PHCM9795: Foundations of Biostatistics

Timothy Dobbins

Invalid Date

Table of contents

Course introduction

Welcome to PHCM9795 Foundations of Biostatistics.

This introductory course in biostatistics aims to provide students with core biostatistical skills to analyse and present quantitative data from different study types. These are essential skills required in your degree and throughout your career.

We hope you enjoy the course and will value your feedback and comment throughout the course.

Course information

Biostatistics is a foundational discipline needed for the analysis and interpretation of quantitative information and its application to population health policy and practice.

This course is central to becoming a population health practitioner as the concepts and techniques developed in the course are fundamental to your studies and practice in population health. In this course you will develop an understanding of, and skills in, the core concepts of biostatistics that are necessary for analysis and interpretation of population health data and health literature.

In designing this course, we provide a learning sequence that will allow you to obtain the required graduate capabilities identified for your program. This course is taught with an emphasis on formulating a hypothesis and quantifying the evidence in relation to a specific research question. You will have the opportunity to analyse data from different study types commonly seen in population health research.

The course will allow those of you who have covered some of this material in your undergraduate and other professional education to consolidate your knowledge and skills. Students exposed to biostatistics for the first time may find the course challenging at times. Based on student feedback, the key to success in this course is to devote time to it every week. We recommend that you spend an average of 10-15 hours per week on the course, including the time spent reading the course notes and readings, listening to lectures, and working through learning activities and completing your assessments. Please use the resources provided to assist you, including online support.

Units of credit

This course is a core course of the Master of Public Health, Master of Global Health and Master of Infectious Diseases Intelligence programs and associated dual degrees, comprising 6 units of credit towards the total required for completion of the study program. A value of 6 UOC requires a minimum of 150 hours work for the average student across the term.

Course aim

This course aims to provide students with the core biostatistical skills to apply appropriate statistical techniques to analyse and present population health data.

Learning outcomes

On successful completion of this course, you will be able to:

1. Summarise and visualise data using statistical software.
2. Demonstrate an understanding of statistical inference by interpreting p-values and confidence intervals.
3. Apply appropriate statistical tests for different types of variables given a research question, and interpret computer output of these tests appropriately.
4. Determine the appropriate sample size when planning a research study.
5. Present and interpret statistical findings appropriate for a population health audience.

Module 1

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

Kirkwood and Sterne (2001); Chapters 4 and 6. [UNSW Library Link]

Bland (2015); Sections 3.3 and 3.4, 8.1 to 8.3. [UNSW Library Link]

1.1 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.

1.2 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 might occur when random sampling is used, simply by chance. However, non-random sampling methods, such as using a study population that does not represent the whole 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 PHCM9796: Foundations of Epidemiology.

1.3 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.

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 value 1.78 is 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 proportional to the standard deviation of the variable, and inversely proportional 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.

1.4 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:

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.

When the population is assumed to be normal:

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

1.5 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.

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

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.

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.

The t-distribution and when should I use it?

The population standard deviation (σ) is required for calculation of the standard error. Usually, 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 underestimates the true value 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 underestimation. It looks very much like the standardised normal distribution, only that it has fatter tails (Figure 1.1). 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.

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.

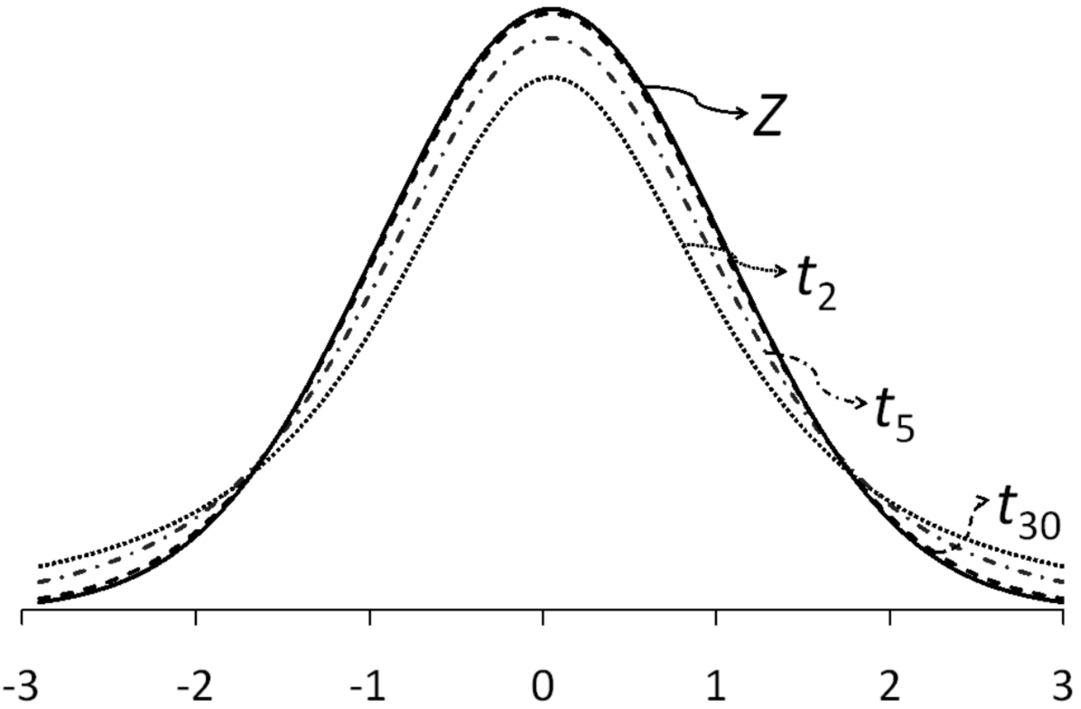


Figure 1.1: The normal (Z) and the student's t-distribution with 2, 5 and 30 degrees of freedom

