CHAPTER 4:

QUANTITATIVE DATA ANALYSIS

Patrick Callaghan and Penny Bee

CHAPTER OVERVIEW

Quantitative data analysis makes sense of numerical data. We often refer to quantitative data analysis as statistical analysis, and you may see this term used in published research papers. We can use numbers to summarise the experiences or characteristics of a group of participants, for example their average age or the number of symptoms they report. We can also use numbers to look at people's behaviours, experiences and views, for example the number of people using mental health services or the proportion of people who are satisfied with their care. Perhaps most importantly, we can use numbers to look at differences between groups of people or the same group over time. This can help us understand the effect of new treatment or policy initiatives, both in terms of the type of effect (e.g. does a new policy make things better, worse or leave things unchanged?) and the size of its impact (e.g. are any changes big enough to be meaningful or could they have happened just by chance?). For some people, numbers and statistics are reassuring, but for others they can be baffling. In this chapter, we will explore some of the different approaches to analysing numerical data, explore the difference between descriptive and inferential statistics, and highlight some of the ways in which you can begin to interpret research data presented as numbers.

LEARNING OBJECTIVES

By the end of this chapter you should be able to:

1. Describe the difference between descriptive and inferential statistics
2. Understand key concepts in analysing quantitative (numerical) data
3. Discuss whether results are statistically and/or clinically significant

INTRODUCTION

The difference between descriptive and inferential statistics

There are two broad categories of numerical data analyses that researchers are likely to find themselves doing: descriptive and inferential statistics.

Descriptive statistics are used to organise and describe a dataset. Often, descriptive statistics are used to describe the characteristics of a group of research participants (e.g. the range, or most common, score that study participants achieved on a scale measuring anxiety symptoms), but they can also be used to describe other things, such as the characteristics of a health service. A good example here might be the average waiting time for treatment or the rate of staff turnover on a hospital ward. Descriptive statistics are important because they help us to visualise or understand what our data is showing. However, they apply only to the data we have collected and do not allow us to draw any bigger conclusions.

Inferential statistics are used to compare two or more datasets and to explore whether and how they differ. Inferential statistics are used to infer something. They allow us to generalise beyond our own dataset to draw conclusions about a bigger population. They can help us to understand, for example, the impact of a change in health policy or the effect of a new treatment.

Descriptive statistics

Many different statistics can be used to describe a dataset, but the main ones that you are likely to see reported are the median, mode, mean, standard deviation and range. The mean, median and mode are often referred to as measures of central tendency. They give us an indication of the central position in a dataset. Standard deviation and range are measures of spread – they tell us how much the data varies, or spreads out around this point.

Let's use **case example A** from the previous chapter to help us work through these concepts.

Case example A

Paul is interested in looking at how common clinical depression is in people living in Manchester. He sends out a questionnaire to everyone in Manchester measuring their levels of depression and collecting basic demographics such as age, gender, ethnicity, sexuality, accommodation and employment status.

The first 11 people to return their questionnaires report their age (in years) as follows:

22, 49, 33, 41, 87, 18, 33, 54, 40, 33, 72

The mean is the average of all the data. We calculate this by adding up all the values in our list and then dividing by the number of values we have. In our example, the mean is $(18+22+33+33+33+40+41+49+54+72+87)$ divided by $11 = 43.8$. This means the average age for our 11 participants is 43.8 years. Note that the mean does not have to be a number from the original list, and often it isn't.

The median is the middle value of the set, when all values are placed in size order. To find the median in the example above, first write all the numbers in order (e.g. 18, 22, 33, 33, 33, 40, 41, 49, 54, 72, 87). The middle value, the median, is 40.

The mode is the most common value or most frequent response in a set. In our example, the mode is 33.

The standard deviation (SD) indicates how much the values in our dataset are clustered (or not clustered around the mean). The mean score on a depression symptom scale might be 7, but not everyone will score 7. Some will score higher and others will score lower. The standard deviation tells us how spread out people's scores are likely to be. A low standard deviation indicates that most data points tend to be close to the mean, while a high standard deviation indicates that the data points are spread out wider. This helps us to decide how variable people's responses are. Usually, we calculate standard deviation with the help of a calculator or a computer package, such as Excel.

The range refers to the difference between the largest and smallest value for a measure, so in the age example above, the range would be $87 - 18 = 69$ years. Often research reports will just write the range out in full, e.g. 18-87 years.

So what do we use and when?

Although not impossible, it is more unusual to see a research report that includes all of the above statistics. Typically, a published study will report either the mean and the standard deviation (mean, SD), or the median and the range (median, range).

The choice of what to present will partially be based on the type of data we have.

There are four different types of numerical data (**Figure 13**):

Figure 13 Levels of measurement with examples

Nominal data refers to data that can be categorised into groups but which does not have an obvious order. One example of nominal data is gender. Often, we ask research participants to indicate if they are 'male', 'female', 'transgender' or 'without gender'. A good summary measure for these data is the mode, because it will tell us which one of these categories most participants identify with.

