Dissertation Module: Research Skills Program Topic 7: STATISTICS WITH CONFIDENCE

Why Statistics?

As health professionals we apply statistics because it provides us with (1) numbers - an unambiguous language understood by everybody worldwide in the same way; and (2) tools to measure the uncertainty or **random error** involved when inferring from a sample to the wider target population.

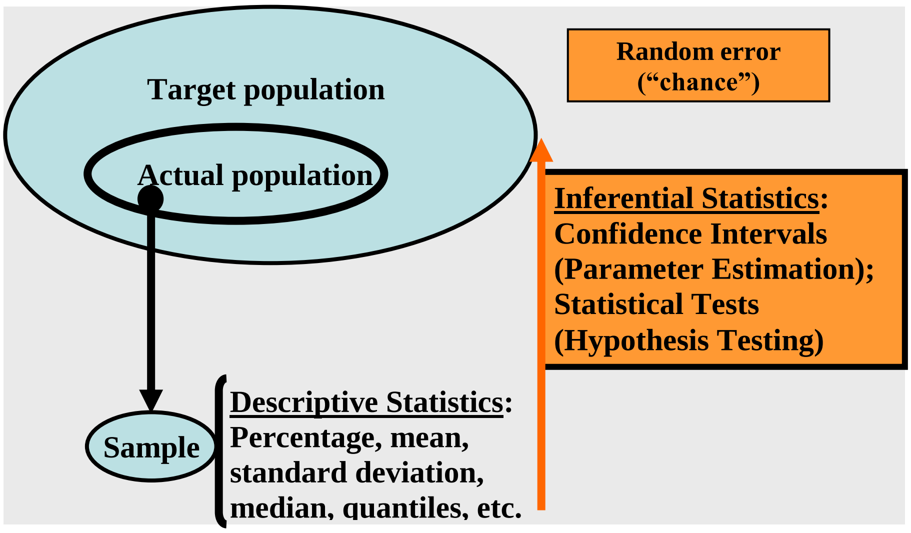
The universal language of numbers allows us to describe data in a precise manner. Consider for instance the following clear-cut statement (based on McGervey, J.D., 1986): For every 10 Million inhabitants, in 1979, 4 persons in Japan, 16 in Canada, 87 in Israel, 3 in Sweden, 5 in Germany, 1 in England, and 383 in the United States were killed by handguns." The numbers in this sentence tell the full story. If we tried to use words instead of numbers we would end up with something rather vague as in: "*Some* people got killed by handguns in Japan, *more* in Canada, *still more* in Israel, *far fewer* in Sweden and Germany, *really few* in England, and *very many* in the United States."

However, using statistics does more: it also enables us to generalize from the sample to the target population. This part of statistics is called **inferential statistics**. The ability to assess the uncertainties involved when deducing from a sample to the target population is one of the main reasons for the ongoing success story of statistics in the health sciences. Please note that statistics allows us to confirm or reject our research hypothesis with statistical confidence. Inferential statistics provide tools that allow us to answer our research questions.

Statistics are a core tool of the **scientific method** which we introduced briefly in Chapter 1 and which helps with acquiring knowledge in a standardized and universally accepted manner. Statistics form the basis of today's gold standard of evidence-based practice. Statistics not only influence the planning phase of an epidemiological study by providing an optimal sample size (for a stated quantitative research hypothesis) but will also later in the course of the study inform the collection, analysis, presentation, and interpretation of the data.

According to Figure 1 there are two main parts of applied statistics: descriptive and inferential statistics. Descriptive statistics are the tools that allow us to describe our sample. On the other hand, inferential statistics enable us to draw conclusions about the target population, based on results gathered in the sample.

Figure 1: The use of statistics in epidemiology: descriptive and inferential statistics.


Target population
Actual population
Sample
Random error ("chance")
Inferential Statistics: Confidence Intervals (Parameter Estimation); Statistical Tests (Hypothesis Testing)
Descriptive Statistics: Percentage, mean, standard deviation, median, quantiles, etc.

Types of quantitative data

All quantitative research collects quantitative data, that is, information of characteristics which can be directly measured in numbers or coded into numbers. These characteristics are called **variables.** For example, in a study on pre-pregnancy physical activity and its effect on gestational diabetes mellitus (based on van der Ploeg, H.P. et al., 2010) the characteristics could include: age of the mother, parity, height of mother, weight of mother, pre-pregnancy level of physical activity, and gestational diabetes. The information collected during the study is called the **data**. The correct choice of a statistical method to analyse the data is dependent on the type(s) of the variable(s) collected.

Box 7.1: The types of quantitative data

We broadly differentiate between categorical and numerical data.

Categorical data arise when people fall into one of separate categories.
For example: Gender has the categories male and female; Gestational diabetes has the : categories yes and no; Blood group usually has the categories A, B, AB and 0; Level of education can be recorded as less than 12; Year 12 completed; apprenticeship or TAFE, tertiary education.

Categorical data can be further classified as being either **nominal** or **ordinal**. In ordinal categorical variables the categories follow a natural order as in level of education: up to year 12; Year 12 completed; apprenticeship or TAFE, tertiary education.
Nominal categorical variables are without such an inherent order as in gender and blood group: the categories male and female or O, A, B, and AB do not have a natural order.

Numerical data arise from counts or measurements. For example: Age; Number of children; Height; Weight; Level of physical activity.

Numerical data can be further classified into **discrete** or **continuous**. Discrete numerical variables are usually natural counts and can only take on natural, whole numbers. For example, number of children: 0, 1, 2, 3,….or number of sexual partners: 0, 1, 2, 3,……
Continuous variables are measurements that can take on any real number, including numbers with decimal places, within a meaningful range. The values recorded for a continuous variable are only limited by our choice of precision. Examples for continuous variables are reaction time, age, or weight.

Comments:

(1) The codes given to the categories of nominal variables for the purpose of statistical analysis are arbitrary. For example, gender codes could be 1 = male and 2 = female; or vice versa or any other dichotomous code combination! With ordinal variables, the coding should reflect the natural order.

(2) In stark contrast, the observed numbers in numerical variables are measurements with an intrinsic meaning! For example, height = 175 means 175 cm. Numerical variables therefore retain quantitative information of measurements.

(3) Genuinely continuous variables are often recorded as discrete or even categorical variables. For example, the age of a person might be 25 years, 1 month, 16 days, 14 hours, etc. However 25 years is recorded (“age at last birthday”) thus creating a discrete variable. In another example we might even turn the - naturally continuous - body mass index into a categorical variable, by recoding the body mass index as underweight, normal weight, overweight, and obese. This categorization of a continuously recorded variable however should only be done during the analysis phase. If we only record the categories of body mass index, then we might not be able to compare our findings with published average values or a differing category system used in another country.

Please note: always collect and record numerical information as such and as precise as possible. Do not only record categorized information for a continuous variable! Categorisation can easily be carried out during statistical analysis.

Descriptive statistics

The first step in statistical analysis is a thorough description of the data providing the researcher with a good understanding of the results and allowing the reader to appreciate the sample. As already discussed in some length, results from a study might not be automatically transferable to other populations or situations. For example, results from a study conducted with pregnant women, aged 23 to 42 years, living in Melbourne might not be transferable to pregnant women, aged 18 to 35 years from rural Western Australia. A concise description of the sample in a publication will allow us to assess whether the conclusions reached by the research are likely to be applicable to our own situation.

Correct descriptive statistics summarise the collected data in a meaningful way. Which descriptive statistics are used is dependent on the type of variables (see Box 7.1). Descriptive statistics for categorical variables usually only describe the frequency of each observation as percentages, for example, “32% of the study participants were female”. Numerical variables are summarised using a **measure of central tendency** together with a **measure of dispersion**. The most frequently used measures of central tendency are the **arithmetic mean** and the **median**. These measures try to identify the centre of the distribution of the numerical data. The arithmetic mean is the sum of the recorded values divided by the number of values; the median is the value in the middle of the ordered list of all observations.

Box 7.2a: How to calculate an arithmetic mean?

The arithmetic mean of a sample of n values x_1, x_2, x_3, …., x_n is:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

For example:

Assume we want to calculate the average age of 9 participants from the study by van der Ploeg, H.P. et al. (2010) on pre-pregnancy level of activity and gestational diabetes mellitus. The ages of the 9 mothers were 21, 32, 35, 29, 27, 27, 24, 30, and 28.

The mean age is therefore:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{21 + 32 + 35 + 29 + 27 + 27 + 24 + 30 + 28}{9} = \frac{253}{9} = 28.1$$

Box 7.2b: How to calculate a median?

