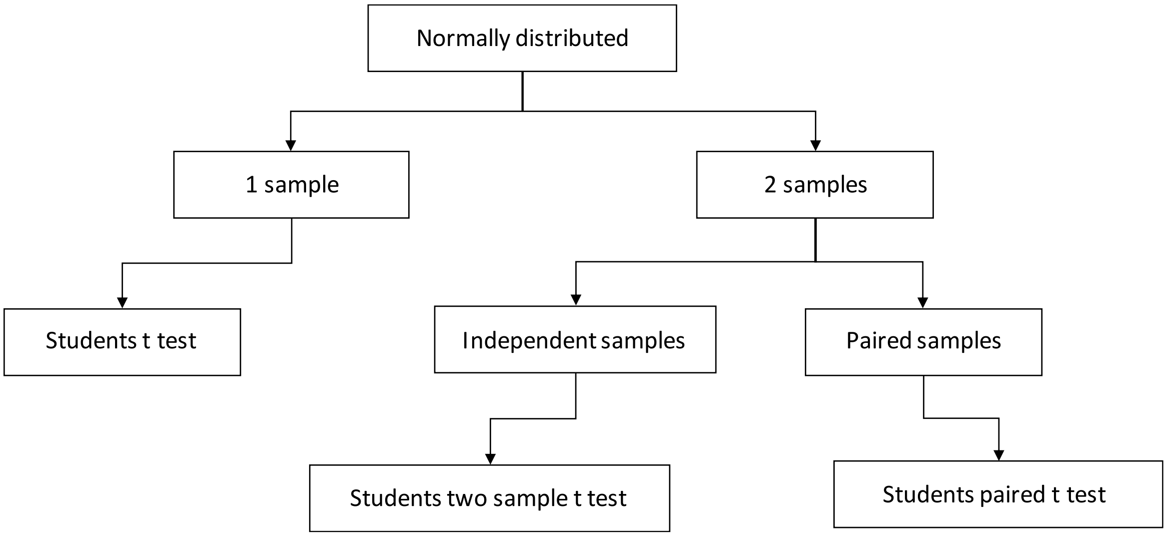
P-values or Confidence intervals?

We saw earlier that a confidence interval is a range of values that the parameter of interest is likely to lie in the population. The parameter might be the population mean or median, or the mean difference between two groups, or a proportion. Presenting study findings directly as confidence intervals, provides information on the imprecision due to sampling variability and has advantages over just giving P values which dichotomies results into significant or non-significant.

With a confidence interval, we can determine whether a parameter is or is not likely to be different to something. If the confidence interval contains a specific number (i.e. the number is between the lower and upper values of the interval), then there is no evidence that the parameter is different from that number. If the number is not within the interval, then there is evidence that the parameter is different from that number.

Parametric tests for numeric data

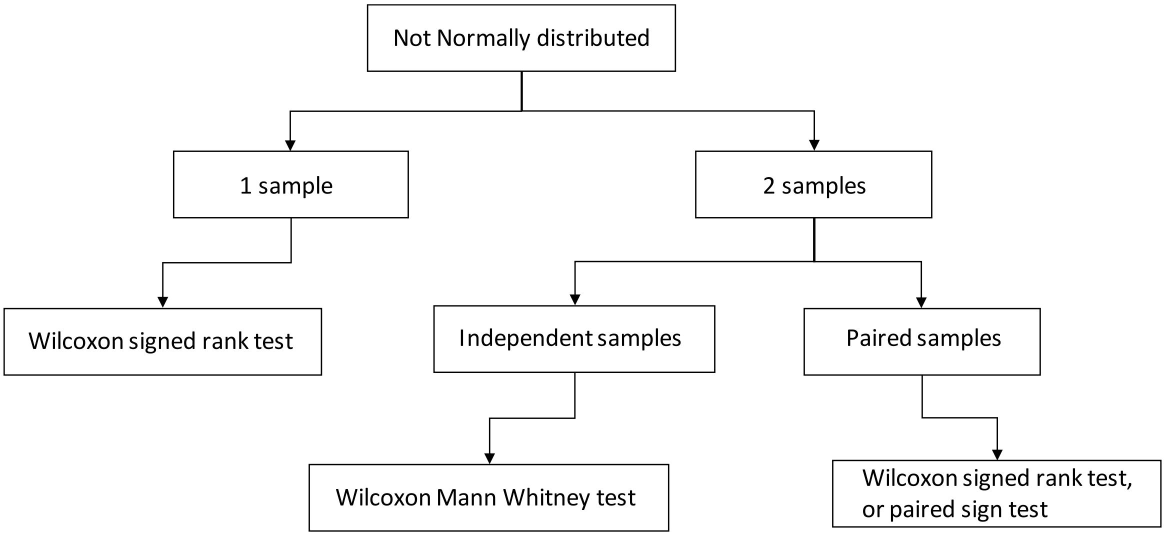
If data are numeric and come from a Normal distribution we can use parametric tests to test whether the population mean equals a specific value, or whether the means from two samples are equal.



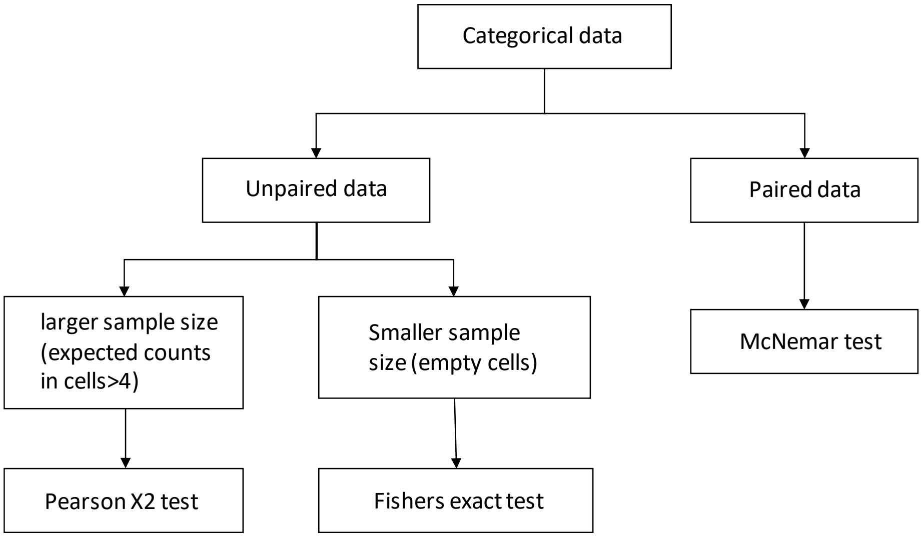
Parametric tests need the assumption that the data derive from a Normal distribution. If this assumption cannot be met (even after transformation) then non-parametric tests must be used.