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) (Section 1.7):

```
. cii means 30 70 6

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

In R we use the `ci_mean` function provided in Section 1.9:

```
ci_mean(n=30, mean=70, sd=6, width=0.95)
```

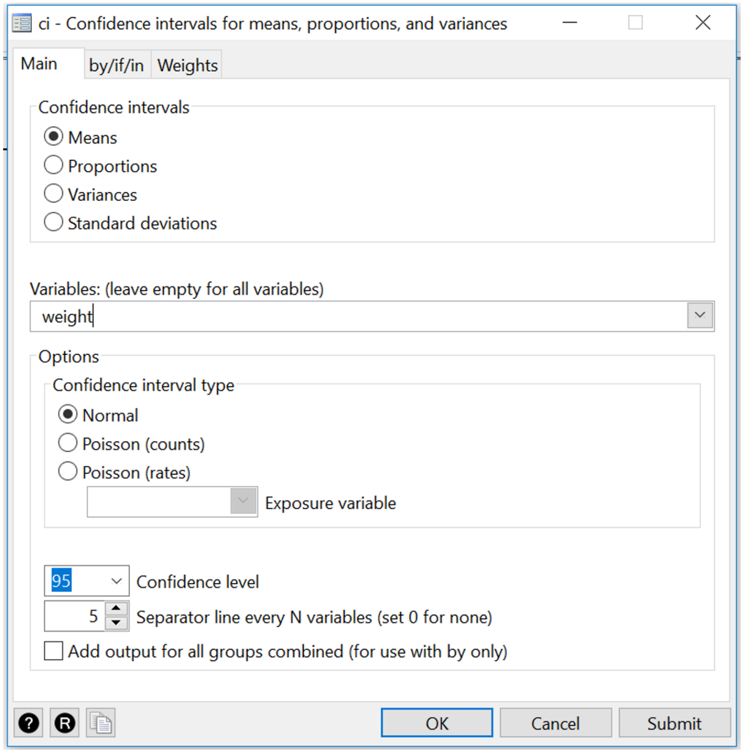
```
[1] "95% CI: 67.760 to 72.240"
```

Both commands use the t-distribution, and the output can be interpreted as: we are 95% confident that the true mean weight of the population from which the sample was drawn lies between 67.8 kg and 72.2 kg.

Stata notes

1.6 Calculating a 95% confidence interval of a mean: Individual data

To demonstrate the computation of the 95% confidence interval of a mean we have used data from `Example_1.3.dta`. To calculate the 95% confidence interval, go to **Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Confidence intervals**. In the **ci** dialog box, select `weight` as the **Variable**.



ci - Confidence intervals for means, proportions, and variances

Confidence intervals

● Means
○ Proportions
○ Variances
○ Standard deviations

Variables: (leave empty for all variables)

weight

Options

Confidence interval type

● Normal
○ Poisson (counts)
○ Poisson (rates)

Exposure variable

95 Confidence level

5 Separator line every N variables (set 0 for none)

□ Add output for all groups combined (for use with by only)

Figure 1.2: Calculating a confidence interval from individual data

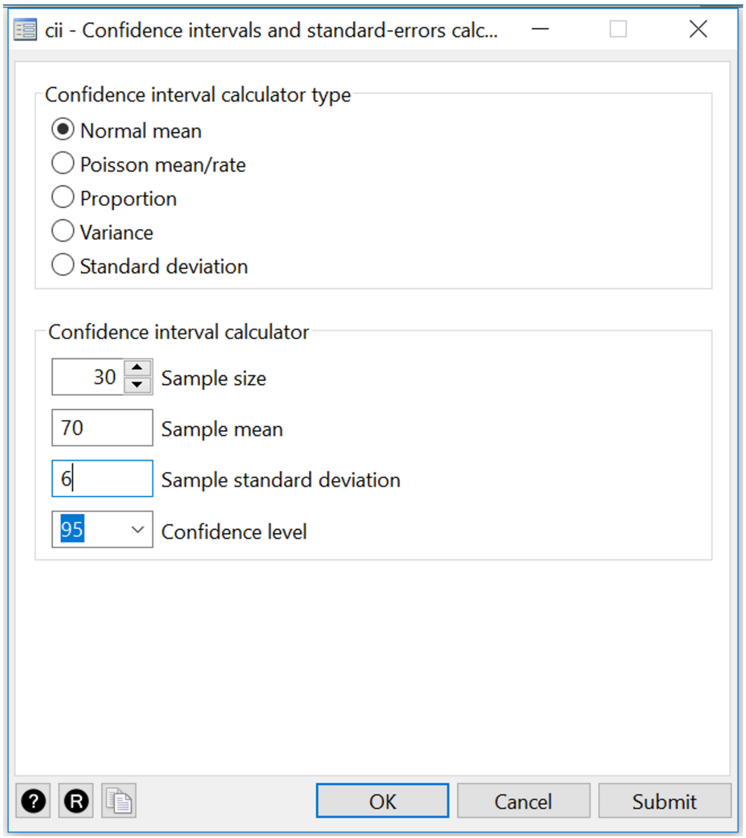
Click **OK** or **Submit** to obtain Output 3.1.

[Command: `ci means weight`]

1.7 Calculating a 95% confidence interval of a mean: Summarised data

For Worked Example 3.2 where we are given the sample mean, sample standard deviation and sample size, we use the `cii means` command. To calculate the 95% CI, go to **Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Normal mean CI calculator**. In the **cii** dialog box, check that the **Normal mean** button is selected, and enter `30` as the **Sample size**, `70` as the **Sample mean**, `6` as the **Sample standard deviation** and check that `95` in entered as the **Confidence level**.

Click **OK** or **Submit** to obtain the following output:



Confidence interval calculator type

● Normal mean
○ Poisson mean/rate
○ Proportion
○ Variance
○ Standard deviation

Confidence interval calculator

30 Sample size
70 Sample mean
6 Sample standard deviation
95 Confidence level

Figure 1.3: Calculating a confidence interval from summarised data

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