The median ($x_{0.5}$) of a sample of n values x_1, x_2, x_3, …., x_n sorted in ascending order is dependent on whether the sample size (n) is odd or even.

(1) If the sample size n is odd, then $x_{0.5} = \frac{(n+1)}{2}$ **th largest observation**

For example:

In order to calculate the median age of the 9 mothers we first have to sort their ages: 21, 24, 27, 27, 28, 29, 30, 32, 35.

Nine is an odd number. Hence the median value is the $\frac{9+1}{2} = 5^{th}$ observation, which is 28 years which is the value in the middle of the nine ordered observations.

(2) If the sample size n is even, then $\mathbf{x_{0.5}}$ **= average of ($\mathbf{\frac{n}{2}}$)th + ($\mathbf{\frac{n}{2}}$+1)th largest observation**

For example:
Let us assume that there were 10 mothers and the “new” mother is aged 25 years. In order to calculate the median age of the 10 mothers we have to sort their ages: 21, 24, 25, 27, 27, 28, 29, 30, 32, and 35.
Ten is an even number. Hence the median value is the average of the 5th and 6th observation, which is (27 + 28)/2 = 27.5 years. This value is again in the middle of the distribution.

Comment:
The median is the middle value of a distribution; that is, 50% of the values are below and 50% of the values are above the median. Statisticians call the median the 50%-quantile of a distribution.

A measure of central tendency is usually accompanied by a measure of dispersion indicating the spread of the data. If the arithmetic mean is used as the measure of central tendency, then the **standard deviation** (SD) is the measure of variability of choice. If the median is used, the 25% and 75% quantiles - also called the **inter-quartile range** (IQR) - are used as the measures of variability. Providing minimum and maximum values, that is, the **range** is sometimes informative, in particular when the observed values show little variation.

Box 7.3a: How to calculate a standard deviation?

The basic idea of the standard deviation is measuring the average distance between the mean value and each individual observation.
The standard deviation (SD) of a sample of n values $x_1, x_2, x_3, \ldots, x_n$ is:

$$\mathbf{SD} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{X} - x_i)^2}{n - 1}}$$

That is, the distances of each individual value from the mean value are squared and then summed up; the Greek letter Σ stands for sum up. The result is divided by the sample size minus 1 and then square rooted.

For example:
Assume we want to calculate the standard deviation of age of the 9 participants from the study based on van der Ploeg, H.P. et al. (2010). The ages of the 9 mothers were

21, 32, 35, 29, 27, 27, 24, 30, and 28; and the mean age was 28.1 years.
The standard deviation of age is therefore:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(\bar{X} - x_i)^2}{n - 1}} = \sqrt{\frac{(28.1-21)^2 + (28.1-32)^2 + (28.1-35)^2 + ... + (28.1-28)^2}{8}} = 4.137$$

We can summarise the age of the participating mothers as mean age of 28.1 (SD 4.1) years.

Box 7.3b: How to calculate the 25% and 75% quantiles?

In general, the calculation of the p-quantile of a distribution is dependent on whether n x p is an integer (= whole number) or not. In this context n is the sample size and p is a number between 0 and 1; p=0.5 for the median.

(1) **If n x p is an integer, then the p-quantile x_p is the average of the (n x p)th and the (n x p + 1)th largest observation.** The observations are sorted in ascending order.

<u>For example:</u>
Let us assume that there were 8 mothers, age sorted in ascending order: 21, 24, 27, 27, 28, 29, 30, and 35 years.
In order to calculate the 25% and the 75%-quantile for age of these 8 mothers we need to look at n x p.
For the 25th-quantile (denoted as $x_{0.25}$), n x p = 8 x 0.25 = 2 is a whole number. Hence the 25th-quantile is the average of the 2nd and 3rd largest observation: (24+27)/2 = 25.5
For the 75th-quantile ($x_{0.75}$), n x p = 8 x 0.75 = 6 is a whole number. Hence the 75th-quantile is the average of the 6th and 7th largest observation: (29+30)/2 = 29.5

The inter-quartile range of age for this sample of 8 women is (25.5 -29.5); the median age is 27.5 years. We can summarize age in this example as median age = 27.5 years (IQR = 25.5 - 29.5).

(2) **If n x p is not an integer, then the p-quantile x_p is the kth largest observation with k being the first integer larger than n x p.** The observations are sorted in ascending order.

<u>For example:</u>
Let us assume that there were 9 mothers age sorted in ascending order: 21, 24, 27, 27, 28, 29, 30, 32, and 35 years.
In order to calculate the 25% and the 75%-quantile for age of these 9 mothers we need to look at n x p.
For the 25th-quantile ($x_{0.25}$), n x p = 9 x 0.25 = 2.25, not a whole number. Hence the 25th-quantile is the 3rd largest observation (k=3 is the first integer larger than 2.25): 27.
For the 75th-quantile ($x_{0.75}$), n x p = 9 x 0.75 = 6.75, not a whole number. Hence the

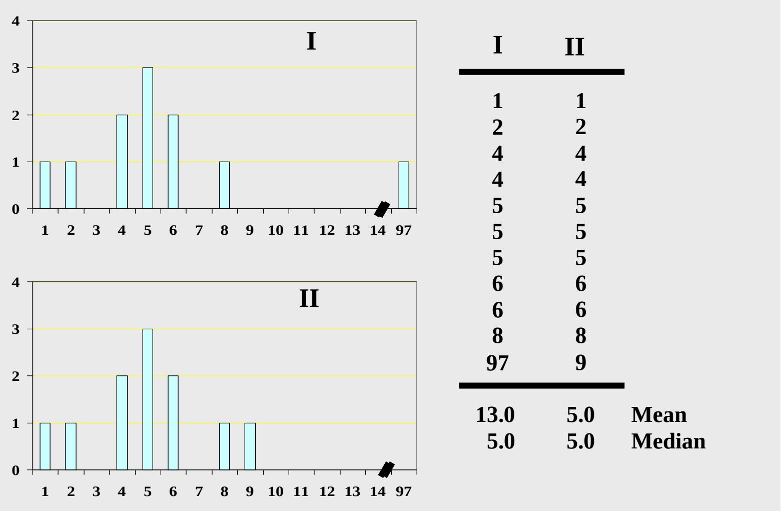
75th-quantile is the 7th largest observation (k=7 is the first integer larger than 6.75): 30.

The inter-quartile range of age for this sample of 9 women is (27 – 30); the median age is 28 years. We can summarize age in this example as median age = 28 years (IQR = 27 - 30).

When do we use the mean and when the median to describe a numerical characteristic? This question can be answered by checking the distribution of the numerical variable. If the distribution is **convex** and **symmetrical**, the mean will be similar to the median and the arithmetic mean and standard deviation are used for descriptive purposes. If the distribution is convex but **asymmetrical**, or there are **too few observations** to judge the distribution, then median and inter-quartile range are the descriptive measures of choice. Please note that the mean is notoriously sensitive to outliers; the median in contrast is a very robust measure.

Consider the following small example of two series of observations (Figure 2). Series II is a convex symmetrical distribution. Series I is identical to series II except for one outlier, that is, 97 instead of 9. The consequence of this outlier is that the means of the two series are quite different (13 compared to 5). A mean of 13 is not a good measure of central tendency for series I as you can tell from Figure 2. On the other hand, the median is identical for the two series and seems appropriate for both the symmetrical and the skewed distribution.

Figure 2: Mean and median of two series of observations



I	II	
1	1	
2	2	
4	4	
4	4	
5	5	
5	5	
5	5	
6	6	
6	6	
8	8	
97	9	
13.0	5.0	Mean
5.0	5.0	Median

There are formal statistical processes, such as the Kolmogorov-Smirnov test for normality, to decide whether to use mean or median and subsequently parametric or non-parametric statistics (see, for example, Zar 2010 for more information). However, checking the shape of the distribution under question is a reasonable start and generally provides a good decision base for the choice of appropriate descriptive measures and statistical test procedures.

Box 7.4: Descriptive statistics

Appropriate descriptive statistics for a variable are dependent on the type of variable and – in case of numerical variables – also on characteristics of their distribution.

Type of variable	Descriptive statistics used
Categorical variable	**Frequency of categories in percent** For example: Occurrence of gestational diabetes mellitus: 9.2%
Numerical variable Distribution is convex and symmetrical:	**Mean and standard deviation (SD)** For example: Mean age 28.1 (SD 4.1) years
Distribution is convex but skewed or there are too few observations to assess the distribution:	**Median and inter-quartile range (IQR)** For example: Median number of daily steps taken pre-pregnancy (pedometer counts) is 6544 (IQR 4250 - 11970).