Non-parametric tests for numeric data

Non-parametric tests compare the median to a specific value, or test the medians between samples to see if they would be equal in the wider population. Non-parametric tests are not as powerful (i.e. the probability of rejecting the null hypothesis when it is false will be smaller) as parametric tests.


Not Normally distributed
1 sample
2 samples
Wilcoxon signed rank test
Independent samples
Paired samples
Wilcoxon Mann Whitney test
Wilcoxon signed rank test, or paired sign test

Tests for categorical data



Tests for categorical data are concerned with the comparisons of proportions in each category of a variable. Just as for numeric data, a special analysis is required if paired data are involved.

Sample size & power considerations

How many subjects should be included in a study is a common consideration. If a study has too few people, the power to detect a statistically significant effect will be low. On the other hand, obtaining a sample size that is large or larger than required can be difficult to achieve and expensive. Recruiting patients to a study which will be too small to detect the minimum effect we are looking for or recruiting more patients than necessary (over-powered) can be considered unethical.

To establish the sample size needed for a study the following factors should be considered.

1) The minimum size of the effect to be detected
2) The variability (standard deviation)
3) The power required
4) The significance level.

For a chosen significance level, power, minimum size of effect to be detected and standard deviation the sample size needed can be calculated.

Power is the probability of rejecting the null hypothesis when it is false. In general …

as the sample size increases the power increases
as the variability (standard deviation) increases the power decreases
as the minimum size of effect to be detected is increased the power increases (i.e. small effects are more difficult to detect)

To increase the power of a study the sample size can be increased, and the minimum size of the effect you are trying to detect can be increased.

The significance level is not affected by choice of power or sample size. It is the decision rule that you employ in the study.

Correlation and Linear regression

Correlation and linear regression are techniques for describing the relationships between variables.

Correlation looks for a linear association between two variables. The strength of the association is summarised by the correlation coefficient.

Regression looks for the dependence of one variable (the dependant variable) on another (the independent) variable. It quantifies the best linear relation between the variables and allows the prediction of the dependent variable when only the independent variable is known.

Correlation

Correlation is used to measure the degree of linear association between two continuous variables.

There are two main types of correlation coefficients **Pearson's correlation coefficient**, and **Spearman's rank correlation coefficient**.

Pearson's correlation relies on assumptions of Normality of the data. Spearman's correlation is a non-parametric alternative. This should be used if data are not approximately normally distributed, have extreme values (outliers), or the sample size is small.

The correlation coefficient (*r*) can takes any value in the range -1 to +1.

The **sign** of the correlation coefficient indicates whether, one variable increases as the other variable increases (positive r) or whether one variable decreases as the other increases (negative r)

The **magnitude** of the correlation coefficient indicates the strength of the linear association.

If r = +1 or −1, then there is perfect correlation. If both variables were plotted on a scatter graph all the points would lie on a straight line; if r = 0, then there is no linear correlation. The closer r is to -1 or 1, the greater the degree of linear association.

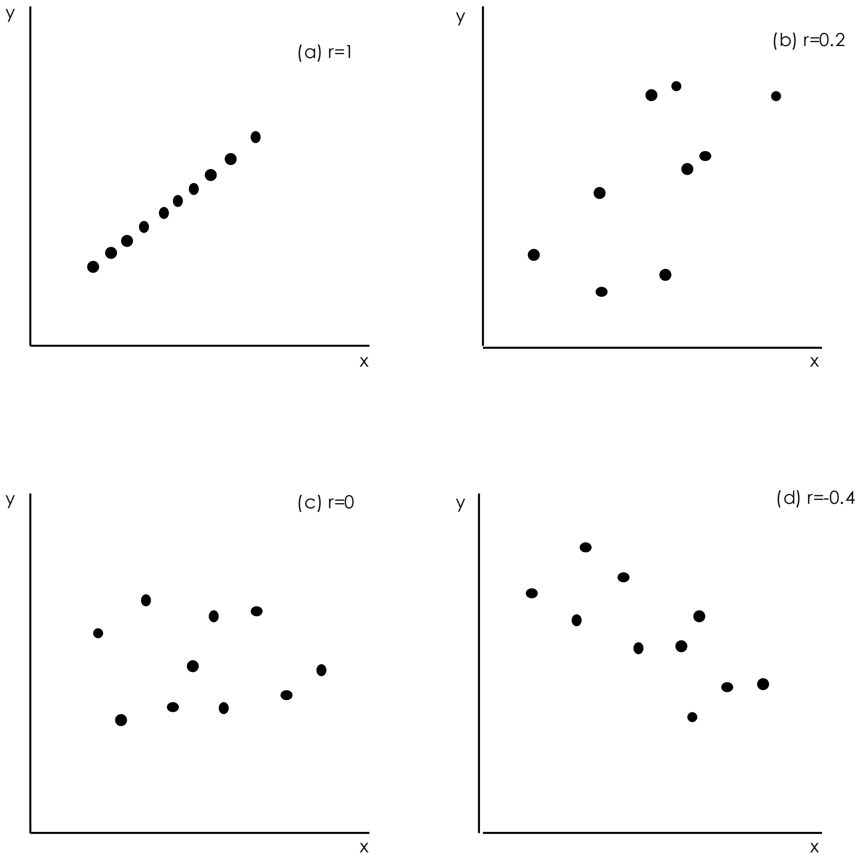


Figure 9 Scatterplots showing datasets with different correlations - (a) strong positive, (b) weak positive, (c) uncorrelated and (d) weak negative

It is important to remember that a correlation between two variables does not necessarily imply a 'cause and effect' relationship.

