PHCM9795 Foundations of Biostatistics

Course notes

Term 2, 2022

Contents

Module 1

Introduction to statistics and presenting data

Learning objectives

By the end of this module, you will be able to:

- Define the term statistics;
- Describe and identify the underpinning concepts of descriptive and inferential statistics;
- Distinguish between different types of variables (i.e. quantitative – discrete and continuous; and qualitative – ordinal and nominal);
- Construct appropriate frequency tables from raw data;
- Compute summary statistics to describe the centre and spread of data;
- Describe the (centre and spread of the) data using appropriate graphs (histogram and box plot);
- Present and interpret graphical summaries of variables using a variety of graphs (bar charts, line-graphs, histograms, boxplots, pie charts and others).

Readings

[**?**]; Chapters 2 and 3.

[**?**]; Chapter 4.

[Acock, 2010]; Chapter 5.

1.1 An introduction to statistics

1 Learning Activities

Activity 1.1

25 participants were enrolled in a 3-week weight loss programme. The following data present the weight loss (in grams) of the participants.

Table 1.1: Weight loss (g) for 25 participants

255	198	283	312	283
57	85	312	142	113
227	283	255	340	142
113	312	227	85	170
255	198	113	227	255

a) Enter these data into Stata.
b) What type of data are these?
c) Construct an appropriate graph to display the relative frequency of participants' weight loss. Your graph should start at 50 grams, with weight loss grouped into 50 gram bins. Provide appropriate labels for the axes and give the graph an appropriate title.

Activity 1.2

Researchers at a maternity hospital in the 1970s conducted a study of low birth weight babies. Low birth weight is classified as a weight of 2,500g or less at birth. Data were collected on age and smoking status of mothers and the birth weight of their babies. The Stata file `Activity_S1.2.dta` contains data on the participants in the study. The file is located on Moodle in the Learning Activities section.

Use Stata to create a 2 by 2 table to show the proportions of low birth weight babies born to mothers who smoked during pregnancy and those that did not smoke during pregnancy. Answer the following questions:

a) What was the total number of mothers who smoked during pregnancy?
b) What proportion of mothers who smoked gave birth to low birth weight babies? What proportion of non-smoking mothers gave birth to low birth weight babies?
c) Use Stata to construct a stacked bar chart of the data to examine if there a difference in the proportion of babies born with a low birth weight in relation to mother's age? Provide appropriate labels for the axes and give the graph an appropriate title.
d) Using your answers to the question a) and b), write a brief conclusion about the relationship of low birth weight and mother's age and smoking status.

Activity 1.3

Using Stata, estimate the mean, median, mode, standard deviation, range and interquartile range for the data Activity_S1.3.dta, available on Moodle.

Activity 1.4

Data of diastolic blood pressure (BP) of a sample of study participants are provided in the dataset Activity_S1.4.dta. Compute the mean, median, range and SD of diastolic BP.

Activity 1.5

In a study of 100 participants data were missing for 5 people. The missing data points were coded as '99'. The mean of the data was estimated as 45.0 with a standard deviation of 5.6; the smallest and greatest values are 16 and 65 respectively.

If the researcher analysed the data as if the 99s were real data, would it make the following statistics larger, smaller, or stay the same?

a) Mean
b) Standard Deviation
c) Range

Activity 1.6

Which of the following statements are true? The more dispersed, or spread out, a set of observations are:

a) The smaller the mean value
b) The larger the standard deviation
c) The smaller the variance

Activity 1.7

If the variance for a set of scores is equal to 9, what is the standard deviation?

Module 2

Probability and probability distributions

Learning objectives

By the end of this module you will be able to:

- Describe the concept of probability;
- Describe the characteristics of a binomial distribution and a Normal distribution;
- Compute binomial probabilities using Stata;
- Compute and use Z-scores to obtain probabilities;
- Decide when to use parametric or non-parametric statistical methods;
- Briefly outline other types of distributions.

Readings

?; Chapters 5, 14 and 15.

?; Chapters 6 and 7.

2.1 Section

Module 2: Stata notes

2 Learning Activities

Activity 2.1

In a Randomised Controlled Trial, the preference of a new drug was tested against an established drug by giving both drugs to each of 90 people. Assume that the two drugs are equally preferred, that is, the probability that a patient prefers either of the drugs is equal (50%). Use one of the binomial functions in Stata to compute the probability that 60 or more patients would prefer the new drug. In completing this question, determine:

a) The number of trials (n)
b) The number of successes we are interested in (k)
c) The probability of success for each trial (p)
d) The form of the Stata function: binomialp, binomial or binomialtail
e) The final probability.

Activity 2.2

A case of Schistosomiasis is identified by the detection of schistosome ova in a faecal sample. In patients with a low level of infection, a field technique of faecal examination has a probability of 0.35 of detecting ova in any one faecal sample. If five samples are routinely examined for each patient, use Stata to compute the probability that a patient with a low level of infection:

a) Will not be identified?
b) Will be identified in two of the samples?
c) Will be identified in all the samples?
d) Will be identified in at most 3 of the samples?

Activity 2.3

If weights of men are Normally distributed with a population mean μ = 87, and a population standard deviation, σ = 8 kg:

a) What is the probability that a man will weigh 95 kg or more? Draw a Normal curve of the area represented by this probability in the population (i.e. with μ = 87 kg and σ = 8 kg).
b) What is the probability that a man will weigh more than 75 kg but less than 95 kg? Draw the area represented by this probability on a standardised Normal curve.

Activity 2.4

Using the health survey data (`health-survey.xlsx`) described in the Stata notes of this module, create a new variable, BMI, which is equal to a person's weight (in kg) divided by their height (in metres) squared (i.e. BMI $= \frac{\text{weight (kg)}}{[\text{height (m)}]^2}$. Categorise BMI using the WHO categories provided in Section XX. Create a two-way table to display the distribution of BMI categories by sex. Does there appear to be a difference in categorised BMI between males and females?

Activity 2.5

The data in the file `LengthOfStay.dta` (available on Moodle) has information about birth weight and length of stay collected from 117 babies admitted consecutively to a hospital for surgery. Complete the following table to make a decision about whether each of the variables is symmetric, and which measures of the centre and spread of the data should be reported.

Activity 2.6

The data set of hospital stay data for 1323 hypothetical patients is available on Moodle in csv format (`activity2.5.csv`). Import this dataset into Stata. There are two variables in this dataset:

- female: female=1; male=0
- los: length of stay in days

a) Use Stata to examine the distribution of length of stay: overall; and separately for females and males. Comment on the distributions.
b) Use Stata to calculate measures of central tendency for hospital stay to obtain information about the average duration of hospital stay. Which summary statistics should you report and why? Report the appropriate statistics of the spread and measure of central tendency chosen.
c) Calculate the measures of central tendency for hospital duration separately for males and females. What can you conclude from comparing these measures for males and females?

Module 3

Precision, standard errors and confidence intervals

Learning objectives

By the end of this module you will be able to:

- Explain the purpose of sampling, different sampling methods and their implications for data analysis;
- Distinguish between standard deviation of a sample and standard error of a mean;
- Recognise the importance of the central limit theorem;
- Calculate the standard error of a mean;
- Calculate and interpret confidence intervals for a mean;
- Be familiar with the t-distribution and when to use it.

Readings

?; Chapters 4 and 6.

?; Sections 3.3 and 3.4, 8.1 to 8.3.

Juul and Frydenberg [2014]; Sections 11.5 to 11.7.

3.1 Introduction

To describe the characteristics of a population we can gather data about the entire population (as is undertaken in a national census) or we can gather data from a sample of the population. When undertaking a research study, taking a sample from a population is far more cost-effective and less time consuming than collecting information from the entire population. When a sample of a population is selected, summary statistics that describe the sample are used to make inferences about the total population from which the sample was drawn. These are referred to as inferential statistics.

However, for the inferences about the population to be valid, a random sample of the population must be obtained. The goal of using random sampling methods is to obtain a sample that is representative of the target population. In other words, apart from random error, the information derived from the sample is expected to be much the same as the information collected from a complete population census as long as the sample is large enough.

3.2 Sampling methods

Methods have been designed to select participants from a population such that each person in the target population has an equal probability of being chosen. Methods that use this approach are called random sampling methods. Examples include simple random sampling and stratified random sampling.

In simple random sampling, every person in the population from which the sample is drawn has the same random chance of being selected into the sample. To implement this method, every person in the population is allocated an ID number and then a random sample of the ID numbers is selected. Software packages can be used to generate a list of random numbers to select the random sample.

In stratified sampling, the population is divided into distinct non-overlapping subgroups (strata) according to an important characteristic (e.g. age or sex) and then a random sample is selected from each of the strata. This method is used to ensure that sufficient numbers of people are sampled from each stratum and therefore each subgroup of interest is adequately represented in the sample.

The purpose of using random sampling is to minimise selection bias to ensure that the sample enrolled in a study is representative of the population being studied. This is important because the summary statistics that are obtained can then be regarded as valid in that they can be applied (generalised) back to the population.

A non-representative sample can occur when random sampling is used, simply by chance. However, non-random sampling methods, such as using a convenient study population, will often result in a non-representative sample being selected so that the summary statistics from the sample cannot be generalised back to the population from which the participants were drawn. The effects of non-random error are much more serious than the effects of random error. Concepts such as non-random error (i.e. systematic bias), selection bias, validity and generalisability are discussed in more detail in PHCM9476: Foundations of Epidemiology.

3.3 Standard error and precision

Module 1 introduced the mean, variance and standard deviation as measures of central tendency and spread for continuous measurements from a sample or a population. As described in Module 1, we rarely have data on the entire population but we can infer information about the population (e.g. the mean weight of people in the population) based on a sample. However, a sample taken from a population is usually a small proportion of the total population. If the sample is very small, we would not expect our estimate of the population mean value to be precise. If the sample is very large, we would expect a more precise estimate of the population mean, i.e. the estimated mean value would be much closer to the true mean value in the population.

3.3.1 The standard error of the mean

A point estimate is a single best guess of the true value in the population. Instead of trying to guess the true value, it may be preferable to give a range of values in which we think the true value lies. For example, suppose we want to estimate the average weight of a population, and found a sample mean of 65 kg. Rather than saying we believe the true mean to be 65 kg, we could say we believe it is somewhere between, say, 58 kg and 72 kg.

Often in papers, one will see something like “the mean is 70.24 $\pm$ 1.78 kg”. The 1.78 is called the standard error of the mean (sometimes shortened to S.E.M. or S.E.). The standard error of the mean measures the extent to which we expect the means from different samples to vary because of chance error in the sampling process. The standard error is a measure of precision of the point estimate. This statistic is directly related to the size of the sample. The standard error of the mean for a continuously distributed measurement for which the SD is an accurate measure of spread is computed as follows:

$$\text{SE}(\bar{x}) = \frac{\text{SD}}{\sqrt{n}}$$

For our sample of weight data from 30 patients in Module 1:

$$\text{SE}(\bar{x}) = \frac{5.04}{\sqrt{30}} = 0.92$$

Because the calculation uses the sample size (n) (i.e. the number of study participants) in the denominator, the SE will become smaller when the sample size becomes larger. A smaller SE indicates that the estimated mean value is more precise.

The standard error is an important statistic that is related to sampling variation. When a random sample of a population is selected, it is likely to differ in some characteristic compared with another random sample selected from the same population. Also, when a sample of a population is taken, the true population mean is an unknown value.

Just as the standard deviation measures the spread of the data around the population mean, the standard error of the mean measures the spread of the sample means. Note that we do not have different samples, only one. It is a theoretical concept which enables us to conduct various other statistical analyses.

3.4 Central limit theorem

Even though we now have an estimate of the mean and its standard error, we might like to know what the mean from a different random sample of the same size might be. To do this, we need to know how sample means are distributed. In determining the form of the probability distribution of the sample mean ($\bar{x}$), we consider two cases:

3.4.1 When the population distribution is unknown:

The central limit theorem for this situation states:

> In selecting random samples of size n from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large.

The sample size n = 30 and above is a rule of thumb for the central limit theorem to be used. However, larger sample sizes may be needed if the distribution is highly skewed.

3.4.2 When the population is assumed to be normal:

In this case the sampling distribution of $\bar{x}$ is normal for any sample size.

3.5 95% confidence interval of the mean

In Module 2, we showed that the characteristics of a Standard Normal Distribution are that 95% of the data lie within 1.96 standard deviations from the mean (Figure 2.2). Because the central limit theorem states that the sampling distribution of the mean is approximately Normal in large enough samples, we expect that 95% of the mean values would fall within 1.96 × SE units above and below the measured mean population value.

For example, if we repeated the study on weight 100 times using 100 different random samples from the population and calculated the mean weight for each of the 100 samples, approximately 95% of the values for the mean weight calculated for each of the 100 samples would fall within 1.96 × SE of the population mean weight.

This interpretation of the SE is translated into the concept of precision as a 95% confidence interval (CI). A 95% CI is a range of values within which we have 95% confidence that the true population mean lies. If an experiment was conducted a very large number of times, and a 95%CI was calculated for each experiment, 95% of the confidence intervals would contain the true population mean.

The calculation of the 95% CI for a mean is as follows:

$$\bar{x} \pm 1.96 \times \text{SE}(\bar{x})$$

This is the generic formula for calculating 95% CI for any summary statistic. In general, the mean value can be replaced by the point estimate of a rate or a proportion and the same formula applies for computing 95% CIs, i.e.

$$95\% \text{ CI} = \text{point estimate} \pm \text{SE}(\text{point estimate})$$

The main difference in the methods used to calculate the 95% CI for different point estimates is the way the SE is calculated. The methods for calculating 95% CI around proportions and other ratio measures will be discussed in Module 6.

The use of 1.96 as a general critical value to compute the 95% CI is determined by sampling theory. For the confidence interval of the mean, the critical value (1.96) is based on normal distribution (true when the population SD is known). However, in practice, Stata and other statistical packages will provide slightly different confidence intervals because they use a critical value obtained from the t-distribution. The t-distribution approaches a normal distribution when the sample size approaches infinity, and is close to a normal distribution when the sample size is ≥30.The critical values obtained from the t-distribution are always larger than the corresponding critical value from the normal distribution. The difference gets smaller as the sample size becomes larger. For example, when the sample size n=10, the critical value from the t-distribution is 2.26 (rather than 1.96); when n= 30, the value is 2.05; when n=100, the value is 1.98; and when n=1000, the critical value is 1.96.

The critical value multiplied by SE (for normal distribution, 1.96 × SE) is called the maximum likely error for 95% confidence.