Inferential Statistics

In epidemiological research often findings from a sample are intended to be generalized to a wider target population. The **generalizibility** of the findings of an epidemiological study is dependent on two issues: **reliability** and **validity**. The question of validity – that is, the presence or absence of systematic error - is mainly a question of a proper design and conduct of the study and is the domain of epidemiology. The question of reliability – that is, the judgment of random error - is the domain of statistics.

Statistical procedures are employed because random error is inherent to all epidemiological research due to **natural or biological variation**. Biological variation means that, within "normal" boundaries, differences exist between and within people. Take as an example systolic and diastolic blood pressures which vary between but also within healthy people quite substantially (Marshall, T., 2004). The biological variation inflicts random error onto studies.

> No two individuals or groups of individuals are ever exactly alike, yet decisions affecting people or the community (= target population) are based on experience with other people and communities (= sample) of similar biological and social characteristics. Because of these inherent differences these decisions cannot be exact; they are always accompanied by some uncertainty. This uncertainty is the random error which we aim to judge by using statistics.

Researchers from Melbourne are conducting a randomised controlled trial encouraging healthy eating and exercise in pre-school children to prevent obesity (based on Skouteris, H. et al., 2010). Let us assume that the study randomised 100 non-obese children in the healthy eating & exercise intervention group and another 100 non-obese children into normal care. At follow-up after one year 5% of children in the control group and 2% of children in the intervention group are identified as obese. Thus, the difference in the cumulative incidences of obesity is 3% - in this study, that is literally three children.

If this study was repeated in an identical way, but with another 200 children one would expect that the results would slightly differ, maybe 4% to 3%, or 6% to 1% - just by chance alone! That is, because the difference in the first study was "only" three children, hence it could be easily just one child or five children….. One task of statistics is to assess whether the result of a study, for example the observed difference between two groups, is likely to be attributable to chance, that is random error, alone.

Common sense tells us that the smaller the observed difference and the smaller the sample size, the more likely it is that the observed difference occurred by chance alone. In the above example, 5 children out of 100 children were compared with 2 children from another 100 children. We can easily imagine that this difference could happen by chance alone; that in repeat studies obesity in either group could change by one or two children. On the other hand, if this study had 1,000 children in each group and found at follow-up 50 obese children in the control and 20 obese children in the intervention group, we would be more inclined to "believe" in the intervention success; although the difference between the two groups is still 3%. In this case, even if the incidence of obesity in both groups changed by one or two children in a repeat study, the difference would still be about 3%.

> Statistics allow us to **quantify** the probability that a result occurred by chance alone, and thereby saves researchers from having to repeat a study over and over again in order to elucidate the impact of random error.

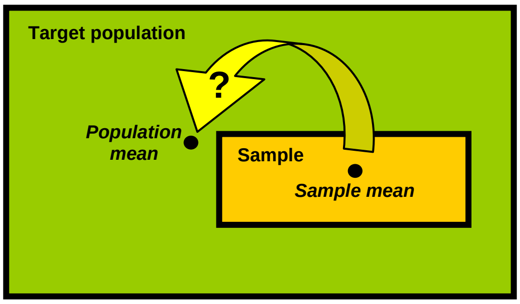
Inferential statistics are applied in two stages of an epidemiological study: (1) during the study design to estimate the appropriate sample size which is bound to the research hypothesis (see Chapter 10); and (2) during the analysis to confirm or reject the research hypothesis. During the statistical analysis we have two options to confirm or reject our research hypothesis; we can either use confidence intervals or carry out statistical hypothesis testing.

Confidence Interval

Researchers based in Sydney hypothesized that young unemployed Australians report worse health in times of low unemployment (based on Scanlan, J.N. and Bundy, A.C., 2009). In order to address the hypothesis, the researchers conducted a cross-sectional survey at a time of low unemployment. The researchers used the SF36 health survey, a validated tool to assess physical and mental health, in a sample of 200 unemployed Australians aged 18 to 25 years. Results of the SF36 were reported as the physical component summary score (PCS) and the mental component summary score (MCS). PCS and MCS range between 0 (= poor health) and 100 (= excellent health). The mean values for PCS was 52.1 (SD 7.9) and for the MCS it was 37.8 (SD 13.9) in the sample of 200 young unemployed Australians.

The first question that inferential statistics will help us addressing is: "What does the sample results tell us about the target population?" Please note that inferential statistics only deal with random error. If the results of a study are unbiased, then inferential statistics will allow us to relate back to the target population. For example, what does the sample results of mean PCS 52.1 (SD 7.9) and mean MCS 37.8 (SD 13.9) tell us about the – unknown (that's why the study was conducted in the first place) - population means? This question is answered by the **confidence interval** (**CI**).

Figure 3: Population and sample



Please note that "**population**" in the statistical sense is either the actual population or the target population, depending on whether bias is present or absent.

We know the sample mean because we measured it. The confidence interval tells us the range within which we can be reasonably certain the true (but unknown, unmeasured) target population mean is. Typically, we use the 95% confidence interval (so we can be 95% certain) but the level of certainty can be set at any level.

Box 7.5: The confidence interval for the population mean

The following formula gives the (1-α)-confidence interval for the population mean:

Formula for lower limit: sample mean - quantile(α) $\times \dfrac{\textbf{standard deviation}}{\sqrt{\textbf{sample size}}}$

Formula for upper limit: sample mean + quantile(α) $\times \dfrac{\textbf{standard deviation}}{\sqrt{\textbf{sample size}}}$

Thus, the confidence interval for the population mean is constructed around the sample mean. It has the following interpretation: **The true but unknown population mean lies within the (1-α)-confidence interval with a probability of 1-α.**

Comments:

(1) The confidence interval allows a statement about the unknown population mean by taking exclusively information from one sample into account!

(2) The population mean lies outside the confidence interval with a probability of α. **Alpha (α)** can take on any value between 0 to 1, however, is most often chosen to be 0.05 (or 5%), leading to a 95%-confidence interval. Choosing alpha to be 5% is completely arbitrary but has been accepted internationally. Please note when a 95%-confidence interval is stated, we are 9% sure that the confidence interval includes the population mean.

For more detailed information on confidence intervals please refer to an applied biostatistics textbook such as Altman, D. G. (1991), Bland, M. (2000), Zar, J.H. (2010) or any other biostatistics book

Let us now calculate confidence intervals for the mean PCS and the mean MCS from the example concerning the unemployed young Australians (based on Scanlan, J.N. and Bundy, A.C., 2009). We assumed that the researchers conducted a cross-sectional study with 200 participants. The sample mean value for PCS was 52.1 (SD 7.9) and for the MCS it was 37.8 (SD 13.9).

Before starting the calculations, one more bit of information is needed: the quantile which is dependent on alpha and the sample size and generally should be obtained from a t-distribution table. Please note that the quantile in the formula for the confidence interval of the population mean is dependent on alpha and on the sample size. If we choose alpha = 0.05 and have a sample size above 60, then 2 is a good approximation. Please refer to biostatistics books for more details.

95%-confidence interval for PCS:

$$\text{Lower limit: sample mean - quantile}(\alpha) \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} = 52.1 - 2 \times \frac{7.9}{\sqrt{200}} = 52.1 - 1.117 = 51.0$$

$$\text{Upper limit: sample mean + quantile}(\alpha) \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} = 52.1 + 2 \times \frac{7.9}{\sqrt{200}} = 52.1 + 1.117 = 53.2$$

95%-confidence interval for MCS:

$$\text{Lower limit: sample mean - quantile}(\alpha) \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} = 37.8 - 2 \times \frac{13.9}{\sqrt{200}} = 37.8 - 1.966 = 35.8$$

$$\text{Upper limit: sample mean + quantile}(\alpha) \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} = 37.8 + 2 \times \frac{13.9}{\sqrt{200}} = 37.8 + 1.966 = 39.8$$

Thus, we are 95% confident that the true yet unknown population mean for PCS lies between 51.0 and 53.2, and that the population mean for MCS can be found within 35.8 and 39.8.

Please note that confidence intervals for the population mean are symmetrically constructed around the sample mean, thus obviously including the sample mean. However, confidence intervals **do not** refer to the sample – they refer to the target (or actual) population!

We can calculate confidence intervals for all types of parameters such as mean values, proportions, medians, odds-ratios, and relative risks to name just a few. The respective formulas do vary in details, but are all based on (1) an α-quantile expressing the uncertainty, (2) an estimation of the sample variability, and (3) the sample size.

