Statistics Primer

Aaron McMurray

2023-02-22

Contents

Statistics Primer

Introduction

Today, more than ever, statistics is part of our lives. Statistics can be used to describe the interactions of fundamental particles or to measure biodiversity in our oceans. It can be used to predict election outcomes through opinion polling in politics or to predict the prices of assets in finance. Whatever quantitative research questions we may have the answers will ultimately come from statistics.

This e-book is intended to provide an introduction to statistics and describe the quantitative methods that are useful in research.

The book has been written for a broad audience and as such most of the mathematical techniques used in statistics are described in plain English (as far as possible) and worked examples are provided to illustrate the process of calculating various statistics. There is no need to be mathematically inclined however to understand and use descriptive statistics in day-to-day research, especially with tools such as Excel, SPSS and Power BI which simplify many of the calculations involved in statistics.

Contents

Chapter 1 details the two main branches of statistics and how they might be useful.

Chapter 2 details sampling methodologies and may be useful to readers interested in data collection.

Chapter 3 details measures of uncertainty and how we describe the differences between sample statistics and population parameters through sampling error.

Chapter 4 describes the various data types and levels of measurement and how these classifications inform which statistics we use.

Chapters 7 and 8 will be most useful to those interested in summarising data sets as they outline the most commonly used descriptive statistics with worked examples.

Chapters 9, 10 and 11 provide an introduction to inferential statistics which may be of use to readers interested in hypothesis testing.

Chapter 12 provides a brief introduction to data visualisation, the tools used and guidance on when and how to use visuals.

Throughout this text, blue boxes will be used to provide depth, context or additional information on mathematical notation for interested readers.

Information

Throughout this e-book blue boxes like this one will be used to provide additional information, context or mathematical details that go beyond the scope of an introduction to statistics. The information provided in these boxes is provided for interested readers but is not necessary to have a basic understanding of statistics.

Red boxes will be used at the end of each chapter to summarise important information.

Summary

These are intended to provide a recap of all the most important information provided in a chapter needed to gain a basic understanding of statistics.

Questions or Corrections

Feel free to forward any feedback, questions or corrections to: xxxxxxxxx@ xxxxxxxxxxxx.uk

Chapter 1

Introduction

1.1 What is Statistics?

Statistics is a collection of methods for collecting, organising and analyzing data, interpreting data and drawing conclusions from and visualising data (Witte R. S. and Witte J. S., 2017). Statistical knowledge helps us to determine the best methods to collect data and to employ appropriate analyses and effectively present and communicate the results. Statistical practices should begin long before the analysis phase to avoid problems with the data collected, like biased samples and over generalisation (J., 2019).

There are two main branches of statistics: descriptive statistics and inferential statistics.

1.1.1 Descriptive Statistics

A descriptive statistic is a summary statistic that is used to describe or summarise data while descriptive statistics is the process of using those statistics to describe the characteristics of a sample or a population (Witte R. S. and Witte J. S., 2017). Descriptive statistics may describe a data set but they do not attempt to generalise beyond the set of observations or measurements made from the data set (J., 2019).

The distributions detailing the frequencies of values, measures of central tendency such as (mean, median and the mode) and measures of dispersion (range, interquartile range and standard deviation) are the most important tools. Descriptive statistics can also be used to describe the data set in terms of skew and correlation.

1.1.2 Inferential Statistics

Inferential statistical is used to make inferences about the properties of a population (Witte R. S. and Witte J. S., 2017). Sometimes these inferences are referred to as predictions. Inferential statistics involves hypothesis testing, confidence intervals and regression analysis. Since the goal is to take a sample and generalise its properties to a population we need to have confidence that the sample accurately reflects the population. We must ensure that the population is well defined, that we draw a representative sample from that population and that we use analyses that incorporate the sampling error (J., 2019).

Random sampling is the primary method used for obtaining samples that reflect the population however it can often be difficult to obtain a random sample.

Summary

There are two main branches of statistics: descriptive statistics and inferential statistics.

Descriptive statistics describes the characteristics of a sample or a population using measures of central tendency and measures of dispersion.

Inferential statistics is used to make inferences about the properties of a population.

Chapter 2

Sampling

Samples and populations are important in both inferential and descriptive statistics. Some descriptive statistics are calculated differently depending on whether they are being calculated for a sample or a population.

2.1 Populations and Samples

2.1.1 Population

The population includes all the elements from a set of data and the population characteristics that we calculate (e.g. mean, median, mode… etc) are called parameters.

Populations can include people but other examples include objects, events, businesses and so on. Before starting a research study it is important to define the population that is being studied. Many populations can additionally be stratified into sub populations where sub-populations share additional attributes (J., 2019) for instance, the population of a country can be subdivided into the male and female or subdivided into age bands. Sometimes the differences between sub-populations are unimportant and other times they are crucial.

2.1.2 Sample

A sample differs from a population in that it is comprised of one or more observations which are drawn from the population. It is a subset of the population. The sample characteristics (e.g. mean, median, mode… etc) that we calculate

are called statistics. Where possible, w use samples to draw inferences about wider populations. For instance, the voting intentions of 1,000 people (a sample) might be used to predict the outcome of a vote held for a general election involving 68,000,000 people (the population).

Surveying a sample rather than an entire population is often more financially cost effective but using a sample comes with a different kind of cost. Using a sample rather than the population means that the reported statistics come with associated measures of uncertainty which describe how the estimate might differ from the true value of the population (Office for National Statistics, 2022). If we were to measure the heights of 1,000 individuals in Northern Ireland, the average height of the sample is typically not going to be the same as the average height of all 1.8 million people in the country. The difference between the sample and the population values is considered a sampling error. If the sample mean is 176 cm and the population mean is 178 cm then the sampling error is 2 cm.

The exact measurement of sampling error is generally not feasible, since the true population values are not usually known. Sampling error however can be estimated by techniques such as the calculation of confidence intervals.

The various measures of uncertainty used to describe how estimates differ from the true value of the population include (Office for National Statistics, 2022):

- the standard error
- the coefficient of variation
- confidence intervals
- statistical significance

Calculating parameters of a population or statistics for a sample often falls within the remit of descriptive statistics but if we want to convert sample responses to population estimates we need to use inferential statistics.

To generalise the results from a sample to the full population, the sample must be representative of the population (this is known as representative sampling). For a sample to be representative of the population it must accurately represent the characteristics of the population.

In practice, this means that if we conducted a survey to gauge the views of the Northern Ireland population on an upcoming election and only residents of Antrim were surveyed then the sample results could not be used to infer attitudes of all Northern Ireland residents.

2.2 Sampling Methods

2.2.1 Probability Sampling

Probability sampling involves random selection allowing you to make statistical inferences about the whole group.

This chapter focuses on simple random sampling as it is one of the most common forms of probability sampling however there a a variety of different methods of probability sampling including: systematic sampling, stratified random sampling, cluster sampling, multi-stage sampling and multi-phase sampling.

2.2.1.1 Simple Random Sampling

Simple random sampling (SRS) is a procedure for selecting samples from a population. Under this procedure samples are taken from a population where each sample has equal probability of being selected. A very simple example of this is removing marbles from a bag of ten marbles one at a time, recording some characteristic (their colour or weight for instance) and replacing them. All the marbles have equal probability of being selected (p=0.1) regardless of how many samples we have taken.

Not replacing the marbles we sampled results in simple random sampling without replacement (SRSWOR). The first time we take a sample, each marble has a p=0.1 chance of being sampled. If we don't replace the marble we sampled before sampling a second marble however then the chance we select the first marble again is 0 and the other 9 marbles now have a p=0.11 chance of being selected. By the time we get to the bottom of the bag and two marbles remain they each have a p=0.5 chance of being selected.

In real world scenarios SRS can be difficult to achieve either due to the scale of the population or because SRS can introduce resampling (inadvertently giving a survey to the same person twice). SRSWOR is often used instead instead for these reasons.

2.2.2 Non-Probability Sampling

Non-probability sampling involves non-random selection. To draw conclusions about a population from a sample it is a requirement that the sample be representative of the population however when using non-probability sampling this is not guarenteed and it is important to consider this when using it as a sampling method in a study.

Non-probability sampling has seen some use within official statistics as a result of the growing popularity of non-probability data sources such as social media (webscraping Twitter or Facebook to conduct sentiment analysis for instance)

and the desire for real time statistics. There are a range of methods which can be used to conduct non-probability sampling including: convenience or haphazard sampling, volunteer sampling, judgement sampling, quota sampling, snowball or network sampling, crowd-sourcing and web panels.

One of the more widely used forms of non-probability sampling is convenience sampling. Convenience sampling consists of drawing from a source that is conveniently accessible to us (Andrade C., 2020). A convenience sample of students may be drawn from a physics department in Queen's University but these students may not be representative of all students in Northern Ireland.

The findings of a study based on convenience sampling can normally only be generalized to the sub population from which the sample is drawn and not to the entire population (Andrade C., 2020).

2.3 Independent and Dependent Samples

Samples are independent if the subjects in one sample do not determine which subjects are chosen for a second sample. Each group contains different subjects with no meaningful way to pair them. Independent groups are more commonly seen in hypothesis testing, for instance medical drug trials typically have a control group and a treatment group with different subjects. These studies typically use inferential statistical tests to determine if there are differences between the groups.

Dependent samples (sometimes referred to as matched pairs) differ from independent samples in that subjects in one sample can be matched with a corresponding subject in another sample. This can get confusing because sometimes matched pair consists of just one subject. For instance, if a sample is drawn of people who have hip replacement surgery with the NHS and the people in the sample are each interviewed before and after the surgery to assess their mobility before and after surgery then the study is engaged in dependent sampling. The same person was interviewed at two points in time.

Summary

Populations and Samples

The population includes all the elements from a set of data. The population characteristics (e.g. mean or standard deviation) that we measure are called parameters.

A sample is comprised of one or more observations which are drawn from the population. Sample characteristics that we measure are called statistics.

Sampling

To generalise the results from a sample to the full population, the sample must be representative of the population.

Surveys are based on a sample rather than the whole population so they are subject to sampling error.

The sampling error is the difference between the sample estimate and the 'true' value (which would have been obtained if a census of the whole population were undertaken.

Probability sampling involves random selection.

Simple random sampling (SRS) is a method for selecting samples from a population where all possible samples are equally likely to occur. Taking marbles from a bag, detailing a property (like weight), and replacing it in the bag is an example of SRS.

Not replacing a sample would constitute simple random sampling without replacement (SRSWOR).

Non-probability sampling involves non-random selection based on convenience or other criteria.

In independent samples subjects in one group provide no information about subjects in another.

Chapter 3

Measures of Uncertainty

The difference between a population parameter and a sample statistic is known as sampling error. There are various measures of uncertainty used to describe how estimates differ from the true value of the population include (Office for National Statistics, 2022):

- the standard error
- the coefficient of variation
- confidence intervals
- statistical significance

3.1 Standard Error

The standard error is a commonly used measure of sampling error.

The standard deviation is a descriptive statistic that details variability in a single sample statistic while the standard error is an inferential statistic that estimates the variability across multiple samples of a population (Lee D. K., In J. and Lee S., 2015).

The standard error shows how close the estimate based on sample data might be to the value that would have been taken from the whole population (Office for National Statistics, 2022).

The standard error of the mean (SEM) is the most commonly reported type of standard error but the standard error can be calculated for other statistics as well.

The standard error is calculated by dividing the standard deviation of a set of measurements by the square root of the number of measurements.

Information

The standard error is given by (Office for National Statistics, 2022):

$$SE = \frac{\sigma}{\sqrt{n}}$$

,

where SE is the standard error, σ is the population standard deviation and n is the number of elements in the sample.