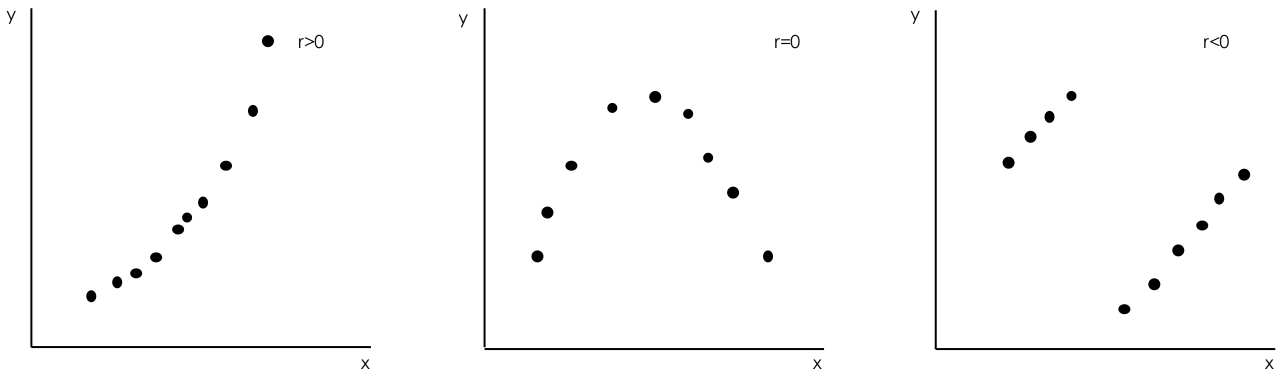


Figure 10 Examples where the use of the correlation coefficient is inappropriate

Correlation refers only to linear relationships. A correlation of 0 means there is no linear relationship between the two variables. However, relationships between variables may still exist but be non-linear (as shown in the figure above). It is always important to look at plots of data.

Linear regression

Linear regression is used when we believe a variable y is linearly dependent on another variable x. This means that a change in x will lead to a change in y. We use linear regression to determine the linear line (the regression of y on x) that best describes the straight line relationship between the two variables.

The equation which estimates the simple linear regression line is:

$$Y= a + bx$$

x is usually called the **independent**, **predictor** or **explanatory variable**;
For a given value of x, Y is the value of y (usually called the **dependent**, **outcome** or **response variable**) which lies on the estimated line. It is an estimate of the value we expect for y if we the value of x is known. Y is called the fitted value of y.

a and b are called the regression coefficients of the estimated line
a is the intercept of the estimated line; it is the average value of Y when x = 0 ;
b is the slope of the estimated line; it represents the average amount by which Y increases if we increase x by one unit.

The residual is the difference between the actual response y and the predicted response Y from the regression line. The method of least squares regression works by minimizes the sum of squared residuals. The residuals are assumed to be Normal and to have an average value of zero.

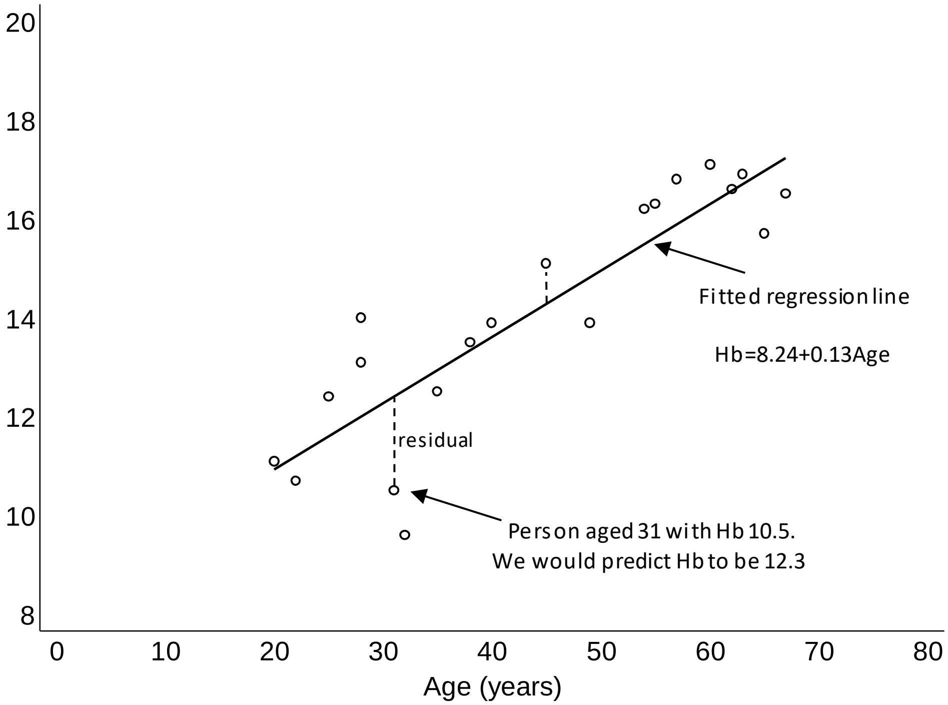


Figure 11Scatterplot of haemoglobin and age in 20 women with fitted regression line

The intercept and slope are determined by the method of least squares (often called ordinary least squares, OLS). This method determines the line of best fit so that the sum of the squared residuals is at a minimum.

Coefficients, confidence intervals, p-values, and R-squared

As mentioned previously the intercept (a) is the average value of the response when the predictor is 0. The slope (b) is the average change in the response when the predictor increases by 1 unit. If the predictor was a binary variable (for example indicating men and women) the slope (b) would indicate the average difference in response between the two groups.

The intercept (a) and the slope (b) are sample estimates of corresponding population parameters. These estimates have an inherent variability which is used to provide 95% confidence intervals for where the true population parameters may lie. The interval for the slope indicates the range, for the wider population, that the change in the response is likely to lie between as the predictor increases by 1 unit. If this interval includes 0, the coefficient is not statistically different from 0.

Each coefficient has a p-value. The p-value relates to a test of the null hypothesis that the coefficient equals 0, versus the alternative hypothesis that the coefficient does not equal 0. If p<0.05 the null hypothesis is rejected in favour of the alternative hypothesis. If p>0.05 the null hypothesis cannot be discounted.