We want to give here one more example of a confidence interval, this time for a proportion. Assume that in the example of the Melbourne based randomised controlled trial aiming at encouraging healthy eating and exercise in pre-school

children to prevent obesity (based on Skouteris, H. et al., 2010), at follow-up after one year 5% of children in the control group and 2% of children in the intervention group were identified as obese. Thus, the difference in the cumulative incidences of obesity was 3%.
The 95% confidence interval for this difference of 3% is [-2.1, 8.1](see Bland, M., 2000, for the formula). Thus, we are 95% confident that the true yet unknown difference in incidence of obesity in the population is between -2.1% and 8.1%. Note that this confidence interval includes the zero; healthy eating could be better for children, but it could also be worse, and in fact we cannot confidently exclude the possibility that there is no difference at all between the two groups!

If the confidence interval would not include the zero (such as in [1.0, 5.0] for example), then we could be 95% sure that the true difference between control and intervention group was unequal to 0, and hence could assume a statistical difference between the groups. However, our result does not allow such a conclusion.

Box 7.6: The width of a confidence interval

It is obvious from the discussion above that the width of a confidence interval decides on the **precision** of the statement that is made by the interval. The narrower a confidence interval, the more precise the assertion about where the true population parameter lays. Thus, it is of interest to discuss how the width of a confidence interval can be changed.

In general, the width of a confidence interval for a specific parameter is dependent on:
(1) the level of uncertainty α;
(2) the observed sample variability of the parameter; and
(3) the sample size.

Have a look back at the formula of the CI for the population mean to identify these quantities.
Thus, changing any of these quantities will have ramifications on the width of the CI.

(1) Changing the level of uncertainty α
If **certainty is decreased,** for example, the uncertainty level of alpha increases from say 5% to 10%, while everything else remains unchanged, then the width of the CI decreases, that is, the **precision increases**!

To achieve 100% certainty (α = 0), the confidence interval would cover the entire possible range of the parameter; for a proportion for instance from 0% to 100%, and no information at all would be gained from the sample. Therefore, some uncertainty has to be allowed in the construction of a CI and 5% is the internationally accepted level. As a consequence the level of certainty is usually not decreased to achieve a narrower confidence interval.

(2) Changing the sample variability
If the **variability is decreased** while everything else remains unchanged, the width of the CI decreases, that is, **precision increases**.

However, changing the variability of a characteristic in epidemiological terms means recruiting a more homogenous sample. To achieve this, increasingly restrictive inclusion and exclusion criteria have to be employed and the generalizibility to the original target population – the overall aim of the study - is lost. Therefore a reduction in variability is also not an option to achieve a narrower confidence interval.
On the other hand, some laboratory experiments are able to achieve high precision even with small sample sizes by using very homogenous "participants", such as genetically engineered mice.

(3) Changing the sample size
If the **sample size is increased** while everything else remains unchanged, the width of the confidence interval decreases, that is, **precision increases**.

By increasing the sample size, we will most likely also increase the required resources to conduct the study. However, **increasing the sample size is in fact the only practical and acceptable way of increasing the precision, that is narrowing, the confidence interval.**

In the example of the RCT to prevent obesity in pre-school children, the difference in cumulative incidences of obesity was 3% (95% confidence interval = -2.1% to 8.1%). This result was based on 100 children per group. If the study had recruited 1,000 children per group instead while everything else remained the same, then the respective 95%-confidence interval for the 3% difference would have been 1.4% to 4.6%.

Note (1) that the width of the CI for the population mean decreased with increasing sample size; from 10.2% to 3.2%. (2) The 95% CI when there are 1,000 children in each group no longer includes the zero, that is, the "no difference" result. Hence we are now confident that there is a statistical difference between intervention and control group!
If the initial research hypothesis for this study was to detect a 3% difference in cumulative incidence, then the study with 200 children would have been "under-powered". In actual fact, 590 children per group is the optimal sample size to detect a difference of 3% with adequate statistical power.

Statistical Hypothesis Testing

Statistical hypothesis testing is used to formally confirm or reject a comparative operational research hypothesis. Statistical hypothesis testing is based on the very same reasoning outlined above for confidence interval estimation. However, usually statistical hypothesis testing involves comparing two or more groups directly.

For example, children with Fetal Alcohol Spectrum Disorders (FASD) may have significant neuro-behavioural problems persisting into adulthood (Peadon, E., et al., 2009). A trial was conducted to investigate the effect of a language intervention on basic literacy and numeracy skills in children with FASD (based on Adnams, C.M., et al., 2007). Language and literacy training was administered to a group of children with FASD. The children were followed up for 9 months. The language and literacy intervention focussed on phonological awareness and other pre- and early literacy skills needed for reading and spelling.

A comparative research hypothesis of this study such as: “In children with FASD, there is a difference between the general scholastic test result for reading at baseline and after language and literacy training ” can be tested by judging how likely it is that any observed difference in the test result is due to random error alone.
This judgement can be based on either of two statistical tools:
(1) By calculating the 95%-confidence interval for the mean difference between baseline scores and scores after intervention (in the example above it was actually one group and the research included before and after intervention assessments, but the same approach could be used to compare two groups). If the 95%-confidence interval does not include zero or more generally the null-value, then the difference in results is (or groups) is called “statistically significantly different”.
(2) By conducting a statistical test which directly gives the probability that the difference between the two results (or two groups) is attributable to chance alone.

There is quite a large number of different statistical tests and depending on the type(s) of variables involved, the parameters under study and the research hypothesis, one particular test is usually most appropriate. We now come to the main principle - and necessary terminology - of statistical testing and will then provide some guidance how to choose a suitable bivariate test for a specific study situation. “Bivariate” means that only two variables are involved in the analysis; usually the study factor and the outcome.

Box 7.7: Statistical tests, p-values and statistical significance

A **statistical test** is a decision making tool used to confirm or reject a research hypothesis. A statistical test judges how likely it is that an observed difference between groups, or an association between characteristics, is likely to be due to random error (chance) alone. A statistical test makes inferences from findings of the sample to the wider population.
In the following the term “**observed difference**” is used in the wider sense of also incorporating any observed associations or correlations between variables.

The result of a statistical hypothesis test is called “**p-value**”. The p-value gives the probability of obtaining in a sample a difference as large as the actually observed one (or an even larger one) if in reality, that is, in the target (or actual) population, there is no difference. Thus the p-value is the probability that an observed difference is attributable to chance alone! The smaller the p-value the less likely that an observed difference occurred by chance alone.

Comments:
(1) By convention, a p-value below 0.05 is considered **statistically significant**. As with confidence intervals, allowing 5% uncertainty is completely arbitrary, however, internationally accepted.

For example, p = 0.006 implies, that there is a 6 in 1,000 probability (0.6%) that the observed (or an even larger) difference occurred by chance alone. Since this p-value of 0.006 is smaller than the 0.05 threshold, the result is termed statistically significant. The result probably is not the result of random error; it probably accurately describes the target population.

(2) In the scientific literature the word “significant” should be used exclusively in a statistical context, that is, only if a statistical test was conducted resulting in a p-value of less than 0.05. Conventionally the calculated p-value is stated when we use the word “significant” in a manuscript.

(3) Statistical significance does not automatically imply clinical relevance! Small differences can be statistically significant if only the sample size is large enough, but may not be clinically relevant. Please note that first the clinical relevance of an observed difference has to be established before any statistical test is conducted! If the observed difference is clinically irrelevant, no statistical test should be conducted.

The study of children with FASD introduced above found that the language and literacy intervention was able to improve the manipulation of syllables (p=0.034) and written letters (p=0.004) significantly, but not the general scholastic test results for reading (p=0.152) and mathematics (p=0.651)(based on Adnams, C.M. et al., 2007).

For example, at nine months of follow-up the mean score for manipulating syllables was 79.6 (SD 34.3) compared to 69.4 (SD 38.0) at baseline (p=0.034). The p-value of 0.034 implies that the probability of obtaining this (or an even grater) difference of 10.2 by chance alone is 3.4%.

The mean results for reading words were 64.2 (SD 39.1) at baseline and 69.4 (SD 39.5) after 9 months of follow-up (p=0.152). The observed difference of 5.2 was not statistically significant; the p-value of 0.152 is greater than the threshold level of 0.05. A p-value of 0.152 implies that there is a 15.2% probability of this difference (or an even greater difference) being found in the sample even if in reality there is no difference at all in the target population. A probability of 15.2% is too large for us to accept significance.

Errors in Statistical Testing and the Problem of Multiple Testing

The p-value assesses how likely it is that an observed difference is attributable to chance alone. If the p-value is below the threshold (usually 5%) we call the difference statistically significant and assume that there is in fact a difference in the wider population. But we can’t be 100% sure; there is still a small chance (below 5%) that the observed difference is attributable to chance alone and that we commit an error - called **type I or alpha error** - by accepting a difference which in reality does not exist.

