MANCHESTER 1824
The University of Manchester

COURSE SAMPLE

DATA ANALYTICS
AND SOCIAL STATISTICS ONLINE

# INTRODUCTION TO STATISTICAL MODELLING

## Section 3: Statistical Inference: Confidence and Significance

# CONTENTS

# SECTION 3: STATISTICAL INFERENCE: CONFIDENCE AND SIGNIFICANCE

## INTRODUCTION

Welcome to Section 3 of Introduction to Statistical Modelling!

Please watch this video which provides an introduction to this section:


## LEARNING OBJECTIVES

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

- Quantify uncertainty of point estimates using confidence intervals for the mean and for a proportion in R
- Conduct hypothesis testing for the mean and for a proportion in R
- Explain the differences between Type I and Type II error and their implications in statistical inference
- Differentiate between confidence and significance in statistics and appreciate the practical limitations of significance tests
- Apply appropriate parametric tests to compare means and proportions in R
- Detect patterns of and measure association between categorical variables in R

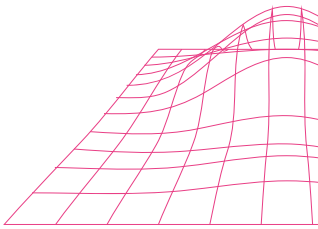
# 3.3 CONFIDENCE INTERVALS

To learn about populations from samples, we must be able to concede to the fact that samples will carry some degree of error. To estimate population parameters from sample statistics and draw appropriate conclusions from samples about populations, we must be able to adequately quantify this error. To quantify our uncertainty around the sample statistics we calculate and specify with some degree of confidence how close we think these approximate the population statistics, we can use interval estimates called confidence intervals. Confidence intervals are constructed using point estimates and the margin of error.

You have already encountered point estimates earlier in the course when you learned about descriptive statistics. Point estimates are single summary statistics values calculated for variables of interest in the sample such as the sample mean. These estimates represent the best possible approximations of the summary statistics we would calculate if we had access to the entire population, but they will vary from sample to sample. Since we are dealing with samples, these point estimates are insufficient to be reported on their own precisely because they are not the true population values so they will carry some error. As a result, point estimates are reported as part of an interval estimate (or confidence interval) which is a plausible interval around the sample point estimate within which it is expected that the true population parameter lies. Since the confidence interval itself is an estimate, we also need to specify the degree of confidence we have in the interval (and by extension in the point estimate). This is called the confidence level and must be set prior to calculating the point and interval estimate.

Point estimates can be different types of summary statistics depending on the type of variable of interest. For categorical variables, the point estimate is usually a proportion which represents the relative frequency of the categories of the variable in a given sample. For quantitative variables which take a normal distribution, the point estimate can be the mean or the median (with symmetric normal distributions, these two summary statistics are expected to be equal) which represents the most typical value of a variable in a population based on the sample data. Different types of estimators can be selected and those that are considered good estimators (or in statistical terms unbiased and efficient estimators) of population parameters must have a sampling distribution that is centred around the parameter (unbiased) and have the smallest error possible (efficient). The confidence interval is directly related to the point estimate and its sampling distribution because the normal distribution of the estimate determines the probability that the estimator falls within a certain distance of the parameter.

To learn more about how confidence intervals work in practice and how they are reported and interpreted, watch the video lecture below.



## 3.3.1 ACTIVITY QUIZ: TASK A: INTERACTIVE QUESTION AND ANSWER

When the 2000 General Social Survey (GSS) asked Americans whether human beings developed from earlier species of animals, 53.8% of 1095 respondents answered that this was probably or definitely not true. Find a 99% confidence interval for the corresponding population proportion and indicate whether you can conclude that a majority of Americans felt this way.