We can assess how well the line fits the data by calculating the coefficient of determination **R-squared** (usually expressed as a percentage ranging from 0-100%), which is equal to the square of the correlation coefficient. This represents the percentage of the variability of the response that can be explained by the predictor. The higher the R-squared, the better the model.

Assumptions of linear regression

Many of the assumptions which underlie regression analysis relate to the distribution of the residuals. The assumptions are

1 The relationship between the response and predictor is approximately linear.
2 The observations in the sample are independent.
3 The distribution of residuals is Normal.
4 The residuals have constant variance.

Assumptions can be check by examining plots of the residuals. The most common method is to plot the residuals against the fitted values. This plot can show systematic deviations from a linear relationship and highlight non-constant variance. A Normal probability plot or histogram can be used to assess the Normality assumption of residuals.

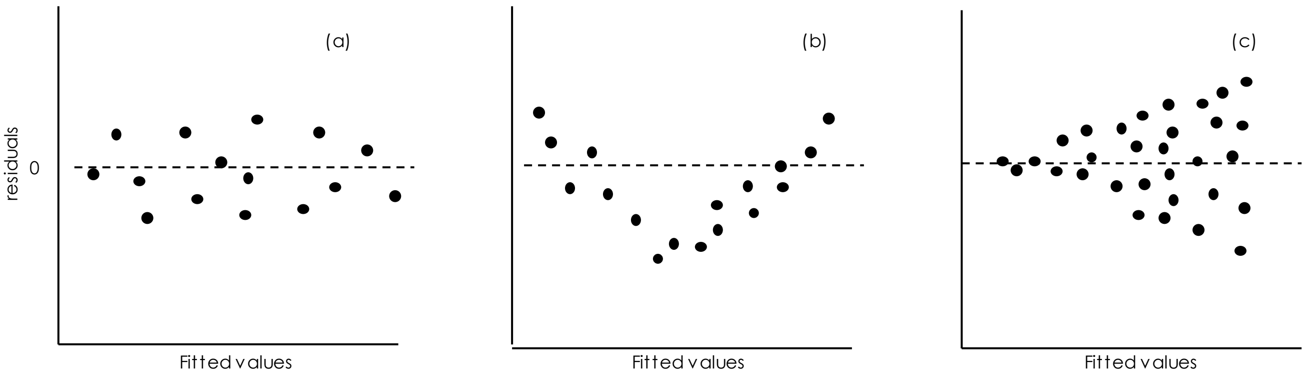


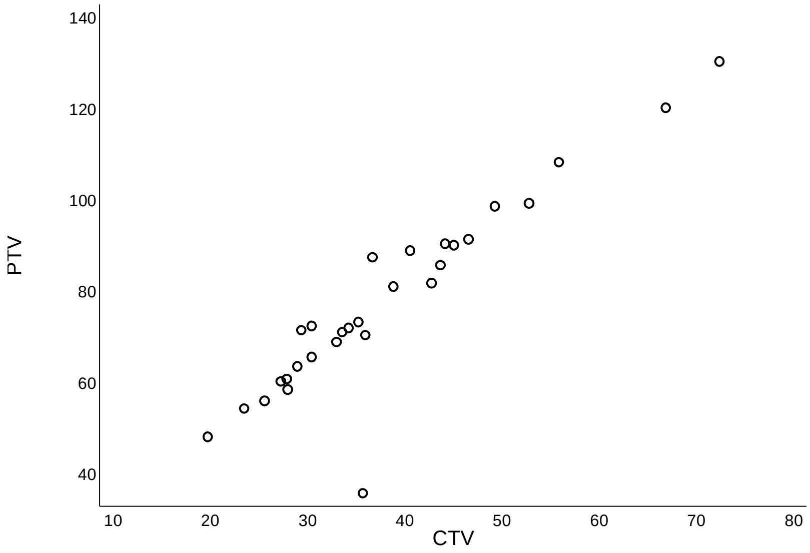
Figure 12 Scatterplots of residuals against fitted values (b) & (c) indicate assumptions of linearity and constant variance do not hold

A linear relationship means that across the range of fitted values, the residuals are spread equally above and below 0. The assumption does not hold in Figure 12 (b).

Constant variance of the residuals means that in a plot of residuals against fitted values, the spread of the residuals doesn't change. The assumption does not hold if the vertical spread of the residuals changes across the plot as in Figure 12 (c).

Example

The clinical target volume (cm^3) (CTV) and planning target volume (PTV) were recorded for 29 patients receiving stereotactic radiotherapy for prostate cancer. The question of interest was what the relationship between PTV and CTV was, and could the average PTV be predicted when the CTV is known.



The equation of the line is

$$PTV=16.96 + 1.58xCTV$$

This means that as the CTV increases by 1cm^3 the PTV increases by 1.58cm^3.
The 95% confidence interval for the slope is (1.32, 1.84). In a wider population of similar prostate cancer patients, as the CTV increases by 1cm^3, the PTV is likely to increase by between 1.32 and 1.84cm^3. The R-squared for the model was 0.85. This means that 85% of the variation in PTV was explained by CTV alone. 15% remains unexplained.

Multiple linear regression

Multiple linear regression is an extension of simple linear regression. We would use multiple linear regression when we want to predict a response using several predictor variables. For example, we may wish to predict respiratory muscle strength from weight, age, height and sex.

The assumptions of multiple linear regression are the same as those for simple linear regression. Interpretation of coefficients is very slightly different:

The intercept (a) is the average value of the response when all the predictors have values equal to zero.

If a predictor is continuous it's slope (b) indicates the average change in the response when all the other predictors are held constant.

If a predictor is binary (for example indicating men and women) the slope (b) represents the average difference in the response between the groups when all other predictors are held constant.

When multivariable regression is used a variation of the R-squared called the **adjusted R-squared** is employed to assess the fit of the model. The **adjusted R-squared** takes account of the number of predictors used in the model, but its interpretation is the same as for R-squared.

Odds, Odds ratios and logistic regression

Odds are simply another way of describing probability. Odds are calculated by dividing the number of times an event happens by the number of times it does not happen.