On the other hand however, there is also room for error if we fail to recognize a factually existing difference as significant. This potential error is called **type II or beta error**. A beta error arises for instance if an exposure is in reality linked to an outcome, but the study was too small to detect it as significant! To minimize this type II error, it is important that the study is large enough, in other words, that it has enough statistical **power** to detect an existing difference as significant. Please note that the **power** of a study is the probability that the study will detect an association or an expected difference if it truly exists in the target population.

Thus, two types of potential errors have to be faced when conducting a statistical test: (1) **Alpha error** (α; type I error) and (2) **Beta error** (ß; type II error). Table 1 summarizes the situation. Type I error implies that one falsely assumes a significant difference between groups, when in reality there is none. Type II error implies that one fails to detect an existing difference as significant.

Table 1: The errors in statistical hypothesis testing

	Reality	
Statistical test	**No difference**	**Difference**
No difference	**1-Alpha**	**Beta error** **(false negative)**
Difference	**Alpha error** **(false positive)**	**1-Beta** **Power of test**

Because of the convention that only a p-value less than 0.05 will be called significant, the Type I error (alpha) cannot exceed 5% and in this sense is under control. But this 5% threshold is only true if one single statistical test was performed in a data set! Usually, however, numerous statistical tests are performed during data analysis and the overall alpha error increases with each single test. This problem is called **multiple testing** and a small list of the increasing alpha error with increasing numbers of tests performed is detailed in Table 2.

Table 2: Overall alpha error and number of statistical tests conducted

Number of statistical tests	**Overall alpha error**
1	**0.05**
2	**0.098**
3	**0.143**
4	**0.185**
5	**0.226**
10	**0.401**
20	**0.642**
50	**0.923**
100	**0.994**
infinite	**1**

Naturally we do not know which of the statistical tests conducted is (are) significant by chance alone. However, we can and should take care of the issue of multiple testing by listing during the design stage all the research hypotheses we want to investigate so that the sample size can be adequately adjusted. Alternatively, at the analysis stage, the alpha level for the single tests can be adjusted, for instance by the Bonferroni-Holm procedure (Bland, M., 2000), in a way so that the overall alpha level still holds at 5%.

The Type II error is controlled by conducting an appropriate sample size calculation during the design phase of the study. The sample size calculation will ensure that there is adequate power to detect an expected difference as significant if it truly exists in the target population. By convention, studies should be designed large enough to have at least a power in excess of 80%.

Please note that an “under” powered epidemiological study, that is, a study with inadequate sample size for a specific research question is likely to commit a Type II error! That is, the study will fail to detect an existing difference because the sample size is too small. Drawing negative, that is, non-significant, results from an under-powered epidemiological study is inappropriate.

A study will only be able to confirm or reject the pre-specified quantified research hypotheses with statistical confidence for which it was designed. All other statistical tests performed during the analysis phase of the study, that is unplanned “*post hoc”* analyses, are subject to uncontrolled Type I and Type II error. If any of these “uncontrolled” statistical tests suggest a “significant” finding, then such a finding requires confirmation by an adequately planned independent study.

Selecting an appropriate bivariate statistical test

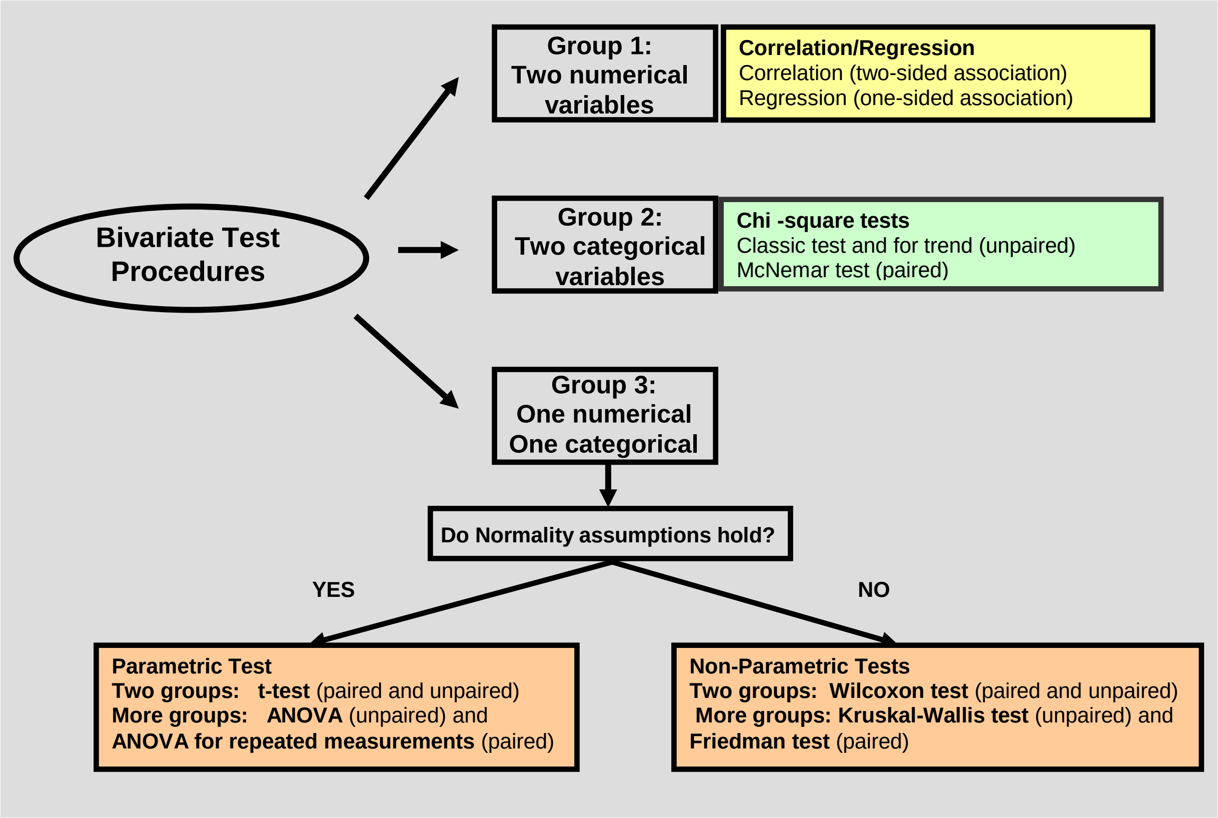
The choice of the correct statistical test procedure for a specific bivariate (two variables) test situation is dependent on the type(s) of the two variables involved. Once identified, this information leads to the right group of statistical tests as outlined in Figure 4.

(1) If both variables are numerical (for example, age and level of physical activity) then the correct statistical test procedures can be found in group1, the **correlation/regression group**;

(2) If both variables are categorical (for example, gender and type of employment) then the correct statistical test belongs to group2, the **Chi-square group**; and

(3) If one variable is numerical and the other is categorical (for example, gender and level of physical activity) then the correct statistical test procedure comes from group3 and is either a parametric **t-test/Analysis of Variance** or a non-parametric **Wilcoxon type test**.

Figure 4: Classification of bivariate statistical tests


Bivariate Test Procedures
Group 1: Two numerical variables
Correlation/Regression
Correlation (two-sided association)
Regression (one-sided association)
Group 2: Two categorical variables
Chi -square tests
Classic test and for trend (unpaired)
McNemar test (paired)
Group 3: One numerical One categorical
Do Normality assumptions hold?
YES
NO
Parametric Test
Two groups: t-test (paired and unpaired)
More groups: ANOVA (unpaired) and
ANOVA for repeated measurements (paired)
Non-Parametric Tests
Two groups: Wilcoxon test (paired and unpaired)
More groups: Kruskal-Wallis test (unpaired) and
Friedman test (paired)

Please refer to the chapter on “Choosing the statistical method” in Martin Bland’s (2000) book for further details. A description of how to **calculate** bivariate statistical tests by hand is provided in most standard statistical textbooks (see, for example, Zar, J., 2010 or Bland, M., 2000).

Box 7.8: How to choose a specific bivariate statistical test

Example 1: In a study on pre-pregnancy physical activity and its effect on gestational diabetes mellitus (based on van der Ploeg, H.P. et al., 2010) the researchers investigated whether age of the mother, level of education, parity, and pre-pregnancy level of physical activity were related with gestational diabetes. Assume we also want to investigate the association between age of the mother and her pre-pregnancy level of physical activity.