In practice the population standard deviation is rarely known so instead the formula takes the sample standard deviation as a point estimate for the population standard deviation (Office for National Statistics, 2022):

$$SE = \frac{s}{\sqrt{n}},$$

where SE is the standard error, s is the sample standard deviation and n is the number of elements in the sample.

The standard error decreases as sample size increases as the extent of chance variation is reduced. This idea underpins sample size calculations of drug trials in medical research (Altman D. G. and Bland J. M., 2005). In contrast, the standard deviation will not tend to change as we increase the size of our sample.

3.2 Coefficient of Variation

The coefficient of variation makes it easier to understand whether a standard error is large compared with the estimate itself. It allows researchers to measure variation in a way which enables comparisons between data with different means (Martin J. D. and Louis N. G., 1997).

The coefficient of variation is also known as the relative standard error as it is a relative measure of dispersion (compared with standard deviation and interquartile range which are absolute measures). The coefficient of variation is calculated by dividing the standard error of an estimate by the estimate itself and the result indicates the relative spread of the data. An advantage to the coefficient of variation is that unlike other dispersion measures it takes central tendency into account (Martin J. D. and Louis N. G., 1997).

Similar to the standard error, the closer the coefficient of variation is to zero, the more precise the estimate is. Higher values indicate the standard deviation is large compared to the estimate. Where it is above 50%, the estimate is considered to be lacking in precision (Office for National Statistics, 2022).

The coefficient of variation should not be used for estimates of values that are close to zero or for percentages.

3.2.1 Example

Imagine a study that measures household expenditures where we want to compare the variability of spending in households of different incomes.

Table 3.1: Income Study Data

Expenditure	Q1 (least deprived)	Q2	Q3	Q4	Q5 (most deprived)
Mean	£120,000	£60,000	£40,000	£25,000	£12,000
Standard Error	£24,000	£12,000	£6,000	£7,500	£4,800

The variability is very high in the high income households compared to the low income households which is unsurprising given the substantial differences in the means. In order to account for the differences in the means a measure of relative variability like the coefficient of variation is used.

Calculating the coefficient of variation shows that when we account for differences in expenses the first two quintiles (Q1 and Q2) actually have equal variability and the greatest variability in seen in the most deprived quntile (Q5).

Table 3.2: Income Study Data

Expenditure	Q1 (least deprived)	Q2	Q3	Q4	Q5 (most deprived)
Coefficient of Variation	20%	20%	30%	30%	40%

3.3 Confidence Intervals

In inferential statistics the key goal is to estimate population parameters. Confidence intervals incorporate the uncertainty and sample error to create a range of values the true population value is likely to fall within (J., 2019).

Consider a study to estimate the mean weight of all 10 year old boys in Northern Ireland. It would be impractical to weigh them all so a sample of 16 might be taken. The mean weight of the sample might be 45 kg. This is a point estimate of the population mean.

Point estimation uses sample data to calculate a single value as a best guess of an unknown population parameter.

This point estimate has limited utility because it does not reveal uncertainty associated with the estimate. Is there confidence that the population mean is

within 5 kg of 45 kg? It's not possible to know with this information. That is why confidence intervals are calculated.

A 95% confidence level is frequently used in official reporting. If we drew twenty random samples and calculated a 95% confidence interval for each sample, we would expect that, on average, 19 out of the 20 (95%) resulting confidence intervals would contain the true population value while 1 in 20 (5%) would not (Office for National Statistics, 2022).

Example 3.3.1 illustrates how confidence intervals can be interpreted. Example 3.3.2. illustrates how confidence intervals are calculated. Some of the concepts (standard deviation and z-scores) have yet to be introduced but the example is provided regardless to show where these intervals come from and how we can improve the accuracy of our measurements through sampling. It isn't necessary to understand it in great detail to achieve a good foundation in statistics although the concepts necessary to understand it will be introduced in later chapters.

3.3.1 Example

Confidence intervals for religious composition of the economically active (Working age) 2011 are shown below:

Table 3.3:

Religious Denomination	Gender	Rate (%)	Confidence Interval (%)	Lower Limit (%)	Upper Limit (%)
Protestant	Male	53.3	+/- 2.6	50.7	55.9
Roman Catholic	Male	46.7	+/- 2.6	44.1	49.3
Protestant	Female	52.6	+/- 2.7	49.9	55.3
Roman Catholic	Female	47.4	+/- 2.7	44.7	50.1
Protestant	All	53.0	+/- 1.9	51.1	54.9
Roman Catholic	All	47.0	+/- 1.9	45.1	48.9

Based on a sample, the table above shows that 52.6 % of Protestant females (C.I. = +/- 2.7) were estimated to be economically active in 2011.

This means that there is 95% confidence that the 'true value' lies somewhere between 49.9% and 55.3%.

To calculate a confidence interval we need to know how many measurements we have, the mean and standard deviation of those measurements and a z-score.

3.3.2 Example

We measure the heights of ten people in the office and get a mean height of 172 cm and a standard deviation of 15 cm.

We need to decide the confidence interval we want. 95% is the most common choice.

We need to know the Z-score for that confidence interval. For a 95% confidence interval the Z score is 1.96.

The confidence interval is then given by multiplying the Z-score by the standard deviation and dividing by the square root of the number of observations or measurements.

Information

The Z-score describes how far a value is from the mean in terms of standard deviations. The z-score indicates how many standard deviations an element is from the mean. A standard score can be calculated from the following formula:

$$z = \frac{(X - \mu)}{\sigma},$$

where z is the z-score, X is the value of the element, μ is the mean of the population, and σ is the standard deviation.

A Z-score of zero would indicate that a value is identical to the mean value while a Z-score of 1 would indicate a distance of one standard deviation from the mean.

The formula for calculating a confidence interval is given by:

$$CI = \pm Z \frac{\sigma}{\sqrt{n}},$$

where the Greek letter sigma (σ) is the standard deviation, n is the number of observations or measurements and Z is the Z-score.

For now, assume the standard deviation is known and takes the value, $\sigma = 15$. The mean and its associated confidence interval can then be calculated:

$$172 \pm 1.96 \frac{15}{\sqrt{10}},$$

Plugging in the numbers gives:

$$172 \text{ cm} \pm 9.30 \text{ cm}.$$

In other words, the lower bound of the confidence interval is 162.7 cm and the upper bound is 181.3 cm. The true mean is likely between these two values. The confidence interval can be narrowed by increasing the number of measurements taken. With 100 measurements of height (and the same mean and standard deviation) the mean and its associated confidence interval would be stated as:

$$172 \pm 1.96 \frac{15}{\sqrt{100}},$$
$$172 \text{ cm} \pm 2.94 \text{ cm}.$$

This would make the range 169.1 cm to 174.9 cm.

The more observations that are collected the more accurate the measurement of the mean height becomes.

If it was somehow possible to measure the heights of a million people the mean and the associated confidence interval would become:

$$172 \text{ cm} \pm 0.03 \text{ cm}.$$

3.4 Statistical Significance

Statistical significance measures how likely it is that differences in outcomes between different groups are not due to chance. p values and confidence intervals are the most commonly used. The p values give the probability that any particular outcome would have arisen by chance while the confidence interval incorporates the uncertainty and sample error to create a range of values the true population value is likely to fall within (Leung W. C., 2001).

Statistical significance can be used to help decide whether a difference between two survey-based estimates reflects a true change in the population rather than random variation in our sample selection. A result is statistically significant if it is not likely to be caused by chance. A 5% standard is often used when testing for statistical significance. The observed change is statistically significant at the 5% level if there is less than a 1 in 20 chance of the observed change being calculated by chance if there is actually no underlying change (Office for National Statistics, 2022).

Summary

Standard Error

The standard deviation details variability in a single sample statistic while the standard error estimates the variability across multiple samples of a population.

It is calculated by dividing the standard deviation by the number of elements in a sample.

Coefficient of Variation

The coefficient of variation makes it easier to understand whether a standard error is large compared with the estimate itself. It is calculated by dividing the standard error of an estimate by the estimate itself.

Confidence Intervals

Confidence Intervals describe a range of values that likely contain the true value for a measurement.

A 95% confidence interval is calculated by multiplying the Z-score by the standard deviation and dividing by the square root of the number of observations or measurements.

Statistical Significance

Statistical significance measures how likely it is that differences in outcomes between different groups are real and not due to chance. p values and confidence intervals are the most commonly used.

Chapter 4

Data Types and Levels of Measurement

Understanding data is key to analysing its contents. The types of data we work with will inform our use of inferential and descriptive statistics.

4.1 Variables

A variable is an attribute that describes a person, place or thing. The value of the variable can vary from one observation to the next. For example, a person's hair colour is a variable that can take values like "blond" or "brown".

Variables can be classified as qualitative (categorical) or quantitative (numerical). Sometimes ranked data consisting of numbers (1st, 2nd,… 30th place) is included as a third category (Witte R. S. and Witte J. S., 2017).

To obtain data we have to observe or measure something, the something we observe or measure is a variable. For example, height, shoe size, weight and nationality are all variables as we can obtain observations or measurements for each of them (Campbell M. J., 2021):

Table 4.1: Variables and Measurements

Variable	Measurement or Observation	Type
Height	170 cm, 173 cm, 182 cm, 179 cm	Quantitative
Shoe Size	6, 6.5, 7, 7.5	Quantitative

Variable	Measurement or Observation	Type
Weight	35 kg, 40 kg, 75 kg, 0.57 kg	Quantitative
Nationality	Canadian, German, Spanish	Qualitative

Variables can be dependent or independent variables.

4.1.1 Independent Variables

When looking at variables it is common to consider whether there are relationships between them. An independent variable isn't changed by other variables. It is the variable that a researcher might change or control in a scientific experiment to test the effects on the dependent variable (Witte R. S. and Witte J. S., 2017).

4.1.2 Dependent Variables

When a variable is believed to have been influenced by changes in an independent variable, it is called a dependent variable. This is the variable that usually observed or measured by a researcher. This variable is not changed or manipulated during the course of a study (Witte R. S. and Witte J. S., 2017).

Statistical data is often classified according to the number of variables which are being studied.

4.1.3 Univariate and Bivariate Data

Univariate data involves only one variable. An example of this might be conducting a survey to estimate the average height of primary school children.

Bivariate data involves data with two variables. An example of this would be a study to determine if there was a relationship between the height and weight of primary school children.

Multivariate data involves data with more than two variables. This type of data is also sometimes referred to as multidimensional.

4.2 Types of Data

Data can be broadly categorised as qualitative (data relating to qualities or characteristics) or quantitative (numerical data relating to sizes or quantities of things).

4.2.1 Qualitative and Quantitative

Qualitative data is data for which the value of the qualitative variable is a name or a label. The colour of hair (blonde, brown, red) or a location (Belfast, Bangor, Lisburn) are examples of qualitative (or categorical) variables that take values which are names or labels.

Quantitative data is data for which the value of the quantitative variable is a number. For example, the population of a country is the number of people in that country. It is a numerical attribute of the country. Population is a quantitative variable that takes numerical values.

We can further categorise quantitative data as being continuous or discrete.

4.2.2 Continuous or Discrete

Discrete data involves whole numbers that can not be divided because of what the numbers represent (the number of students in a class, the number of cars owned, the number of fish in a lake). The number of students in a class cannot be 10.5 or 3.14.

Continuous data can be divided and measured to some number of decimal places (height, weight, speed). Height is an example of continuous data. Height can be any number (provided it lies within the range of possible human heights) and can be reported to any number of decimal places (150 cm or 150.1 cm or 150.12 cm) depending on how accurate the measurement tool is.