## 3.3.2 ACTIVITY QUIZ: TASK B: INTERACTIVE QUESTION AND ANSWER

Find and interpret the 95% confidence interval for $\mu$ if $\hat{y} = 70$ and $s = 10$, based on a sample size of:

- (a) 5
- (b) 20

## 3.4 HYPOTHESIS TESTING AND STATISTICAL SIGNIFICANCE

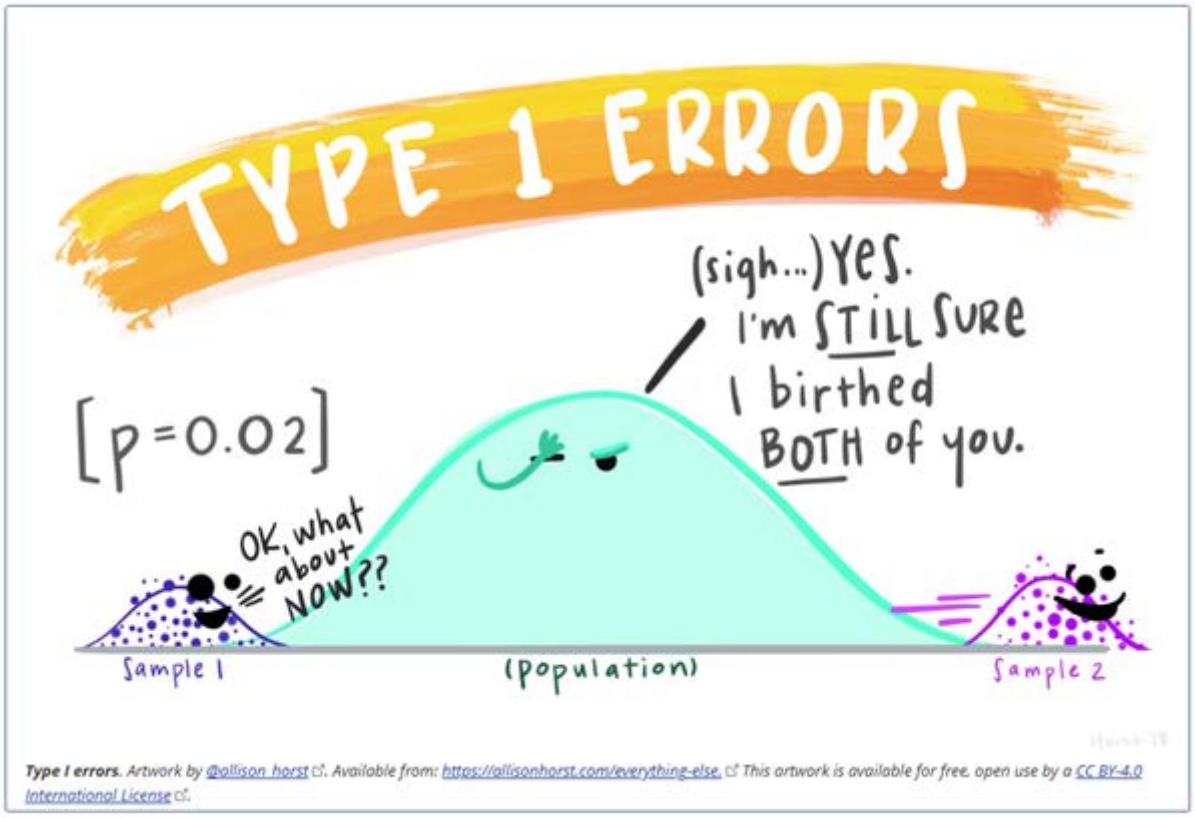
In Topic 1, we explored how to estimate population parameters and quantify our uncertainty with respect to these estimates.

Yet, how about if we want to make predictions or in statistical terms, hypothesise about population outcomes or behaviours using sample data? In this case, we would apply a process referred to as hypothesis testing. Whilst confidence intervals are used to estimate population parameters, hypothesis testing involves developing and testing justified statements (or in statistical terms, hypotheses) to find out whether there is evidence with respect to the prediction. This evidence is summarised through statistical significance which is represented by a measure called the p-value.