If one in every 100 patients suffers a side-effect from a treatment, the odds are

$$1:99 = 1/99 = 0.0101$$

Risk, on the other hand indicates the probability that an event will happen. It is calculated by dividing the number of events by the number of people at risk. In the example above the risk would be

$$1/100=0.01$$

Odds ratios are calculated by dividing the odds in one group of patients (e.g. experimental group) with the odds in a comparison group of patients (control group).

An odds ratio of 1 indicates no difference in risk between the groups, i.e. the odds in each group are the same.
If the odds ratio of an event is >1, the rate of that event is increased
If <1, the rate of that event is reduced.
Odds ratios are frequently given with 95% confidence intervals – if the confidence interval for an odds ratio does not include 1 (no difference in odds), it is statistically significant.

Logistic regression is similar to linear regression but is used when the outcome variable is binary (e.g. having a disease or not) as opposed to continuous. The coefficients in a logistic regression are interpreted as odds ratios. The coefficients indicate the percent change in the odds of the event when a unit change in the explanatory variable occurs.

Screening & Diagnostic tests

In clinical practice it is desirable to have a simple test which, depending on the presence or absence of an indicator (for example, fecal occult blood), provides a good prediction to whether or not a patient has a particular condition (for example, colorectal cancer).

To evaluate a potential diagnostic test, we apply the test to a group of individuals whose true disease status is known. We then draw up a 2 × 2 table of frequencies

	True Disease Status		
	Disease	No disease	Total
Test result			
positive	a (true positive)	b (false positive)	a+b
negative	c (false negative)	d (true negative)	c+d
Total	a+c	b+d	n=a+b+c+d

Of the n individuals studied, a + c individuals have the disease and b+d do not.

Sensitivity, **specificity** and **predictive values** are measures for assessing the effectiveness of the test

Sensitivity is the proportion of individuals with the disease who are correctly identified by the test.

$$sensitivity = \frac{a}{a+c}$$

Specificity is the proportion of individuals without the disease who are correctly identified by the test.

$$specificity = \frac{d}{b+d}$$

Sensitivity and specificity quantify the diagnostic ability of the test.

Positive predictive value is proportion of individuals with a positive test result who have the disease

$$positive\ predictive\ value = \frac{a}{a+b}$$

Negative predictive value is proportion of individuals with a negative test result who do not have the disease

$$negative\ predictive\ value = \frac{d}{c+d}$$

The predictive values indicate how likely it is that the individual has or does not have the disease, given the test result.

Predictive values are dependent on the **prevalence** of the disease in the population being studied. Prevalence is the proportion of the population who have the disease.

The **prevalence** of the disease in the sample is $\frac{(a+c)}{n}$

In populations where the disease is common, the positive predictive value of a given test will be higher than in populations where the disease is rare.

The **likelihood ratio** (LR) for a positive test result is the ratio of the probability of a positive result if the patient has the disease (sensitivity) to the probability of a positive result if the patient does not have the disease (1-specificity).

$$likelihood\ ratio = \frac{sensitivity}{1 - specificity}$$

For example, a LR of 4 for a positive result indicates that a positive result is four times as likely to occur in an individual with the disease compared to one without it.

Cut off values

Sometimes a diagnostic test needs to be performed on the basis of a continuous numerical measurement. Often there is no threshold above (or below) which the disease definitely occurs. In this situation, a cut-off value is identified at which it is believed an individual has a very high chance of having the disease.

The receiver operating characteristic (ROC) curve provides a way of assessing an optimal cut-off value for a test. A ROC curve plots sensitivity against 1- specificity at all potential cut-off points. It essentially compares the probabilities of a positive test result in those with and without disease. The overall accuracy can be assessed by the area under the curve (AUC).

Example

	True Disease Status		
	Prostate cancer	No prostate cancer	Total
PSA Test result			
positive (≥2.1ng/ml)	167	508	675
negative (<2.1ng/ml)	282	1993	2275
Total	449	2501	2950

Table 2 Prevalence of prostate cancer among men with PSA<4.0ng/ml

$$sensitivity = \frac{167}{167 + 282} = 0.37;\ specificity = \frac{1993}{508 + 1993} = 0.80;$$

Using this test, if prostate cancer is present there is a 37% chance of detecting it. If there is no prostate cancer, there is a 80% chance of a negative result. 20% of people will have a false positive result.

$$positive\ predicitive\ value = \frac{167}{167 + 508} = 0.25;\ negative\ predicitive\ value = \frac{1993}{282 + 1993} = 0.88$$

There is a 25% chance that if the test is positive the patient actually has prostate cancer. There is a 88% chance, if the test is negative, that the patient does not have prostate cancer. This means there is a 12% chance of a false negative result.

$$likelihood\ ratio = \frac{0.37}{1 - 0.80} = 1.85$$

If the test is positive, the patient is 1.85 times (almost twice) as likely to have prostate cancer as not have it.

Survival analysis

Survival analysis involves the use of data which measures the time to an ‘event’. In the context of cancer, the event could be death and the survival data might record time from cancer diagnosis to death. Other survival events of interest can include time until disease recurrence, or time to re-hospitalisation following discharge.

Survival data are concerned with recording the length of time for a patient to reach the specific endpoint, rather than simply recording whether the end point was reached. It therefore involves two variables – time to the event and whether the event occurred.

One problem with survival data is that there is often incomplete follow-up for patients. This is known as **censoring.** Censoring arises when a study is finished before all patients experience the event (which could be death) or when patients have to be excluded from the study due to other reasons (migration, lost to follow-up, other adverse event).

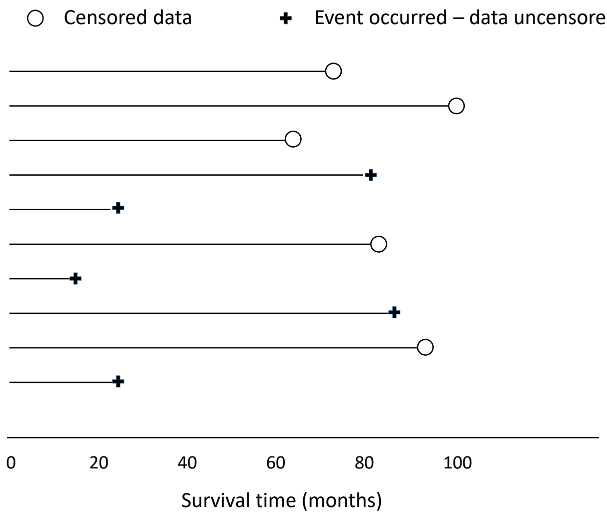


Figure 13 Schematic of a survival analysis study. Each line represents a subject.