```
. cii means 30 70 6

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

R notes

1.8 Calculating a 95% confidence interval of a mean: individual data

To demonstrate the computation of the 95% confidence interval of a mean we have used data from `Example_1.3.rds` which contains the weights of 30 students:

```
students <- readRDS("data/examples/Example_1.3.rds")
```

We can examine the data set using the `summary` command:

```
summary(students)
```

```
     weight          gender   
 Min.   :60.00   Male  :16  
 1st Qu.:67.50   Female:14  
 Median :70.00              
 Mean   :70.00              
 3rd Qu.:74.38              
 Max.   :80.00              
```

The mean and its 95% confidence interval can be obtained many ways in R. We will use the `t.test()` function installed in R to calculate the confidence interval:

```
t.test(students$weight)
```

```

	One Sample t-test

data:  students$weight
t = 76.029, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 68.11694 71.88306
sample estimates:
mean of x 
       70 
```

The output of the `t.test()` function gives us the sample mean (70.0 kg) as well as the 95% confidence interval around the mean: 68.1 to 71.9 kg.

Note: the `descriptives()` function within the `jmv` package also calculates a 95% confidence interval around the mean. **It is recommended not to use this function** as it currently (as of June 2022) uses a *z* value to calculate the confidence interval, rather than a *t* value.

1.9 Calculating a 95% confidence interval of a mean: summarised data

For Worked Example 3.2 where we are given the sample mean, sample standard deviation and sample size. R does not have a built-in function to calculate a confidence interval from summarised data, but we can write our own.

Note: writing your own functions is beyond the scope of this course. You should copy and paste the code provided to do this.

```
### Copy this section
ci_mean <- function(n, mean, sd, width=0.95, digits=3){
  lcl <- mean - qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)
  ucl <- mean + qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)

  print(paste0(width*100, "%", " CI: ", format(round(lcl, digits=digits), nsmall = digits),
               " to ", format(round(ucl, digits=digits), nsmall = digits) ))

}
### End of copy

ci_mean(n=30, mean=70, sd=6, width=0.95)
```

```
[1] "95% CI: 67.760 to 72.240"
```

```
ci_mean(n=30, mean=70, sd=6, width=0.99)
```

```
[1] "99% CI: 66.981 to 73.019"
```

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 2

Hypothesis testing

Learning objectives

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

- Formulate a research question as a hypothesis;
- Understand the concepts of a hypothesis test;
- Consider the difference between statistical significance and clinical importance;
- Use 95% confidence intervals to conduct an informal hypothesis test;
- Perform and interpret a one-sample t-test;
- Explain the concept of one and two tailed statistical tests.

Readings

Kirkwood and Sterne (2001); Chapter 8. [UNSW Library Link]

Bland (2015); Sections 9.1 to 9.7; Sections 10.1 and 10.2. [UNSW Library Link]

Acock (2010); Section 7.4.

2.1 Introduction

In earlier modules, we examined sampling and how summary statistics can be used to make inferences about a population from which a sample is drawn. In this module, we introduce hypothesis testing as the basis of the statistical tests that are important for reporting results from research and surveillance studies, and that you will be learning in the remainder of this course.

We use hypothesis testing to answer questions such as whether two groups have different health outcomes or whether there is an association between a treatment and a health outcome. For example, we may want to know:

- whether a safety program has been effective in reducing injuries in a factory, i.e. whether the frequency of injuries in the group who attended a safety program is lower than in the group who did not receive the safety program;
- whether a new drug is more effective in reducing blood pressure than a conventional drug, i.e. whether the mean blood pressure in the group receiving the new drug is lower than the mean blood pressure in the group receiving the conventional medication;
- whether an environmental exposure increases the risk of a disease, i.e. whether the frequency of disease is higher in the group who have been exposed to an environmental factor than in the non-exposed group.