3.5.1 Worked Example 3.1: 95% CI of a mean

For our sample of weights data with standard error of 0.92:

$$\begin{aligned} 95\% \text{ CI}(\bar{x}) &= \bar{x} \pm 1.96 \times \text{SE}(\bar{x}) \\ &= 70.0 \pm 1.96 \times 0.92 \\ &= 68.2 \text{ to } 71.8\text{kg} \end{aligned}$$

We interpret this confidence interval as: we are 95% confident that the true mean of the population from which our sample was drawn lies between 68.2 kg and 71.8 kg.

This calculation takes into account both the sample mean of 70.0 kg and the sampling error that has arisen by chance due to the sample size of 30 people.

For a 95% CI to be reported around a mean value, the data values need to be approximately normally distributed, as discussed in Module 2.

Note: Had we used the t-distribution to calculate the critical value, the 95%CI would have been slightly wider, 68.1 kg to 71.9 kg as shown in Output 3.1 below using Example_1.3.dta with the ci mean command in Stata.

Output 3.1 Mean and 95%CI of weight

```
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      weight |         30          70    .9207069        68.11694    71.88306
```

3.5.2 The t-distribution and when should I use it?

The population standard deviation (σ) is required for calculation of the standard error. Often, σ is not known and the sample standard deviation (s) is used to estimate it. It is known, however, that the sample standard deviation of a normally distributed variable is a downward-biased estimator of σ, particularly when the sample size is small.

Someone by the pseudonym of Student came up with the Student's t distribution with $(n-1)$ degrees of freedom to account for this bias. It looks very much like the standardised normal distribution, only that it has fatter tails (Figure 3.1). As the degrees of freedom increase (i.e. as n increases), the t-distribution gradually approaches the standard normal distribution. With a sufficiently large sample size, the Student's t-distribution closely approximates the standardised normal distribution.

Figure 3.1 The normal (Z) and the student's t-distribution with 2, 5 and 30 degrees of freedom

If a variable X is normally distributed and the population standard deviation σ is known, using the normal distribution is appropriate. However, if σ is not known then one should use the student t-distribution with $(n–1)$ degrees of freedom.

3.5.3 Worked example 3.2

The publication of a study using a sample of 30 patients reported a sample mean of 70 kg and a sample standard deviation of 6 kg. Find the 95% confidence interval estimate for the mean weight from this sample.

In Stata we use the `cii means` command to compute the 95% confidence interval given the sample mean, sample standard deviation and the sample size (i.e. without using individual data from a dataset). This command uses the t-distribution, and the output is shown below:

Output 3.2 95%CI for a given sample mean, sample standard deviation and sample size

```
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         30          70    1.095445        67.75956    72.24044
```

We are 95% confident that the true mean weight lies between 67.8 kg and 72.2 kg.

Module 3: Stata notes

3 Learning Activities

Activity 3.1

An investigator wishes to study people living with agoraphobia (fear of open spaces). The investigator places an advertisement in a newspaper asking for volunteer participants. A total of 100 replies are received of which the investigator randomly selects 30. However, only 15 volunteers turn up for their interview.

1. Which of the following statements is true?

a) The final 15 participants are likely to be a representative sample of the population available to the investigator
b) The final 15 participants are likely to be a representative sample of the population of people with agoraphobia
c) The randomly selected 30 participants are likely to be a representative sample of people with agoraphobia who replied to the newspaper advertisement
d) None of the above

2. The basic problem confronted by the investigator is that:

a) The accessible population might be different from the target population
b) The sample has been chosen using an unethical method
c) The sample size was too small
d) It is difficult to obtain a sample of people with agoraphobia in a scientific way

Activity 3.2

A dental epidemiologist wishes to estimate the mean weekly consumption of sweets among children of a given age in her area. After devising a method which enables her to determine the weekly consumption of sweets by a child, she conducted a pilot survey and found that the standard deviation of sweet consumption by the children per week is 85 gm (assuming this is the population standard deviation, σ). She considers taking a random sample for the main survey of:

- 25 children, or
- 100 children, or
- 625 children or
- 3,000 children.

a) Estimate the standard error and maximum likely (95% confidence) error of the sample mean for each of these four sample sizes.
b) What happens to the standard error as the sample size increases? What can you say about the precision of the sample mean as the sample size increases?

Activity 3.3

The dataset for this activity is the same as the one used in Activity 1.4 in Module 1. The file is Activity1.4.dta on Moodle.

a) Plot a histogram of diastolic BP and describe the distribution.
b) Use Stata to obtain an estimate of the mean, standard error of the mean and the 95% confidence interval for the mean diastolic blood pressure.
c) Interpret the 95% confidence interval for the mean diastolic blood pressure.

Activity 3.4

Suppose that a random sample of 81 newborn babies delivered in a hospital located in a poor neighbourhood during the last year had a mean birth weight of 2.7 kg and a standard deviation of 0.9 kg. Calculate the 95% confidence interval for the unknown population mean. Interpret the 95% confidence interval.

Module 4

Hypothesis testing

Learning objectives

By the end of this module you will be able to:

- blah
- blah
- blah

Readings

[**?**]

[**?**]

4.1 Introduction

4 Learning Activities

4.1.1 Activity 4.1

In each of the following situations, what decision should be made about the null hypothesis if the researcher indicates that:

a) P < 0.01
b) P > 0.05
c) 'ns' indicating not significant
d) significant differences exist

4.1.2 Activity 4.2

For the following hypothetical situations, formulate the null hypothesis and alternative hypothesis and write a conclusion about the study results:

a) A study was conducted to investigate whether the mean systolic blood pressure of males aged 40 to 60 years was different to the mean systolic blood pressure of females aged 40 to 60 years. The result of the study was that the mean systolic blood pressure was higher in males by 5.1 mmHg (95% CI 2.4 to 7.6; P = 0.008).
b) A case-control study was conducted to investigate the association between obesity and breast cancer. The researchers found an OR of 3.21 (95% CI 1.15 to 8.47; P = 0.03).
c) A cohort study investigated the relationship between eating a healthy diet and the incidence of influenza infection among adults aged 20 to 60 years. The results were RR = 0.88 (95% CI 0.65 to 1.50; P = 0.2).

4.1.3 Activity 4.3

A pilot study was conducted to compare the mean daily energy intake of women aged 25 to 30 years with the recommended intake of 7750 kJ/day. In this study, the average daily energy intake over 10 days was recorded for 12 healthy women of that age group. The data are in the the Excel file Activity_4.3.xls. Import the file into Stata for this activity.

a) State the research question
b) Formulate the null hypothesis
c) Formulate the alternative hypothesis
d) Analyse the data in Stata and report your conclusions

4.1.4 Activity 4.4

Which procedure gives the researcher the better chance of rejecting a null hypothesis?

a) comparing the data-based p-value with the level of significance at 5%
b) comparing the 95% CI with a nominated value
c) neither procedure

4.1.5 Activity 4.5

Setting the significance level at P < 0.10 instead of the more usual P < 0.05 increases the likelihood of:

a) a Type I error
b) a Type II error
c) rejecting the null hypothesis
d) Not rejecting the null hypothesis

4.1.6 Activity 4.6

For a fixed sample size setting the significance level at a very extreme cutoff such as P < 0.001 increases the chances of: a) obtaining a significant result b) rejecting the null hypothesis c) a Type I error d) a Type II error

Module 5

Comparing the means of two groups

Learning objectives

By the end of this module you will be able to:

- Decide whether to use an independent samples t-test or a paired t-test to compare two groups for a continuous outcome variable;
- Conduct and interpret the results from an independent samples t-test;
- Describe the assumptions of an independent samples t-test;
- Conduct and interpret the results from a paired t-test;
- Describe the assumptions of a paired t-test;
- Conduct an independent samples t-test and a paired t-test in Stata;
- Report results and provide a concise summary of the findings of statistical analyses.

Readings

[**?**]; Sections 7.1 to 7.5.

[**?**]; Section 10.3.

Acock A. A Gentle Introduction to Stata, 3rd Edition (2012): Section 7.7, 7.8.

Juul S and Frydenberg M, An Introduction to Stata for Health Researchers, 4th Edition (2014): Section 11.5.

5.1 Introduction

In Module 4, a one-sample t-test was used for comparing a mean value with a hypothesised value. In this module, we show how to compare the mean values of two groups for which the outcome variable is normally distributed. In health research, we often want to compare the mean value of a measurement between two groups in an observational study or between a control and intervention group in an experimental study. For example, in an observational study, we may want to compare cholesterol levels in people who are normal weight to the levels in people who are overweight. In a clinical trial, we may want to compare cholesterol levels in people who have been randomised to a dietary modification or to usual care.

From the decision tree presented in the Appendix, we can see that if we have a continuous outcome measure and two categorical groups that are not related, i.e. a binary exposure measurement, the test for such data is an independent samples t-test. The test is also sometimes called a 2-sample t-test.

However, in research, data are often ‘paired’ or ‘matched’, that is the two data points are related to one another. This occurs when measurements are taken:

- From each participant on two occasions, e.g. at baseline and follow-up in an experimental study or in a longitudinal cohort study;
- From related people, e.g. a mother and daughter or a child and their sibling;
- From related sites in the same person, e.g. from both limbs, eyes or kidneys;
- From matched participants e.g. in a matched case-control study;
- In cross-over clinical trials where the patient receives both drugs, often in random order.

An independent samples t-test cannot be used for analysing paired or matched data because the assumption that the two groups are independent is violated. Treating paired or matched measurements as independent samples would artificially inflate the sample size and lead to inaccurate analyses and biased P values. When the data are related in a paired or matched way and the outcome is continuous, a paired t-test is the appropriate statistic to use if the data are normally distributed.

5.2 Independent samples t-test

An independent samples t-test is a parametric test that is used to assess whether the mean values of two groups are different from one another. Thus, the test is used to assess whether two mean values are similar enough to have come from the same population or whether the difference between them is so large that the two groups can be considered to have come from separate populations with different characteristics.

The null hypothesis is that the mean values of the two groups are not different, that is:

H0: (Mean2 – Mean1) = 0

Rejecting the null hypothesis using an independent samples t-test indicates that the difference between the means of the two groups is large in relation to the variation in the samples and is unlikely to be due to chance or to sampling variation.

5.2.1 Assumptions for an independent samples t-test

The assumptions that must be met before an independent samples t-test can be used are:

- The two groups are independent
- The measurements are independent
- The outcome variable must be continuous and must be normally distributed in each group
- The variance in the two groups is similar (homogenous)

The first two assumptions are determined by the study design. The two samples must be independent, i.e. if a person is in one group then they cannot be included in the other group, and the measurements within a sample must be independent, i.e. each person must be included in their group once only.

The third assumption of normality is important although t-tests are robust to some degree of non-normality as long as there are no influential outliers and, more importantly, if the sample size is large. We examined how to assess normality in Module 2. If the data are not normally distributed, it may be possible to transform them using a mathematical function such as a logarithmic transformation. If not, then we may need to use non-parametric tests. This is examined in Module 9.

The final assumption is homogeneity of variance between the groups. This can be verified by checking that the standard deviation (square root of the variance) of each group is similar. If the variances are different, then Welch’s t-test, an alternative version of the t-test can be used.

5.2.2 Worked Example

In an observational study of a random sample of 100 full term babies from the community, birth weight and gender were measured. There were 44 male babies and 56 female babies in the sample. The research question asked whether there was a difference in birth weights between boys and girls. The two groups are independent of each other and therefore an independent samples t-test can be used to test the null hypothesis that there is no difference in weight between the genders.

Some preliminary descriptive statistics of the distribution of the variable of interest in each group should always be obtained before a t-test is undertaken to ensure that the assumptions are met. Box plots and histograms are ideal for this. Histrograms and box plots of the data obtained in Stata using **Graphics > Box plot** is shown in Figure 5.1. The dataset Example_5.1.dta is available on Moodle.

The plots show that the data are approximately normally distributed: the histograms are relatively bell shaped and symmetric, and the boxes are fairly symmetrical, there are no outliers as indicated by dots, and the spread is similar in both groups as the similar length of the whiskers suggesting that the variance is homogenous.

We can obtain statistics using the summarize command by gender with the detail option to check the data (e.g. skewness, plausibility of the minimum and maximum values).

```
. by gender, sort: summarize birthweight , detail

---------------------------------------------------------------------
-> gender = Female

                         Birthweight
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.95           2.95
 5%         3.03           2.97
10%         3.14           3.03       Obs                  56
25%        3.325           3.07       Sum of Wgt.          56

50%         3.53                      Mean           3.587411
                        Largest       Std. Dev.      .3629788
75%         3.88            4.2
90%         4.15            4.2       Variance       .1317536
95%          4.2            4.2       Skewness       .2453238
99%         4.25           4.25       Kurtosis       1.962126

---------------------------------------------------------------------
-> gender = Male

                         Birthweight
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.75           2.75
 5%         2.82           2.79
10%         2.85           2.82       Obs                  44
25%         3.15           2.85       Sum of Wgt.          44

50%         3.43                      Mean           3.421364
                        Largest       Std. Dev.      .3536165
75%        3.635           3.94
90%          3.9           3.97       Variance       .1250446
```

```
95%          3.97            4.06       Skewness      -.0895932
99%           4.1             4.1       Kurtosis       2.325761
```

The table shows that girls have a mean weight of 3.59 kg (SD 0.36) and boys have a mean weight of 3.42 kg (SD 0.35) with females being heavier than males. The variabilities of birth weight, as indicated by the standard deviations, are similar.

5.2.3 Conducting and interpreting an independent samples t-test

An independent samples t-test provides us with a t statistic from which we can compute a P value. The computation of the t statistic is as follows:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{SE(\overline{x}_1 - \overline{x}_2)}$$

with $n_1 + n_2 - 2$ degrees of freedom.

Given that the standard error is estimated from the variance, the t value is an estimate of how different the mean values are compared to their variability. Thus, the t value will become larger as the difference in means increases with respect to the variability.

In Stata, both the t and P values are provided. If the t-value falls outside a critical range, the P value will be small and we can reject the null hypothesis of no difference between the groups.

Output 5.3 shows the Stata results of the example dataset obtained using **Statistics - Summaries, tables, and tests - Classical tests of hypotheses - t test (mean-comparison test)** and choosing the two-sample test in the ttest dialog box.

5.2.3.1 Output 5.3: Independent samples t-test results from Stata

```
. ttest birthweight, by(gender)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Female |      56    3.587411    .0485051    .3629788    3.490204    3.684617
    Male |      44    3.421364    .0533097    .3536165    3.313854    3.528873
---------+--------------------------------------------------------------------
combined |     100     3.51435    .0366567    .3665666    3.441615    3.587085
---------+--------------------------------------------------------------------
    diff |            .1660471    .0723027                .0225648    .3095293
------------------------------------------------------------------------------
    diff = mean(Female) - mean(Male)                              t =   2.2966
Ho: diff = 0                                     degrees of freedom =       98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9881         Pr(|T| > |t|) = 0.0238          Pr(T > t) = 0.0119
```

The output table reports the mean, standard deviation, 95% confidence interval etc of weights of the two groups separately as well as that of their difference. It shows the mean difference in weights between the genders is 0.17 kg (95% CI 0.02, 0.31). We are 95% confident that the true mean difference lies between 0.02 and 0.31, this interval does not contain the null value of 0.

