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

Module 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

Module 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 Section

Module 3: Stata notes

Module 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

Module 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:

- blah
- blah
- blah

Readings

[?]

[?]

5.1 Introduction

Module 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:

- blah
- blah
- blah

Readings

[?]

[?]

6.1 Introduction

Module 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 6.1: 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

Module 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:

- blah
- blah
- blah

Readings

[?]

[?]

8.1 Introduction

Module 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
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.

Module 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.

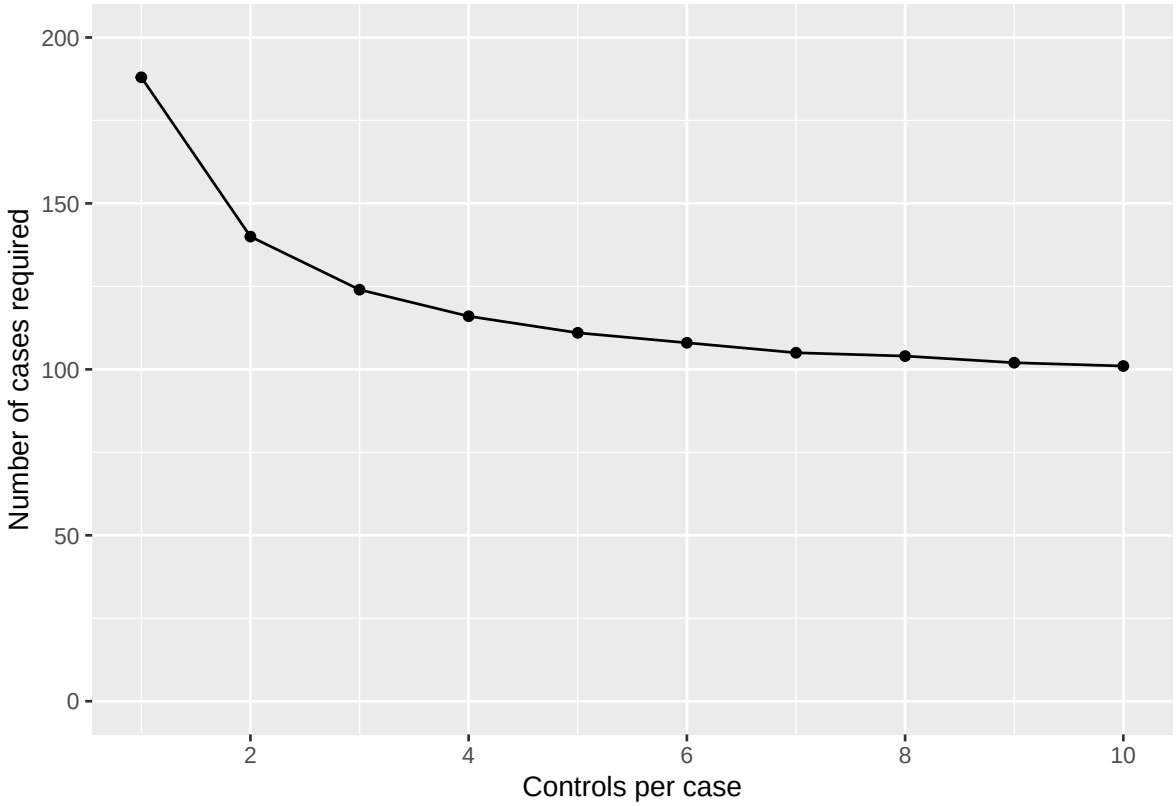
However, there is a trade-off between increasing the ratio of group size and the total number that needs to be enrolled. Consider Worked Example 10.5: selecting an equal number of controls and cases would require 188 cases and 188 controls, a total of 376 participants.

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.

Module 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.