We may also want to know something about a single group. For example, whether the mean blood pressure of a sample is the same as the general population.

These questions can be answered by setting up a null hypothesis and an alternative hypothesis, and performing a hypothesis test (also known as a significance test).

2.2 Hypothesis testing

Hypothesis testing is a statistical technique that is used to quantify the evidence against a null hypothesis. A null hypothesis (H_0) is a statement that there is no difference in a summary statistic between groups. For example, a null hypothesis may be stated as follows:

H_0 = there is no difference in mean systolic blood pressure between a group taking a conventional drug and a group taking a newly developed drug

We also have an alternative hypothesis that is opposite or contrasting to the null hypothesis. In our example above, the alternative hypothesis above we be that there is a difference between groups. The alternative hypothesis is usually of most interest to the researcher but in practice, formal statistical tests are used to test the null hypothesis (not the alternative hypothesis). The hypotheses are always in reference to the population from which the sample is drawn, not the sample itself.

After setting up our null and alternative hypotheses, we use the data to generate a test statistic. The particular test statistic differs depending on the type of data being analyses (e.g. continuous or categorical), the study design (e.g. paired or independent) and the question being asked.

The test statistic is then compared to a known distribution to calculate the probability of observing a test statistic which is as large or larger than the observed test statistic, if the null hypothesis was true. The probability is known as the P-value. Informally, the P-value can be interpreted as the probability of observing data like ours, or more extreme, if the null hypothesis was true.

If the P-value is small, it is unlikely that we would observe data like ours or more extreme if the null hypothesis was true. In other words, our data are not consistent with the null hypothesis, and we conclude that we have evidence against the null hypothesis. If the P-value is not small, the probability of observing data like ours or more extreme is not unlikely. We therefore have little or no evidence against the null hypothesis. In hypothesis testing, the null hypothesis cannot be proven or accepted; we can only find evidence to refute the null hypothesis.

To summarise:

- a small P-value gives us evidence against the null hypothesis;
- a P-value that is not small provides little or no evidence against null hypothesis;
- the smaller the P-value, the stronger the evidence against the null hypothesis.

Historically, a value of 0.05 has been used as a cut-point for finding evidence against the null hypothesis. A P-value less than 0.05 would be interpreted as “statistically significant”, and would allow us to “reject the null hypothesis”. A P-value greater than 0.05 would be interpreted as “not significant”, and we would “fail to reject the null hypothesis”. This arbitrary dichotomy is overly simplistic, and a more nuanced view is now recommended. Possible interpretations for P-values are given in Table 2.1.

P-values are usually generated using statistical software although other methods such as statistical tables or Excel functions can be used to generate test statistics and determine the P-value. In traditional statistics, the probability level was described as a lower-case p but in many journals today, probability is commonly described by upper case P. Both have the same meaning.

2.3 Effect size

In hypothesis testing, P-values convey only part of the information about the hypothesis and need to be accompanied by an estimation of the effect size, that is, a description of the magnitude of the difference between the study groups. The effect size is a summary statistic that conveys the size of the difference between two groups. For continuous variables, it is usually calculated as the difference between two mean values.

Table 2.1: Interpretation of P-values

Size of P value	**Strength of evidence**
$<$0.001	Very strong evidence
0.001 to $<$0.01	Strong evidence
0.01 to $<$0.05	Evidence
0.05 to $<$0.1	Weak evidence
≥0.1	Little or no evidence

If the variable is binary, the effect size can be expressed as the absolute difference between two proportions (attributable risk), or as an odds ratio or relative risk.

Reporting the effect size enables clinicians and other researchers to judge whether a statistically significant result is also a clinically important finding. The size of the difference or the risk statistic provides information to help health professionals decide whether the observed effect is large and important enough to warrant a change in current health care practice, is equivocal and suggests a need for further research, or is small and clinically unimportant.

2.4 Statistical significance and clinical importance

When applying statistical methods in health and medical research, we need to make an informed decision about whether the effect size that led to a statistically significant finding is also clinically important (see Figure 2.1)). The decision about whether a statistically significant result is also clinically important depends on expert knowledge and is best made by practitioners with experience in the field.