It is important to note that when performing hypothesis testing, hypotheses are always reported in pairs. The null hypothesis is reported first, followed by the alternative hypothesis which represents the researcher's prediction. The null hypothesis is simply a hypothesis which negates the alternative hypothesis in the sense that it is a statement which claims that there is no change, association, relationship, etc. in the data. It is the null hypothesis that is actually under investigation in a hypothesis test, not the researcher's prediction (i.e. the alternative hypothesis)! The reason for this is that the hypothesis testing process is a 'falsification' process which aims to reduce the chance of the researcher committing confirmation bias or in other words, seeking to interpret the results according to their pre-existing beliefs without considering the evidence. Therefore, through hypothesis tests, we are looking to reject OR fail to reject the null hypothesis (or that there is no difference/relationship etc. in the data).

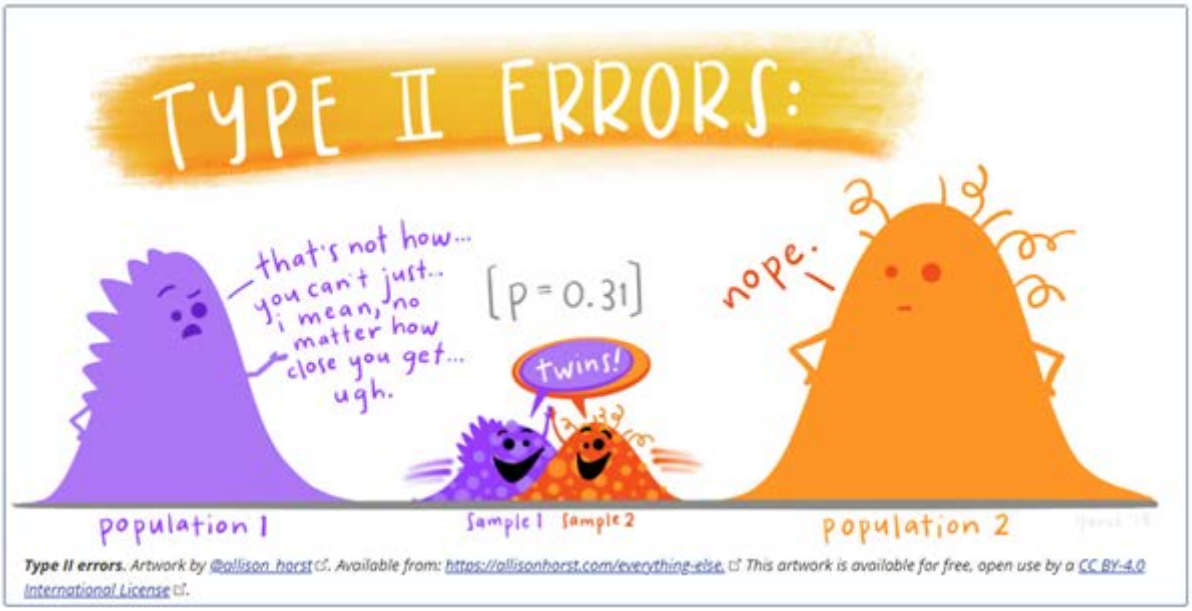
Yet, why would it be incorrect to say that we accept the null hypothesis or that we accept or reject the alternative hypothesis? The principal reason for this is that, since we are using samples, we do not know the true population values. Therefore, we cannot be 100% sure about the evidence for a prediction; rather, we quantify our probability of correctly rejecting or failing to reject the null hypothesis using the p-value. Even so, we can still commit errors. We can commit a Type I error when we reject the null hypothesis even though it is true, or we can commit a Type II error when we fail to reject the null hypothesis even though it is false. We can quantify these errors to a certain extent, but we cannot know with certainty if we did commit them in our research or not. This is one reason why it is important for studies to be repeated in order to validate or refute previous findings and therefore contribute to a better understanding of the world in which we live.

# TYPE 1 ERRORS




**Type I errors.** Artwork by @allison_horst. Available from: https://allisonhorst.com/everything-else. This artwork is available for free, open use by a CC BY-4.0 International License.

# TYPE 2 ERRORS




**Type II errors.** Artwork by @allison_horst. Available from: https://allisonhorst.com/everything-else. This artwork is available for free, open use by a CC BY-4.0 International License.