First variable	Second variable	Bivariate statistical test
Age of mother - numerical	Gestational diabetes - categorical	Group 3 – either t-test or Wilcoxon test
Level of education - categorical	Gestational diabetes - categorical	Group 2 – Chi-square test
Parity – numerical, discrete	Gestational diabetes - categorical	Group 3 – either t-test or Wilcoxon test

Pre-pregnancy level of physical activity - numerical	Gestational diabetes - categorical	Group 3 – either t-test or Wilcoxon test
Age of mother - numerical	Pre-pregnancy level of physical activity - numerical	Group 1 – Regression

Example 2: Researchers from Melbourne are conducting a randomised controlled trial encouraging healthy eating and exercise in pre-school children to prevent obesity (based on Skouteris, H. et al., 2010). Let us assume that the study randomised 100 non-obese children in the healthy eating & exercise intervention group and another 100 non-obese children into normal care. At follow-up after one year 5% of children in the control group and 2% of children in the intervention group are identified as obese.

First variable	Second variable	Bivariate statistical test
Intervention or control - categorical	Obesity - categorical	Group 2 – Chi-square test

Example 3: Researchers based in Sydney hypothesized that young unemployed Australians report worse health in times of low unemployment (based on Scanlan, J.N. and Bundy, A.C., 2009). In order to address the hypothesis, the researchers conducted a cross-sectional survey at a time of low unemployment and repeated the study at a time of high unemployment The researchers used the SF36 health survey as the outcome measure. Results of the SF36 were reported as the physical component summary score (PCS) and the mental component summary score (MCS).

First variable	Second variable	Bivariate statistical test
Unemployment high or low - categorical	PCS and MCS – both numerical	Group 3 – either t-tests or Wilcoxon tests

Example 4: A trial was conducted to investigate the effect of a language intervention on basic literacy and numeracy skills in children with FASD (based on Adnams, C.M., et al., 2007). Language and literacy training was administered to a group of children with FASD. The children were assessed at baseline and then followed up for 9 months. The language and literacy intervention focussed on phonological awareness and other pre- and early literacy skills needed for reading and spelling.

First variable	Second variable	Bivariate statistical test
Before and after intervention - categorical	Assessment of literacy skills - all numerical	Group 3 – either paired t-tests or Wilcoxon tests

Thus the identification of the type of the two variables involved brings us already to the right group of tests. As we can see from Figure 4, however, some more concepts need to be introduced before an unambiguous decision for a specific test can be made:

One-sided versus two-sided scientific question and tests

This decision is relevant for all three test groups and refers to the distinction between a situation where the researcher is interested (1) in both directions (e.g. group A *differs* from group B, that is, group A may be *better or worse* than group B) in which case we have a two-sided scientific question and use accordingly a two-sided statistical test; or (2) in only one direction (e.g. group A is *better* than group B) in which case the scientific question and the correct test is one-sided.

For example, the Sydney based study which hypothesized that young unemployed Australians report worse health in times of low unemployment compared to times of high unemployment (based on Scanlan, J.N. and Bundy, A.C., 2009), stated a one-sided scientific question which will require a one-sided statistical test,

Please note that one-sided statistical tests are *more likely* to return statistically significant results than two-sided tests if the data follow the pre-defined one-sided hypothesis. Therefore one-sided testing has to be thoroughly justified. Most research hypotheses are formulated as two-sided questions.

Paired versus unpaired test

If the research hypothesis under consideration involves the comparison of two (or more groups) who are independent from each other (in most cases consist of different individuals) then the **unpaired** version of the statistical test procedure is used. If the research hypothesis involves comparing the same people who were measured twice or more often then the **paired** version of the statistical is correct. This distinction is relevant in statistical test groups 2 and 3 only.

For example, in the study on pre-pregnancy physical activity and its effect on gestational diabetes mellitus (based on van der Ploeg, H.P. et al., 2010), the researchers compared women with gestational diabetes with women without gestational diabetes. These two groups of women were independent (different women) and hence an unpaired test procedure is adequate. On the other hand, the experiment which was conducted to investigate the effect of a language intervention on basic literacy and numeracy skills in children with FASD (based on Adnams, C.M., et al., 2007), the children were assessed at baseline, followed up for 9 months and then measured again. That is, the skills of the same children were assessed twice and the results compared; thus we clearly face a paired scenario.

Parametric versus non-parametric test

In bivariate test groups 1 and 3 a decision between **parametric** and **non-parametric** statistical test is required. This distinction is linked to the distribution of the numerical variable(s) involved. If the numerical variable is **approximately normally distributed** (see below) in all categories of the categorical variable (group 3), or both numerical variables are approximately normally distributed (group 1), then a parametric test is used. Otherwise a non-parametric test and procedures will be conducted.

As a rule of thumb, a numerical variable is **approximately normally distributed** if (1) the distribution is convex and symmetrical; (2) mean and median differ by less than 10%; and (3) the standard deviation is less than a third of the mean value (this last criterion is only appropriate for distributions which are not centrally located around zero). One can also test for Normality formally by using the Kolmogorov-Smirnov test (Zar, 2010).

Regression and Correlation: one-sided versus two-sided association

Statistical test group 1 differs from the two other groups as regression and correlation procedures are not testing for a difference, but rather investigate whether an **association** between two numerical variables exceeds in strength what would be expected by chance alone. This association can either be one- or two-sided.

When the association between the two variables is **one-sided**, that is, one variable may influence the other but the reverse is not possible, then a regression approach is used and in the graphical display, a scatter plot, the independent variable is depicted on the x-axis. The association between age of mother and pre-pregnancy level of physical activity is clearly a regression example: age may influence the level of physical activity (as we get older, we might be less active), however, the level of physical activity cannot influence age. Age is the independent variable and in a scatter plot is depicted on the x-axis.

When the association is **two-sided**, that is, both variables can influence each other, then a correlation approach is used and the allocation of the variables to the axes of the scatter plot is arbitrary. An example of a two-sided association is the association between level of physical activity and body mass index. Either characteristic could influence the other; there is no clear independent variable hence this is an example of a correlation problem.

Survival analysis

One additional bivariate test, the **logrank test**, is worth mentioning when introducing the most common bivariate test procedures. When the occurrence of a specific event, such as death, is observed in a cohort of people who are followed up for a period of time, special statistical procedures (**survival analysis**) may become necessary since the individual survival time, that is the individual follow-up time to the event, is usually unknown for at least some people in the cohort, because not everybody will have the event during the study period. The participants who do not have the specific event during the follow-up period or who withdraw from the study are called **censored cases**. These participants cannot be deleted from the data set because this would lead to an underestimation of the **survival probability** – as these people were observed and did survive for some time. In this situation we cannot use a t-test or Wilcoxon test to compare groups, but we need to calculate the logrank test to compare survival probabilities between groups. For further details on survival analysis please refer to Altman, D. (1991).

The pitfalls of agreement and equivalence

All of the above introduced statistical test procedures refer to statistical tests for difference. However some comparative research questions might not ask for difference, but for **equivalence** of effects or for **agreement** between observers or instruments. Both situations require special statistical test procedures. For example, researchers based in Mackay conducted a randomised controlled trial to compare the standard management of keeping wounds dry and covered with allowing wounds to be uncovered and wet in the first 48 hours following minor skin excision (Heal, C. et al., 2006). The research hypothesis for the study was that both the dry and the wet management of wounds will lead to very similar infection rates. Hence the study was not asking for whether infection rates between the wet and the dry patient group were different, but whether they were equivalent. It is tempting to use a statistical test of difference (i.e. a Chi-square test in this example) and interpret a non-significant result as “no difference”. However, this interpretation is wrong! As outlined before, any difference even the smallest one becomes statistically significant given a large enough sample size. In other words, if a statistical test results in “no significant difference”, it mainly implies that the sample size was not large enough to detect the difference.

When testing for equivalence, first the maximal difference which can still be considered clinically “equivalent” must be defined. A simple test of equivalence is then to assess whether the two-sided confidence interval of the difference is completely covered by the maximal allowable difference interval; if so the observed difference is significantly equivalent. We can also calculate p-values for equivalence tests, however, this is not straight forward and if you are interested in more technical information about statistical hypothesis testing for equivalence, we suggest the book by Wellek, S. (2003).