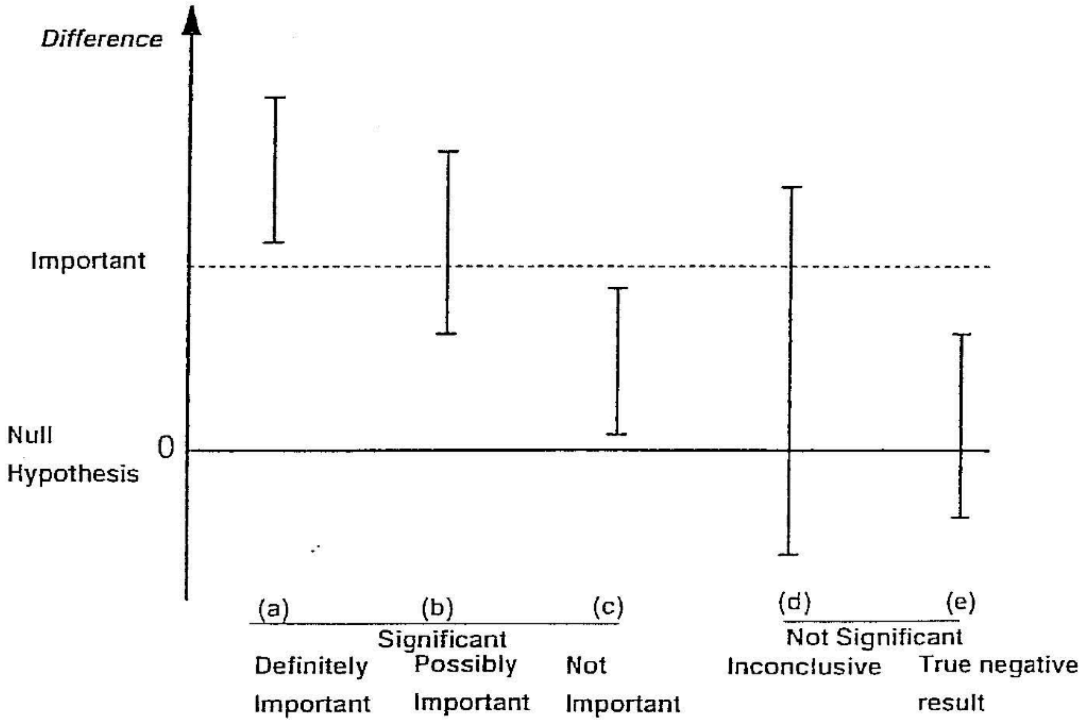


Figure 2.1: Statistical significance vs. clinical importance (Source: Armitage P, Berry G, Matthews JNS. (2001)

It is possible when conducting significance tests, particularly in very large studies, that a small effect is found to be statistically significant. For example, say in a large study of over 1000 patients, a new medication was found to lower blood pressure on average by 1 mmHg more than a currently accepted drug and this was statistically significant (P < 0.05). However, such a small decrease in blood pressure would probably not be considered clinically important. The cost and side effects of prescribing the new medication would need to be weighed against the very small average benefit that would be expected. In this case, although the null hypothesis would be rejected (i.e. the result is statistically significant), the result would not be clinically important. This is the situation described in scenario (c) of Figure Figure 2.1.

Conversely, it is possible to obtain a large, clinically important difference between groups, but a P value that does not demonstrate a statistically significant difference.

For example, consider a study to measure the rate of hospital admissions. We may find that 80% of children who present to the Emergency Department are admitted before an intervention is introduced compared to only 65% of children after the intervention. However, the P value may be calculated as 0.11 and is non-significant. This is because only 60 children were surveyed in each period. Here, the reduction in the admission rate by 15% represents a clinically important difference, but not statistically significant. This situation is represented in scenario (d) of Figure 2.1.

The important thing to remember is that statistical significance does not always correspond to clinical importance. A statistically significant result may be clinically unimportant, and a statistically non-significant results may be clinically important.

2.5 Errors in significance testing

There are two conclusions we can draw when conducting a hypothesis test: if the P-value is small, there is strong evidence against the null hypothesis and we reject the null hypothesis. If the P-value is not small, there is little evidence against the null hypothesis and we fail to reject the null hypothesis. As discussed above, the “small” cut-point for the P-value is often taken as 0.05. We refer to this value as α (alpha).

We can conduct a thought experiment and compare our hypothesis test conclusion to reality. In reality, either the null hypothesis is true, or it is false. Of course, if we knew what reality was, we would not need to conduct a hypothesis test. But we can compare our possible hypothesis test conclusions to the true (unobserved) reality.

If the null hypothesis was true in reality, our hypothesis test can fail to reject the null hypothesis – this would be a correct conclusion. However, the hypothesis test could lead us to rejecting the null hypothesis – this would be an incorrect conclusion. We call this scenario a Type I error, and it has a probability of α.

The other situation is where, in reality, the null hypothesis is false. A correct conclusion would be where our hypothesis test rejects the null hypothesis. However, if our hypothesis test fails to reject the null hypothesis, we have made a Type II error. The probability of making a Type II error is denoted β (beta). We will see in Module 10 that β is determined by the size of the study.

The error in falsely rejecting the null hypothesis when it is true (type I error), or in falsely accepting the null hypothesis when it is not true (type II error) is summarised in Table 2.2. We will return to these concepts in Module 10, when discussing how to determine the appropriate sample size of a study.

Table 2.2: Comparison of study result with the truth

Study result	**Truth**	
	Effect	No effect
Evidence	✓	α
No evidence	β	✓

2.6 Confidence intervals in hypothesis testing

In Module 3, the 95% confidence interval around a mean value was calculated to show the precision of the summary statistic. The 95% confidence intervals around other summary statistics can also be calculated.

For example, if we were comparing the means of two groups, we would want to test the null hypothesis that the difference in means is zero, that there is no true difference between the groups.