To learn how to perform hypothesis testing and interpret the results, watch the following short video lecture:


## 3.4.1 OPTIONAL READING

If you want to learn more about hypothesis testing and statistical significance, browse Chapter 6 in Agresti, A. (2018). Statistical Methods for the Social Sciences, Global Edition. Pearson.

Note: This further reading is optional. It is an additional resource that you may explore if you wish but it is not a compulsory part of the course materials.

## 3.4.2 ACTIVITY QUIZ: INTERACTIVE QUESTION AND ANSWER

We want to find out whether there is evidence that the proportion of adults in Manchester who are against reforming the British secondary school curriculum is different from 50%. Let's say we have a random sample of n = 1,353 people. Of these, 637 were opposed to the reforms: the rest favoured it.

Conduct a hypothesis test to find out if there is evidence to suggest that the proportion opposing the reforms is different from 50% and interpret the results using a 0.95 confidence level.

## 3.4.3 ACTIVITY

To learn how to perform significance tests for a mean and for a proportion in R. go through the example exercises below.

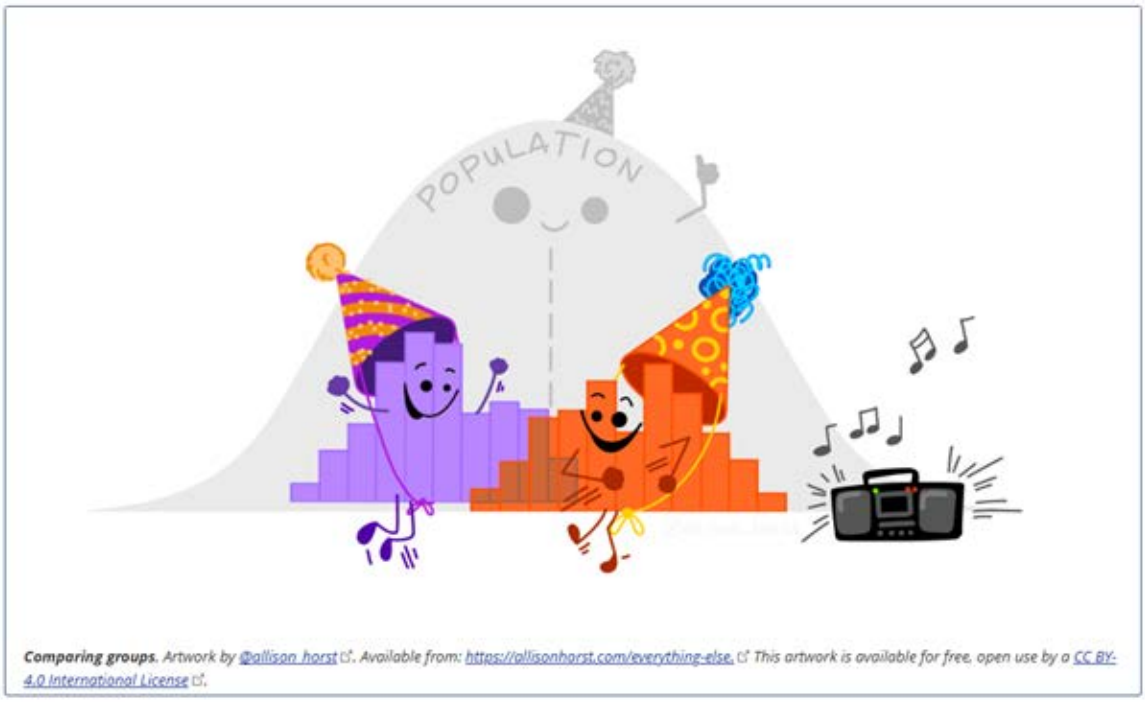
The Practical Worksheet can be downloaded here: Practical Worksheet: Section 3 - Practical 1

# 3.5 COMPARING TWO GROUPS

Let us now consider how we can compare differences between two groups, or more precisely, how we can apply specific statistical inference methods to analyse whether associations exist between a response variable and an explanatory variable that is binary, using sample statistics. In other words, we will explore whether the probability distribution of the response variable is influenced in some way by the values of the explanatory variable. Note here that the existence of an association does not imply that the changes in the response variable are caused by the explanatory variable; rather, the question is whether the explanatory variable has some influence on the response variable (we will explore causality in a different section of this course).