Stata reports 3 different P-values. Among them the middle one reports the P-value for a two-sided test evaluating the null hypothesis "Difference =0" and is our desired test. The test has a t-value of 2.297 with 98 degrees of freedom, and a two-sided P value of 0.024 which is less than 0.05 and is statistically significant. Thus, we can reject the null hypothesis of no difference in weights between the genders.

5.3 Paired t-tests

If the outcome of interest is the difference in the continuously distributed outcome measurement between each pair or between each case and its matched control, i.e. the within-pair differences a paired t-test is used. In effect, a paired t-test is used to assess whether the mean of the differences between the two related measurements is significantly different from zero. In this sense, a paired t-test is very closely aligned with a one sample t-test.

When using a paired t-test, the variation between the pairs of measurements is the most important statistic and the variation between the participants, which is critical for the interpretation of a two-sample t-test, is of little interest.

For related measurements, the data for each pair of values must be entered on the same row of the spreadsheet. Thus, the number of rows in the data sheet is the number of participants or the number of participant-pairs when cases and controls are matched. Thus, the effective sample size is the total number of pairs and not the total number of measurements.

5.3.1 Assumptions for a paired t-test

For a paired samples t-test, it is not important to test whether the measurements are normally distributed for each of the time points in the two matched samples, but it is important to test whether the differences between the two measurements are normally distributed.

The assumptions for a paired t-test are:

- the outcome variable is continuous
- the differences between the pair of the measurements are normally distributed

If the assumptions for a paired t-test cannot be met, a non-parametric equivalent is a more appropriate test to use (Module 9).

5.3.2 Computing a paired t-test

The null hypothesis for using a paired t-test is as follows:

H_0: Mean (Measurement1 – Measurement2) = 0

To compute a t-value, the size of the mean difference between the two measurements is compared to the standard error of the paired differences, i.e.

$$t = \frac{\overline{d}}{SE(\overline{d})}$$

with n–1 degrees of freedom, where n is the number of pairs.

Because the standard error becomes smaller as the sample size becomes larger, the t-value increases as the sample size increases for the same mean difference.

5.3.3 Worked Example 5.2

A total of 107 people were recruited into an experimental trial to assess whether ankle blood pressure measured in two different sites would be the same. For each person, systolic blood pressure (SBP) was measured in two sites: dorsalis pedis and tibialis posterior.

The dataset Example_5.2.dta is available on Moodle. First, we need to compute the pairwise difference between SBP measured in the two sites in Stata using the generate command. This is shown in the Stata manual at the end of this module (Checking the assumptions for a Paired t-test) and in the Foundations module. The distribution of the difference between SBP measured in dorsalis pedis and tibialis posterior is shown in Figure 5.2. It approximates a normal distribution and therefore a paired t-test can be used.

5.3.3.1 Figure 5.2: Distribution of differences in ankle SBP between two sites of 107 participants

[INSERT FIGURE]

The paired t-test is performed using the ttest command in Stata (see the Stata Notes section for details). We specify the data is paired with sbp_dp as First variable and sbp_tp as the Second variable. Output 5.4 shows the summary statistics for both sites. From this we can see that the mean SBP is very similar in the two sites.

5.3.3.2 Output 5.4: Paired t-test results from Stata

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  sbp_dp |     107     116.729    3.460296    35.79358    109.8686    123.5893
  sbp_tp |     107    117.9907    3.431356    35.49422    111.1877    124.7937
---------+--------------------------------------------------------------------
    diff |     107   -1.261682    1.311368    13.56489   -3.861596    1.338232
------------------------------------------------------------------------------
     mean(diff) = mean(sbp_dp - sbp_tp)                           t =  -0.9621
 Ho: mean(diff) = 0                              degrees of freedom =      106

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.1691         Pr(|T| > |t|) = 0.3382          Pr(T > t) = 0.8309
```

The next line for “diff” shows the statistics for the mean of within-pair difference. It indicates that average SBP measured in dorsalis pedis is 116.7 mmHg and that in tibialis posterior is 118.0 mmHg. The difference is −1.26 (95% CI: −3.86 to 1.34).

The t-value of −0.96 yields a two-sided P-value of 0.34 (under Ha: mean(diff) ≠ 0) confirms that these data provide no evidence against the null hypothesis, and conclude that the SBP measured in the two sites are not different.

As with any statistical test, it is important to decide what mean difference between measurements would be considered clinically important in addition to considering statistical significance.

5 Learning Activities

Activity 5.1

Indicate what type of t-test could be used to analyse the data from the following studies and provide reasons:

a) A total of 60 university students are randomly assigned to undergo either behaviour therapy or Gestalt therapy. After twenty therapeutic sessions, each student earns a score on a mental health questionnaire.
b) A researcher wishes to determine whether attendance at a day care centre increases the scores of three year old twins on a motor skills test. Random assignment is used to decide which member from each of 30 pairs of twins attends the day care centre and which member stays at home.
c) A child psychologist assigns aggression scores to each of 10 children during two 60 minute observation periods separated by an intervening exposure to a series of violent TV cartoons.
d) A marketing researcher measures 100 doctors' reports of the number of their patients asking them about a particular drug during the month before and the month after a major advertising campaign.

Activity 5.2

A study was conducted to compare haemoglobin levels in the blood of children with and without cystic fibrosis. It is known that haemoglobin levels are normally distributed in children. The study results are given below:

Table 5.1: Summary of haemoglobin (g/dL)

Statistic	**Children without CF**	**Children with CF**
n	12	15
Mean	19.9	13.9
SD (SE)	5.9 (1.70)	6.2 (1.60)

a) State the appropriate null hypothesis and alternate hypothesis
b) Use Stata to conduct an appropriate statistical test to evaluate the null hypothesis. Are the assumptions for the test met for this analysis to be valid?

Activity 5.3

A randomised controlled trial (RCT) was carried out to investigate the effect of a new tablet supplement in increasing the hematocrit (%) value in anaemic participants. In the study, hematocrit was measured as the proportion of blood that is made up of red blood cells. Hematocrit levels are often lower in anaemic people who do not have sufficient healthy red blood cells. In the RCT, 33 people in the intervention group received the new supplement and 31 people in the control group received standard care (i.e. the usual supplement was given). After 4 weeks, hematocrit values were measured as shown in the Stata file ActivityS5.3.dta. In the community, hematocrit levels are normally distributed.

a) State the research question and frame it as a null hypothesis.
b) Use Stata to conduct an appropriate statistical test to answer the research question. Before using the test, check the data to see if the assumptions required for the test are met. Obtain a box plot to obtain an estimate of the centre and spread of the data for each group.
c) Run your statistical test.
d) Construct a table to show how you would report your results and write a conclusion.

Activity 5.4

A total of 41 babies aged 6 months to 2 years with haemangioma (birth mark) were enrolled in a study to test the effect of a new topical medication in reducing the volume of their haemangioma. Parents were asked to apply the medication twice daily. The volume (in mm3) of the haemangioma was measured at enrolment and again after 12 weeks of using the medication.

a) What is the research question in this study? State the null and alternative hypotheses.
b) Use the data in the Stata file ActivityS5.4.dta to answer the research question. Which statistical test is appropriate to answer the research question and why? Conduct the test in Stata and write your conclusion.
c) What are the limitations of this study?

Module 6

Summary statistics for binary data

Learning objectives

By the end of this module you will be able to:

- Compute and interpret 95% confidence intervals for proportions;
- Conduct and interpret a significance test for a one-sample proportion;
- Use Stata to compute 95% confidence intervals for a difference in proportions, a relative risk and an odds ratio.

Readings

[**?**]; Chapter 16

[**?**]; Section 8.6, Section 13.7

[Juul and Frydenberg, 2014]; Section 11.4.

[Acock, 2010]; Section 7.5.

6.1 Introduction

In Modules 4 and 5, we discussed methods used to test hypotheses when the data are continuous. In Modules 6 and 7, we will focus on hypothesis testing for binary categorical data.

In health research, we often collect information that can be put into two categories, e.g. male and female, disease present or disease absent etc. Binary categorical variables such as these are summarised using proportions.

6.2 Calculating proportions and 95% confidence intervals

6.2.1 Calculating a proportion

Calculating a proportion is based on the binomial distribution, which was introduced in Module 2. We need two pieces of information to calculate a proportion: n, the number of trials, and k, the number of 'successes'. Note that we use the term 'success' to describe the outcome of interest, recognising that a success may be a adverse outcome such as death or disease.

The following formula is used to calculate the proportion, p:

$$p = k/n$$

The proportion, p, is a number that lies between 0 and 1. Proportions and their confidence intervals can easily be converted to percentages by multiplying by 100 once computed.

As for all summary statistics, it is useful to compute the precision of the estimate as a 95% confidence interval (CI) to indicate the range of values in which are 95% confident that the true population value lies. In this module, we present two methods for computing a 95% confidence interval around a proportion.

6.2.2 Calculating the 95% confidence interval of a proportion (Wald method)

The Wald method for calculating the 95% confidence interval is based on assuming that the proportion, p, is Normally distributed. This assumption is reasonable if the sample is sufficiently large (for example, if $n > 30$) and if $n \times (1 - p)$ and $n \times p$ are both larger than 5.

The Wald method for calculating a 95% confidence interval is given by:

$$\text{95\% CI} = p \pm (1.96 \times \text{SE}(p))$$

where the standard error of a proportion is computed as:

$$\text{SE}(p) = \sqrt{\frac{p \times (1 - p)}{n}}$$

6.2.3 Worked Example

In a cross-sectional study of children living in a rural village, 47 children from a random sample of 215 children were found to have scabies. Here $n = 215$ and $k = 47$, so the proportion of children with scabies is estimated as:

$$p = \frac{47}{215} = 0.2186$$

Given the large sample size and the number of children with the rarer outcome is larger than 5, the Wald method is used to calculate the standard error of the proportion as:

$$\text{SE}\,(p) = \sqrt{\frac{0.2186 \times (1 - 0.2186)}{215}} = 0.02819$$

Then, the 95% confidence interval is estimated as:

$$\text{95\% CI} = 0.2816 \pm 0.02819$$

$$= 0.1634\, to\, 0.2739$$

The prevalence of scabies among children in the village is 21.9% (95% CI 16.3%, 27.4%). These values tell us that we are 95% confident that the true prevalence of scabies among children in the village is between 16.3% and 27.4%.

NB: This can also be computed in Stata using the `cii proportions` command as below.

6.2.3.1 Output 6.1: 95% confidence interval for the prevalence of scabies using normal approximation to the binomial distribution

```
. cii proportions 215 47, wald

                                                          -- Binomial Wald ---
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        215    .2186047    .0281868        .1633595    .2738498
```

6.2.4 Calculating the 95% confidence interval of a proportion (Wilson method)

Another method to calculate the confidence interval of a proportion is the Wilson (sometimes also called the 'score') method. We can use it in situations where it is not appropriate to use the normal approximation to the binomial distribution as described above i.e. if the sample size is small ($n < 30$) or the number of subjects with the rarer outcome is 5 or fewer. This method is a bit more difficult to implement by hand than the standard confidence interval, and so we will not discuss the hand calculation using the mathematical equation in this course. Instead, we will use the Stata command `cii proportions` specifying `wilson` as an option to do this.

Using the data from the study given in Worked Example 6.1, we obtain the following:

6.2.4.1 Output 6.2: 95% confidence interval for prevalence of scabies using the Wilson method

```
. cii proportions 215 47, wilson

                                                          ------ Wilson ------
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        215    .2186047    .0281868        .1685637    .2785246
```

6.2.5 Wald vs Wilson methods

We have presented two methods for calculating the 95% confidence interval for a proportion. The Wald method, which assumes that the underlying proportion follows a Normal distribution, is easy to calculate and follows the form of other confidence intervals. The Wilson method, which is more difficult to calculate by hand, has nicer mathematical properties. There are also a number of other methods for calculating confidence intervals for proportions, but we do not discuss these in this course.

A paper by Brown, Cai and DasGupta (Statistical Science, 2001) has compared the properties of the Wald and Wilson methods (among others) and concluded that the Wilson method is preferred over the Wald method.

6.3 Hypothesis testing for one sample proportion

We can carry out a hypothesis test to compare a sample proportion to a hypothesised proportion. In much the same way as a one sample t-test was used in Module 5 to test a sample mean against a hypothesised mean, we can perform a one-sample test to test a sample proportion against a hypothesised proportion. The significance test will provide a P value to assess the evidence against the null hypothesis, while the 95% confidence interval will provide the range in which we are 95% confident that the true proportion lies.

For example, we can test the following null hypothesis:

H0: sample proportion is not different from the hypothesised proportion

Much like constructing a 95% confidence interval, there are two main options when performing a hypothesis test on a single proportion: the first assumes that the proportion follows a Normal distribution, while the second relaxes this assumption.

6.3.1 z-test for testing one sample proportion

The first step in the z-test is to calculate a z-statistic, which is then used to calculate a P-value. The z-statistic is calculated as the difference between the population proportion and the sample proportion divided by the standard error of the population proportion, i.e.

$$z = \frac{(p_{population} - p_{sample})}{\text{SE}(p_{population})}$$

This z-statistic is then compared to the standard Normal distribution to calculate the P-value.

6.3.2 Worked Example

$$\begin{aligned} z &= \frac{(0.20 - 0.18)}{\sqrt{\frac{0.20 \times (1 - 0.20)}{300}}} \\ &= 0.87 \end{aligned}$$

This Z value does not meet or exceed the critical value of 1.96 for a two tailed test. This indicates that there is insufficient evidence to conclude that there is a difference between the population proportion of 20% smokers and the sample proportion of 18% smokers which is consistent with our hypothesis testing using 95% CIs.

The P-value for the test above can be obtained from a Normal distribution table as $P = 2 \times 0.192 = 0.38$ (using Table A2.1 in the Appendix), or using the hand-calculator in Stata. **INCLUDE NORMAL CURVE DIAGRAM TO ILLUSTRATE 2-TAILED TEST**

6.3.3 Binomial test for testing one sample proportion

We can use the binomial distribution to obtain an exact P-value for testing a single proportion. Historically, this was a time consuming process with much hand calculation. These days, Stata and other statistical software performs the calculations quickly and efficiently, and is the preferred method.

6.3.4 Worked example

A national census in a country shows that 20% of the population are smokers. A survey of a community within the country that has received a public health anti-smoking intervention shows that 54 of 300 people sampled are smokers (18%). The research question is whether this proportion of smoking in the community is lower than the population prevalence of smoking of 20%. The Stata file Example_6.3.dta contains the data for this example. In the data file, smokers are coded as 1 and non-smokers are coded as 0.

In Stata, we can use the `prtest` command to perform a z-test, or the `bitest` command to perform the exact binomial test:

6.3.4.1 Output 6.3: z-test and binomial test for prevalence of smoking