Similarly, questions of **agreement** need special consideration. When agreement between let’s say two observers is under investigation it is obviously not good enough if the two observers agree by more than chance alone; but this is exactly what statistical tests for difference test. Agreement however needs more than showing that an association beyond chance exists. We need to apply specially developed measures of agreement. For example, researchers from Queensland assessed whether parents were able to identify and count the number of moles on their children’s skin correctly (Harrison, S.L., et al., 2002). The number of moles a person has is the strongest risk factor for melanoma. Moles are visible to everybody and therefore it is of interest to investigate whether a risk assessment can be done by lay persons. In Harrison’s study, counts of moles from parents were compared with counts from dermatologists and counts from photographs. The question was whether these different methods of counting moles were in concordance.

There are agreement measures available for numerical data as well as for categorical data. For a numerical characteristic, graphical assessment (Bland, J.M. and Altman D.G., 1986) and special correlation coefficients (Lin, L., 1989) can be used to assess agreement. For a categorical characteristic, percentage of overall agreement and the kappa statistic (Fleiss, J.L., 1981) are available.

MULTIVARIABLE STATISTICAL ANALYSIS

Studying health phenomena in human populations is usually more complicated than just relating two characteristics, such as age and quality of life. Quality of life might be influenced by other characteristics such as gender, severity of disease, or socio-economic status. That is, bivariate statistical analysis is often insufficient to adequately address the research question under study. The main advantage of multivariable techniques in the health sciences is that they enable the researcher to assess more than one single study factor at a time and also allow adjustments for the influence of factors other than the study factor. These “other factors” are called confounders and are a common problem in quantitative studies.

Please note that in the health sciences multivariable analysis is often called multivariate.

Confounding occurs when extraneous variables (called confounders) which are associated with both the study factor and the outcome distort the bivariate association under study. In randomised studies confounding should be largely controlled and as a consequence statistical analysis of randomised controlled studies can be mostly restricted to bivariate tests. Multivariable analyses usually become necessary in studies in which we cannot randomise or in which randomisation failed.

For example, in the study on pre-pregnancy physical activity (= study factor) and its effect on gestational diabetes mellitus (= outcome) (based on van der Ploeg, H.P. et al., 2010) the researchers might consider the body mass index of the mother as a potential confounder. Obesity of the mother is a risk factor for gestational diabetes, and body mass index might be negatively associated with physical activity. Hence, one could argue that part of the association found between pre-pregnancy physical activity and gestational diabetes may be due to the body mass index of the mother. Therefore, we face a multivariable situation since we need to take the body mass index of the mother into account in order to assess the relationship between pre-pregnancy physical activity and gestational diabetes.

Before multivariable statistical analysis can be conducted, several preparatory steps are usually required. The first step for any multivariable approach is always a thorough descriptive analysis. At least two additional steps are necessary: a detailed examination of all bivariate associations between all variables involved and the identification of an appropriate coding for all variables. More detailed advice on preparation for multivariable analysis can be found in Feinstein’s (1996) book.

Selecting an appropriate multivariable model

The selection of a suitable multivariable model is dependent on a number of considerations which include the research hypothesis, the type of the target variable (outcome) and, to a lesser degree, the types of independent variables, and model-specific statistical assumptions. Table 3 provides a commented overview of the multivariable models most frequently used in the health sciences.

Table 3: Overview of multivariable models most frequently used in the health sciences.

Target variable	**Multivariate model and use**	**Main assumptions**	**Comments**
Numerical - continuous	**Multiple linear regression analysis** Assesses strength and direction of relationship	Normality, linearity, homoscedasticity, no outliers	(1) Classically all independent variables are continuous, in reality all types can be used if dummy coded; (2) Discrete or ordinal target variables with many categories can also be analysed; (3) If linearity is violated transform data or use non-linear regression analysis.
Numerical - continuous	**Analysis of Variance** Assesses whether relationship exists	Normality, homoscedasticity, equal sample sizes, random sampling	(1) Classically all independent variables are nominal (ANOVA), generalizations for continuous variables available (ANCOVA, MANCOVA); (2) Equal sample size requirement per cell, might be fulfilled in experimental designs.
Categorical - binary	**Logistic regression** Assesses strength and direction of relationship	Linearity, Homogeneity of variances, homoscedasticity of residuals, no outliers	(1) Preferred over discriminant function analysis because normality of independent variables not required; (2) Transform data if linearity is violated.
Categorical - binary	**Probit analysis** Assesses strength and direction of relationship	Linearity, Homogeneity of variances, homoscedasticity of residuals, no outliers	(1) Specific regression analysis mainly used in toxicology to analyse dose-response and binomial response experiments; (2) Approach very similar to logit regression.
Categorical – more than 2 categories	**Polytomous and ordered logit analysis** Assesses strength and direction of relationship	Linearity, Homogeneity of variances, homoscedasticity of residuals, no outliers	(1) Direct generalization of logistic model; (2) Transform data if linearity is violated.
Numerical or Categorical or "moving binary"	**CART*** Assesses size and direction of effects; Identifies interactions and collinearities	Free of assumptions; Explorative data analysis – not truly a multivariable model	(1) Different versions available for different types of target variables; (2) Delivers estimates of the effects assessed and a graphical tree displaying effect sizes; (3) Defines risk or prognostic groups.
Survival ("moving binary")	**Cox proportional hazard analysis** Assesses strength and direction of relationship with mortality	Proportional hazards and linearity	(1) Robust method to analyse binary outcome moving in time; (2) Transform data if linearity is violated; (3) Sample size requirements are dependent on expected differences and the event rate.

*CART = Classification and Regression Tree analysis

Box 7.9: Multivariable models

A multivariable model is used to investigate the relationship between study factor(s) and outcome by simultaneously allowing adjustment for confounding. Multivariable models assess the effects of several study factors together. Multivariable models are required in randomised controlled trials when randomisation failed. The choice of an appropriate model is foremost dependent on the type of the outcome variable.

Examples:

(1) Researchers from Brisbane conducted a cross-sectional study of 132 healthy adult Australians to investigate the contributions of the HTR2A gene, chronic psychological stress, and impulsivity to the prediction of cigarette smoking status (White, M.J., et al. 2010). The outcome was current cigarette smoker ("no"/"yes"). Therefore, the researchers conducted a multivariable logistic regression analysis to estimate the genetic effect on smoking adjusted for gender, severity of depressive symptoms and chronic stress.

(2) Another research group based in Brisbane aimed to identify factors contributing to reduced quality of life (= outcome) in an older population referred to a community rehabilitation team (Comans, T.A. et al., 2010). The outcome variable "quality of life" is numerical. Hence the researchers used multiple linear regression analysis to investigate the impacts of participation in functional activities, history of falls, number of medications, number of co-morbidities, depression, environmental hazards, physical function, and nutrition on quality of life.

Presenting results from multivariable modelling

Lastly we would like to give you an impression of how results from multivariable analysis are presented by an example from a study conducted by Australian researchers examining pregnancy outcomes for Indigenous people in a regional setting (Panaretto, K. et al., 2006). The results presented in Table 4 are based on a prospective cohort of 386 singleton births to women attending Townsville Aboriginal and Islander Health Services (TAIHS) for shared antenatal care between January 1 2000 and December 31 2003.

The model presented in Table 4 is an hierarchical model, however the effects of urinary track infection and of hazardous drinking habits are not stated since the single effects cannot be interpreted when the interaction between them is included. The model was adjusted for the confounding effects of age of mother and drug use.

Table 4: Descriptive and multivariate logistic regression results of a cohort study of predictors of low birth weight for gestational age of 386 Indigenous in Townsville between 2000 and 2003 (based on Panaretto, K. et al., 2006).

Predictor	**Low birth weight for gestational age**			
	no (n = 352)	**yes (n = 34)**	**RR* [95%-CI]****	**p-value**
Body mass index [kg/m^2]	continuous		0.92 [0.85, 0.99]	= 0.0311
Mother was active smoker				
No (= baseline)	140	5	1	
Yes	212	29	3.7 [1.2, 11.7]	= 0.0245
Pregnancy induced hypertension				
No (= baseline)	328	29	1	
Yes	24	5	6.6 [1.9, 22.7]	= 0.0025
Interaction between urinary track infection and hazardous drinking				
Baseline***	343	25	1	
With urinary track infection AND hazardous drinking	9	9	7.0 [1.01, 46.6]	= 0.0453

*RR = Relative risk; **95%-CI = 95% confidence interval; ***Baseline = no urinary track infection OR no hazardous drinking habits.

Interpretation of the results from Table 4:

(1) The higher the body mass index the less likely a baby was born with low birth weight for gestational age. The relative risk was 0.92, that is, the effect was protective (RR < 1). Body mass index was assessed in the model as a continuous variable, that is, the 0.92 relates to a change of 1 kg/m^2. This is why the RR shows an effect relatively small in comparison to the other characteristics. Please note that body mass index should have been linearly related with low birth weight for gestational age in order to allow this interpretation.