From the data from the two groups, we could estimate the difference in means, the standard error of the difference in means and the 95% confidence interval around the difference. To estimate the 95% confidence interval, we use the formula given in Module 3, that is:

$$95\% \text{ CI} = \text{Difference in means} \pm 1.96 \times \text{SE}(\text{Difference in means})$$

It is important to remember that the 95% CI is estimated from the standard error, and that the standard error has a direct relationship to the sample size. For small sample sizes, the standard error is large and the 95% CI becomes wider. Conversely, the larger the sample size, the smaller the standard error and the narrower the 95% CI becomes indicating a more precise estimate of the mean difference.

The 95% CI tells us the region in which we are 95% confident that the true difference between the groups in the population lies. If this region contains the null value of no difference, we can say that we are 95% confident that there is no true difference between the groups and therefore we would not reject the null hypothesis. This is shown in the top two estimates in Figure 2.2. If the zero value lies outside the 95% confidence interval, we can conclude that there is evidence of a difference between the groups because we are 95% confident that the difference does not encompass a zero value (as shown in the lower two estimates in Figure 2.2.

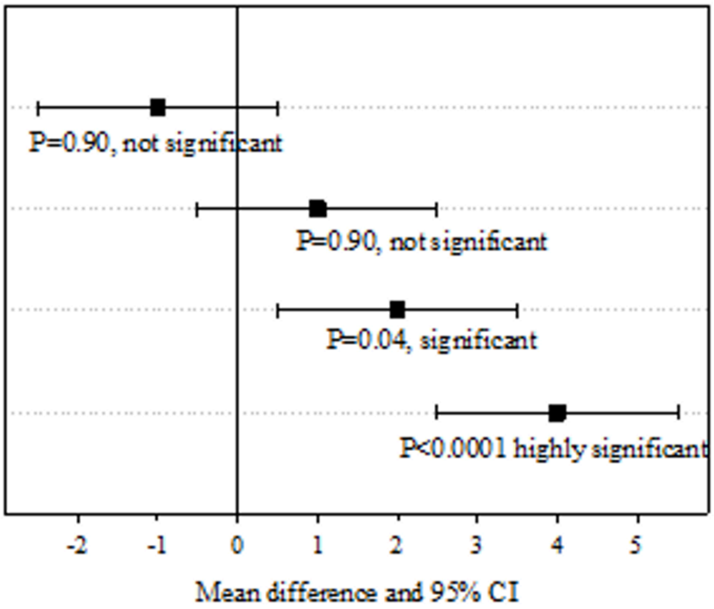


Figure 2.2: Using confidence intervals as informal hypothesis tests

For relative risk and odds ratio measures, when the 95% CI includes the value of 1 it indicates that we can be 95% confident that the true RR or OR of the association between the study factor and outcome factor includes 1.0 in the source population. This indicates little evidence of an association between the study factor and the outcome factor, e.g. if the results of a study were reported as RR = 1.10 (95% CI 0.95 to 1.25). The P-value can be calculated to assess this (discussed in Module 7).

Table 2.3: Values indicating no effect

Type of outcome	Measure of effect	Null value (indicating no difference)
Continuous	Difference in means	0
	Difference in proportions	0
Binary	Relative risk	1
	Odds ratio	1

2.7 One-sample t-test

A one-sample t-test tests whether a sample mean is different to a hypothesised value. The t-distribution and its relation to normal distribution has been discussed in detailed in Module 3.

In a one-sample t-test, a t-value is computed as the sample mean divided by the standard error of the mean. The significance of the t-value is then computed using software, or can be obtained from a statistical table.

The principles of this test can be used for applications such as testing whether the mean of a sample is different from a known population mean, for example testing whether the IQ of a group of children is different from the population mean of 100 IQ points or testing whether the number of average hours worked in an adult sample is different from the population mean of 38 hours.

Worked Example

The mean diastolic blood pressure (BP) of the general US population is known to be 71 mm Hg. The diastolic blood pressure of 733 female Pima indigenous Americans was measured and a histogram showed that the data were approximately normally distributed. The mean diastolic blood pressure in the sample was 72.4 mm Hg with a standard deviation of 12.38 mm Hg.

We can use Stata or R to conduct a one sample t-test using the data available on Moodle (`Example_4.1.csv`). The results from this test are summarised below.

Table 2.4: Summary of blood pressure from female Pima indigenous Americans

n	Mean	Standard deviation	Standard error	95% confidence interval of the mean
733	72.4	12.38	0.46	71.5 to 73.3

The test statistic for the one-sample t-test is calculated as t_{732}=3.07, with a P-value of 0.002.

The mean diastolic blood pressure of females from Pima is estimated as 72.4 mmHg (95% CI: 71.5 to 73.3 mmHg), which is higher than that of the general US population. Note that this interval does not contain the mean of the general US population (71 mm Hg), providing some indication that the mean diastolic blood pressure of female Pima people is higher than that of the general US population.

The result from the formal hypothesis test gives strong evidence that the mean diastolic BP of the female Pima people is higher than that of the general US population (t_{732}=3.07, P=0.002).