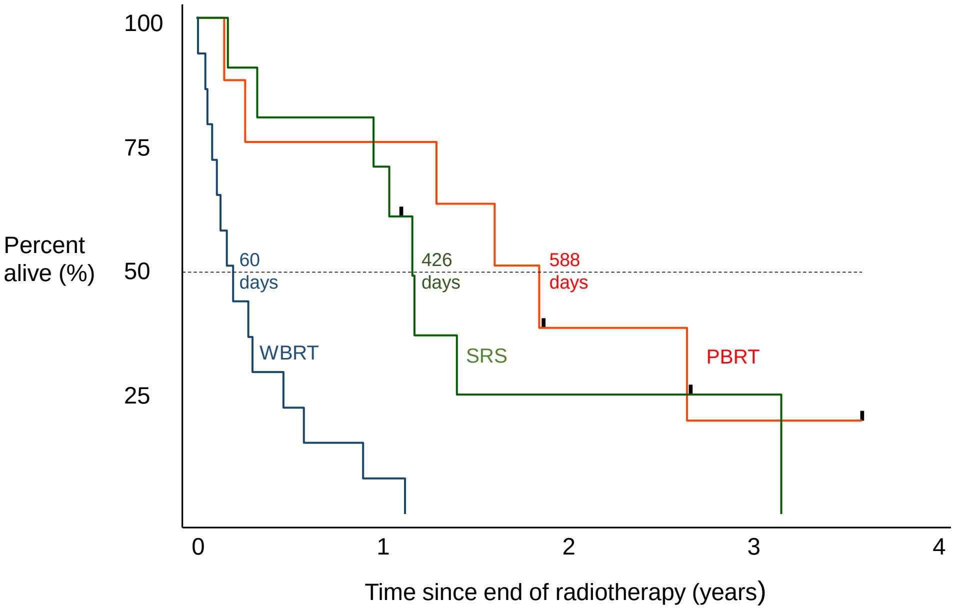
A comparison of the mean time to reach the endpoint in patients can give misleading results because of censored data. Therefore, a number of statistical techniques, known as survival methods, have been developed to deal with these situations

The aim of these techniques is to statistically describe survival times, and compare survival over several therapy groups (the idea being the longer the survival times the better the therapy).

It is also important to find relations between survival times and other prognostic variables (for example, age, stage, severity of disease).

Kaplan-Meier survival curves

Survival data can be displayed graphically in a Kaplan-Meier plot. The vertical axis displays the proportion of patients remaining free of the endpoint at any time after baseline. Each event (death) is indicated by a step in the curve and censored data are indicated by (+). The resulting curve is therefore a series of steps, starting at a survival probability of 1 (or 100%) at time 0 and drops towards 0 as time increases. In effect it shows the proportion of people who are still at risk of experiencing the event while the horizontal axis shows the time that the subjects were followed up for. The plot can be used to describe time-to the event in one group, or to compare time-to-event among different groups (for example, by treatment, or sex).



WBRT whole brain radiotherapy (14 patients), SRS stereotactic radiotherapy (10) , PBRT partial brain radiotherapy (8)

Figure 14 Kaplan Meier curves for overall survival among renal cancer patients with brain metastases treated with radiotherapy

Figure 9 shows overall survival among 32 renal cancer patients with brain metastases who were treated with different types of radiotherapy. At the start of the study, no patients have died and the proportion at risk is 1.0 (100%). At each time point where a line drops, at least 1 patient experiences the event (death).

The dashed horizontal line at 50% indicates the **median survival time** in each treatment group. The median survival time is a useful summary measure that indicates the time at which 50% of patients have experienced the event. In this example, the median survival time for patients selected to receive WBRT was 60 days, and 426 days for patients selected to receive stereotactic radiotherapy.

Log rank test

The most common (non-parametric) method of comparing survival between independent groups is the **log rank test**. In the example above the null hypothesis would be that the groups have equal median survival times (i.e. there is no difference in median survival between the groups). The alternative hypothesis is that at least one of the groups is different. In this example the p-value was <0.0001. Since this is less than 0.05, the null hypothesis is rejected and we can conclude that there is evidence of at least one difference in the median survival times in patients selected to receive these treatments.

Extensions to the log rank test exist. The **stratified log rank** test can adjust for categorical prognostic covariates (for example age group or sex). The **log rank test for trend** is used for ordered groups (for example, cancer stage, or number of metastases). This is a more appropriate test for considering a trend in survival across the groups.

Life tables

Life tables describe survival data, where the results have been grouped into time intervals, often of equal length. This method is often described as **actuarial**. The method of calculation is similar to the Kaplan-Meier method, but differences arise because of the lack of precision of recording of survival times. In general the Kaplan-Meier analysis is recommended.

Cox proportional hazards model

The log rank test is solely a hypothesis test, comparing survival in two or more groups. It does not allow the relationships between one or more categorical or continuous factors and survival to be quantified. To do this we employ the hazard rate (sometimes called the failure rate) which is closely related to survival curves. The hazard rate represents the risk of dying in a very short time interval after a given time **t**, assuming survival to time **t**. We want to know whether there are any systematic differences between the hazards, over all time points, of individuals with different characteristics.

This can be achieved using the **Cox proportional hazards model** to test the independent effects of a number of explanatory variables on the hazard. The method assumes that the effect of a predictor on survival does not change over time.

For each variable the Cox proportional hazards model produces an output which includes a hazard ratio with a confidence interval, and a p-value.

Hazard ratios above one (>1) indicate a raised hazard, and suggest shorter survival times. The hazard increases by 100*(Hazard ratio-1)%.

Values below one (<1) indicate a decreased hazard, and suggest longer survival times. The hazard decreases by 100*(1-Hazard ratio)%.

Values equal to one indicate that there is no increased or decreased hazard of the endpoint. The p value relates to a significance test performed to test whether that hazard ratio is different from one.