```
. prtest smoking_status == 0.2

One-sample test of proportion                      Number of obs      =       300
```

```
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
smoking_st~s |        .18   .0221811                      .1365259    .2234741
------------------------------------------------------------------------------
    p = proportion(smoking_st~s)                                  z =  -0.8660
Ho: p = 0.2

     Ha: p < 0.2                 Ha: p != 0.2                 Ha: p > 0.2
 Pr(Z < z) = 0.1932         Pr(|Z| > |z|) = 0.3865          Pr(Z > z) = 0.8068

bitest smoking_status == 0.2

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
smoking_st~s |      300         54           60       0.20000      0.18000

  Pr(k >= 54)           = 0.825531  (one-sided test)
  Pr(k <= 54)           = 0.215202  (one-sided test)
  Pr(k <= 54 or k >= 66) = 0.427280  (two-sided test)
```

The z-test provides a two-sided P-value of 0.39, while the binomial test gives a two-sided P-value of 0.43. Both tests provide little evidence against the hypothesis that the prevalence of smoking in the community is 20%.

6.4 Contingency tables

As introduced in PHCM9794: Foundations of Epidemiology, 2-by-2 contingency tables can be used to examine associations between two binary variables, most commonly an exposure and an outcome. There are two commands in Stata to construct and analyse 2-by-2 contingency tables: `cs` (for cross-sectional or cohort studies) and `cc` (for case-control studies).

It is important to note that Stata presents the exposure or intervention (present, absent) in the columns and the outcome or disease (present, absent) in the rows. This is opposite to the way most epidemiological tables are presented, with exposure in rows and outcome in columns. Care must be taken when reading 2-by-2 tables generated from Stata.

6.4.0.1 Table 6.1: Contingency tables for estimating associations between two binary variables

Traditional format

	Outcome present	Outcome absent	Total
Exposure present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	N

Stata format

	Exposure present	Exposure absent	Total
Outcome present	a	c	a+c
Outcome absent	b	d	b+d

Stata format			
Total	a+b	c+d	N

When using a statistics program such as Stata, it is recommended that the outcome and exposure variables are coded by assigning ‘absent’ as 0 and ‘present’ as 1, for example ‘No’ = 0 and ‘Yes’ = 1. This is needed for some of the commands to work (e.g. the epidemiology table commands). This coding ensures that measures of association, such as the odds ratio or relative risk, are computed correctly by Stata.

6.5 Calculation of the 95%CI for relative risk, odds ratio and other measures of association

We can measure the strength of the association between an exposure and an outcome as either a relative risk or odds ratio. The relative risk is a direct comparison of the risk in the exposed group with the risk in the non-exposed group, and can only be calculated for a cohort study (including a randomised controlled trial) or a cross-sectional study (where it is also called a prevalence ratio).

6.5.1 Worked Example 6.4

A randomised controlled trial was conducted among a group of patients to estimate the side effects of a drug. Fifty patients were randomly allocated to receive the active drug and 50 patients were allocated to receive a placebo drug. The outcome measured was the experience of nausea. The data is given in the file Example_6.4.dta.

The relative risk (RR=3.75) and its 95% confidence interval (1.34, 10.51) shown in Output 6.4 can be obtained using the cs command in Stata. This tells us that nausea is 3.75 times more likely to occur in the active drug group compared with the placebo group. Because this is a randomised controlled trial, relative risk would be an appropriate measure of association.

6.5.1.1 Output 6.4 Relative risk and 95% CI from the cs command in Stata

```
                 | Group                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       15           4   |         19
        Noncases |       35          46   |         81
-----------------+------------------------+------------
           Total |       50          50   |        100
                 |                        |
            Risk |       .3         .08   |        .19
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |             .22        |    .0723899    .3676101
      Risk ratio |            3.75        |     1.33754     10.5137
 Attr. frac. ex. |         .7333333       |    .2523589     .904886
 Attr. frac. pop |         .5789474       |
                 +-------------------------------------------------
                              chi2(1) =     7.86  Pr>chi2 = 0.0050
```

From Output 6.4, you can check the relative risk estimate:

$$\begin{aligned}\text{RR} &= \frac{a/(a+b)}{c/(c+d)}\\ &= \frac{15/(15+35)}{4/(4+46)}\\ &= \frac{0.3}{0.08}\\ &= 3.75\end{aligned}$$

6.5.2 Worked Example 6.5

A case-control study investigated the association between human papillomavirus and oropharyngeal cancer (D'Souza, et al. NEJM 2007), and the results appear in Table 6.2.

6.5.2.1 Table 6.2: Association between human papillomavirus and oropharyngeal cancer

	Cases (Oropharyngeal cancer)	Controls (No oropharyngeal cancer)	Total
HPV Positive	57	14	71
HPV Negative	43	186	229
Total	100	200	300

The odds ratio is the odds of being HPV positive in those with oropharyngeal cancer compared to the odds of being HPV positive in those without oropharyngeal cancer.

You can use the Stata command cci with the Cornfield option to obtain odds ratio and its 95% CI as shown in Output 6.5. The Cornfield option is used to provide a better estimate of the 95% confidence interval.

6.5.2.2 Output 6.5: Odds ratio and 95% CI from the cc command in Stata

```
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     exposed
-----------------+------------------------+------------------------
           Cases |        57          43  |        100       0.5700
        Controls |        14         186  |        200       0.0700
-----------------+------------------------+------------------------
           Total |        71         229  |        300       0.2367
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         17.6113        |    9.043258    34.25468 (Cornfield)
 Attr. frac. ex. |        .9432183        |    .8894204    .9708069 (Cornfield)
 Attr. frac. pop |        .5376344        |
                 +-------------------------------------------------
                               chi2(1) =    92.26  Pr>chi2 = 0.0000
```

The odds ratio (OR) and its 95% CI can be read directly from the output as: OR = 17.6 (95% CI 9.0, 34.3).

From the cross-tabulated output, you can check the odds ratio estimate as follows:

$$
\begin{aligned}
\text{OR} &= \frac{a/c}{b/d)} \\
&= \frac{57/14}{43/186} \\
&= 17.6
\end{aligned}
$$

Identical Stata output to Outputs 6.4 and 6.5 can be obtained for either individual record data or aggregate data. The steps for computing RR and OR using both individual record and aggregate data is described in the Stata notes section.

Also estimated in Output 6.4 is the risk difference, and in both Outputs 6.4 and 6.5, the population attributable fraction (or proportion) and the attributable fraction among the exposed and their corresponding 95% CI. While the value of 1 indicates no effect for both OR and RR, the value of 0 indicates no effect for risk difference and the attribution fractions.

Risk statistics are usually only reported with one or two decimal places. The interpretation of the confidence intervals for both the relative risk and the odds ratio is the same as for the confidence intervals around other summary measures in that it shows the region in which we are 95% confident that the true population estimate lies.

6 Learning Activities

Activity 6.1

In a clinical trial involving a dietary intervention, 150 adult volunteers agreed to participate. The investigator wanted to know whether this sample was representative of the general population. One interesting finding was that 90 of the participants drink alcohol regularly compared to 70% of the general population.

a) State the null hypothesis
b) Calculate the 95% CIs for the proportion of regular drinkers in the sample using Stata.
c) Use the Stata file Activity_S6.1.dta to decide if the sample of volunteers is representative of the population?

Activity 6.2

A survey was conducted of a random sample of upper primary school children to measure the prevalence of asthma using questionnaires completed by the parents. A total of 514 children were enrolled. Use the Stata dataset Activity_S6.2.dta for this activity.

a) Calculate the relative risk and odds ratio with 95% confidence interval using Stata for children to have asthma symptoms if they are male? Which risk estimate would be the correct statistic to report?
b) Use the tabulated data on the frequency of cases and exposure you obtained in Stata output in part a to calculate RR and OR with their 95% confidence interval using Stata.

Activity 6.3

In a study to determine the cause of mortality, 89 people were followed up for 5 years. The participants are classified into two groups of those who did or did not have a heart attack. At the end of the follow-up 15 people died among them 10 had a heart attack. Among the 74 survivors 35 had a heart attack. Present the data in a 2-by-2 table and calculate relative risk of death from heart attack with 95% confidence interval using Stata.

Activity 6.4

A study is conducted to test the hypothesis that the observed frequency of a certain health outcome is 30%. If the results yield a CI around the sample proportion that extends from 23.8 to 30.2, what can you say about the evidence against the null hypothesis?

Activity 6.5

In an experiment to test the effect of vitamin C on IQ scores, the following confidence intervals were estimated around the percentage with improved scores for five different populations:

a) Which CI is the most precise?

Table 6.4: Summary of improvement in IQ

Population	% with improved IQ	95% confidence interval
1	30.0	32.0 to 38.0
2	29.5	25.0 to 34.0
3	43.5	42.0 to 45.0
4	30.5	20.0 to 41.0
5	24.5	21.0 to 28.0

b) Which CI implies the largest sample size?
c) Which CI is the least precise?
d) Which CI most strongly supports the conclusion that vitamin C increases IQ score and why?
e) Which would most likely to stimulate the investigator to conduct an additional experiment using a larger sample size?

Module 7

Hypothesis testing for categorical data

Learning objectives

By the end of this module you will be able to:

- blah
- blah
- blah

Readings

[?]

[?]

7.1 Introduction

7 Learning Activities

Activity 7.1

Use Stata and the Stata file `Activity_S7.1.dta` to further investigate whether there is a gender difference in asthma in a random sample of 514 upper primary school children:

a) Use a contingency table (cross-tabulation) to determine the observed and expected frequencies. Which cell has the lowest expected cell count?
b) Use a chi-squared test to evaluate the hypothesis and interpret the result. Are the assumptions for a chi-squared test met? Calculate the 95% CI of the difference in proportions.

Activity 7.2

The Stata file `Activity_S7.2.dta` summarises 5-year mortality (the outcome) for 89 people who did or did not have a heart attack (the exposure).

a) State the null hypothesis.
b) Using Stata, carry out the appropriate significance test to evaluate the hypothesis. Do the data fulfil the assumptions of the statistical test you have used?
c) Estimate the appropriate risk estimate for mortality. Are the confidence intervals of the risk estimates consistent with the P value?
d) Summarise your results and state your conclusion.

Activity 7.3

The effect of two penicillin allergens B and G was tested in a random sample of 500 people. All people were tested with both allergens. For each person, data were recorded for whether or not there was an allergic reaction to the allergen.

Use the Stata data set `Activity_S7.3.dta` to test the null hypothesis that the proportion of participants who react to allergen G is the same as the proportion who react to allergen B. Are the 95% CI around the difference consistent with the P value?

Activity 7.4

We examined a survey of 200 live births in an urban region in which 2 babies were born prematurely. We also surveyed 80 live births in a rural region and found that 5 babies were born prematurely. Conduct an appropriate statistical analysis to find out whether the proportion of premature births is higher in the rural region.

Module 8

Correlation and linear regression

Learning objectives

By the end of this module you will be able to:

- Explore the association between two continuous variables using a scatter plot;
- Estimate and interpret correlation coefficients;
- Estimate and interpret parameters from a simple linear regression;
- Decide whether a regression model is valid;
- Test a hypothesis using regression coefficients;
- Outline the concept of multiple regression and its role in investigative epidemiology.

Readings

[**?**]; Chapter 10

[**?**]; Chapter 11

[Acock, 2010]; Chapter 8.

[Juul and Frydenberg, 2014]; Section 12.1.

8.1 Introduction

In Module 5, we saw how to test whether a categorical and a continuous variable are related. However, we often want to know how closely two continuous variables are related. For example, we may want to know how closely blood cholesterol levels are related to dietary fat intake in adult men. To measure the strength of association between two continuously distributed variables, a correlation coefficient is used.

We may also want to know how well one continuous measurement predicts the value of another continuous measurement. For example, we may want to know how well height predicts values of lung capacity in a community of adults. A regression model allows us to use one measurement to predict another measurement.

Although both correlation coefficients and regression models can be used to describe the degree of association between two continuous variables, the two methods provide very different statistical information. For both methods, a significant statistical association only implies an association between the variables and does not imply a causal relationship.

8.2 Correlation

We use correlation to measure the strength of a linear relationship between two variables. Before calculating a correlation coefficient, a scatter plot should first be obtained to give an understanding of the nature of the relationship between the two variables.

8.2.1 Worked Example

The Stata file Example_8.1.dta has information about height and lung function collected from a sample of 120 adults. A random sample of adults was approached to take part in the research study, but the response rate was low at 45%. Information was collected on height (cm) and lung function, which was measured as forced vital capacity (FVC). Using the twoway command in Stata we can obtain the plot shown in Figure `\@ref(fig:scatter-plot))`. This shows that as height increases, lung function also increases, which is as expected. One or two of the data points are separated from the rest of the data but are not so far away as to be considered outliers because they do not seem to stand out of other observations.

8.2.2 Correlation coefficients

A correlation coefficient (r) describes how closely the variables are related, that is the strength of linear association between two continuous variables. The range of the coefficient is from +1 to −1 where +1 is a perfect positive association, 0 is no association and −1 is a perfect inverse association. In general, an absolute (disregarding the sign) r value below 0.3 indicates a weak association, 0.3 to < 0.6 is fair association, 0.6 to < 0.8 is a moderate association, and $\geq$ 0.8 indicates a strong association.

The coefficient is positive when large values of one variable tend to occur with large values of the other, and small values of one variable (y) tend to occur with small values of the other (x) (Figure `\@ref(fig:scatter-plot-four)` (a and b)). For example, height and weight in healthy children or age and blood pressure.

The coefficient is negative when large values of one variable tend to occur with small values of the other, and small values of one variable tend to occur with large values of the other (Figure `\@ref(fig:scatter-plot-four)` (c and d)). For example, percentage immunised against infectious diseases and under-five mortality rate.

The P value associated with an r value is an estimate of whether the correlation coefficient is significantly different from zero. However, a correlation coefficient that does not have a significant P value does not imply that there is no relationship because the correlation coefficient only tests for a linear association and there may be a non-linear relationship such as a curved or irregular relationship.

The assumptions for using a Pearson's correlation coefficient are that:

- observations are independent;
- both variables are continuous variables;
- the relationship between the two variables is linear.

There is a further assumption that the data follow a bivariate normal distribution. This assumes: *y* follows a normal distribution for given values of *x*; and *x* follows a normal distribution for given values of *y*. This is quite a technical assumption that we do not discuss further.

There are two types of correlation coefficients– the correct one to use is determined by the nature of the variables as shown in Table 8.1).

Spearman's rho is calculated using the ranks of the data, rather than the actual values of the data. We will see further examples of such methods in Module 9, when we consider non-parametric tests, which are often based on ranks.

For the data in the Worked Example 8.1, using the pwcorr command in Stata gives the information shown in Output 8.1.

Table 8.1: A table

Type.of.correlation.coefficient	Application
Pearson’s correlation coefficient: r	Both variables are continuous and a bivariate normal distribution can be assumed
Spearman’s rank correlation: rho	Bivariate normality cannot be assumed. Also useful when at least one of the variables is ordinal

8.2.2.1 Stata output from the pwcorr command