In social statistics, there are many instances when we would want to explore differences between two groups. For example, we may want to find out whether males and females respond differently when asked whether a career is more important than starting a family first and so we would compare differences in proportions. We may also be interested in gender-based differences in the amount of time spent (in hours per week) performing volunteer work in the community, in which case we would be comparing differences in means. Acknowledging differences between groups is important; for instance, women are more likely to want to start a family first whilst men would likely be more career-driven. Currently, this discrepancy may no longer be as prominent in developed countries as in the past, but it will still exist to some extent. If these differences exist, then the two groups would have their own population distribution rather than 'share' the distribution of a single population, as exemplified in the image shown. This difference has important implications for conclusions drawn about the population from the sample and must be acknowledged when conducting statistical inference.

When comparing groups from different studies/surveys, it is also important to identify whether the samples are dependent or independent because the standard error which quantifies how precisely the sample statistics estimate the population parameters, will be calculated differently. Independent samples refer to those which have observations that are not paired in both samples (e.g. from cross-sectional studies). Dependent samples will contain observations that are paired across samples, such as observations obtained in two different surveys from the same participants at two or more points in time (e.g. longitudinal studies). Longitudinal studies are very useful in social statistics because behaviours, outcomes, and characteristics can be measured over time. For example, you may want to investigate unemployment changes over time between males and females in particular professions and use these data to draft policies that addresses these existing problems.




***Comparing groups.*** *Artwork by @allison_horst. Available from: https://allisonhorst.com/everything-else. This artwork is available for free, open use by a CC BY-4.0 International License.*


## 3.5.1 OPTIONAL READING

If you want to learn more about comparing groups, you can read Chapter 7 in Agresti, A. (2018). *Statistical Methods for the Social Sciences,* Global Edition. Pearson.

Note: This further reading is optional. It is an additional resource that you may explore if you wish but it is not a compulsory part of the course materials.

# 3.6 ANALYSING CATEGORICAL VARIABLES: TESTS OF INDEPENDENCE AND ASSOCIATION

In Topic 3, we explored differences between the groups of a binary explanatory variable and categorical and continuous response variables. When considering associations between two categorical variables, there is another special statistical test that tests for statistical independence, or in other words, whether the conditional probability distribution of one of the categorical variables differs or not at each category of the other categorical variable. This test is called the Chi-Squared Test of Independence.

Of note is that the Chi-squared test does not provide any information other than if there is an association or not by means of statistical significance. If you want to draw more explicit conclusions about the pattern of association, the Chi-squared test result can be accompanied by an additional standardised measure, the standardised residual which functions much like a z-score and tells us whether independence is likely to be due to chance. If we are interested to find out the strength of association, there are other measures such as odds ratios for a pair of binary variables.

The Chi-squared test assumes that the variables being investigated are nominal. Measuring association between ordinal categorical variables differs because ordinal variables have ranked categories. As a result, the association between ordinal variables is of a positive or negative nature. For example, survey respondents that score high on variable X may also score high on variable Y (positive association) or respondents that score high on variable X may score low on variable Y or vice versa (negative association). For such cases, we can use measures such as Gamma.
To learn how to use statistical techniques to measure association between nominal and ordinal variables and interpret the results, watch the video below.


## 3.6.1 OPTIONAL READING

If you want to learn about association between categorical variables in more depth, you can have a look at Chapter 8 in Agresti, A. (2018). *Statistical Methods for the Social Sciences*, Global Edition. Pearson.