For a continuous predictor the hazard ratio describes what happens to the hazard as the predictor changes by 1 unit of measurement.

For levels of a categorical predictor the hazard ratio describes the relative difference in the hazard between a level and a chosen reference level.

Using data from the previous example on brain metastases among renal cancer patients a proportional hazard model produces

Patient characteristics		**Hazard ratio (95%CI)**		**p-value**
Sex	*men*	1		
	women	1.33	(0.24- 2.36)	0.54
Age (years)		1.02	(0.97-1.08)	0.40
Number of metastases	*1*	1		
	2+	1.68	(0.45-6.19)	0.43
Treatment	*SRS*	1		
	WBRT	7.11	(1.85-27.3)	0.004
	PBRT	0.75	(0.23-2.36)	0.62

The reference group is men for the sex variable. After accounting for age, treatment and number of brain metastases, the hazard for women is higher than the hazard for men by 33% (1.33-1). The p-value is greater than 0.05 and the confidence interval includes 1, so there is no evidence to suggest an association between sex and hazard in the wider population.

Age is a continuous variable. After accounting for sex and treatment increasing age is associated with increasing hazard. A one year increase in age is associated with a 2% (1.02-1) increase in hazard. However, the p-value is above 0.05 and the confidence interval includes 1, so there is no evidence to suggest an association between age and hazard in the wider population.

The treatment reference group is SRS. After accounting for age and sex, the hazard for WBRT is statistically significantly higher than the hazard for SRS. Hazard for WBRT patients is 611% (7.11-1) higher than for SRS. This increase is statistically significantly different from 0. In a wider population of similarly chosen patients, the increased hazard is likely to be between 85% and 2630%. For patients chosen to receive PBRT the hazard is 25% (1-0.75) lower than SRS, but is not significantly different from 0. In a population of similar patients, the difference is likely to be between 77% and 136%.

Epidemiological Studies

Epidemiology is the study of the occurrence and determinants of ill health in the population. Epidemiological studies, assess the relationship between factors of interest (these may include biological, social, behavioral, and environmental) and the occurrence of disease in the population (for example the incidence of cancer, heart disease, hypertension). Epidemiological studies are mostly **observational** in design (in contrast to **experimental** studies which involve interventions to affect an outcome).

Routinely collected administrative data such as death certification, cancer registration and hospital discharge records can be used for epidemiological purposes and provide insights into the health of the population. Death certification provides estimates of the annual death rates in the population. These rates are usually age standardized to take account of differences in the age structure of the population over time or between places.

Observational epidemiological studies fall into three broad groups – **cross-sectional**, **cohort studies** and **case control** studies. The unit of observation is usually individuals but can also be groups of individuals (for example – different populations defined by a shared geography – these are called **ecological** studies). Studies in which the health event of interest has yet to happen are called **prospective**. Studies in which the health event has already occurred are called **retrospective**.

Cross-sectional study

A cross-sectional study is carried out at a single point in time. A health survey is a type of cross-sectional study where the aim is to describe health behaviours or health status in a large sample of the population. A cross-sectional study is suitable for estimating the **prevalence** of a condition in the population. **Prevalence** is the proportion (or percent) of individuals with a particular condition in the population at a point in time

Cohort study

A cohort study takes a group of individuals and follows them forward in time. These studies are usually prospective. The aim is to assess whether exposure to a particular factor affects the **incidence** of disease in the future. **Incidence** is the number of new cases of a condition occurring in a population over a set time period. The **incidence rate** is the number of new cases divided by the person time at risk, and is usually expressed in terms of person years.

The analysis of cohort studies can be summarised by the ratio of incidence rates in the exposed and non-exposed groups (**incidence rate ratios** *or* **relative risk)**.

The incidence rate in each group (exposed and unexposed) can be calculated from the following table

	Disease of interest			
	Yes	No	Total	Incidence rate
Exposed to factor				
Yes	a	b	a+b	a/(a+b)
No	c	d	c+d	c/(c+d)
Total	a+c	b+d	n=a+b+c+d	

The **relative risk** (RR) indicates the increased (or decreased) risk of disease associated with exposure to the factor of interest.

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Relative Risk (RR)	Interpretation
>1	an increased risk in the exposed group
1	risk is the same in the exposed and unexposed groups
<1	a reduced risk in the exposed group

A relative risk of one indicates that the risk is the same in the exposed and unexposed groups. A relative risk **greater than one** indicates that there is ***an increased risk*** in the exposed group compared with the unexposed group; a **relative risk less than one** indicates a reduction in the risk of disease in the exposed group.

	Disease of interest		
	Anaemic	Not Anaemic	Total
Exposed to factor			
Serum ferritin<20	7	8	15
Serum ferritin>20	2	13	15

Table 3 Cohort study of serum ferritin and anaemia

The table shows the results of a cohort study investigating low serum ferritin and development of anaemia.

$$RR = \frac{7/(7+8)}{2/(2+13)} = 3.5$$

This means that the risk of developing anaemia among women with low serum ferritin is 3.5 times the risk among women who do not have low serum ferritin. Another way of presenting this is that the risk is 250% (3.5-1) higher.

Case-control study

A case–control study compares the characteristics of a group of patients with a particular disease (the cases) to a group of individuals without the disease (the controls), to see whether exposure to a factor occurred more or less frequently in the cases than the controls

Because patients are selected on the basis of their disease status, it is not possible to estimate the risk of disease. For cases and controls we can estimate the odds of being exposed to the risk factor.

	Disease of interest		
	Case (disease)	Control (no disease)	Total
Exposed to factor			
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n=a+b+c+d
Odds of exposure	a/c	b/d	

The **odds ratio** (OR) gives an indication of the increased (or decreased) odds associated with exposure to the factor of interest. An odds ratio of one indicates that the odds is the same in the exposed and unexposed groups; an odds ratio greater than one indicates that the odds of disease is greater in the exposed group than in the unexposed group,

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Odds ratio	Interpretation
<1	a reduced odds of disease in the exposed group
1	odds is the same in the exposed and unexposed groups
>1	an increased odds of disease in the exposed group

	Breast cancer	
	Case	Control
Oral contraceptives		
Ever used	537	554
Never used	639	622
Total	1176	1176

Table 4 Case-control study of oral contraceptives and breast cancer

The above table shows results from a case-control study of oral contraceptives and breast cancer. The cases were women recently diagnosed with breast cancer in a certain hospital. The controls were women inpatients in the same hospital.