```
. pwcorr Height FVC, sig

             |   Height      FVC
-------------+------------------
      Height |   1.0000
             |
             |
         FVC |   0.6976   1.0000
             |   0.0000
             |
```

This table shows that the Pearson’s correlation coefficient between height and lung function is 0.698 with P<0.001 indicating very strong evidence of a linear association between height and FVC.

This r value was calculated for the full data set of 120 adults who had heights ranging from 160 to 172cms. If the r value is calculated for the 60 adults with a height less than 165cms, it is much lower at 0.433 although significant at P=0.001. In general, r values are higher for a wider range of values on the x axis even though the relationship between the two variables remains the same.

Correlation coefficients are rarely used as important statistics in their own right because they do not fully explain the relationship between the two variables and the range of the data has an important influence on the size of the coefficient. In addition, the statistical significance of the correlation coefficient is often over interpreted because a small correlation which is of no clinical importance can become statistically significant even with a relatively small sample size. For example, a poor correlation of 0.3 will be statistically significant if the sample size is large enough.

8.3 Linear regression

The nature of a relationship between two variables is more fully described using regression. There are two principal purposes for building a regression model. The most common is to build a predictive model, for example in situations in which age and gender are used to predict normal values of characteristics such as lung size or body mass index. Normal values are the range of values that occur naturally in the general population.

The second purpose for using a regression model is for testing the hypothesis that there is a linear relationship between one or more explanatory variables and an outcome variable. For example, a regression model can be used to test the extent to which age predicts BMI or to test the hypothesis that two groups with a different dietary regime have significantly different BMI values after adjusting for age differences.

From Worked Example 8.1, Stata can be used to plot a regression line through the scatter. Figure `\@ref(fig:scatter-plot-line)` shows the data with the line fitted.

The line through the plot is called the line of 'best fit' because the size of the deviations between the data points and the line is minimised in the calculation. The distance between each data point and the regression line is called a 'residual'.

8.3.1 Regression equations

The mathematical equation for the line explains the relationship between the two variables. The equation of the regression line is as follows:

$$y = \beta_0 + \beta_1 x$$

This line is shown in Figure `\@ref(fig:regression-parameters)` using the notation shown in Table 8.2.

Table 8.2: Notation for linear regression equation

Symbol	**Interpretation**
y	Observed value of the outcome variable
x	Observed value of the explanatory variable
β_0	Intercept of the regression line
β_1	Slope of the regression line

The intercept is the point at which the regression line intersects with the y-axis when the value of 'x' is zero. In most cases, the intercept does not have a biologically meaningful interpretation. The slope of the line is the unit change in the outcome variable 'y' with each unit change in the explanatory variable 'x'. For any data set, the fitted regression line passes through the mean values of both the explanatory variable 'x' and the outcome variable 'y'.

When using regression, the research question must be framed so that the explanatory variable 'x' and outcome variable 'y' are classified correctly. An important concept is that regression predicts a mean value of 'y' given an observed value of 'x' so that any error around the explanatory variable is not taken into account. For this reason, measurements that can be taken accurately, such as age and height, make good explanatory variables.

8.3.2 Fit of a linear regression model

After fitting a linear regression model, it is important to know how well the model fits the observed data. One way of assessing the model fit is to compute a statistic called coefficient of determination, denoted by R^2. It is the square of the Pearson correlation coefficient $r : r^2 = R^2$. Since the range of r is from −1 to 1, R^2 must lie between 0 and 1.

R^2 can be interpreted as the proportion of variability in y that can be explained by variability in x. Hence, the following conditions may arise:

If $R^2 = 1$, then all variation in y can be explained by variation of x and all data points fall on the regression line.

If $R^2 = 0$, then none of the variation in y is related to x at all, and the variable x explains none of the variability in y.

If $0 < R^2 < 1$, then the variability of y can be partially explained by the variability in x. The larger the R^2 value, the better is the fit of the regression model.

8.3.3 Assumptions for linear regression

Regression is robust to moderate degrees of non-normality in the variables, provided that the sample size is large enough and that there are no influential outliers. Also, the regression equation describes the relationship between the variables and this is not influenced as much by the spread of the data as the correlation coefficient is.

The assumptions that must be met when using linear regression are as follows:

- observations are independent;
- the relationship between the explanatory and the outcome variable is linear;
- the residuals are normally distributed.

A residual is defined as the difference between the observed and predicted outcome from the regression model. If the predicted value of the outcome variable is denoted by $\widehat{y}$ then:

$$\text{Residual} = \text{observed} - \text{predicted} = y - \widehat{y}$$

It is important for regression modelling that the data are collected in a period when the relationship remains constant. For example, in building a model to predict normal values for lung function the data must be collected when the participants have been resting and not exercising and people taking bronchodilator medications that influence lung capacity should be excluded. In regression, it is not so important that the variables themselves are normally distributed, but it is important that the residuals are. Scatter plots and specific diagnostic tests can be used to check the regression assumptions. Some of these will not be covered in this introductory course but will be discussed in detail in the **Advanced Biostatistics** course.

8.4 Obtaining a regression equation in Stata

To measure whether height is a significant predictor of forced vital capacity (FVC), we use the `regress` command in Stata.

Output 8.2 shows the model summary in the first part of the Stata output. The R-squared is 0.487, indicating that 48.7% of the variation in FVC is explained by height. The square root of R-squared gives us the (absolute value of) Pearson's correlation coefficient of 0.698 as obtained in Section 8.2.

8.4.0.1 Output 8.2: Model summary from the regress command in Stata

```
. regress FVC Height

      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(1, 118)       =    111.88
       Model |  17.5914327         1  17.5914327   Prob > F        =    0.0000
    Residual |  18.5540027       118  .157237311   R-squared       =    0.4867
-------------+----------------------------------   Adj R-squared   =    0.4823
       Total |  36.1454355       119  .303743155   Root MSE        =    .39653

------------------------------------------------------------------------------
         FVC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Height |   .1407567   .0133075    10.58   0.000     .1144042    .1671092
```

```
       _cons |  -18.87347   2.193651    -8.60   0.000    -23.2175   -14.52944
------------------------------------------------------------------------------
```

The adjusted R-squared is only used when comparing multivariable models (i.e. those with different numbers of explanatory variables, including confounders), and will not be used in this course.

The coefficients table, the second part of the Stata output, provide the estimated regression coefficients. Stata labels the regression slope with the name of the explanatory variable and the intercept `_cons`.

From this output, we see that the slope is estimated as 0.141 with an estimated intercept of -18.873. Therefore, the regression equation is estimated as:

FVC (L) = − 18.873 + (0.141 × Height in cm)

This equation can be used to predict FVC for a person of a given height. For example, the predicted FVC for a person 165 cm tall is estimated as:

FVC = − 18.87347 + (0.1407567 × 165.0) = 4.40 L.

Note that for the purpose of prediction we have kept all the decimal places in the coefficients to avoid rounding error in the intermediate calculation.

The t-values are calculated by dividing the coefficients by their SEs and are tests of whether each coefficient is significantly different from zero. A coefficient that is significantly different from zero indicates a significant linear relationship between the variables. In this model, both the intercept and the coefficient are significantly different from zero at $P < 0.001$.

In the above example, the response rate for the survey was low and there may be selection bias in that people who were healthier may have been more likely to attend so the predictive equation may not be considered representative of the general population of adults from which the sample was drawn.

The distribution of the residuals should always be checked. Outlying residuals can have a large effect on the slope of the model and need to be censored or brought closer to the remainder of the data to reduce their influence. The residuals can be generated using the predict command in Stata.

The histogram of residuals from the model is shown in Figure 8.5. They are normally distributed and indicate that there are no influential outliers so that the assumptions for regression are met.

8.4.1 Critical appraisal

When reading the literature, it is important to be critical about how correlation coefficients are interpreted. It is a good idea to check if a scatter plot is shown to help interpret the relationship and to indicate if there are any influential outliers. Also, question whether the correlation coefficient has been calculated from a random sample and if not, what selected samples the value can be generalised to.

When regression is reported it is essential that the axes are correctly presented so that the equation is predictive. Thus, the explanatory variable must be presented on the x axis and the outcome on the y axis. It is also a good idea to check that all the assumptions are met. Outliers which result in a non-normal distribution of the residuals can severely bias the regression coefficients.

8.5 Multiple linear regression

In the above example, we have only used a simple linear regression model of two continuous variables. Other more complex models can be built from this e.g. if we wanted to look at the effect of gender (male vs. female) as binary indicator in the model while adjusting for the effect of height. In that case we would include both the variables in the model as explanatory variables. In the same way we can include any number of explanatory variables (both continuous and categorical) in the model: this is called a multivariable model. Multivariable models are often used for building predictive equations, for example by using age, height, gender and smoking history to predict lung function, or to adjust for confounding and detect effect modification to investigate the association between an exposure and an outcome factor.

Multiple regression has an important role in investigating causality in epidemiology. The exposure variable under investigation must stay in the model and the effects of other variables which can be confounders or effect-modifiers are tested. The biological, psychological or social meaning of the variables in the model and their interactions are of great importance for interpreting theories of causality.

Other multivariable models include binary logistic regression for use with a binary outcome variable, Cox regression for survival analyses, or Poisson regression for count data. These models, together with multiple regression, will be taught in *PHCM9517: Advanced Biostatistics*.

8 Stata notes

8.6 Creating a scatter plot

We will demonstrate using Stata for correlation and simple linear regression using the dataset `Example_8.1.dta`.

To create a scatter plot to explore the association between height and FVC click: **Graphics > Twoway graph (scatter, line, etc.)**. In the `twoway` dialog box, click **Create…**

A new dialog box will open. Select the **Basic plots** radio button and highlight **Scatter** under **Basic plots: (select type)**. Choose **FVC** for the **Y variable** and **Height** for the **X variable**.

Click the **Accept** button in the **Plot 1** dialog box to return to the **twoway** dialog box, then click the **OK** or **Submit** button to produce the scatter plot shown in **Figure 8.1**.

[Command: `twoway (scatter FVC Height)`]

To add a fitted line, go back to the `twoway` dialog box. If you clicked the **OK** button, you can go to **Graphics > Twoway graph (scatter, line, etc.)** to bring it back again.

Click **Create…**, then select the **Fit plots** radio button and **Linear prediction** under **Fit plots: (select type)**. Choose **FVC** for the **Y variable** and **Height** for the **X variable**.

Click the **Accept** button, then the **OK** or **Submit** button to produce the scatterplot below.

[Command: `twoway (scatter FVC Height) (lfit FVC Height)`]

Notice that a legend now appears, and the y-axis title is missing. To add a y-axis title, go to the **Y axis** tab in the **twoway** dialog box to enter your title as shown below.

You can click the **Submit** button to check how the scatter plot looks like. Next go the **Legend** tab and select the **Hide legend** radio button.

Click the **OK** or **Submit** button when you are finished to produce **Figure 8.3**.

[Command: `twoway (scatter FVC Height) (lfit FVC Height), ytitle(Forced vital capacity (L)) legend(off)`]

To save your graph, go to **File > Save** in the **Graph** window, and be sure to save your file as a PNG file:

8.7 Calculating a correlation coefficient

To calculate the Pearson's correlation using the dataset `Example_8.1.dta` go to: **Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Pairwise correlations**

Select the two variables, **FVC** and **Height** in the **Variables** box. You can click the **Submit** button to check the output. Next, tick the box for **Print significance level for each entry** to obtain the P-value and the box for **Print number of observations for each entry** to obtain the number of observations used as shown below.

Click the **OK** or the **Submit** button when you are done to produce **Output 8.1**,

[Command: `pwcorr Height FVC, obs sig`]

8.8 Fitting a simple linear regression model

We will fit a simple linear regression model with `Example_8.1.dta` to quantify the relationship between FVC and height.

Choose **Statistics > Linear models and related > Linear regression**

In the `regress` dialog box, select `FVC` as the **Dependent variable**, and `Height` as the **Independent variable**.

Click the **OK** or the **Submit** button when you are done to produce **Outputs 8.2 and 8.3**.

[Command: `reg FVC Height`]

8.9 Plotting residuals from a simple linear regression

To obtain the residuals, go to **Statistics > Post estimation after running the regress command.**

[Graphical user interface, application Description automatically generated][136]

In the Postestimation Selector dialog box, select Predictions and their SEs, leverage statistics, distance statistics, etc. in the list under Predictions as shown below.

[Graphical user interface, text, application Description automatically generated][137]

In the predict dialog box, choose the Residuals button and enter a New variable name (e.g. FVC_resid) for the residuals from the regression model.

[Graphical user interface, application Description automatically generated][138]

Click the **OK** button when you are done.

$$Command : predictFVC_resid, residuals$$

You can now check the assumption that the residuals are normally distributed by creating a histogram with the normal curve using **Graphics > Histogram** as shown in **Stata Notes** section for **Module 2**. Below is the **histogram** dialog box used to produce the graph in **Figure 8.5**.

$$Command : histogramFVC_resid, bin(12)frequencynormal$$

[Graphical user interface, text, application, email Description automatically generated][139]

8 Learning Activities

Activity 8.1

To investigate how body weight (kg) effects blood plasma volume (mL), data were collected from 30 participants and a simple linear regression analysis was conducted. The slope of the regression was 68 (95% confidence interval 52 to 84) and the intercept was −1570 (95% confidence interval −2655 to −492).

[*You do not need Stata for this Activity*]

a) What is the outcome variable and explanatory (exposure) variable?
b) Interpret the regression slope and its 95% CI
c) Write the regression equation
d) If we randomly sampled a person from the population and found that their weight is 80kg, what would be the predicted value of plasma volume for this person?

Activity 8.2

To examine whether age predicts IQ, data were collected on 104 people. Use the data in the Stata file `Activity_8.2.dta` to answer the following questions.

a) What are the outcome variable and the explanatory variable?
b) Create a scatter plot with the two variables. What can you infer from the scatter plot?
c) Using Stata, obtain the correlation coefficient between age and IQ and interpret it.
d) Conduct a simple linear regression using Stata and report the relationship between the two variables including the interpretation of the R2 value. Are the assumptions for linear regression met in this model?
e) What could you infer about the association between age and IQ in the population, based on the results of the regression analysis in this sample?

Activity 8.3

Which of the following correlation coefficients indicates the weakest linear relationship and why?

a) r = 0.72 [SHOULD I INCLUDE P-VALUES AS WELL?]
b) r = 0.41
c) r = 0.13
d) r = −0.33
e) r = −0.84

Activity 8.4

Are the following statements true or false?

a) If a correlation coefficient is closer to 1.00 than to 0.00, this indicates that the outcome is caused by the exposure.
b) If a researcher has data on two variables, there will be a higher correlation if the two means are close together and a lower correlation if the two means are far apart.

Module 9

Analysing non-normal data

Learning objectives

By the end of this module you will be able to:

- Transform non-normally distributed variables;
- Explain the purpose of non-parametric statistics and key principles for their use;
- Calculate ranks for variables;
- Conduct and interpret a non-parametric independent samples significance test;
- Conduct and interpret a non-parametric paired samples significance test;
- Calculate and interpret the Spearman rank correlation coefficient.

Readings

[Kirkwood and Sterne, 2001] Chapter 13.

[Bland, 2015] Chapter 12.

[Juul and Frydenberg, 2014] Section 11.5.

[Acock, 2010] Section 7.11.

9.1 Introduction

In general, parametric statistics are preferred for reporting data because the summary statistics (mean, standard deviation, standard error of the mean etc) and the tests used (t-tests, correlation, regression etc) are familiar and the results are easy to communicate. However, non-parametric tests can be used if data are not normally distributed. Non-parametric tests make fewer assumptions about the distribution of the data.

9.2 Transforming non-normally distributed variables

When a variable has a skewed distribution, one possibility is to transform the data to a new variable to try and obtain a normal or near normal distribution. Methods to transform non-normally distributed data include logarithmic transformation of each data point, or using the square root or the square or the inverse (i.e. 1/x) etc.

9.2.1 Worked Example

We have data from 132 patients who had a hospital stay following admission to ICU available on Moodle (Example_9.1.dta). The distribution of the length of stay for these patients is shown in the histogram in Figure 9.1. As is common with variables that record time, the data are skewed with many patients having relatively short stays and a few patients having very long hospital stays. Clearly, it would not be possible to use parametric statistical methods for these data.

INSERT FIGURE Figure 9.1 Length of hospital stay in 132 patients

When data are positively skewed, as shown in Figure 9.1, a logarithmic transformation can often make the data closer to being normally distributed. This is the most common transformation used. You should note, however, that the logarithmic function cannot handle 0 or negative values. One way to deal with zeros in a set of data is to add 1 to each value before taking the logarithm. In Stata, we can use the generate command to obtain a new variable, as shown in the Stata Notes section. As the minimum length of stay in these sample data was 0, we have added 1 to each length of stay before taking the logarithm. The distribution of the logarithm of (length of stay + 1) is shown in Figure 9.2.

INSERT FIGURE Figure 9.2 Distribution of log transformed (length of stay + 1)

The distribution now appears much more bell shaped. Output 9.1 shows the descriptive statistics for length of stay before and after logarithmic transformation. Before transformation, the SD is almost as large as the mean value which indicates that the data are skewed and that these statistics are not an accurate description of the centre and spread of the data.

Output