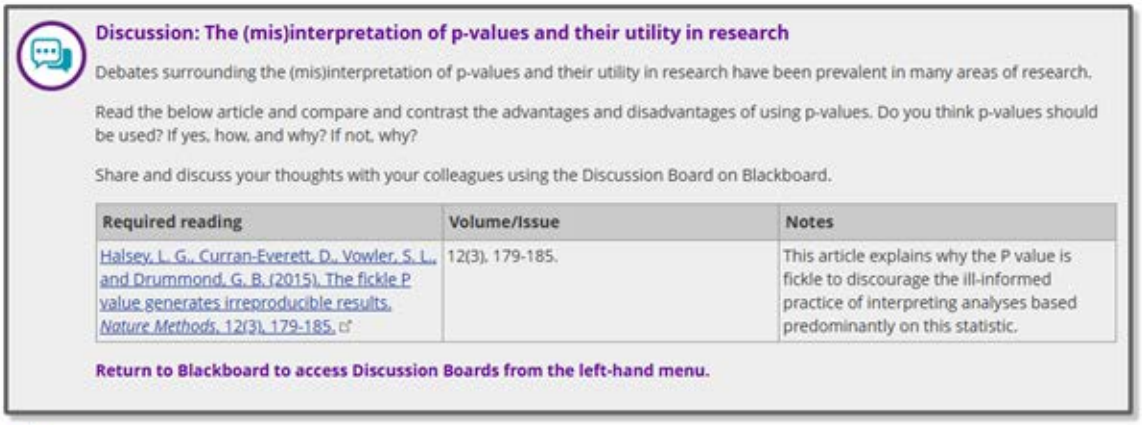
Note: This further reading is optional. It is an additional resource that you may explore if you wish but it is not a compulsory part of the course materials.

# 3.6.2 ACTIVITY QUIZ: INTERACTIVE QUESTION AND ANSWER

For a 2 by 4 cross-classification of gender and religiosity (very, moderately, slightly, not at all), for recent US General Social Survey data. the standardised residual was 3.2 for females who are very religious, -3.2 for males who are very religious, -3.5 for females who are not at all religious, and 3.5 for males who are not at all religious. All other standardised residuals fell between -1.1 and 1.1. Interpret these results.

# 3.7 SOCIAL ACTIVITY

During your studies you will be encouraged to discuss your learning with your colleagues via the Discussion Board in the VLE:



**Discussion: The (mis)interpretation of p-values and their utility in research**

Debates surrounding the (mis)interpretation of p-values and their utility in research have been prevalent in many areas of research.

Read the below article and compare and contrast the advantages and disadvantages of using p-values. Do you think p-values should be used? If yes, how, and why? If not, why?

Share and discuss your thoughts with your colleagues using the Discussion Board on Blackboard.

| Required reading | Volume/Issue | Notes |
|---|---|---|
| Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12(3), 179-185. | 12(3), 179-185. | This article explains why the P value is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic. |

**Return to Blackboard to access Discussion Boards from the left-hand menu.**

# 3.8 QUIZ

You will also be able to undertake interactive quizzes to test your understanding of the subject matter:

**Quiz**

To test your understanding of the materials from this section, please complete this short formative quiz. Don't worry, it doesn't form part of the unit assessment, but it will help you to check your understanding of the topics that we have covered. If you don't know the answers to any of the questions, you may need to refer back to the videos or your notes and read about the concepts again.

**Return to Blackboard to access Section Activities from the left-hand menu.**

# 3.9 TUTORIAL

There are also online tutorials, giving you the opportunity to discuss the course materials in more detail with your peers and instructors:

This online tutorial will help you develop your understanding of the topics covered so far and provide opportunities for questions. The tutorial will be recorded and all course participants will have access to the recording of the tutorial. If you have any concerns about being recorded, please contact your Student Support Advisor. This recording is to allow you to revisit the tutorial, and for those who are unable to attend to be able to benefit from the discussion within the tutorial.


# SECTION SUMMARY

In this section, you have learned how to quantify uncertainty using confidence intervals, conduct hypothesis testing and report significance of results, and explore relationships between variables using different types of tests.

In Section 4, you will learn how to analyse association between continuous quantitative variables using linear regression

# DATA ANALYTICS AND SOCIAL STATISTICS

Through studying this fully online, part-time course, you will learn to process and analyse complex social data effectively, improving your skills and professional outcomes in the process.

manchester.ac.uk/socialdata

studyonline@manchester.ac.uk