(2) Mothers who were active smokers were 3.7 times more likely to give birth to a child with low birth weight for gestational age compared to mothers who were not actively smoking at that time. We can be 95% confident that the true relative risk is between 1.2 and 11.7. The result was statistically significant, as the 95% confidence interval of the relative risk does not include 1. The p-value reflects this observation. P=0.0245 implies that the likelihood of observing a relative risk of 3.7 (or even larger) by chance alone (that is, without a factual difference in the wider population) is 0.0245 (or 2.45%).

(3) Mothers who suffered from pregnancy induced hypertension were 6.6 times more likely to give birth to a child with low birth weight for gestational age compared to mothers who did not have pregnancy induced hypertension. We can be 95% confident that the true relative risk in the

wider population is between 1.9 and 22.7. The result is statistically significant, as the 95% confidence interval of the relative risk does not include 1. This statement is again corroborated by the p-value of 0.0025 which means that the likelihood to observe a relative risk of 6.6 (or even larger) by chance alone is 0.0025 (or 0.25%).

(4) There was an interaction between urinary track infection and hazardous drinking: Mothers who suffered from a urinary track infection and who drank alcohol at an hazardous rate were 7.0 times more likely to give birth to a child with low birth weight for gestational age compared to mothers who had no urinary track infection OR had no hazardous drinking habits. We can be 95% confident that the true relative risk is between 1.01 and 46.6. The result is statistically significant, as the 95% confidence interval of the relative risk did not include 1. The p-value again corroborates this observation and the interpretation follows those listed above.

Please note the "combined" effect of two or more characteristics on the outcome is called interaction if it differs from the "sum" of the independent effects. Interactions can be amplifying (synergism) or reversing (antagonism). We talk about **synergism** if the simultaneous effect of two or more variables on the outcome exceeds the "sum" of the single effects. In the health sciences, synergism is often a biologic response to the simultaneous exposure to two or more agents that exceeds the combined action of the agents when acting independently. On the other hand, **antagonism** is the antithesis of synergism, and refers to the situation when the simultaneous effect of variables on the outcome is less than the "sum" of the individual effects.

The logistic model is a multiplicative model: That is, for example, mothers who were active smokers and had pregnancy induced hypertension were 3.7 x 6.6 = 24.4 times more likely to give birth to a child with low birth weight for gestational age compared to mothers who did not smoke and had no pregnancy induced hypertension.

Please note that sample sizes are very informative in tables presenting the results of multivariate models. In Table 4 for instance, the small sample sizes for mothers with pregnancy induced hypertension and both hazardous drinking and urinary tract infections explain the large confidence intervals for those results.

Summary

- Statistics are widely used in the health sciences because they provide us with:
  (1) numbers - an unambiguous language understood by everybody worldwide in the same way; and
  (2) tools to measure the uncertainty (that is the random error) present when inferring from a sample to the wider target population.
- The correct choice of an appropriate statistical method to present and analyse the data is dependent on the type(s) of the variable(s) involved. The main differentiation is between categorical and numerical data.
- Descriptive statistics for categorical variables usually simply describe the frequency of each category in percentages. Numerical variables are summarised using a measure of central tendency together with a measure of dispersion. Mean and standard deviation are used when the distribution of the numerical data is convex and symmetrical. If the distribution of numerical data is convex but skewed or there are too few observations to assess the distribution, then median and inter-quartiles are used for descriptive purposes.
- Inferential statistics allow us to assess the random error involved when generalising from a sample to the target population.
- A 95% confidence interval of a parameter, such as a mean or relative risk, implies that we are 95% confident that the true yet unknown parameter in the target population lies within this interval.
- If the sample size is increased while everything else remains unchanged, the width of a confidence interval decreases, that is, the precision of the statement improves.
- A statistical test is a decision making tool. It is used to confirm or reject a research hypothesis. A statistical test judges how likely it is that an observed difference between groups, or an association between characteristics, is due to random error (chance) alone. A statistical test is based on data from a sample, but delivers inferences about the target population.
- The result of a statistical hypothesis test is called a “p-value”. The p-value gives the probability that the observed difference or an even larger difference is attributable to chance alone, given that in the target population there is really no difference. A p-value below 0.05 is considered as statistically significant. Statistical significance does not necessarily imply clinical relevance!
- During the conduct of a statistical test two types of errors can occur: (1) Alpha error (α; type I error) and (2) Beta error (ß; type II error). Type I error occurs if a statistical test falsely identifies a difference between groups as significant (but in reality there is no difference in the target population). Type II error occurs when a test fails to identify a difference as significant when the difference exists in the target population. Alpha error is measured by the statistical test via the p-value. Beta error can only be controlled by a pre-defined quantified research hypothesis and a corresponding appropriate sample size calculation.
- The choice of the correct statistical test group for a specific bivariate (two variables) test situation is dependent on the type(s) of the two variables

involved. To come to an unambiguous decision for a specific test procedure, more detailed test-group specific differentiations are necessary.

- For many study situations bivariate statistical tests are insufficient because several study factors are assessed simultaneously and / or because potential confounding is involved in the analysis.
- The choice of an appropriate multivariable model is primarily determined by the type of the outcome variable.

References

Adnams, C.M., Sorour, P., Kalberg, W.O., Kodituwakku, P., Perold, M.D., Kotze, A., September, S., Castle, B., Gossage, J., May, P.A. (2007) Language and literacy outcomes from a pilot intervention study for children with fetal alcohol spectrum disorders in South Africa. Alcohol 41:403–14.

Altman, D.G. (1991) Practical statistics for medical research. Chapman & Hall, London.

Bland, J.M. and Altman D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–10.

Bland, M. (2000) An introduction to medical statistics. 3rd edn. Oxford University Press, Oxford.

Comans, T.A., Currin, M.L., Brauer, S.G., Haines, T.P. (2011) Factors associated with quality of life and caregiver strain amongst frail older adults referred to a community rehabilitation service: Implications for service delivery. Disabil Rehabil 33(13–14):1215–21.

Feinstein, A.R. (1996) Multivariable analysis: An introduction. Yale University Press, New Haven.

Fleiss, J.L. (1981) Statistical methods for rates and proportions. 2nd edn. John Wiley & Sons, New York.

Harrison, S.L., Buettner, P.G., MacLennan, R., Kelly, J.W., Rivers, J.K. (2002) How good are parents at assessing melanocytic nevi (moles) on their children? Am J Epidemiol 155(12):1128–36.

Heal, C., Buettner, P., Raasch, B., Browning, S., Graham, D., Bidgood, R., Campbell, M., Cruikshank, R. (2006) Can sutures get wet? A randomised controlled trial of wound management in general practice. Br Med J 332(7549):1053–6.

Lin, L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1):255–68.

Marshall, T. (2004) When measurements are misleading: Modelling the effects of blood pressure misclassification in the English population. Br Med J 328:933.

McGervey, J.D. (1986) Probabilities in everyday life. Chicago, Nelson-Hall.

Panaretto, K., Lee, H., Mitchell, M., Larkins, S., Manessis, V., Buettner, P., Watson, D. (2006) Risk factors for preterm, low birth weight and small for gestational age birth in urban Aboriginal and Torres Strait Islander women in Townsville. Aust N Z J Public Health 30(2):163–70.

Peadon, E., Rhys-Jones, B., Bower, C., Elliott, E.J. (2009) Systematic review of interventions for children with Fetal Alcohol Spectrum Disorders. BMC Pediatr 9:35.

Scanlan, J.N. and Bundy, A.C. (2009) Is the health of young unemployed Australians worse in times of low unemployment? Aust N Z J Public Health 33(1):79–82.

Skouteris, H., McCabe, M., Swinburn, B., Hill, B. (2010) Healthy eating and obesity prevention for preschoolers: A randomised controlled trial. BMC Public Health 10:220.

van der Ploeg, H.P., van Poppel, M.N., Chey, T., Bauman, A.E., Brown, W.J. (2010) The role of pre-pregnancy physical activity and sedentary behaviour in the development of gestational diabetes mellitus. J Sci Med Sport 14(2):149–52.

Wellek, S. (2003) Testing statistical hypotheses of equivalence. Chapman & Hall, CRC Press LLC, Virginia Beach.

White, M.J., Young, R.M., Morris, C.P., Lawford, B.R. (2010) Cigarette smoking in young adults: The influence of the HTR2A T102C polymorphism and punishment sensitivity. Drug Alcohol Depend 114(2–3):140–6.

Zar, J.H. (2010) Biostatistical analysis. 5th edn. Prentice Hall, Upper Saddle River.