```
                       Length of stay
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%           13              1
10%           15              9       Obs                 132
25%         20.5             11       Sum of Wgt.         132

50%           27                      Mean           38.05303
                        Largest       Std. Dev.      35.78057
75%           42            138
90%           60            153       Variance       1280.249
95%          117            211       Skewness       3.175108
99%          211            244       Kurtosis       15.15463

                   log(Length of stay + 1)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .6931472              0
 5%     2.639057       .6931472
10%     2.772589       2.302585       Obs                 132
25%     3.067783       2.484907       Sum of Wgt.         132

50%     3.332205                      Mean           3.407232
                        Largest       Std. Dev.      .7149892
75%       3.7612       4.934474
90%     4.110874       5.036952       Variance       .5112096
95%     4.770685       5.356586       Skewness      -.4932881
99%     5.356586       5.501258       Kurtosis       7.847303
```

The mean and standard deviation of the transformed length of stay are in log base e (i.e. ln) units. If we raise the mean of the log of length of stay to the power of e, it returns a value of 30.2 days ($e^{3.41} = 30.2$). To do this in Stata, you can use the display command that was shown in Module 2 and with the exponential function, exp().

Technically, this is called the geometric mean of the data, and it has a different interpretation to the usual mean, the arithmetic mean. This is a much better estimate in this case of the “average” length of stay than the mean of 38.1 days (95% CI 31.9, 44.2 days) obtained from the non-transformed positively skewed data. Note that, if you have added 1 to your data to deal with 0 values, the back-transformed estimate is *approximately* equal to the geometric mean.

If we were testing the hypothesis that there was a difference in length of stay between groups (status of nosocomial infection), t-tests could not be used with length of stay but could be used for the log transformed variable, which is approximately normally distributed. The output from the t-test of the log-transformed length of stay is shown in Output 9.2. This is done using the t-test shown in Module 5.

[Command: `ttest ln_los, by(infect)`] How do you interpret the test statistics (i.e. the t-value and p-value)?

Output 9.2: Independent samples t-test on log-transformed length of stay data

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |     106    3.328976     .068083    .7009579     3.19398    3.463972
     Yes |      26    3.726274    .1363363    .6951816    3.445484    4.007064
---------+--------------------------------------------------------------------
combined |     132    3.407232    .0622318    .7149892    3.284122    3.530341
---------+--------------------------------------------------------------------
    diff |           -.3972974    .1531626               -.7003113   -.0942835
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -2.5940
Ho: diff = 0                                     degrees of freedom =      130

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0053         Pr(|T| > |t|) = 0.0106          Pr(T > t) = 0.9947
```

As explained above, the estimated statistics would need to be converted back to the units in which the variable was measured. From Output 9.2, we can take the exponential of the corresponding log-transformed values:

- the geometric mean of the infected group is approximately 41.5 days with a 95% confidence interval from 31.4 to 55.0 days.
- the geometric mean of the uninfected group is approximately 27.9 days with a 95% confidence interval from 24.4 to 31.9 days.

9.3 Non-parametric significance tests

It is often not possible or sensible to transform a non-normal distribution, for example if there are too many zero values or when we simply want to compare groups using the unit in which the measurement was taken (e.g. length of stay). For this, non-parametric significance tests can be used but the general idea behind these tests is that the data values are replaced by ranks. This also protects against outliers having too much influence.

9.3.1 Ranking variables

Table 9.1 shows how ranks are calculated for the first 21 patients in the length-of-stay data (Example_9.1.dta). First the data are sorted in order of their magnitude (from the lowest value to the highest) ignoring the group variable. Each data point is then assigned a rank. Data points that are equal are assigned the mean of their ranks. Thus, the two lengths of stay of 11 days share the ranks 4 and 5, and have a mean rank of 4.5. Similarly, there are 5 people with a length of stay of 14 days and these share the ranks 9 to 13, the mean of which is 11. Once ranks are computed they are assigned to each of the two groups and summed within each group.

Table 9.1: Transforming data to ranks: first 21 participants

ID	Infection	Length of stay	Rank	Infection=no	Infection=Yes
32	No	0	1	1	
33	No	1	2	2	
12	No	9	3	3	
22	No	11	4.5	4.5	
16	No	11	4.5	4.5	
28	Yes	12	6		6
27	No	13	7.5	7.5	
20	No	13	7.5	7.5	
24	No	14	11	11	
11	No	14	11	11	
130	No	14	11	11	
10	No	14	11	11	
25	No	14	11	11	
19	No	15	15.5	15.5	
30	No	15	15.5	15.5	
23	No	15	15.5	15.5	
14	No	15	15.5	15.5	
15	No	17	20.5	20.5	
13	No	17	20.5	20.5	
21	Yes	17	20.5		20.5
17	No	17	20.5	20.5	

By assigning ranks to individuals, we lose information about their actual values and this makes it more difficult to detect a difference. However, outliers and extreme values in the data are brought back closer to the data so that they are less influential. For this reason, non-parametric tests have less power than parametric tests and they require much larger differences in the data to show statistical significance between groups.

9.4 Non-parametric test for two independent samples (Wilcoxon ranked sum test)

The non-parametric equivalent to an independent samples t-test (Module 5) is the Wilcoxon ranked sum test, also known as the Mann-Whitney U test. In Stata, this can be obtained using the ranksum command.

The assumption for this test is that the distributions of the two populations have the same general shape. If this assumption is met, then this test evaluates the null hypothesis that the medians of the two populations are equal. This test does not assume that the populations are normally distributed, nor that their variances are equal.

For the length of stay data in the Worked Example 9.1, we first get a ranks table as shown in Output 9.3. The rank sum table gives us a direction of effect that the ranks are higher than expected in patients who had nosocomial infection. While the positive infection group has a lower sum of ranks because there were fewer people who contracted an infection, it is higher than expected, i.e. they have a longer length of stay compared with the negative infection group. This ranks table does not provide any summary statistics of direction of effect, central tendency or spread that describe the data.

Output 9.3 Results output from Wilcoxon rank-sum test

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      infect |      obs    rank sum    expected
-------------+---------------------------------
          No |      106        6620        7049
         Yes |       26        2158        1729
-------------+---------------------------------
    combined |      132        8778        8778

unadjusted variance    30545.67
adjustment for ties      -53.87
                     ----------
adjusted variance      30491.80

Ho: los(infect==No) = los(infect==Yes)
             z =  -2.457
    Prob > |z| =   0.0140
    Exact Prob =   0.0135
```

The test statistics are shown under the rank sum table in Output 9.3. The variance shown immediately under the table are used to conduct the test, and are not reported on.

From the output 9.3, there are two P-values shown: one assuming normality of the ranks (not the underlying data), and an “Exact” P-value. The exact P-value is calculated when the sample size is not too large (less than 200), and is preferred. The rounded exact P value is 0.014 which indicates that the there is evidence of a difference in length of stay between the groups. This P-value should be provided alongside non-parametric summary statistics such as medians and inter-quartile ranges.

Using the summary command with the detail option in Stata and splitting the LOS variable by the Infect variable (as shown in Module 5), we can obtain the median length of stay values of 24 (Interquartile Range: 19 to 40 days) in the group with no infection and 37 (Interquartile Range: 24 to 50 days) in the group with infection.

9.5 Non-parametric test for paired data (Wilcoxon signed-rank test)

There are two types of non-parametric tests for paired data, called the Sign test and the Wilcoxon signed rank test. In practice, the Sign test is rarely used and will not be discussed in this course.

If the differences between two paired measurements are not normally distributed, a non-parametric equivalent of a paired t-test (Module 5) should be used. The equivalent test is the Wilcoxon matched-pairs signed rank test, also simply called the Wilcoxon matched-pairs test. This test is resistant to outliers in the data, however the proportion of outliers in the sample should be small. This test evaluates the null hypothesis that the median of the paired differences is equal to zero.

In this test, the absolute differences between the paired scores are ranked and the difference scores that are equal to zero (i.e. scores where there is no difference between the pairs) are excluded. Thus, the test is not suitable when a large proportion of the differences are zero because the effective sample size is reduced considerably.

9.5.1 Worked Example

A crossover trial is done to compare symptom scores for two drugs in 11 people with arthritis (higher scores indicate more severe symptoms). The data are contained in Stata datafile file Example_9.2.dta. The data are shown in Table 9.2. The descriptive statistics indicate that the differences are not normally distributed. You can use the Explore function in Stata to determine this.

Table 9.2: Arthritis symptom scores for 11 patients after administering two drugs

Patient ID	Score: Drug 1	Score: Drug 2	Difference (Drug 2 – Drug 1)
1	3	4	1
2	2	7	5
3	3	4	1
4	8	10	2
5	6	8	2
6	6	1	-5
7	2	6	4
8	3	7	4
9	5	8	3
10	9	10	1
11	7	8	1

Before doing the analysis let us examine the distribution of the difference of symptom scores between the two drugs. As in Module 5, we first need to compute the difference between the symptom scores. To examine the distribution, we plot a histogram as shown in Figure 9.3.

Figure 9.3: Distribution of difference in symptom scores between Drug 1 and Drug 2 **UPDATE**

The histogram that the differences are not normally distributed. The data looks negatively skewed with a gap in the histogram between the values of -5 and 0. Therefore, it would not be appropriate to conduct a paired t-test. Hence, we conduct a non-parametric paired test (Wilcoxon matched-pairs signed-rank test).

A non-parametric paired test can be obtained in Stata using the signrank command and the results of the test are shown in Output 9.4.

Output 9.4: Results from Wilcoxon matched-pairs signed-rank test

```
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        1        10.5          33
    negative |       10        55.5          33
        zero |        0           0           0
-------------+---------------------------------
         all |       11          66          66

unadjusted variance      126.50
adjustment for ties       -1.63
adjustment for zeros       0.00
                     ----------
adjusted variance        124.88

Ho: drug_1 = drug_2
             z =  -2.013
    Prob > |z| =   0.0441
    Exact Prob =   0.0459
```

The table in Output 9.4 shows that there is 1 person who has a positive difference, where the symptom score on drug 2 that is smaller than that for drug 1 (i.e., drug 2 is better than drug 1); and 10 people who have a negative difference. No one has the same score for both drugs. The difference scores are ranked and the observed and expected sum of the ranks are shown in the output. This provides no intuitive summary statistics except to indicate which drug has higher ranks.

The test statistics are also shown under the table in Output 9.4. From the output, the exact P value of 0.046 indicates that there is evidence of a difference in symptom score between the two drugs.

9.6 Non-parametric estimates of correlation

Estimating correlation using Pearson's correlation coefficient can be problematic when bivariate Normality cannot be assumed, or in the presence of outliers or skewness. There are two commonly used non-parametric alternatives to Pearson's correlation coefficient: Spearman's rank correlation (ρ or rho), and Kendall's rank correlation (τ or tau).

When estimating the correlation between x and y, Spearman's rank correlation essentially replaces the observations x and y by their ranks, and calculates the correlation between the ranks. Kendall's rank correlation compares the ranks between every possible combination of pairs of data to measure concordance: whether high values for x tend to be associated with high values for y (positively correlated) or low values of y (negatively correlated).

In terms of which is the more appropriate measure to use, the following passage from An Introduction to Medical Statistics, 4th Edition (Bland 2015) provides some guidance:

> "Why have two different rank correlation coefficients? Spearman's ρ is older than Kendall's τ, and can be thought of as a simple analogue of the product moment correlation coefficient, Pearson's r. Kendall's τ is a part of a more general and consistent system of ranking methods, and has a direct interpretation, as the difference between the proportions of concordant and discordant pairs. In general, the numerical value of ρ is greater than that of τ. It is not possible to calculate τ from ρ or ρ from τ, they measure different sorts of correlation. ρ gives more weight to reversals of order when data are far apart in rank than when

> there is a reversal close together in rank, τ does not. However, in terms of tests of significance, both have the same power to reject a false null hypothesis, so for this purpose it does not matter which is used.”

We will illustrate estimating rank correlation using the Stata file Example_8.1.dta, which has information about height and lung function collected from a sample of 120 adults.

Output 9.5: Results from rank correlation analysis