2.8 One and two tailed tests

Most statistical tests are two tailed tests, that is, we conduct a test that allows for the summary statistic in the group of interest to be either higher or lower than in the comparison group. For a t-test, this requires that we obtain a two-tailed P value which gives us the probability of the t-value being in either one of the two tails of the t-distribution as shown in Figure 2.3. The shaded regions show the t values that indicate a P value less than 0.05.

Occasionally, one tailed tests are conducted in which the summary statistic in the group of interest can only be higher or lower than the comparison group, i.e. a difference is specified to occur in one direction only. This makes it easier to reject the null hypothesis because the consequence is that the P value is essentially halved. The P value for a one tailed test would be 0.025 i.e. the shaded region for a one-tailed test would be doubled on one side of the distribution and eliminated from the other side of the distribution as shown in Figure 2.4.

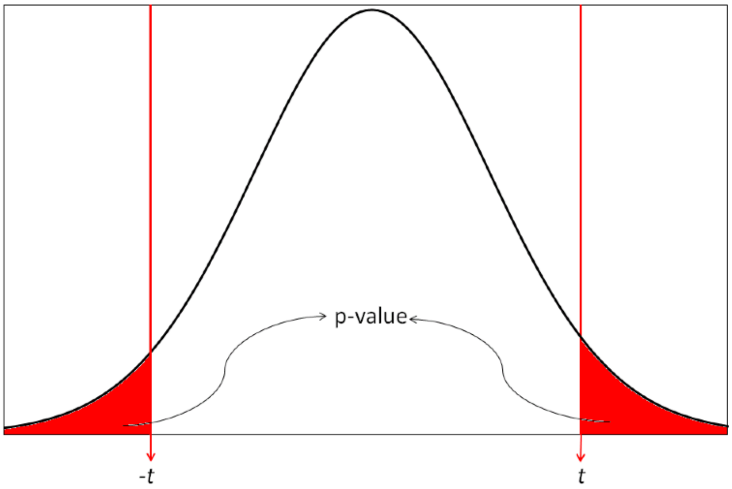


Figure 2.3: P-value for a 2-tailed test

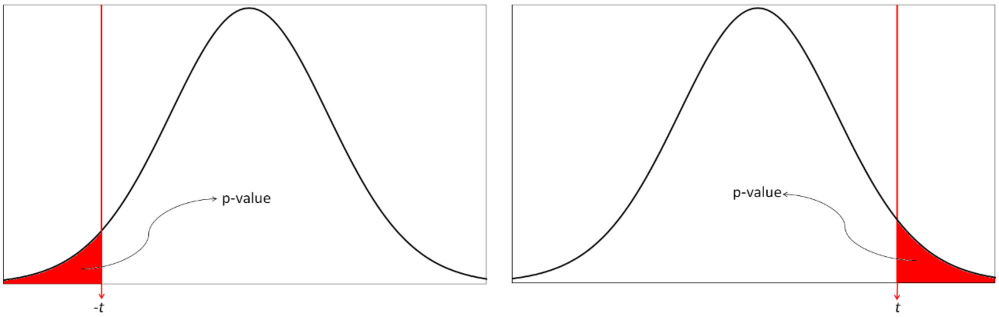


Figure 2.4: P-value for 1-tailed tests

Obviously, the choice of whether to use a one or two tailed test is not as important when the P value is highly significant or clearly non-significant but can make a difference to the conclusions when the P value is on the margins of significance.

In most health research, the use of a one tailed test is rarely justified because it is unusual to be certain of the direction of effect prior to the research study being undertaken. It has been suggested that if the researchers were sure enough to consider using a one-tailed test, the research study would not be needed.

In most studies, two tailed tests of significance are used to allow for the possibility that the effect size could occur in either direction. In clinical trials, this would mean allowing for a result that can indicate a benefit or an adverse effect in response to a new treatment. In epidemiological studies, two tailed tests are used to allow for the fact that exposure to a factor of interest may be adverse or may be beneficial. This conservative approach is usually adopted to prevent missing important effects that occur in the opposite direction to that expected by the researchers.

2.9 A note on P-values displayed by software

You will often see P-values generated by statistical software (including Stata) presented as 0.000 or 0.0000. As P-values can never be equal to zero, any P-value displayed in this way should be converted to <0.001 or <0.0001 respectively (i.e. replace the last 0 with a 1, and use the less-than symbol).

R can display P-values in a very cryptic way: 6.478546e-05 for example. This is translated as:

$$\begin{aligned} 6.478546e-05 &= 6.478546 \times 10^{-5} \\ &= 6.478546 \times 0.00001 \\ &= 0.00006478546 \end{aligned}$$

As for the Stata output, such a P-value would be better presented as P<0.0001.

2.10 Decision Tree

In the following modules in this course, several formal statistical tests will be described to analyse different types of data sets that have been collected to test set null hypotheses. It is important that the correct statistical test is selected to generate P-values and estimate effect size. If an incorrect statistical test is used, the assumptions of the test may be violated, the effect size may be biased and the P value generated may be incorrect.