Information

Accuracy and precision are terms used to refer to the quality of measurements. The accuracy of a measurement describes how close the measurement is to its true value. Precision describes the degree to which an instrument or process will repeat the same value or how close measurements of the same item or quality are to one another.

There are also different levels of measurement.

4.3 Levels of Measurement

The levels of measurement describe how precisely variables are recorded. The different levels of measurement limit which statistics can be used to summarise data and which inferential statistics can be performed. These levels are:

- Nominal
- Ordinal

- Interval
- Ratio

4.3.1 Nominal

To be measured at the nominal level, a variable must be able to be divided up into separate or discrete categories, which are then named (McHugh M. L. and Villarruel A. M., 2003). The nominal level of measurement is the simplest form of a scale of measurement.

Nominal data is also called categorical data since the subjects are allocated to different categories. It can be categorised but not ranked (eye colour and gender for instance). The values grouped into these categories have no meaningful order. It is not possible to form a meaningful hierarchy of gender, hair colour, eye colour or marital status for instance.

It is fairly common to represent nominal data using bar charts like the chart below. The bar chart below shows the marital status of people in Northern Ireland based on the Census 2011 data (Census, 2011b).

Nominal data can be analysed by grouping variables into categories. For each category, the frequency or the relative frequency can be calculated. The data can be presented visually and usually is illustrated using bar charts or pie charts. The only measure of central tendency used with nominal data is the mode.

Inferential statistics can be used with nominal data. Chi-square tests are non-parametric tests for categorical variables.

4.3.2 Ordinal

The ordinal level of measurement is the second level of measurement. Ordinal data is another type of qualitative data that groups variables into descriptive categories. The categories used for ordinal data are ordered in some kind of hierarchical scale although the distance between those categories may be uneven or even unknown. For example, measuring economic status using a social class hierarchy involves the use of categories with no clearly identifiable or evenly spaced interval between them.

The pie chart below shows the highest level of qualification of usual residents in households in Northern Ireland aged 16-64 in 2011 (Census, 2011a). Ordinal data is often illustrated through pie charts or bar charts.

Ordinal variables often include ratings about opinions that can be categorised (strongly agree, agree, don't know, disagree, strongly disagree).

The descriptive statistics which can be used with ordinal data are the mode and the median. Ordinal data can also be described with a measure of dispersion, namely, range.

There are a number of possible statistical tests that can be used with ordinal data. Which one is used depends on the aims of the researcher and the number and type of samples.

Table 4.2: Inferential Statistics

Non-parametric test	Aim	Samples or variables
Mood's median test	Compares medians	2 or more samples
Mann-Whitney U test (also referred to as the Wilcoxon rank sum test)	Compares sums of rankings of scores	2 independent samples
Wilcoxon matched-pairs signed-rank test	Compares the magnitude and direction of the difference between distributions of scores	2 dependent samples
Kruskal-Wallis H test (also referred to as the one-way ANOVA on ranks)	Compares the mean rankings of scores	3 or more samples
Spearman's rho or Kendall's Tau	Measures correlation between 2 variables	2 ordinal variables

4.3.3 Interval

The next level of measurement is interval measurement. Interval measures have categories and magnitude, just like nominal and ordinal measures but this measure adds the concept that the intervals between each measure are exactly equal (McHugh M. L. and Villarruel A. M., 2003). Interval data groups variables into categories where the values are ordered and separated by equal distances.

Interval data is a type of quantitative data that groups variables into categories. Values can be ordered and separated using an equal measure of distance.

An example of interval level data is temperature data recorded in Celsius or Fahrenheit. The values on either scale are ordered and separated using an equal measure of distance (the distances between notches on a thermometer are always equally spaced).

Temperature in Celsius is interval data. The values are ordered and separated by an equal interval. The distance between 0°C and 1°C is the same as the distance between 2°C and 3°C.

The line chart below shows some simulated temperature data for Belfast. Interval data is often visualised using line charts.

Mathematical operations can be carried out on this type of data, for instance, subtracting one value from another to find the difference.

Interval data lacks a true zero. True zero indicates a lack of whatever is being measured. The Celsius scale doesn't qualify as having a true zero since the zero point in a thermometer is arbitrary.

Information

When the Celsius scale was first created by Anders Celsius 0°C was selected to match the boiling point of water and a value of 100 °C was the freezing point of water - it could have as easily been a different liquid with a different boiling point and freezing point making it arbitrary. The scale was later reversed. Thermometers measure heat and at 0°C there is still heat, maybe not a great deal of it but heat is still measurable meaning 0°C is not a true zero. The thermodynamic Kelvin Scale has a true zero - where particles have no motion and can become no colder (there is a true absence of heat).

A range of descriptive statistics can be used to describe interval data. The measures of central tendency applicable to interval data are the mode, median and the mean. The measures of dispersion applicable to interval data are the range, standard deviation and the variance.

A number of parametric tests can be applied to interval data.

Table 4.3: Inferential Statistics

Parametric test	Aim	Samples or variables
T-test	Compares means of two samples	2 samples
Analysis of Variance (ANOVA)	Compares means of several samples	3 or more samples
Pearson's r	Measures correlation between two variables	2 variables
Simple linear regression	Estimate the relationship between a dependent variable and an independent variable using a straight line	2 variables

4.3.4 Ratio

Ratio data measures variables on a continuous scale and has a true zero.

Ratio data is a form of quantitative data. It measures variables on a continuous scale with an equal distance between adjacent values (weight, height). Ratio data has a true zero unlike interval data. Ratio data is the most complex of the four data types.

Ratio data can be analysed with descriptive statistics including the mode, median and mean. Range, standard deviation, variance and the coefficient of variation can all be used to describe the dispersion of ratio data.

The tests used with interval data can be used with ratio data as well.

Summary

Data can be broadly categorised as qualitative (data relating to qualities or characteristics) or quantitative (numerical data relating to sizes or quantities of things).

Qualitative data deals with names or labels.

Quantitative data is numerical.

Discrete data involves whole numbers that can not be divided because of what the numbers represent (the number of people in a class, the number of cars owned, the number of fish in a lake).

Continuous data can be divided and measured to some number of decimal places (height, weight, speed).

There are four levels of measurement:

- Nominal - Used to label variables. Frequencies can be calculated, as can the mode.
- Ordinal - Groups variables into descriptive categories with some sort of heirarchy. The mode and the median can be calculated, as can the range.
- Interval - Groups variables into categories where the values are ordered and separated by equal distances. The mode, median and mean can all be calculated as can the range, standard deviation and variance.
- Ratio - Measures variables on a continuous scale and has a true zero. The mode, median and mean can all be calculated as can the range, standard deviation, variance and the coefficient of variation.

Chapter 5

Levels of Measurement Quiz

If you would like to try and test your knowledge of the various levels of measurement outlined in this chapter you can take the Levels of Measurement Quiz below. This quiz isn't scored or recorded anywhere.

Chapter 6

Describing Data

There are lots of different ways to describe data and some of them have been mentioned already (mode, mean and median). This chapter details how various descriptive statistics are calculated.

6.1 Measures of Central Tendency

Measures of central tendency are used to find the middle, or the average, of a data set. The measures of central tendency are the mean, median and mode.

6.1.1 Mean

The mean (sometimes referred to as the arithmetic mean) is the sum of the recorded values divided by the number of values recorded. The formula for the mean then is given by:

$$\text{Mean} = \frac{\text{Sum of recorded values}}{\text{Number of values recorded}}.$$

6.1.1.1 Example

Find the mean of the following list of numbers:

$$2, 3, 3, 4, 20.$$

$$\text{Mean} = \frac{\text{Sum of recorded values}}{\text{Number of values recorded}},$$

$$\text{Mean} = \frac{2 + 3 + 3 + 4 + 20}{5},$$

$$\text{Mean} = \frac{32}{5},$$

$$\text{Mean} = 6.4.$$

The list of numbers could have been a population or a sample from a population and the steps involved in the calculation would remain unchanged.

6.1.2 Median

The median is the middle number in a sorted, ascending or descending, list of values.

If there are an odd number of values the median is simply the middle value.

For an even number of values there will be two values in the center. Those values are summed and divided by two.

The median is sometimes used as opposed to the mean when there are outliers that might skew the average of the values. For ordinal data, the median is usually the best indicator of central tendency (McHugh M. L. and Hudson-Barr D., 2003).

6.1.2.1 Example

Find the median of this list of numbers:

$$2, 3, 3, 4, 20.$$

There are 5 values listed in ascending order and the middle value is the third value in the list so the median is 3.

Information

In the previous example the mean was 6.4. It was skewed by the outlier (20). The median remains closer to what might be considered to be the middle of the data set if the outlier was not present. This illustrates one of the main uses of the median. It is often used when there are outliers in a data set that might skew the average of the values.

6.1.2.2 Example

Find the median of this list of numbers:

$$3, 5, 4, 4, 2, 8, 7, 1.$$

The list should be sorted:

$$1, 2, 3, 4, 4, 5, 7, 8.$$

There are an even number of values so there will be two middle values. The middle values are 4 and 4. Sum them and divide by two to get the median: 4.

6.1.3 Mode

The mode of a set of data values is the value that appears most often. It is the value that is most likely to be sampled. There can be multiple modes or no modes. The mode is the only useful measure of central tendency when a variable is measured on a nominal scale (McHugh M. L. and Hudson-Barr D., 2003).

6.1.3.1 Example

Find the mode of this list of numbers:

$$1, 2, 2, 2, 3, 3, 4.$$

A simple way to find the mode is to make a frequency table with the unique values on the left hand side and their frequency on the right hand side. We can tally up how many times each number occurs. Whichever has the greatest frequency is our mode.

Table 6.1:

Value	Frequency
1	1
2	3
3	2
4	1

The mode is 2.

6.1.3.2 Example

Find the mode of this list of numbers:

$$7, 3, 5, 3, 4, 3, 5, 6, 8, 5.$$

Table 6.2:

Value	Frequency
3	3
4	1
5	3
6	1
7	1
8	1

This is bimodal, it has two modes, 3 and 5.

6.1.3.3 Example

Find the mode of this list of numbers:

$$1, 2, 3, 4, 5, 6.$$

Table 6.3:

Value	Frequency
1	1
2	1
3	1
4	1
5	1
6	1

Every value is unique and occurs only once so this data has no mode.

6.1.4 Mean or Median?

The median may be a better indicator of the most typical value if a set of scores has outliers. Outliers are extreme values that differ greatly from other values. When the sample size is large and does not contain outliers the mean score usually provides a better measure of central tendency.

6.1.5 Using Excel

It is useful to calculate descriptive statistics by hand for understanding but for larger data sets it is not always possible to arrange data and perform calculations by hand.

Excel has a number of functions designed to perform descriptive statistics.

Frequency

=FREQUENCY(start:end,bins_array)

The frequency() function will return a frequency table describing your data. It takes two arguments, the first being the array of values and the second being an array describing the upper boundary of the bins used.

Average

=AVERAGE(start:end)

The mean is calculated using the average() function. There are several other functions relating to means: geomean(), harmean() and trimmean(). Take care not to use these as they are quite different from calculating the mean that has been described here.

Median

=MEDIAN(start:end)

The median is calculated using the median() function.

Mode

=MODE.SNGL(start:end)

=MODE.MULT(start:end)

There are several functions for calculating the mode: mode(), mode.sngl() and mode.mult().

mode() was used in Excel 2007 and may still appear as an option in some versions of Excel.

mode.sngl() will return one mode and mode.mult() will return multiple modes (if there are multiple modes).

Neither mode() nor mode.sngl() will provide a warning if there are multiple modes so mode.mult() is usually the safest option.