```
. spearman Height FVC

 Number of obs =     120
Spearman's rho =       0.7476

Test of Ho: Height and FVC are independent
    Prob > |t| =       0.0000

. ktau Height FVC

  Number of obs =     120
Kendall's tau-a =       0.5431
Kendall's tau-b =       0.5609
Kendall's score =    3878
    SE of score =     439.463   (corrected for ties)

Test of Ho: Height and FVC are independent
     Prob > |z| =       0.0000  (continuity corrected)
```

The Spearman rank correlation coefficient is estimated as 0.75, demonstrating a positive association between height and FVC. Stata provides two versions of the Kendall rank correlation coefficient: we would use tau-b (τ_b) as it allows for tied observations. The Kendall rank correlation coefficient is estimated as 0.56, again demonstrating a positive association between height and FVC.

9.7 Summary

In this module, we have presented methods to conduct a hypothesis test with data that are not normally distributed. Non-parametric methods do not assume any distribution for the data and use significance tests based on ranks or sign (or both). A non-parametric test is always less powerful than its equivalent parametric test if the data are normally distributed and so whenever possible parametric significance tests should be used. In some cases when data are not normally distributed with a reasonably large sample size, the data can be transformed (most commonly by log transformation) to make the distribution normal. A parametric significance test should then be used with the transformed data to test the hypothesis.

9 Learning Activities

Activity 9.1

There is a hypothesis that university students who live and dine in the university hall consume less vitamin C than the students who live and dine at home. To test the hypothesis, 30 students were randomly selected and their urinary ascorbic acid level was measured in mg over 3 hours. Urinary excretion of ascorbic acid is a measure of vitamin C nutrition in humans. The data is given in the following table and a copy of the data set, `Activity9.1.dta` is also available on Moodle.

Table 9.3: Urinary level of ascorbic acid (mg per 3 hours) of university students

Living and dining in Hall (n.=.17)	**Living and dining at Home (n = 13)**
34	163
62	205
37	83
27	372
38	50
20	22
7	47
53	255
22	30
37	89
14	96
28	48
28	25
70	163
16	
9	
121	

a) Examine the distribution of the data using a box-plot and histogram, and obtain descriptive statistics. How would you describe the distribution of ascorbic acid?
b) Which statistical test would be appropriate to test the hypothesis mentioned in the question and why?
c) State the hypotheses appropriate to the analytical method you mentioned in (b).
d) Use Stata to carry out the statistical test you have mentioned in (b) and write your conclusion.

Activity 9.2

A drug was tested for its effect in lowering blood pressure. Fifteen women with hypertension were enrolled and had their systolic blood pressure measured before and after taking the drug. The data are available in the Stata file `Activity_9.2.dta` on Moodle.

a) State the research question and the null hypothesis.
b) Use Stata to obtain suitable summary statistics and test the null hypothesis. Describe the reason for choosing the test.
c) Write a brief conclusion.
d) What are the main limitations of this study? Consider both epidemiological and statistical aspects.

Module 10

Sample size estimation

Learning objectives

By the end of this module you will be able to:

- Explain the issues involved in sample size estimation for epidemiological studies;
- Estimate sample sizes for descriptive and analytical studies;
- Compute the sample size needed for planned statistical tests;
- Adjust sample size calculations for factors that influence study power.

Readings

[**?**]; Chapter 35.

[**?**]; Chapter 18.

Further readings (for interest)

Woodward [2013]; Chapter 8.

10.1 Introduction

Determining the appropriate sample size (the number of participants in a study) is one of the most critical issues when designing a research study. A common question when planning a project is "How many participants do I need?" The sample size needs to be large enough to ensure that the results can be generalised to the population and will be accurate, but small enough for the study question to be answered with the resources available. In general, the larger the sample size, the more precise the study results will be.

Unfortunately, estimating the sample size required for a study is not straightforward and the method used varies with the study design and the type of statistical test that will be conducted on the data collected. In the past, researchers calculated the sample size by hand using complicated mathematical formula. More recently, look-up tables have been created which has removed the need for hand calculations. Now, most researchers use computer programs where parameters relevant to the particular study design are entered and the sample size is automatically calculated. In this module, we will use an abbreviated look-up table to demonstrate the parameters that need to be considered when estimating sample sizes for a confidence interval and use Stata for all other sample size calculations. The look-up table allows you to see at a glance, the impact of different factors on the sample size estimation.

10.1.1 Under and over-sized studies

In health research, there are different implications for interpreting the results if the sample size is too small or too large.

An under-sized study is one which lacks the power to find an effect or association when, in truth, one exists. If the sample size is too small, an important difference between groups may not be statistically significant and so will not be detected by the study. In fact, it is considered unethical to conduct a health study which is poorly designed so that it is not possible to detect an effect or association if it exists. Often, Ethics Committees request evidence of sample size calculations before a study is approved.

A classic paper by Freiman et al examined 71 randomised controlled trials which reported an absence of clinical effect between two treatments.[Freiman et al., 1978] Many of the trials were too small to show that a clinically important difference was statistically significant. If the sample size of an analytic study is too small, then only very limited conclusions can be drawn about the results.

In general, the larger the sample size the more precise the estimates will be. However, large sample sizes have their own effect on the interpretation of the results. An over-sized study is one in which a small difference between groups, which is not important in clinical or public health terms, is statistically significant. When the study sample is large, the null hypothesis could be rejected in error and research resources may be wasted. This type of study may be unethical due to the unnecessary enrolment of a large number of people.

10.2 Sample size estimation for descriptive studies

To estimate the sample size required for a descriptive study, we usually focus on specifying the width of the confidence interval around our primary estimate. For example, to estimate the sample size for a study designed to measure a prevalence we need to:

- nominate the expected prevalence based on other available evidence;
- nominate the required level of precision around the estimate. For this, the width of the 95% confidence interval (i.e. the distance equal to 1.96 $\times$ SE) is used.

Table 10.1 is an abbreviated look-up table that we can use to estimate the sample size for this type of study. Note that the sample size required to detect an expected population prevalence of 5% is the same as to detect a prevalence of 95%. Similarly 10% is equivalent to 90% etc. It is symmetric about 50%. From Table 10.1, you can see that the sample size required increases as the expected prevalence approaches 50% and as the precision increases (i.e. the required 95% CI becomes narrower).

10.2.1 Worked Example

A descriptive cross-sectional study is designed to measure the prevalence of bronchitis in children age 0-2 years with a 95% CI of $\pm$ 4%. The prevalence is expected to be 20%. From the table, a sample size of at least 385 will be required for the width of the 95% CI to be $\pm$ 4% (i.e. the reported precision of the summary statistic will be 20% (95% CI 16% to 24%)).

If the prevalence turns out to be higher than the researchers expected or if they decided that they wanted a narrower 95% CI (i.e. increase precision), a larger sample size would be required.

- What sample size would be required if the prevalence was 15% and the desired 95% CI was $\pm$ 3%?
- Answer: 545

Table 10.1: Sample size required to calculate a 95% confidence interval with a given precision

	Width of 95% confidence interval (precision)									
Prevalence	**1%**	**1.5%**	**2%**	**2.5%**	**3%**	**3.5%**	**4%**	**5%**	**10%**	**15%**
5% or 95%	1,825	812	457	292	203	149	115			
10% or 90%	3,458	1,537	865	554	385	283	217	139		
15% or 85%	4,899	2,177	1,225	784	545	400	307	196	49	
20% or 80%	6,147	2,732	1,537	984	683	502	385	246	62	28
25% or 75%	7,203	3,202	1,801	1,153	801	588	451	289	73	33

10.3 Sample size estimation for analytical studies

Analytical study designs are used to compare characteristics between different groups in the population. The main study designs are analytical cross-sectional studies, case-control studies, cohort studies and randomised controlled trials. For analytical study designs, the outcome measure of interest can be a mean value, a proportion or a relative risk if a random sample has been enrolled. For case-control studies the most appropriate measure of association is an odds ratio.

10.3.1 Factors to be considered

The first important decision in estimating a required sample size for an analytic study is to select the type of statistical test that will be used to report or analyse the data. Each type of test is associated with a different method of sample size estimation.

Once the statistical method has been determined, the following issues need to be decided: - Statistical power – the chance of finding a difference if one exists, e.g. 80%; - Level of significance – the P value that will be considered significant, e.g. P<0.05; - Minimum effect size of interest – the size of the difference between groups e.g. the difference in the proportion of parents who oppose immunisation in two different regions or the difference in mean values of blood pressure in two groups of people with different types of cardiac disease; - Variability – the spread of the measurements, e.g. the expected standard deviation of the main outcome variable (if continuous), or the expected proportions; - Resources – an estimate of the number of participants available and amount of funding to run the study.

In addition to deciding the level of power and probability that will be used, the difference between groups that is regarded as being important has to be estimated. The smallest difference between study groups that we want to detect is described as the minimum expected effect size. This is determined on the basis of clinical judgement, public health importance and expertise in the condition being researched, or may it be need to be determined from a pilot study or a literature review. The smaller the expected effect or association, the larger the sample size will need to obtain statistical significance. We also need some knowledge of how variable the measurement is expected to be. For this we often use the standard deviation for a continuous measure. As measurement variability increases, the sample size will need to increase in order to detect the expected difference between the groups. Above all, a study has to be practical in terms of the availability of a population from which to draw sufficient numbers for the study and in terms of the funds that are available to conduct the study.

10.3.2 Power and significance level

The power of a study, which was discussed in Module 4, is the chance of finding a statistically significant difference when one exists, i.e. the probability of correctly rejecting the null hypothesis. The relationship between the power of a study and statistical significance is shown in Table 10.2.

Table 10.2 Comparison of study results with the truth

Study result Truth Effect No effect Effect □ α (Type I error) No effect β (Type II error) □

The power of a study is expressed as 1– β where β is the probability of a false negative (that is, the probability of a Type II error - incorrectly not rejecting the null hypothesis. In most research, power is generally set to 80% (a Type II error rate of 20%). However, in some studies, especially in some clinical trials where rigorous results are required, power is set to 90% (a Type II error rate of 10%).

The significance level, or α level, is the level at which the P value of a test is considered to be statistically significant. The α level is usually set at 5% indicating a probability of <0.05 will be regarded as statistically significant. Occasionally, especially if several outcome measures are being compared, the α level is set at 1% indicating a probability of <0.01 will be regarded as statistically significant.

10.4 Detecting the difference between two means

The test that is used to show that two mean values are significantly different from one another is the independent samples t-test (Module 5). The sample size needed for this test to have sufficient power can be calculated using Stata as shown in Worked Example 10.2.

10.4.1 Worked Example

There is a hypothesis that the use of the oral contraceptive (OC) pill in premenopausal women can increase systolic blood pressure. A study was planned to test this hypothesis using a two sided t-test. The investigators are interesting in detecting an increase of at least 5 mm Hg systolic blood pressure in the women using OC compared to the non-OC users with 90% power at a 5% significance level. A pilot study shows that the SD of systolic blood pressure in the target group is 25 mm Hg and the mean systolic blood pressure of non-OC user women is 90 mm Hg. What is the minimum number of women in each group that need to be recruited for the study to detect this difference?

Solution The effect size of interest is 5 mm Hg and the associated standard deviation is 25 mm Hg. For power of 90% and alpha of 5%, the sample size calculation using the `power twomeans` command in Stata is shown in Output 10.1.

Output 10.1: Two independent samples t-test sample size calculation

```
. power twomeans 90 95, sd(25) power(0.9)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    5.0000
```

```
           m1 =   90.0000
           m2 =   95.0000
           sd =   25.0000

Estimated sample sizes:

            N =     1,054
  N per group =       527
```

From the output, we can see that with 90% power we will need 527 participants in each group, i.e., 1054 participants in total. If the above were carried out by taking baseline measures of systolic blood pressure, and then again when the women were taking the OC pills, it would be a matched-pair study. We can compute the required sample size using the power pairedmeans command.

Output 10.2: Paired samples t-test sample size using Worked Example 10.2

```
. power pairedmeans 90 95, corr(0) power(0.9) sd(25)

Performing iteration ...