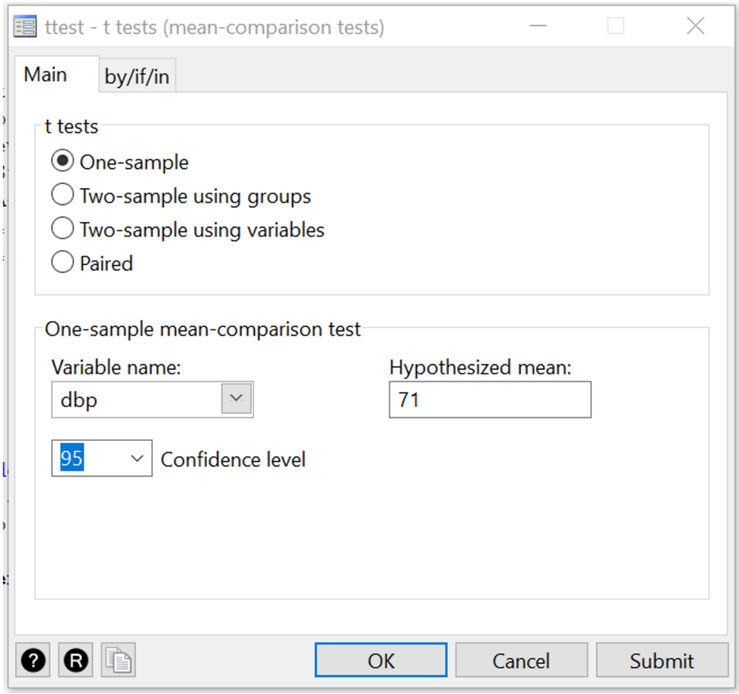
Selecting the correct test to use in each situation depends on the study design and the nature of the variables collected. Figure 1 in the Appendix shows a decision tree which enables you to decide the type of test to select based on the nature of the data.

Stata notes

2.11 One sample t-test

We will use data from `Example_4.1.csv` to demonstrate how a one-sample t-test is conducted in Stata. To perform the test, go to **Statistics > Summaries, tables, and tests > Classical tests of hypotheses > t test (mean-comparison test)**.

Ensure that the **One-sample** option is selected, then choose `dbp` as the **Variable name** from the drop-down list. Enter the **Hypothesised mean** value (`71` in this example) as shown below.



ttest - t tests (mean-comparison tests)

t tests

● One-sample
○ Two-sample using groups
○ Two-sample using variables
○ Paired

One-sample mean-comparison test

Variable name: dbp

Hypothesized mean: 71

95 Confidence level

Click **OK** or **Submit** to obtain the output below.

[Command: `ttest dbp == 71`]

```
. ttest dbp == 71

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     dbp |     733    72.40518    .4573454    12.38216    71.50732    73.30305
------------------------------------------------------------------------------
    mean = mean(dbp)                                              t =   3.0725
Ho: mean = 71                                    degrees of freedom =      732

    Ha: mean < 71                Ha: mean != 71                 Ha: mean > 71
```

```
 Pr(T < t) = 0.9989         Pr(|T| > |t|) = 0.0022          Pr(T > t) = 0.0011
```

The table gives the sample mean and standard deviation, as well as the standard error and 95% confidence interval of the mean. The test statistics are given under the table: t and the degrees of freedom as well as the P values from two-sided (Ha: mean != 71) and one-sided (Ha: mean <71 and Ha: mean > 71) tests. Refer to the previous sections of this module on the appropriate test to use.

R notes

2.12 One sample t-test

We will use data from `Example_4.1.rds` to demonstrate how a one-sample t-test is conducted in R.

```
library(jmv)

bloodpressure <- read.csv("data/examples/Example_4.1.csv")

descriptives(bloodpressure)
```

```
 DESCRIPTIVES

 Descriptives

                        dbp

   N                         733
   Missing                    35
   Mean                 72.40518
   Median                     72
   Standard deviation   12.38216
   Minimum                    24
   Maximum                   122
```

To test whether the mean diastolic blood pressure of the population from which the sample was drawn is equal to 71, we can use the t.test command:

```
t.test(bloodpressure$dbp, mu=71)
```

```

    One Sample t-test

data:  bloodpressure$dbp
t = 3.0725, df = 732, p-value = 0.002202
alternative hypothesis: true mean is not equal to 71
95 percent confidence interval:
 71.50732 73.30305
sample estimates:
mean of x
 72.40518
```

The output provides:

- a test statistic (t=3.07);
- degrees of freedom for the test statistic (df = 732);
- a P-value from the two-sided test (P=0.002);
- the mean of the sample (72.4);
- and the 95% confidence interval of the mean (71.6 to 73.3).

Activities

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”
d) “significant differences exist”

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).

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

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

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

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

References

Acock, Alan C. 2010. *A Gentle Introduction to Stata*. 3rd ed. College Station, Tex: Stata Press.

Bland, Martin. 2015. *An Introduction to Medical Statistics*. 4th ed. Oxford, New York: Oxford University Press.

Kirkwood, Betty, and Jonathan Sterne. 2001. *Essentials of Medical Statistics*. 2nd ed. Malden, Mass: Wiley-Blackwell.