6.2 Frequency

The frequency of an observation is the number of times it occurs or is recorded. A frequency distribution for grouped data, like the one shown below detailing exam grades, is a commonly used method of depicting frequency.

Table 6.4: Frequency Table for Grouped Data

Grade	Frequency
A	15
B	20
C	25
D	21
E	14

A frequency distribution is a collection of observations produced by sorting observations into classes and showing their frequency of occurrence in each class. A frequency distribution helps us discern patterns in data (assuming they exist) by imposing a structure to the data (Witte R. S. and Witte J. S., 2017).

The total of all frequencies so far in a frequency distribution is the cumulative frequency. It is the 'running total' of frequencies.

Table 6.5: Cumulative Frequency Table

Grade	Frequency	Cumulative Frequency
A	15	15
B	20	35
C	25	60
D	21	81
E	14	95

The relative frequency is the ratio of the category frequency to the total number of outcomes. For grade A, the relative frequency is:

$$\text{Relative Frequency} = \frac{15}{15 + 20 + 25 + 21 + 14} = 0.16.$$

The table can be extended to include the relative frequency.

Table 6.6: Relative Frequency Table

Grade	Frequency	Relative Frequency
A	15	0.16
B	20	0.21
C	25	0.26
D	21	0.22
E	14	0.15

The relative frequency relates the count for a particular event to the total number of events using percentages, proportions or fractions and it can be reported as a percentage by multiplying the values by 100%. For grade A, the relative frequency reported as a percentage is: 100% x 0.16 = 16%.

6.2.1 Mean of a Frequency Distribution

While it is common to calculate the mean of a data set sometimes we receive data in the form of a frequency table. To calculate the mean we multiply the value by its frequency, sum the results and divide by the cumulative frequency.

6.2.1.1 Example

Calculate the mean given the values and their respective frequencies in the table below:

Table 6.7: Relative Frequency Table

Value	Frequency	Value x Frequency
1	2	2
2	3	6
3	5	15
4	6	24
5	5	25
6	4	24
7	2	14
8	1	8

The products of the values and their frequencies have been calculated in the table above, all that is left is to sum them and divide by the cumulative frequency:

$$\text{Mean} = \frac{2+6+15+24+25+24+14+8}{2+3+5+6+5+4+2+1} = \frac{118}{28} = 4.21$$

Information

We can write this using mathematical notation:

$$\mu = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i},$$

where x_i are the individual values and f_i their respective frequencies.

6.2.2 Mode of a Frequency Distribution

The modal value (or the modal class in the case of a frequency distribution) is simply the value which corresponds to the largest frequency. In the example above the modal value is 4.

6.2.3 Median of a Frequency Distribution

To find the median of a frequency distribution we need to first calculate the cumulative frequency:

Table 6.8: Relative Frequency Table

Value	Frequency	Value x Frequency	Cumulative Frequency
1	2	2	2
2	3	6	5
3	5	15	10
4	6	24	16
5	5	25	21
6	4	24	25
7	2	14	27
8	1	8	28

We divide the cumulative frequency by 2 to find the midpoint. In this case, it's 14. Then, check each value to see if its corresponding cumulative frequency is greater than that number. The first value which has a cumulative frequency greater than that number is the median value. The first value in the table above which has a cumulative frequency greater than 14 is 4. This is the median.

6.2.4 Mean of a Grouped Frequency Distribution

If the frequency table is a grouped data frequency table, where the values are banded (0-5,5-10,10-15...etc), then the equation for the mean uses the midpoint of the band (which is the upper limit minus the lower limit) in place of a single value.

Take the table below for instance:

Table 6.9: Grouped Frequency Table

Bin	Frequency
10-14	1

Bin	Frequency
15-19	3
20-24	9
25-29	2

To calculate the mean we would rewrite this table as follows:

Table 6.10: Grouped Frequency Table

Midpoint	Frequency
12	1
17	3
22	9
27	2

Previously we created a new column for the product of the value and the frequency. We do the same again but this time the new column will hold values for the product of the midpoint with the frequency:

Table 6.11: Grouped Frequency Table

Midpoint	Frequency	Mf
12	1	12
17	3	51
22	9	198
27	2	54

The process is the same as before. We sum the products of the midpoint and the frequency and divide by the cumulative frequency:

$$\text{Mean} = \frac{12 + 51 + 198 + 54}{1 + 3 + 9 + 2} = 21$$

Information

We can write this in mathematical notation as:

$$\mu = \frac{\sum_{i=1}^{n} M_i f_i}{\sum_{i=1}^{n} f_i},$$

.

where M is the midpoint and f is the frequency.

6.2.5 Median of a Grouped Frequency Distribution

To find the median we need several values, l, the lower limit of the median class, n the total number of observations, c_f, the cumulative frequency of the class preceding the median class, f, the frequency of the median class and c_l the class length. Given these, the median is:

$$\text{Median} = l + c_l \frac{\frac{n}{2} - c_f}{f}$$

Table 6.12: Relative Frequency Table

Bin	Frequency	cf	Mf
10-14	1	1	12
15-19	3	4	51
20-24	9	13	198
25-29	2	15	54

The total number of observations $n = 15$.

Divide this by 2 to get 7.5

From this we can find the lower limit of the median class by finding the cumulative frequency which is just larger than this number. This corresponds to the median class. For us that's the 20-24 class.

The lower limit, l, of this class is 20.

The cumulative frequency of the class preceding the median class, c_f, is 4.

The frequency of the median class, f, is 9.

The class length, c_l, is 4.

The median then is calculated by plugging these values into the formula above:

$$\text{Median} = l + c \frac{\frac{n}{2} - cf}{f},$$

$$\text{Median} = 20 + \frac{4(\frac{15}{2} - 4)}{9},$$

$$\text{Median} = 21.6.$$

Information

The variance (more on this later) of a grouped frequency distribution is given by:

$$V = \frac{\sum_{i=1}^{n} f_i M_i^2 - \mu \sum_{i=1}^{n} f_i}{\mu - 1},$$

where f_i are the frequencies, M_i are the midpoints of the bands (or bins), μ is the mean.

The standard deviation given by the square root of V.

6.3 Measures of Dispersion

Dispersion (or variability) describes how far apart data points lie from each other and the center of a distribution. The range, interquartile range, variance and standard deviation are all measures of dispersion and they describe how far apart data points lie from one another and the center of a distribution.

6.3.1 Range

The range is the difference between the highest and lowest values and is calculated by subtracting the minimum value from the maximum value.

6.3.1.1 Example

Calculate the range for the following set of numbers:

$$23, 42, 75, 19, 74.$$

First, arrange the values in ascending order:

$$19, 23, 42, 74, 75.$$

The maximum value is 75 and the minimum is 19.

$$\text{Range} = 75 - 19,$$

$$\text{Range} = 56.$$

6.3.2 Interquartile Range

The interquartile range (IQR) describes the spread of the middle half of a distribution. How the interquartile range is calculated depends on whether there are an even or an odd number of values in a dataset.

For an even number of values the dataset in split half. The medians for the two new subsets of data are calculated. The positive difference of those medians is the interquartile range.

For an odd number of values either the inclusive or the exclusive method of finding the interquartile range must be used.

The algorithm for the exclusive method is detailed below:

1. Arrange the data in numeric order.
2. Remove the median and split the data about its center.
3. Find the medians of the two newly appended subsets of data.
4. Calculate the difference.

The algorithm for the inclusive method is detailed below:

1. Arrange the data in numeric order.
2. Remove the median and split the data about its center.
3. Append the two new subsets of data with the median.
4. Find the medians of the two newly appended subsets of data.
5. Calculate the difference.

6.3.2.1 Example

Find the interquartile range for the list of numbers below:

$$6, 7, 8, 8, 7, 6, 9, 5, 10, 4.$$

There are an even number of values. Arrange them in numeric order:

$$4, 5, 6, 6, 7, 7, 8, 8, 9, 10.$$

Split the values about their center into two sub sets of data.

$$(4, 5, 6, 6, 7), (7, 8, 8, 9, 10).$$

Find the medians of each of these sub sets. The first subset has a median of 6 while the second has a median of 8.

The interquartile range is:

$$\text{IQR} = 8 - 6 = 2.$$

Note: To calculate the interquartile range the smaller median value is always subtracted from the larger.

6.3.2.2 Example

Find the interquartile range for the list of numbers below:

$$2, 3, 2, 4, 3, 5, 4, 4, 2.$$

Arrange the values in numeric order:

$$2, 2, 2, 3, 3, 4, 4, 4, 5.$$

Remove the median (3) and split the data as before:

$$(2, 2, 2, 3), (4, 4, 4, 5).$$

The interquartile range is:

$$\text{IQR} = \text{Median of sub set 2} - \text{Median of sub set 1},$$

$$\text{IQR} = \frac{4+4}{2} - \frac{2+2}{2} = \frac{8}{2} - \frac{4}{2} = 4 - 2 = 2.$$

Example

Find the interquartile range of the list of numbers below:

$$2, 3, 2, 4, 3, 5, 4, 4, 2.$$

Sort in numeric order as before:

$$2, 2, 2, 3, 3, 4, 4, 4, 5.$$

Split the data as before but append each subset of data with the median (at the end and start of each subset respectively):

$$(2, 2, 2, 3, 3), (3, 4, 4, 4, 5).$$

Find the medians of each of the subsets and calculate the interquartile range. The median of the first subset is 2 and the median of the second subset is 4.

$$\text{IQR} = 4 - 2 = 2$$

The interquartile range is a useful measure of variability for skewed distributions. It can show where most values lie and how clustered they are. It is useful for datasets with outliers as it is based on the middle half of the distribution and less influenced by extreme values. Exclusive calculations result in a wider interquartile range than inclusive calculations.

6.3.3 Variance

The standard deviation describes to what extent a set of numbers lie apart (their spread). It is the square root of variance which is also an indicator of the spread of values.

To calculate the variance:

1. Start by finding the mean of the values in the dataset.
2. Find the difference between each recorded value and the mean.
3. Square those differences.
4. Sum the squared differences.
5. Divide the sum by the number of values recorded for population variance or the sum of the number of values minus 1 for sample variance.

Information

The population variance is given by:

$$V_p = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

where V_p is the population variance, n is the number of observations, x_i are the observations and μ is the population mean.

The sample variance is given by:

$$V_s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where V_s is the population variance, n is the number of observations, x_i are the observations and $\bar{x}$ is the sample mean.

6.3.4 Standard Deviation

Taking square root of the variance corrects for the fact that all the differences were squared, resulting in the standard deviation. It is the square root of the variance which is also an indicator of spread.

A standard deviation can range from 0 to infinity. A standard deviation of 0 means that a list of numbers are all equal and they don't lie apart at all.

To make sense of this through an example, the plot below shows some simulated data for test scores. Three groups given the same test could achieve the same average score but with different spreads of scores.

For the group with a mean test score of 30 and a standard deviation of 5, most of the test scores are tightly packed within the range 25-35.

Information

In statistics there is a rule called the empirical rule that states that 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively (Lee D. K., In J. and Lee S., 2015).

For a mean of 30 and standard deviation of 5: 68% of the values will lie within the range 25-35.

For a mean of 30 and standard deviation of 10: 68% of the values will lie within the range 20-40.

For a mean of 30 and standard deviation of 15: 68% of the values will lie within the range 15-45.

Statisticians will sometimes use a z-score to indicate how far from the mean a particular element in the data set is.