Estimated sample size for a two-sample paired-means test
Paired t test assuming sd1 = sd2 = sd
Ho: d = d0  versus  Ha: d != d0

Study parameters:

        alpha =    0.0500          ma1 =   90.0000
        power =    0.9000          ma2 =   95.0000
        delta =    0.1414           sd =   25.0000
           d0 =    0.0000         corr =    0.0000
           da =    5.0000
         sd_d =   35.3553

Estimated sample size:

            N =       528
```

Assuming a correlation of 0 between the two sets of measurements, we can see that we will need 528 pairs of measurements to achieve a power of 90% (virtually the same as for an independent samples study).

If we do not know the correlation between the two sets of observations, we can enter 0 for the correlation. If the correlation is positive, a zero for correlation would give a more conservative estimate of sample size required (i.e. estimate a sample size larger than necessary). While a negative correlation would require a bigger sample size than a zero correlation, it is relatively uncommon to encounter negative correlations between pairs. Any discussions on the effect of correlation on sample size is beyond the scope of this course. Thus, we will always assume a correlation of zero between paired measurements in this course.

10.5 Detecting the difference between two proportions

The statistical test for deciding if there is a significant difference between two independent proportions is a Pearson's chi-squared test (Module 7). The sample size required in each group to observe a difference in two independent proportions can be calculated using the power twoproportions command in Stata.

Other than the power and alpha required for the test, the expected prevalence or incidence rate of the outcome factor needs to be estimated for each of the two groups being compared, based on what is known from other studies or what is expected. Occasionally, we may not know the expected proportion in one of the groups, e.g. in a randomised control trial of a novel intervention. In the sample size calculation for such a study, we should instead justify the minimum expected difference between the proportions based on what is important from a clinical or public health perspective. Based on the minimum difference, we can then derive the expected proportion for both groups. Note that the smaller the difference, the larger the sample size required.

10.5.1 Worked Example

If we expect that the prevalence of smoking in two comparison groups (e.g. males and females) will be 35% and 20%. The sample size required in each group to show that the prevalences are significantly different at P<0.05 with 80% power is shown in Output 10.3.

Output 10.3: Sample size calculation for two independent proportions

```
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   -0.1500  (difference)
           p1 =    0.3500
           p2 =    0.2000

Estimated sample sizes:

            N =       276
  N per group =       138
```

From Output 10.3, we see that we would need 138 males and 138 females (i.e. a total sample size of 276 participants). What sample size would be required if the prevalence of smoking among men was 30%? Answer = 294 men and 294 women would be needed. [Command: power twoproportions .3 .2, test(chi2)]

10.6 Detecting an association using a relative risk

The relative risk is used to describe the association between an exposure and an outcome variable if the sample has been randomly selected from the population. This statistic is often used to describe the effect or association of an exposure in a cross-sectional or cohort study or the effect/association of a treatment in an randomised controlled trial. To estimate the sample size required for the RR to have a statistically significant P value, i.e. to show a significant association, we need to define: - the size of the RR that is considered to be of clinical or public health importance; - the event rate (rate of outcome) among the group who are not exposed to the factor of interest (reference group); - the desired level of significance (usually 0.05); - the desired power of the study (usually 80% or 90%).

In general, a RR of 2.0 or greater is considered to be of public health importance. However, a smaller RR can be important when exposure is high, for example say the relative risk of respiratory infection among young children with a parent who smokes is very small at approximately 1.2 but 25% of children are exposed to smoking in their home. The high exposure rate leads to a very large number of children who have preventable respiratory infections across the community.

10.6.1 Worked Example

A study is planned to investigate the effect of an environmental exposure on the incidence of a certain common disease. In the general (unexposed) population the incidence rate of the disease is 50% and it is assumed that the incidence rate would be 75% in the exposed population. Thus the relative risk of interest would be 1.5 (i.e. 0.75 / 0.50). We want to detect this effect with 90% power at a 5% level of significance. Using the power twoproportions command, Output 10.4 is obtained.

Output 10.4: Sample size calculation for relative risk

```
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    0.2500  (difference)
           p1 =    0.5000
           p2 =    0.7500
        rrisk =    1.5000

Estimated sample sizes:

            N =       154
  N per group =        77
```

From Output 10.4, we can see that for a control proportion of 0.5 and RR of 1.5, we need a total sample size of 154, that is 77 people would be needed in each of the exposure groups.

10.7 Detecting an association using an odds ratio

If we are designing a case-control study, the appropriate measure of effect is an odds ratio. The method for estimating the sample size required to detect an odds ratio of interest is slightly different to that for the relative risk. However, the same parameters are required for the estimation: - the minimum OR to be considered clinically important; - the proportion of exposed among the control group; - the desired level of significance (usually 0.05); - the desired power of the study (usually 80% or 90%).

10.7.1 Worked Example

A case-control study is designed to examine an association between an exposure and outcome factor. Existing literature shows that 30% of the controls are expected to be exposed. We want to detect a minimum OR of 2.0 with 90% power and 5% level of significance.

```
. power twoproportions .3, test(chi2) oratio(2) power(0.9)
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    0.1615  (difference)
```

```
           p1 =    0.3000
           p2 =    0.4615
   odds ratio =    2.0000

Estimated sample sizes:

            N =       376
  N per group =       188
```

We find that 188 controls and 188 cases are required i.e. a total of 376 participants.

This sample size would be smaller if we increased the effect size (OR) or reduced the study power to 80%. You could try this in Stata (answer: 141 per group).

10.8 Factors that influence power

10.8.1 Dropouts

It is common to increase estimated sample sizes to allow for drop-outs or non-response. To account for drop-outs, the estimated sample size can be divided by (1 minus the dropout rate). Consider the following case:

- n-completed: the number who will complete the study (i.e. n after drop-out)
- n-recruited: the number who should be recruited (i.e. n before drop-out)
- d: drop-out rate (as a proportion - i.e. a number between 0 and 1)

Then n-completed = n-recruited × (1 - d)

Re-arranging this formula gives: n-recruited = n-completed ÷ (1 - d).

10.8.2 Unequal groups

Many factors that come into play in a study can reduce the estimated power of a study. In clinical trials, it is not unusual for recruitment goals to be much harder to achieve than expected and therefore for the target sample size to be impossible to realise within the timeframe planned for recruitment.

In case-control studies, the number of potential case participants available may be limited but study power can be maintained by enrolling a greater number of controls than cases. Or in an experimental study, more participants may be randomised to the new treatment group to test its effects accurately when much is known about the effect of standard care and a more precise estimate of the new treatment effect is required.

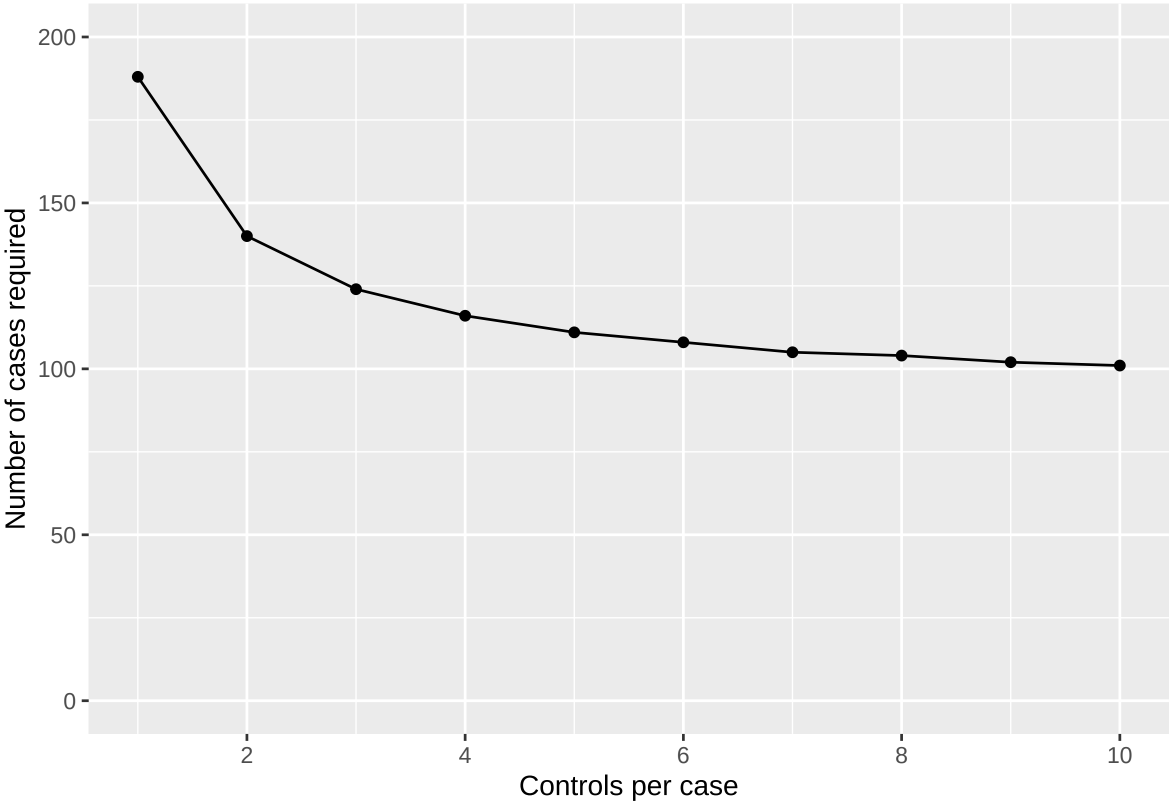
However, there is a trade-off between increasing the ratio of group size and the total number that needs to be enrolled. Consider Worked Example 10.5: selecting an equal number of controls and cases would require 188 cases and 188 controls, a total of 376 participants.

We may want to reduce the number of cases required, by selecting 2 controls for every case. When performing sample size calculations with unequal groups, Stata refers to cases as N2, and controls as N1. Selecting 2 controls (N1) per case (N2) (corresponding to a ratio of N2/N1 0.5 in Stata) would require 140 cases and 280 controls, a total of 420 participants. We can extend this example and investigate the impact of changing the ratio of controls per case.

This can be visualised graphically, as in Figure 10.1.

Table 10.2: Increasing controls per case

Controls per case	**Stata's allocation ratio (N2/N1)**	**Number of cases required**	**Number of controls required**	**Total participants required**
1	1	188	188	376
2	0.5	140	280	420
3	0.3333	124	371	495
4	0.25	116	462	578
5	0.2	111	553	664
6	0.1666	108	644	752
7	0.1429	105	734	839
8	0.125	104	825	929
9	0.1111	102	916	1,018
10	0.1	101	1,006	1,107



We can see that the number of cases required drops off if we go from 1 to 2 controls per case, and again from 2 to 3 controls per case. Once we go from 3 to 4 controls per case, we only reduce the number of cases by 8 (124 vs 116 cases), but at an increase of 91 (371 vs 462) controls. Clearly, this reduction in cases is not offset by the extra controls required.

10.9 Limitations in sample size estimations

In this module we have seen how to use Stata for estimating the sample size requirement of a study given the statistical test that will be used and the expected characteristics of the sample. However, once a study is underway, it is not unusual for sample size to be compromised by the lack of research resources, difficulties in recruiting participants or, in a clinical trial, participants wanting to change groups when information about the new experimental treatment rapidly becomes available in the press or on the internet.

One approach that is increasingly being used is to conduct a blinded interim analysis say when 50% of the total data that are planned have been collected. In this, a statistician external to the research team who is blinded to the interpretation of the group code is asked to measure the effect size in the data with the sole aim of validating the sample size requirement. It is rarely a good idea to use an interim analysis to reduce the planned sample size and terminate a trial early because the larger the sample size, the greater the precision with which the treatment effect is estimated. However, interim analyses are useful for deciding whether the sample size needs to be increased in order to answer the study question and avoid a Type II error.

10.10 Summary

In this module we have discussed the importance of conducting a clinical or epidemiological study with enough participants so that an effect or association can be identified if it exists (i.e. study power), and how this has to be balanced by the need to not enrol more participants than necessary because of resource issues. We have looked at the parameters that need to be considered when estimating the sample size for different studies and have used a look-up table to estimate required sample size for a prevalence study and Stata to estimate appropriate sample sizes in epidemiological research under the most straightforward situations. The common requirement in all the situations is that the researchers need to specify the minimum effect measure (e.g. difference in means, OR, RR etc) they want to detect with a given probability (usually 80% to 90%) at a certain level of significance (usually P<0.05). The ultimate decision on the sample size depends on a compromise among different objectives such as power, minimum effect size, and available resources. To make the final decision, it is helpful to do some trial calculations using revised power and the minimum detectable effect measure.

10 Stata resources

10 Learning Activities

Activity 10.1

We are planning a study to measure the prevalence of a relatively rare condition (say approximately 5%) in children age 0-5 years in a remote community.

a) What type of study would need to be conducted?
b) Use the correct sample size table included in your notes to determine how many children would need to be enrolled for the confidence interval to be

   i. 2%
   ii. 4% around the prevalence?

What would the resulting prevalence estimates and 95% CIs be?

Activity 10.2

We are planning an experimental study to test the use of a new drug to alleviate the symptoms of the common cold compared to the use of Vitamin C. Participants will be randomised to receive the new experimental drug or to receive Vitamin C. How many participants will be required in each group (power = 80%, level of significance = 5%).

a) If the resolution of symptoms is 10% in the control group and 40% in the new treatment group?
b) How large will the sample size need to be if we decide to recruit two control participants to every intervention group participant?
c) If we decide to retain a 1:1 ratio of participants in the intervention and controls groups but the resolution of symptoms is 20% in the control group and 40% in the new treatment group?
d) How many participants would we need to recruit (calculated in c) if a pilot study shows that 15% of people find the new treatment unpalatable and therefore do not take it?

Activity 10.3

In a case-control study, we plan to recruit adult males who have been exposed to fumes from an industrial stack near their home and a sample of population controls in whom we expect that 20% may also have been exposed to similar fumes through their place of residence or their work. We want to show that an odds ratio of 2.5 for having respiratory symptoms associated with exposure to fumes is statistically significant.

a) What statistical test will be needed to measure the association between exposure and outcome?
b) How large will the sample size need to be to show that the OR of 2.5 is statistically significant at $P < 0.05$ with 90% power if we want to recruit equal number of cases and controls?
c) What would be the required sample size (calculated in b) if the minimum detectable OR were 1.5?
d) If there are problems recruiting cases to detect an OR of 1.5 (as calculated in c), what would the sample size need to be if the ratio of cases to controls was increased to 1:3?

Activity 10.4

In the above study to measure the effects of exposure to fumes from an industrial stack, we also want to know if the stack has an effect on lung function which can be measured as forced vital capacity in 1 minute (FEV1). This measurement is normally distributed in the population.

a) If the research question is changed to wanting to show that the mean FEV1 in the exposed group is lower than the mean FEV1 in the control group what statistical test will now be required?
b) Population statistics show that the mean FEV1 and its SD in the general population for males are 4.40 L (SD=1.25) which can be expected in the control group.

We expect that the mean FEV1 in the cases may be 4.0 L. How many participants will be needed to show that this mean value is significantly different from the control group with P < 0.05 with an 80% power if we want to recruit equal number in each group?

c) How much larger will the sample size need to be if the mean FEV1 in the cases is 4.20 L?

Bibliography

Alan C. Acock. *A Gentle Introduction to Stata, Third Edition*. Stata Press, College Station, Tex, 3rd edition edition, August 2010. ISBN 978-1-59718-075-7.

Martin Bland. *An Introduction to Medical Statistics*. Oxford University Press, Oxford, New York, fourth edition edition, July 2015. ISBN 978-0-19-958992-0.

Jennie A. Freiman, Thomas C. Chalmers, Harry Smith, and Roy R. Kuebler. The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. *New England Journal of Medicine*, 299(13):690–694, September 1978. ISSN 0028-4793. doi: 10.1056/NEJM197809282991304.

Svend Juul and Morten Frydenberg. *An Introduction to Stata for Health Researchers, Fourth Edition*. Stata Press, College Station, Texas, 4th edition edition, March 2014. ISBN 978-1-59718-135-8.

Betty Kirkwood and Jonathan Sterne. *Essentials of Medical Statistics*. Wiley-Blackwell, Malden, Mass, 2nd edition edition, April 2001. ISBN 978-0-86542-871-3.

Mark Woodward. *Epidemiology: Study Design and Data Analysis, Third Edition*. Chapman and Hall/CRC, 3rd edition edition, December 2013.