In this example the odds ratio for contraceptive users and breast cancer patients is

$$OR = \frac{537/639}{554/622} = \frac{537x622}{554x639} = 0.94$$

The OR<1 and this indicates that the odds of breast cancer patients using contraceptives is 6% (1-0.94) smaller than among controls. Another interpretation is that the odds of contraceptive users developing breast cancer is 6% smaller compared to those who do not use contraceptive.

Pros and cons of cohort studies and case control studies

Case control studies are useful for investigating rare diseases. The selection of appropriate controls (who we want to be as similar to controls) however can be difficult. When exposures are rare it is not a very efficient type of study. Because they are retrospective in nature, they cannot be used to establish incidence and are subject to recall bias and data inaccuracy.

Prospective cohort studies need to follow up subjects over a long period of time. They are expensive, and prone to subject dropping out (loss to follow up). They are not efficient for rare diseases. Data recording tends to be more accurate, multiple outcomes can be studied, and incidence rates can be established.

Clinical trials

A clinical trial is a planned **experimental study** on humans designed to evaluate new interventions (e.g. type or dose of drug, or surgical procedure) compared to a comparative treatment.

There are several different stages of clinical trials. **Phase I & II** trials are pre-clinical or small studies which investigate treatment effects and safety. They establish dosage, side effects and delivery mechanisms. A **Phase III** trial is a full evaluation of the new treatment compared to a comparative treatment. After a treatment has been approved and licensed for general use **Phase IV** trials observe how the treatment works in a non-trial setting, identify long term effects and rare side effects.

Phase III trials have at least one treatment group and comparator group (the control group). If the condition under investigation has a standard treatment then this is the treatment that the control group may receive. If standard care does not exist, then the control might be given a **placebo** (a treatment which does not consist of an active component) or no treatment if considered ethical. The purpose of the control group is to quantify the effect of the treatment by comparing the outcome of interest in the control group to the outcome in the treatment group.

Treatment allocation

Patients are usually **randomised** to treatment groups. This is to avoid systematic bias, and ensure that each patients has an equal chance of being allocated to each treatment. There are various ways of achieving this which include **simple randomization; block randomization -** which ensures that after the entry of every x patients into the trial, the number of patients on each treatment will be equal; and **stratified randomization -** which uses block randomisation to balance treatment allocation across important prognostic factors (such as age and stage). **Cluster randomization** -randomly allocates a group or cluster of individuals, rather than each individual, to a treatment (for example patients within a GP practice).

Allocation concealment

Randomization is not sufficient to ensure that a trial is unbiased. If outcome assessment is subjective or open to interpretation systematic bias can be introduced by knowledge of the treatment received. Masking or blinding means that people are unaware of the treatment that someone received. There are three levels of blinding - **single blind** – the patient does not know which treatment they have been allocated, **double blind** – neither patient nor doctor/evaluator knows which treatment has been allocated, **triple blind** – neither the patient, nor the doctor, nor those reviewing the interim results know which treatment the patient has been allocated.

Designs for randomized trials

Parallel groups

The simplest form of randomised trial is a parallel group trial. Eligible patients are randomised to two or more groups, treated according to the assigned group, and assessed for their response to treatment.

Factorial designs

In a factorial trial, two (or more) intervention comparisons are carried out simultaneously. For example, in a trial for surgical patients with colorectal cancer those who participate might be randomised to receive peri-operative radiotherapy or not, and also randomised to receive local regional-chemotherapy or not. Most factorial trials have two 'factors' (in this example radiotherapy and chemotherapy), each of which has two levels (in this example they received the treatment or they did not). This is called a 2×2 factorial trial, and gives 4 combinations of treatment - no radiotherapy and no chemotherapy, radiotherapy alone, chemotherapy alone, and radiotherapy and chemotherapy in combination.

Potential problems exist if treatment effects are not additive. It must also be practical to combine the treatments, and the toxicity of combined treatment must be acceptable.

Cross-over trials

In a cross over trial, every patient receives all treatments under investigation, but the order in which they receive them is randomised. In the case of a two-treatment cross-over trial, eligible patients are randomised to receive either treatment A followed by treatment B, or B followed by A. The benefit is that each patient acts as their own control, effectively leading to a **smaller sample size** - halving the number of patients when compared to a conventional parallel design.

Potential problems include drop-out after the first treatment, which may be related to treatment; carry-over of treatment effects from the first period which is not eliminated by a wash-out period; or treatment period interaction – in which the effect of a treatment is substantially different in the two periods.

Analysis of clinical trial data

Randomised controlled trials often suffer from noncompliance, drop outs and missing outcomes. An analysis based only on those patients who completed the study without protocol violations (per-protocol population) can introduce bias and lead to an overestimation of effectiveness. This is because the reason why patients do not comply or drop out may be related to the treatment they received. One potential solution is called **intention-to-treat (ITT)** analysis. ITT analysis includes every subject who was randomised according to the randomised treatment assignment. It ignores noncompliance, protocol deviations, withdrawal, and anything that happens after randomisation. In ITT analysis, the estimate of treatment effect will generally be more conservative than that found with a per-protocol analysis.

Numbers needed to treats (NNT)

The **NNT** is the number of patients that need to be treated for one patient to benefit from a treatment compared to a control treatment. It is defined as 1/(*absolute risk reduction)*. The **absolute risk reduction** is the risk in control group *minus* the risk in treatment group. A NNT of 1, would mean that everyone improves with treatment and no one improves with the control. The higher the NNT, the less effective is the treatment.

Consort statement

CONSORT stands for Consolidated Standards of Reporting Trials. The CONSORT statement is a set of recommendations for the standardised reporting of randomised trials. It is used to aid transparency, critical appraisal and interpretation.

The CONSORT statement contains a 25-item checklist and a flow diagram. The checklist ensures that information relating to trial design, participants, treatment allocation, analysis, interpretation and limitations is included in the study report. The flow diagram displays for each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome, along with, losses and exclusions after randomisation.