6.3.4.1 Example

Calculate the sample estimate of variance and sample estimate of standard deviation for the following list of values:

$$2, 4, 4, 5, 6.$$

Start by finding the mean of the values in the dataset:

$$\text{Mean} = \frac{2+4+4+5+6}{5} = 4.2.$$

Find the difference between each recorded value and the mean.

Table 6.13:

Value	Difference
2	2 - 4.2 = -2.2
4	4 - 4.2 = -0.2
4	4 - 4.2 = -0.2
5	5 - 4.2 = 0.8

Value	Difference
6	6 - 4.2 = 1.8

Square the differences.

Table 6.14:

Value	Difference	Squared Difference
2	-2.2	4.84
4	-0.2	0.04
4	-0.2	0.04
5	0.8	0.64
6	1.8	3.24

Sum the squared differences.

$$\text{Sum} = 4.84 + 0.04 + 0.04 + 0.64 + 3.24 = 8.8.$$

Divide the sum by the number of values recorded minus one to get the sample estimate of variance.

$$\text{Variance}_s = \frac{8.8}{5-1} = 2.2.$$

To get the sample estimate of the standard deviation take the square root of this value:

$$\text{Standard Deviation}_s = \sqrt{\text{Variance}_s} = \sqrt{2.2} = 1.48.$$

6.3.5 Using Excel

Calculating the variance and standard deviation by hand is a long process and due to the number of steps involved it is prone to error. Excel, SPSS, Python and R all have functions which allow users to calculate these descriptive statistics and their use is highly recommended over calculating the statistics by hand.

Range

=MAX(start:end)-MIN(start:end)

There is no single function for calculating the range in Excel but the formula above will subtract the smallest value from the largest value in an array.

Standard Deviation

=STDEV.S(start:end)

=STDEV.P(start:end)

stdev.s() estimates standard deviation based on a sample. stdev.p() calculates standard deviation based on the entire population given as arguments.

Variance

=VAR.S(start:end)

=VAR.P(start:end)

var.s() estimates variance based on a sample. var.p() calculates variance based on the entire population given as arguments.

Summary

Measures of Central Tendency

The mean (sometimes referred to as the arithmetic mean) is the sum of the recorded values divided by the number of values recorded. The formula for the mean then is given by:

$$\text{Mean} = \frac{\text{Sum of recorded values}}{\text{Number of values recorded}}.$$

The median is the middle number in a sorted, ascending or descending, list of values.

If there are an odd number of values the median is simply the middle value.

For an even number of values there will be two values in the center. Those values are summed and divided by two.

The mode of a set of data values is the value that appears most often.

Frequency

The frequency of an observation is the number of times it occurs or is recorded. A frequency table is a commonly used method of depicting frequency.

The total of all frequencies so far in a frequency distribution is the cumulative frequency. It is the 'running total' of frequencies.

The relative frequency is the ratio of the category frequency to the total number of outcomes.

Measures of Dispersion

The range is the difference between the highest and lowest values.

The interquartile range (IQR) describes the spread of the middle half of a distribution. How the interquartile range is calculated depends on whether there are an even or an odd number of values in a dataset.

The standard deviation describes to what extent a set of numbers lie apart (their spread). It is the square root of variance which is also an indicator of the spread of values.

To calculate the variance:

1. Start by finding the mean of the values in the dataset.
2. Find the difference between each recorded value and the mean.
3. Square those differences.
4. Sum the squared differences.
5. Divide the sum by the number of values recorded for population variance or the sum of the number of values minus 1 for sample variance.

Chapter 7

Descriptive Statistics Quiz

If you would like to try and test your knowledge of the various levels of measurement outlined in this chapter you can take the Measures of Central Tendency and Measures of Dispersion Quiz below. This quiz isn't scored or recorded anywhere.

Chapter 8

Comparing Data

Statisticians use several statistical measures like the percentage difference, percentage change and percentage error to evaluate the differences between measured values. All three differ in what they measure.

Percentage difference is the difference between two values divided by the average of two values multiplied by 100%. This is typically used to understand how close two values are to one another.

If the data tracks changes in values over time (comparing old values to new values) then it is nearly always better to calculate the percentage change instead of the percentage difference. This is the key difference between the two.

While percentage change aims to measure change over time, the percentage difference seeks to understand the difference between two values.

8.1 Percentage Difference

The percentage difference is used to compare two values.

$$\text{Percentage Difference} = 100\% \frac{|\text{Value}_1 - \text{Value}_2|}{\frac{\text{Value}_1 + \text{Value}_2}{2}}.$$

The | symbol in the formula below indicates that the absolute value should be taken.

Information

The absolute value of a calculation is the result of the calculation if the numerical answer was always assumed to be positive.

For example:

$$|3-2| = 1,$$

$$|10-5| = 5,$$

$$|7-9| = 2.$$

Seven minus nine is equal to -2 but when we take the 'absolute' value (|7-9|) we ignore the negative sign and report the result as 2.

8.1.1 Example

One researcher produced thirteen research reports in 2022 another produced 11. What is the percentage difference?

$$\text{Percentage Difference} = 100\% \frac{|13-11|}{(\frac{13+11}{2})},$$

$$\text{Percentage Difference} = 100\% \frac{2}{(\frac{24}{2})},$$

$$\text{Percentage Difference} = 100\% \frac{2}{12},$$

$$\text{Percentage Difference} = 100\% \frac{1}{6},$$

$$\text{Percentage Difference} = 16.7\%.$$

8.2 Percentage Change

Percentage change is about comparing old to new values. The formula for calculating a percentage change is given below:

$$\text{Percentage Change} = 100\% \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}}.$$

8.2.1 Example

What is the percentage change in the population in these three places between 2018 and 2019?

Table 8.1:

Location	2018	2019	Percentage Change (%)
Somewhere	50	36	-28
Anywhere	50	50	0
Elsewhere	50	58	16

Use the formula above to calculate the percentage change for Upper Braniel:

$$\text{Percentage Change} = 100\% \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}},$$

$$\text{Percentage Change} = 100\% \frac{36 - 50}{50},$$

$$\text{Percentage Change} = 100\% \frac{(-14)}{50},$$

$$\text{Percentage Change} = -100\% \frac{14}{50},$$

$$\text{Percentage Change} = -100\% \frac{7}{25},$$

$$\text{Percentage Change} = -28\%.$$

A negative percentage change indicates a percentage decrease while a positive percentage change indicates a percentage increase.

8.3 Percentage Point Change

Note that subtracting one percentage from another gives the percentage point change rather than the percentage change.

8.3.1 Example

What is the percentage point change in the population in these locations between 2018 and 2019?

Table 8.2:

Location	2018	2019	Percentage Change (%)
Somewhere	3.5	2.5	-1
Anywhere	3.4	3.3	-0.1
Elsewhere	3.3	3.8	0.5

8.4 Quantiles

Quantiles are subsets of a larger data set that has been split into some number of equal parts. Quartiles and quintiles are commonly used quantiles which divide data in four and five equal parts respectively.

Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. 25% of the measurements or observations in the data set are less than or equal to the first quartiles (Q1); 50% are less than or equal to the second quartile (Q2); and 75% are less than or equal to the third quartile (Q3).

8.5 Deciles and Percentiles

Deciles are another type of quantile that divide the data into 10 equal parts, instead of Q1, Q2 and Q3, we have decile 1 (D1) through to decile 9 (D9) as a result.

Percentiles divide the data into 100 equal parts resulting in percentile 1 (P1) through to percentile 99 (P99).Assume that the elements in a data set are rank ordered from the smallest to the largest. The values that divide a rank-ordered set of elements into 100 equal parts are called percentiles. The observation at the 50th percentile would be denoted would be greater than 50 percent of the observations in the set. An observation at the 50th percentile would correspond to the median value in the set.

Summary

The percentage difference is used to compare two values.

$$\text{Percentage Difference} = 100\% \frac{|Value_1 - Value_2|}{\frac{Value_1 + Value_2}{2}}.$$

The formula for calculating a percentage change is given below:

$$\text{Percentage Change} = 100\% \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}}.$$

Subtracting one percentage from another gives the percentage point change rather than the percentage change.

Chapter 9

Inferential Statistics

9.1 What is Inferential Statistics?

Inferential statistics is mostly used to explain different phenomenon, predict trends or make decisions relating to whether results are statistically significant or not. This is in contrast to descriptive statistics which is restricted to merely describing the important characteristics of data by using measures of central tendency and measures of dispersion.

9.2 Hypothesis Testing

People can form different opinions by looking at data but hypothesis testing provides a consistent framework for making decisions about different assumptions for everyone by providing a statistical using a set of rules rather than relying on subjective impressions (Pereira S. M. C, Leslie G. R. N., 2009).

Hypothesis testing is used to test assumptions relating to a population parameter based on a sample taken from that population. It involves formulating hypotheses, including a null hypothesis and an alternative hypothesis, and collecting data before using statistical methods to determine whether or not the null hypothesis can be rejected.

The null hypothesis describes the assumption that there is no difference between observations while the alternative hypothesis describes the assumption that there is a difference (not due to chance).

The decision to reject the null hypothesis is based on the strength of the evidence provided by the sample data. The strength of the evidence provided by the sample data is described by the p-value which is the probability of observing a test statistic as extreme or more extreme than the one computed from the sample, assuming that the null hypothesis is true.

Researchers will construct hypotheses with the expectation that their findings will contradict the null hypothesis.

When the null hypothesis is rejected it important to avoid stating acceptance of the alternative hypothesis. In general, studies provide evidence for or against a hypothesis rather than conclusively proving one or another to be true.

Similarly, the null hypothesis cannot be rejected it is important not to state that the null hypothesis is accepted instead it should be stated that the null hypothesis cannot be rejected.

This is often compared to how verdicts are made in a court of law (Banerjee A., Chitnis U. B., Jadhav S. L., Bhawalkar J. S., Chaudhury S., 2009). A person can be found guilty or not guilty. A not guilty verdict means that the prosecution was unable not prove beyond a reasonable doubt that the person committed the crime but it doesn't necessarily mean the person is innocent. The court can provide evidence of guilt but it cannot prove innocence.

In the same way, statistical tests cannot prove that either hypothesis is true.

9.3 Statistical Significance

Statistical significance is a concept in hypothesis testing that describes whether the results of a study are meaningful or not. Statistical significance is used to determine if the results are due to random chance or not and is a measure of the probability of the null hypothesis being true (Tenny S. and Abdelgawad I., 2022).

A result is statistically significant if the p-value is less than a pre-defined value (Tenny S. and Abdelgawad I., 2022). Results that are unlikely to have occurred by chance are considered to be statistically significant. The level of statistical significance is typically set by choosing a p-value of 0.05 (Office for National Statistics, 2022). A p-value less than this it means means that there is less than a 5% chance that the results are due to random chance, the null hypothesis is rejected and the result is considered statistically significant.

Statistical significance does not always imply practical significance hover and a result may be statistically significant but may not have a large enough effect to be of any practical use.

9.4 Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors that can occur: Type I errors and Type II errors.

9.4.1 Type I Errors

A Type I error is known as a false positive and occurs when the null hypothesis is rejected despite being true (Banerjee A., Chitnis U. B., Jadhav S. L., Bhawalkar J. S., Chaudhury S., 2009). A significant result is obtained by chance but is interpreted as a real effect. The probability of making a Type I error is represented by alpha, α:

$$\alpha = \text{P(Null hypothesis rejected|Null hypothesis is true).}$$

9.4.2 Type II Errors

A Type II error, known as a false negative, occurs when the null hypothesis is not rejected despite being false (Banerjee A., Chitnis U. B., Jadhav S. L., Bhawalkar J. S., Chaudhury S., 2009). The real effect goes undetected and is interpreted as the result of chance. The probability of making a Type II error is represented by beta, β:

$$\beta = \text{P(Null hypothesis accepted|Null hypothesis is false).}$$

Minimizing the risk of both of these types of errors is important but there is often a trade-off in that increasing the sample size and reducing the significance level can reduce Type I errors but it can also increase Type II errors.

9.4.3 Test Power

Test power is another useful measure in hypothesis testing. It is a measure of the ability of a hypothesis test to detect an effect or a difference if it actually exists. It represents the probability of correctly rejecting the null hypothesis when it is false. The formula for test power is (Swinscow T.D.V, 1997):

$$\text{Power} = 1 - \beta,$$

where β is the probability of a Type II error.

9.5 Parametric and Non-Parametric Testing

Parametric tests and non-parametric tests are statistical tests used to compare groups of data.

They differ in that parametric tests generally make assumptions about the data being normally distributed with equal variances. There are a range of parametric tests including: t-tests, ANOVA, and regression analysis. Parametric tests are typically more powerful but for the tests to be reliable the assumptions must be valid.

Non-parametric tests are often used when data fails to meet the assumptions of parametric tests although non-parametric tests often come with their own sets of assumptions. There are a range of non-parametric tests including: the Wilcoxon rank-sum test, Kruskal-Wallis test, and the Mann-Whitney U test. These tests are generally more flexible but this comes at the expense of being less powerful than parametric tests.

Chapter 10

Correlation

10.1 Pearson Correlation Coefficient

Correlation indicates the strength and direction of a relationship between two variables (Schober P., Christa B., Lothar S. A.,, 2018). Its values range from -1 to 1, where a value of -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally), 1 indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally), and 0 indicates no linear relationship between the two variables (Swinscow, T.D.V, 1997).

The extreme values of -1 and 1 indicate perfect linear relationships and for these relationships all of the data points fall on a single line. In practice it is rare to ever see a correlation this strong. Correlation is measured through the Pearson's Correlation Coefficient which is represented by the Greek letter ρ for a population parameter or the lower case letter r for a sample statistic.

Correlation can be used to summarize the relationship between two variables and can also be used to make predictions about one variable based on the values of the other. Hypothesis tests and confidence intervals can be used to address the statistical significance of the results (Schober P., Christa B., Lothar S. A.,, 2018).

Height and weight are an example of two variables which are correlated. As height increases weight also tends to increase. This means that we can reasonably predict that people who are tall will weigh more than people who are short. Correlation is a quantitative assessment of both the direction and the strength of this tendency for variables to change together.

Values of r which lie between 0 and 0.3 are typically considered to be weak, between 0.3 and 0.5 are considered to be moderate and greater than 0.5 are considered to be strong although there are differences of opinion on this as the scale is somewhat subjective and different fields of study will deem relationships to be strong when other fields would consider them to be moderate or even weak. As a general rule however these guidelines are generally accurate.

10.1.1 Examples

The figure below shows a perfect negative correlation where r=-1. It's clear from the scatter plot that as x increase the values of y decrease.

The figure below illustrates what two variables with no correlation might look like when plotted in a scatter plot. In this case, r = 0. The data has no discernible pattern through which x and y can be related to one another.

10.2 Identifying Correlation

The simplest way to determine whether two variables are correlated is to plot them together. Scatterplots are particularly useful for checking whether there might be some relationship between pairs of continuous data.

The scatterplot below shows some simulated height and weight data with a line of best fit. At a glance it's clear that as height increases weight increases. It's not a perfect relationship however. Looking at various heights you can see there are clusters of weight values associated with them. There are taller people who weigh less than some of their shorter peers and there are shorter people who weigh more than their taller peers.

Pearson's correlation takes all of the data points and represents them with a single summary statistic. In this case the output indicates a correlation of 0.8.

10.3 Calculating the Pearson Correlation Coefficient

Calculating the correlation coefficient is not simple and the best way to calculate it is to use R, Python, Excel, SPSS or some equivalent software. In Excel, calculating the value of r is as simple as typing:

=PEARSON(array1, array2)

This function takes two arguments, each of which should be an array containing the values of the variables for which the Pearson correlation coefficient is to be

calculated. The first array should contain the set of independent variables and the second the set of dependent variables.

The calculations that are performed in the background are described below but it isn't necessary to manually calculate r or its associated test statistics as the process is long and prone to error. The details are provided for context to help understand what goes on in the background.

The formula for calculating the Pearson correlation coefficient is (Swinscow, T.D.V, 1997):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}},$$

where r is the Pearson correlation coefficient, n is the total number of observations and x and y are the recorded data values (in the case of the example above these would be height and weight measurements). This is equivalent to writing:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}.$$

Notice that the denominator in this formula is similar to the formula for variance in that it sums the values of x minus the mean of x and squares the result.

Essentially the denominator in the formula for r describes the product of the two variables' standard deviations and standard deviation is obtained by taking the square root of variance. The numerator in the formula for r is the covariance. Intuitively it makes sense that the spread should matter when trying to determine if two variables are related - when two variables are perfectly correlated they form a straight line in a scatter plot but when they're poorly correlated they appear more like a cloud of points in a scatter plot.

The r value for the height and weight data presented above is approximately 0.8.

The Pearson correlation coefficient can also be used to test whether a relationship between two variables is significant using hypothesis testing where the null hypothesis represents the assumption that the Pearson correlation of the population is equal to zero and the alternative hypothesis represents the assumption that it is not equal to zero.

To test the hypotheses it is necessary to calculate a t-value (Swinscow, T.D.V, 1997):

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

For the data above this gives:

$$t = \frac{0.8}{\sqrt{\frac{1-0.8^2}{80-2}}},$$

$$t = \frac{0.8}{\sqrt{\frac{0.36}{78}}} = 11.7.$$

The next step is to find the critical value of t. The critical value of t can be found in a t-table. To find the critical value the degrees of freedom are needed (df=n-2); the significance level (0.05 by convention) and whether the test is one-tailed or two-tailed (two-tailed is most frequently applied).

If the calculated t value is greater than the critical value, then the relationship is statistically significant and the null hypothesis is rejected. If however the calculated t value is less than the critical value, then the relationship is not statistically significant and the null hypothesis cannot be rejected.

From the t table the critical value of t in this case is 1.990 and the calculated value of t is 11.7 which is much greater than the critical value so the null hypothesis is rejected.

10.4 When to use Pearson Correlation Coefficient

The Pearson correlation coefficient is one of several correlation coefficients to choose from. It is a good choice when both variables being examined are quantitative; when then variables are normally distributed; the data is free from outliers; and the relationship is linear (the relationship between the two variables can be described by a straight line).

Spearman's rank correlation coefficient may be a better choice when the variables are ordinal; aren't normally distributed; the data contains outliers or the relationship between the variables is non-linear and monotonic.

These techniques are also closely related to linear regression however they all serve distinct purposes. A linear regression allows us to fit a line through data points in a scatter plot and estimate the values of the dependent variable from the independent variable but it does not provide any information on how strongly the variables are related. Correlation does not fit the data or allow estimations but it does describe the strength of the relationship (Schober P., Christa B., Lothar S. A.,, 2018). These two analyses often go hand in hand.

10.5 Fitting Data

It is also worth noting that the plots presented above used "linear regression" to establish a line of best fit. Correlation is often used hand in hand with linear regression.

10.5.1 Linear Relationships

Correlation quantifies the strength of a linear relationship between variables whilst linear regression expresses the relationship in the form of an equation which we can use to predict the values of our dependent variable and establish a line of best fit.

When the relationship is not linear we may need to use more complex methods of fitting.

10.5.2 Non-Linear Relationships

A non-linear regression can be used when the relationship between the dependent variable and the independent variable is non-linear (when a straight line will not adequately describe the relationship between the variables). Non-linear regression is generally used when fitting a curve or a function that best describes the relationship between the variables to the data. The plot below shows a fit achieved with non-linear regression and the associated confidence bands. The data is simulated for a fictional political party and represents the percentage change in their share of the vote according to various fictional polls held between the year 2000 and 2005.

The 95% confidence bands were calculated using a technique called "bootstrapping" where the the original data set is resampled with replacement to create a larger number of new data sets allowing us to estimate the uncertainty in the fitted curve and calculate confidence intervals. This is not the only way to calculate confidence intervals and there are pros and cons to different methods.

LOWESS stands for Locally Weighted Scatterplot Smoothing and is another method used to create a smooth curve through a set of data points. It is particularly useful when there is no clear relationship between variables or when the relationship is too complex to be described by a simple linear or non-linear model.

LOWESS works by fitting a regression line to a subset of the data points. This regression line is then shifted along the data points, fitting a new regression line at each point. The final curve is the result of a weighted combination of all the regression lines.

The plot below illustrates and LOWESS fit for simulated data representing the percentage change in a fictional party's share of the vote according to fictional polls held between the year 2000 and 2005.

The faint blue lines show the outcome of the resampling that occurs when bootstrapping.

10.6 Correlation vs. Causation

The expression that "correlation does not imply causation" is fairly well known and it is a warning against interpreting a strong correlation to imply that change in one variable directly causes the change in another. This is rarely the case. Ice cream sales and drownings are positively correlated but it is unlikely that ice cream sales are somehow driving people to drown or vice versa. Instead, it is temperature that causes a change in the other two variables. Higher temperatures cause higher ice cream sales and drives a greater number of swimmers into the sea and swimming pools (where they will be at greater risk of drowning).

Chapter 11

Chi Square Tests

11.1 Description

Chi-square tests are a set of several non-parametric hypothesis tests used for categorical variables with nominal or ordinal measurement scale. There are three tests, the chi square test of goodness of fit; the chi square test of independence; and the chi square test of homogeneity.

The goodness-of-fit test is used to determine whether sample data matches a given distribution.

The test of independence is used to determine whether there is a relationship between two categorical variables in a contingency table.

The test of homogeneity is used to determine whether two or more categorical variables with unknown distributions have the same distribution as each other in several populations.

11.2 Assumptions

Non-parametric tests and parametric tests come with assumptions that ought to hold true if the test is to be performed. Both assume the data was obtained through random selection, however it is not uncommon to find inferential statistics used when data has been collected from convenience samples. Typically, multiple replication studies are performed to ensure confidence in results when this assumption is violated.

The assumptions of the chi square test are (McHugh M. L., 2013):

- The data in the table must be counts or frequency data, not percentages or transformed data.

- The categories of the variables should not overlap (e.g., male or female).
- The chi-square test should not be used if the same subjects are tested at different time points.
- The groups being compared must be independent and a different test should be used if they are related (e.g. paired samples).
- The variables being compared must be categorized at the nominal level, although data that has been transformed from ordinal, interval or ratio data may also be used.
- At least 80% of the cells should have expected values greater than or equal to five. When this occurs, the chi-square test with Yates' continuity correction (if the total sample size is greater than 20) or Fisher's exact test should be used (Gonzalez-Chica D. A., Bastos J. L., Duquia R. P., Bonamigo R. R., Martínez-Mesa J., 2015).
- No cell should have an expected value of zero.

11.3 Chi Square Goodness of Fit (Distribution Test)

The Chi-square Goodness-of-fit test, checks whether the frequencies of the individual characteristic values in the sample correspond to the frequencies of a defined distribution. In most cases, this defined distribution is that of the population.

In the general case the null and alternative hypotheses being tested are:

H_0: The population fits the given distribution.

H_a: The population does not fit the given distribution.

The chi-square value is calculated via:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k},$$

where O is the observed frequency and E is the expected frequency of the population.

The expected value is given by:

$$E = np_i,$$

where n is the total sample size and p_i is the hypothesised population proportion of the ith group.

The degrees of freedom also need to be calculated:

$$df = (p - 1),$$

where p is the number of categories.

A χ^2 table is used to compare the calculated value of χ^2 with critical values listed that correspond to the degrees of freedom and desired significance level.

Chi Square Distribution Table

df	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57

Figure 11.1: Table containing the critical values of the chi-square distribution.

You can also find these values using Excel with the function:

=CHISQ.INV.RT(p,df)

where p is the probability and df is the number of degrees of freedom.

11.3.1 Example

Rock-scissors-paper is a game of chance in which any player should be expected to win, tie and lose with equal frequency. Over the course of 30 games a player could be expected to win approximately ten times, lose ten times and tie times. Imagine a random sample of 24 games:

Table 11.1: Chi Square Goodness of Fit

Outcome	Frequency (Observed)
Win	4
Loss	13

Outcome	Frequency (Observed)
Tie	7

It doesn't seem like the outcomes occur with equal probability. It seems that the player is losing much more frequently than they win or tie.

A χ^2 goodness of fit test could be used to determine if the distribution of these outcomes disagrees with an even distribution. This is essentially a hypothesis test:

H_0: All the outcomes have equal probability.

H_a: The outcomes do not have equal probability.

Since it's expected that the player should win, lose and tie with equal frequency the expected outcomes are 8 wins, 8 losses and 8 ties.

Table 11.2: Chi Square Goodness of Fit

Outcome	Frequency (Observed)	Frequency (Expected)
Win	4	8
Loss	13	8
Tie	7	8

Calculating χ^2 gives:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k},$$

$$\chi^2 = \frac{(4-8)^2}{8} + \frac{(13-8)^2}{8} + \frac{(7-8)^2}{8},$$

$$\chi^2 = \frac{(-4)^2}{8} + \frac{(5)^2}{8} + \frac{(-1)^2}{8},$$

$$\chi^2 = \frac{16}{8} + \frac{25}{8} + \frac{1}{8},$$

$$\chi^2 = \frac{42}{8},$$

$$\chi^2 = 5.25.$$

After calculating chi^2 the number of degrees of freedom, df, is needed. This is given by:

$$df = (p - 1),$$

where p is the number of categories.

There are three categories (win, lose or tie) so the degrees of freedom, df, is given by:

$$df = (3 - 1) = 2.$$

A χ^2 table is used to compare the calculated value of χ^2 with critical values that correspond to three degrees of freedom and a desired significance level.

For a df of 3 and a desired significance of 0.05 the calculated χ^2 value would need to be greater than 7.81 to reject the null hypothesis. In this case, $\chi^2 = 5.25$ which is not greater than 7.81 so the null hypothesis is not rejected.

11.4 Chi-Square Test of Independence

The chi-square test of independence is used when two categorical variables are to be tested for independence. The aim is to analyze whether the characteristic values of the first variable are influenced by the characteristic values of the second variable and vice versa.

In order to calculate the chi-square, an observed and an expected frequency must be given. In the independence test, the expected frequency is the one that results when both variables are independent. If two variables are independent, the expected frequencies of the individual cells are obtained with:

$$\text{Expected Value} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}.$$

Note that this differs from the goodness of fit test.

The calculation of degrees of freedom also differs and is given by:

$$df = (p - 1)(q - 1),$$

where p is the number of rows and q is the number of columns.

11.4.1 Example

The example below calculates chi square for some simulated data.

Table 11.3: Educational Attainment: Observed Values

Education	Male (Observed)	Female (Observed)	Total
GCSE	15	13	28
A Level	5	3	8
Total	20	16	36

Expected values are calculated using:

$$\text{Expected Value} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}.$$

The results are shown below:

Table 11.4: Educational Attainment: Expected Values

Education	Male (Expected)	Female (Expected)
GCSE	16	12
A Level	4	4

Now calculate χ^2:

$$\chi^2 = (15-16)^2/16 + (13-12)^2/12 + (5-4)^2/4 + (3-4)^2/4 = 0.645.$$

After calculating χ^2 the degrees of freedom, df, is needed. This is given by:

$$df = (p-1)(q-1) = (2-1)(2-1) = 1.$$

For a significance level of 5% a χ^2 value greater than 3.84 is needed. Since the calculated χ^2 value is larger, there is a significant difference and the null hypothesis is rejected. A p value close to zero means that our variables are very unlikely to be completely unassociated in some population. However, this does not mean the variables are strongly associated; a weak association in a large sample size may also result in $p<0.05$.

11.5 Strength of Association

A few different statistics can be used for calculating effect size when a Chi squared test has been conducted, depending on the number of categories for each variable, the type of variables (nominal or ordinal) and the nature of the study.

11.5.1 Phi

In the case where there are only two categories for each variable, then effect size can be measured using phi (ϕ). Phi is a measure for the strength of an association between two categorical variables in a 2×2 contingency table. It is calculated by taking the chi-square value, dividing it by the sample size, and then taking the square root of this value (Akoglu, H., 2018).

This is calculated using the Chi-square value (χ^2) and the sample size (n), as follows:

$$\phi = \sqrt{\frac{\chi^2}{n}}.$$

In our case:

$$\phi = \sqrt{\frac{0.3125}{36}} = 0.13.$$

This is considered a small effect (0.1 is considered a small effect size, 0.3 medium and 0.5 and above large).

Note also that an extension of for nominal variables with more than two categories is Cramer's V, while other measures of effect size include relative risk and odds ratio.

11.5.2 Cramer's V

Cramer's V is an alternative to phi in tables bigger than 2×2 tabules (Akoglu, H., 2018).

$$V = \sqrt{\frac{\chi^2}{n * min(r-1, c-1)}},$$

where n in the sample size and $min()$ is the minimum of the two arguments $r-1$ and $c-1$. A Cramer's V value of 0-0.29 indicates a weak association; 0.3-0.59 indicates a moderate association and a Cramer's V of 0.6-1 indicates a strong association. The scale goes from complete independence to perfect association.

Association should never be greater than 1. A relatively weak correlation is all that can be expected when a phenomena is only partially dependent on the independent variable.

An alternative association measure for two nominal variables is the contingency coefficient. However, it's better avoided since its maximum value depends on the dimensions of the contingency table involved.

For two ordinal variables, a Spearman correlation or Kendall's tau are preferable over Cramér's V. For two metric (interval or ratio) variables, a Pearson correlation is the preferred measure.

11.6 Relative Risk

The relative risk (RR) and the odds ratio (OR) are widely used in medical research and are very commonly used measures of association in epidemiology (Schmidt C. O. and Kohlmann T., 2008). Relative risk is a comparison of the chance of an event happening in one group to the chance of it happening in another. For instance, the relative risk of lung cancer in smokers compared to non-smokers is the chance of getting lung cancer for smokers divided by the chance for non-smokers (Tenny S. and Hoffman M. R., 2022). Relative risk only shows the difference in likelihood between the two groups, it does not show the actual risk of the event happening (Tenny S. and Hoffman M. R., 2022).

The Relative Risk is used when comparing the probability of an event occurring to all possible events considered in a study.

For example consider the risk of developing cancer in those exposed and unexposed to second hand smoke. On a study's conclusion we might have a table like the one below:

Table 11.5:

Status	Disease	No Disease
Exposed	A	B
Unexposed	C	D

To calculate the relative risk associated with an exposure we must compare the risk (incidence) among the exposed to those not exposed.

$$\text{Ratio of Risks} = \frac{\text{Disease Risk (incidence) in exposed (A/(A+B))}}{\text{Disease Risk (incidence) in Non-Exposed (C/(C+D))}} = \text{Relative Risk.}$$

Let's add values to do the calculations:

Table 11.6:

Status	Disease	No Disease
Exposed	A	B
Unexposed	C	D

$$\text{Disease Risk (incidence) in exposed} = \frac{366}{366+32} = 0.92,$$

$$\text{Disease Risk (incidence) in unexposed} = \frac{64}{64+319} = 0.17,$$

$$\text{Relative Risk} = \frac{0.92}{0.17} = 5.41.$$

RR=1, the incidence in the exposed is the same as the incidence in the non-exposed. No increased risk, no association.

RR>1, the incidence in the exposed is greater than the incidence in the non-exposed. Increased risk, positive association.

RR<1, the incidence in the exposed is lower than the incidence in the non exposed. Decreased risk, negative association.

The further the RR is from 1 the stronger the association.

The relative risk will be reported alongside a p value or a 95% confidence interval. If the p value is not less than 0.05 or if the confidence interval includes 1 then the RR is not statistically significant.

Information

You cannot calculate relative risk in a case control study. For instance, a study might involve a control group of 100 people without cancer and a group of 100 people with cancer. The disease rate might be 50% because of how the study was designed.

11.7 Odds Ratio

The odds ratio is used in cohort or case-control studies.

Information

Odds ratios consistently overestimate risk.

Odds are not the same as probability:

$$\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}}.$$

For instance, a 60% probability to win gives 1.5 odds to win.

Given some data:

Table 11.7:

Status	Disease (case)	Disease (Control
Exposed	A	B
Unexposed	C	D

The Odds ratio is given by:

$$\text{Odds Ratio} = \frac{\text{odds that a case was exposed (A/C)}}{\text{odds that a control was exposed (B/D)}}.$$

Here's an example with real data:

Table 11.8:

Exposure: Parental smoking in pregnancy	Disease (Cancer)	No Disease (No Cancer)
Yes: Smoking	87	147
No: Non-Smoking	201	508

The odds ratio is:

$$OR = \frac{(\frac{87}{201})}{(\frac{147}{508})} = 1.48.$$

This is different from the relative risk equation.

OR = 1, exposure is not associated with the disease.

OR > 1, exposure is positively associated with the disease.

OR < 1, exposure is negatively associated with the disease.

The further the OR is from 1, the stronger the association.

11.8 Chi Square Homogeneity Test

The Chi-square homogeneity test can be used to check whether two or more samples come from the same population. One question could be whether the subscription frequency of three fictional video streaming services Netflocks, Amazing, Fizzney and Tantamount differ in different age groups. As a fictitious example, a survey is made in three age groups with the following result:

Table 11.9:

Network	15-25	25-35	35-45
Netflocks	25	23	20
Amazing	29	30	33
Fizzney	11	13	12
Tantamount	16	24	26

As with the chi-square independence test, this result is compared with the table that would result if the distributions of Streaming providers were independent of age.

The process is the same as in the Independence test where the expected values are calculated.

Summary

Goodness-of-fit

Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution.

H_0: The population fits the given distribution.

H_a: The population does not fit the given distribution.

Independence

Use the test for independence to decide whether two variables (factors) are independent or dependent.

H_0: The two variables (factors) are independent.

H_a: The two variables (factors) are dependent.

Homogeneity

Use the test for homogeneity to decide if two or more populations with unknown distributions have the same distribution as each other.

H_0: The two populations follow the same distribution.

H_a: The two populations have different distributions.

Chapter 12

Data Visualisation

12.1 Data Visualisation

Data visualisation is formally defined as the encoding of data using visual cues such as variations in the size, shape and colour of geometric objects (points, lines, bars). The encoding is generally informed by the relationships within the data.

The bar chart below shows the marital status of people in Northern Ireland based on the Census 2011 data (Census, 2011b). The frequencies of different marital statuses have been mapped to the heights of the bars.

12.2 Visual Cues

Whether data is visualised using points, lines, bars or something else entirely is largely determined by the relationships within the data. Some of the visual cues and relationships used to inform data visualisation are shown below.

The illustration above shows some of the visual cues used to encode data. Magnitudes are typically mapped to sizes of objects. Colour is often used to represent quantities or highlight data. Shapes can be used to represent qualitative data.

12.3 Relationships in Data

The Government Statistical Service has produced guidance on the relationships in data and how they inform chart choices. The guidance can be useful and

some of the key points are summarised below.

12.3.1 Frequency Distributions

Histograms and bar charts are useful for showing category frequencies. Population by age band for instance could be visualised using a histogram or bar chart. A boxplot can also be useful in visualising additional descriptive statistics such as the mean, median, quartiles, outliers and the range.

The figure below shows the age distributions of GPs in Northern Ireland as of 2020 (Family Practicioner Services, 2020).

12.3.2 Time Series

A line chart is often used to demonstrate the trend of a variable over some time period. For instance, temperature over time can be visualised with a line chart.

The line chart below shows simulated temperature data for Northern Ireland.

12.3.3 Rankings

Data that is ranked usually consists of categories presented in ascending or descending order. A bar chart may be used to show the comparisons between the different categories. Sometimes, change in ranking over time is shown through slope charts but usually only when comparing a start date and an end date without consideration for the time period in between.

The slope chart below shows the change in the percentage of Health Survey respondents reporting a longstanding illness between 2010 and 2020 (Health & Social Care Trust, 2020).

12.3.4 Deviation

Deviation from a reference value can be shown through bar charts.

12.3.5 Correlation

Correlation is usually visualised using scatterplots. Scatterplots are a good way to show comparisons between observations of two variables to determine if there is some correlation because it quickly becomes apparent if there is correlation between the variables or not.

The scatterplot below shows simulated height and weight data.

12.3.6 Magnitude

Comparing differences in the magnitudes of values often relies on bar charts. Comparing the total number of research papers by journal for insance.

12.3.7 Spatial

Cartograms and heat mapping are common ways to show differences between geographical regions.

12.4 Why Visualise Data?

In general, people are better at recognising differences in shapes, colours and sizes than they are at identifying the number of times a value occurs or the differences between values in a large excel spread sheet. For this reason data visualisation can be used to find errors in data quickly. It's much easier to recognise an anomalous value on a bar chart than in an Excel spread sheet. Data visualisation can also be used to see patterns that are difficult to determine by looking at raw data. Data visualisation can also be used to:

- Answer research questions.
- Discover new research questions.
- Explain complex relationships in data visually.
- Aid in decision making.
- Engage and inform.

12.5 Data Visualisation Tools

New programming languages and software products have made data analysis and visualisation vastly more accessible. In addition, many of these facilitate dynamic or interactable visualisations. There is an ever expanding ecosystem of data visualisation tools (many of which have been used in this document) including:

- **Excel** and **SPSS** produce high quality visualisations and while dynamic visuals are not their focus they are often the simplest and most time efficient option for visualising data.
- **Genially** is an online tool for creating interactive and animated content that is particularly effective for presentations.
- **Tableau** and **Power BI** are visual analytics platforms which are well suited to the development of dashboards to visualise complex interconnected data sets.

- **Flourish** can be used to produce interactive visuals although its functionality is more limited than Power BI or Tableau. It can be useful for animated visuals however it struggles with larger data sets.
- **Javascript** facilitates data visualisation through its D3 library. D3 has a steep learning curve as it requires JavaScript skills to use it effectively however it offers a greater degree of customisation and a broader spectrum of visualisation options as a result.
- **Python** libraries such as Matplotlib, Seaborn and Plotly can also be used to visualise data. The learning curve is steep as it requires programming skills to use Python effectively however Python offers customisation options that are not available in Excel or Power BI. Python has been used to produce many of the visuals in this e-book.
- **R** is another useful tool with libraries such as ggplot2 which can be used to visualise data. This is the programming language used to write this e-book.

12.6 Dynamic Visualisations (Dashboards)

There are a number of considerations when developing dynamic data visualisations (sometimes called dashboards) as not all data visualisations need to be dynamic.

Considering the audience, objectives and what visuals will be most appropriate to communicate data can help in determining whether a dynamic or interactive visualisation is needed.

Dashboard style visualisations are best suited to data reporting where there is a need to repeatedly produce the same visuals or reports either daily, monthly, quarterly or annually.

Power BI is well designed for these types of visualisation requirements as it offers automation options enabling data sets to be refreshed at regular time intervals. Automation can be as simple as setting a refresh time in the Power BI dashboard and manually updating the excel file it stores in memory or it can be more complex and involve using programming languages to make API calls and perform automated calculations.

Producing dynamic visualisations is often considerably more time expensive than producing static visuals and time constraints should be considered before developing a dashboard visualisation.

12.6.1 Best Practice

GSS have produced guidance on designing dashboards that covers most aspects of dashboard design. The content below summarises some of the key points in this guidance.

Consider Audience and User Needs

Consider the user needs and whether a dashboard is really needed. Often the simplest solution (bar charts drawn in Excel or SPSS) is the best. Consider the visuals used and whether they're the best way to communicate the data. Sometimes tables or even text can communicate data better than a visual.

Guidance

Providing guidance on how to use a dynamic visual or dashboard is important as many users will not be familiar with interactive dashboards. Guidance can be provided through supplementary documentation, blog text if the visual is being embedded, or it can be provided through tool tips and information pages in the dashboard itself.

Streamline Content

When adding any new data or visuals it is important to ask whether it adds value. Try to group related content and streamlining the content to guide the users through the data.

Automate

Automation can be simple or complex, it can be achieved by setting a refresh date in a Power BI dashboard. It can also involve the use of programming languages to make API calls, web scrape data and perform calculations. Automation typically results in less manual updating and a reduced chance of error and can make the management of the product less resource intensive. It's important to note that automation does not necessarily mean less work, the scripts used to automate processes will need to be updated as languages are developed and updated over time.

Consider Design Principles

Give your dashboard a header and dedicated areas for visuals. Consider other dashboards you have seen in the past and draw inspiration from web design. Most websites have a navigation bar at the top, lists with filters along the left or right hand side and content in the center of the page. Think about things like symmetry, flow and a consistent style or layout. Use white space where possible and try to avoid cluttered visualisations.

Ensure Accessibility

Ensure your product is accessible by checking the colour contrast ratios of text and including alt text in your visualisations where possible. Ensure the fonts are large enough to read and avoid using multiple fonts.

Appendices

Excel Functions

Frequency

=FREQUENCY(start:end,bins)

The frequency() function will return a frequency table describing your data. It takes two arguments, the first being the array of values and the second being an array describing the upper boundary of the bins used.

Average

=AVERAGE(start:end)

The mean is calculated using the average() function. There are several other functions relating to means: geomean(), harmean() and trimmean(). Take care not to use these as they are quite different from calculating the mean that has been described here.

Median

=MEDIAN(start:end)

The median is calculated using the median() function.

Mode

=MODE.SNGL(start:end)

=MODE.MULT(start:end)

There are several functions for calculating the mode: mode(), mode.sngl() and mode.mult().

mode() was used in Excel 2007 and may still appear as an option in some versions of Excel.

mode.sngl() will return one mode and mode.mult() will return multiple modes (if there are multiple modes).

Neither mode() nor mode.sngl() will provide a warning if there are multiple modes so mode.mult() is usually the safest option.

Range

=MAX(start:end)-MIN(start:end)

There is no single function for calculating the range in Excel but the formula above will subtract the smallest value from the largest value in an array.

Square Root

=SQRT(number)

Takes the square root of a number.

Percentile

=PERCENTILE.EXC(array,k)

=PERCENTILE.INC(array,k)

These two percentile functions will return the k-th percentile of values in a range where k is in the range 0 to 1 (exclusive and inclusive respectively).

Standard Deviation

=STDEV.S(start:end)

=STDEV.P(start:end)

stdev.s() estimates standard deviation based on a sample. stdev.p() calculates standard deviation based on the entire population given as arguments.

Variance

=VAR.S(start:end)

=VAR.P(start:end)

var.s() estimates variance based on a sample. var.p() calculates variance based on the entire population given as arguments.

Pearson Correlation Coefficient

=PEARSON(array1, array2)

Calculating the correlation coefficient is not simple and the best way to calculate it is to use R, Python, Excel, SPSS or some equivalent software.

Chi-Square Distribution

=CHISQ.INV.RT(p,df)

Returns the critical values of the chi-square distribution where p is the probability and df is the number of degrees of freedom.

Bibliography

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93.

Altman D. G. and Bland J. M. (2005). Standard Deviations and Standard Errors. *British Medical Journal*, 331:903:331.

Andrade C. (2020). The Inconvenient Truth About Convenience and Purposive Samples. *Indian Journal of Psychological Medicine*, 43(1):86–88.

Banerjee A., Chitnis U. B., Jadhav S. L., Bhawalkar J. S., Chaudhury S. (2009). Hypothesis testing, type I and type II errors. *Ind Psychiatry J.*, 18(2):127–31.

Campbell M. J. (2021). *Statistics at Square One*. John Wiley & Sons.

Census (2011a). Highest Level of Qualification of Usual Residents in households aged 16 to 64 (Northern Ireland).

Census (2011b). Marital and Civil Partnership Status (Northern Ireland).

Family Practicioner Services (2020). General Medical Services for Northern Ireland, Annual Statistics 2019/20.

Gonzalez-Chica D. A., Bastos J. L., Duquia R. P., Bonamigo R. R., Martínez-Mesa J. (2015). Tests of association: which one is the most appropriate for my study? *An Bras Dermatol.*, 90(4):523–528.

Health & Social Care Trust (2020). The Health Survey Northern Ireland.

J., F. (2019). *Introduction to Statistics*. Statistics By Jim Publishing.

Lee D. K., In J. and Lee S. (2015). Standard deviation and standard error of the mean. *Korean Journal of Anesthesiology*, 68(3):220–223.

Leung W. C. (2001). Balancing statistical and clinical significance in evaluating treatment effects. *Postgraduate Medical Journal*, 77(905):201–204.

Martin J. D. and Louis N. G. (1997). Measurement of Relative Variation: Sociological Examples. *American Sociological Review*, 33(3):496–502.

McHugh M. L. (2013). The chi-square test of independence. *Biochem Med (Zagreb)*, 23(2):1:43–9.

McHugh M. L. and Hudson-Barr D. (2003). Descriptive Statistics, Part II: Most Commonly Used Descriptive Statistics. *Journal for Specialists in Pediatric Nursing*, 8(3):111–116.

McHugh M. L. and Villarruel A. M. (2003). Descriptive Statistics, Part I: Level of Measurement. *Journal for Specialists in Pediatric Nursing*, 8(1):35–37.

Office for National Statistics (2022). Uncertainty and how we measure it for our surveys. https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/uncertaintyandhowwemeasureit. Accessed: 21 November 2022.

Pereira S. M. C, Leslie G. R. N. (2009). Hypothesis testing. *Australian Critical Care*, 22(4):187–191.

Schmidt C. O. and Kohlmann T. (2008). When to use the odds ratio or the relative risk? *International journal of public health*, 53(2):165.

Schober P., Christa B., Lothar S. A., (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Swinscow, T.D.V (1997). *Statistics at Square One: Correlation and Regression*. BMJ Publishing Group, 9th edition.

Swinscow T.D.V (1997). *Statistics at Square One: Differences between means: type I and type II errors and power*. BMJ Publishing Group, 9th edition.

Tenny S. and Abdelgawad I. (2022). *Statistical Significance.* StatPearls [Internet] StatPearls Publishing, Treasure Island (FL).

Tenny S. and Hoffman M. R. (2022). *Relative Risk*. StatPearls [Internet] StatPearls Publishing, Treasure Island (FL).

Witte R. S. and Witte J. S. (2017). *Statistics*. John Wiley & Sons.