Ordinal data is data that can be categorised into groups but can also be ordered in a meaningful way. An example would be responses from a questionnaire where participants have been asked to indicate their level of agreement from 'strongly agree', 'agree', 'neither agree nor disagree', 'disagree' or 'strongly disagree'. We can order these response categories (e.g. from most positive to most negative), but we cannot place a value on them. The best measure for ordinal data is the median, because this will give us the response category which half the sample falls above and half the sample falls below. Presenting the median and range together will tell us the spread of responses as well as the middle response.

Interval data is data that can be ordered, and where the difference between two data points is quantifiable and meaningful. An example is air temperature. The difference between 50 and 60 degrees is the same as the difference between 60 and 70 degrees.

Finally, **ratio data** is the same as interval data, but it also has a meaningful zero score. A good example of ratio data is salary level; if person A earns £20,000 per year and person B earns £10,000, person A earns twice as much as person B. Other variables, like weight, are ratio variables, but temperature is not. You can have zero earnings, but zero degrees Celsius does not mean that a room has no temperature! Interval and ratio data are usually summarised by reporting the mean and standard deviation.

Inferential statistics

Inferential statistics are tests that allow researchers to draw conclusions from their data. Whilst descriptive statistics summarise the characteristics of a group of people or things that have been measured (our study sample), inferential statistics allow us to use these data to estimate characteristics for a bigger population. This means we can draw conclusions beyond the actual group of people that have been measured.

Inferential statistics are important in research because it is never possible to obtain a measurement from everyone – not everyone wants to take part in research and even if they did, the time and cost commitment would be enormous. So instead we work with a manageable group of research volunteers. They provide their data and we use this to estimate the values that would be measured in a population.

To ensure our study group is as similar as possible to the wider population, we select or sample our volunteers carefully. However, we have to accept that there will always be some degree of uncertainty in our estimates (because we cannot measure the whole population) and this means our conclusions are to some extent always going to be our 'best guess.'

How do we know when our guess is good enough?

Deciding how many people to collect data from in a research study is not easy. People (and patients) are individuals, which means that their characteristics, outcomes and responses to treatment are likely to vary. If we only had one or two people in our sample, we would probably not be confident that their data was representative of the whole population. Similarly, if we had everyone in our sample, we could not afford to run our study and we would probably be exhausted. We therefore need to strike a balance.

Research studies, and particularly randomised controlled trials, often use a special calculation called a **power calculation** to decide how many people we need to recruit to a study. A study must have sufficient power to infer the correct result. The minimum power level which is normally accepted is 80%, which means that a study would have an 8 in 10 chance of detecting a relationship or difference between groups (assuming a difference exists). If a study has less than 80% power, then these genuine differences may not be picked up. The big risk here is that a research team would conclude that there was no relationship between variables (e.g. no effect of a new treatment) when in fact there was.

Using inferential statistics to interpret research data: Understanding variables

All research studies examine some kind of characteristic or variable. In research, a variable is not only something that we can measure, but also something that we can manipulate or control for (if we want to do so).

An **independent variable**, sometimes called a predictor variable, is a variable that is being manipulated in an experiment. It is manipulated to observe the effect on a **dependent variable**, which can also be called an outcome variable.

Let's return to an example used in the last chapter (Case example B).

Case example B

Dave is interested in testing the effectiveness of a new self-help book for depression. He includes participants who have clinical depression according to their score on a depression questionnaire that he gets participants to complete at the beginning of the study. He then provides them with the self-help book and asks them to use it for three months. Next he asks them to complete the depression questionnaire again to compare scores before participants used the book to after they had used it. This enables Dave to assess the effect of the self-help book on participant depression symptoms.

Discussion point 1:

Read through case example B above. What do you think were the independent and dependent variables in this study?

The aim of Dave's study is to examine whether the use of a new self-help book result in a change in depression symptoms. So, in our example, the independent variable is the use of the self-help book and the dependent variable is depressive symptoms. The dependent variable does exactly what is says – it is the variable that is dependent on the independent variable.

In experimental research, the aim is to manipulate directly the independent variable(s) and measure the effect on the dependent variable(s). Because a change in the former can be directly linked to a change in the latter, experimental research has the advantage of enabling a researcher to identify cause and effect.

In non-experimental research, the researcher does not manipulate the independent variable(s) themselves. Often this is because it is impractical or unethical to do so. For example, we might be interested in the effect of sudden trauma on people's mental health. It would be unethical to expose study participants to trauma just to study its effects. In this case we might identify a group of adults who have already experienced trauma and compare their questionnaire responses with another group who have not had this experience. Exposure to trauma is still the independent variable and mental health the dependent variable but because we have not directly manipulated the trauma variable, it is not possible to fully establish cause and effect. Instead, we focus on the strength of association or correlation between the two variables.

Using inferential statistics to interpret research data: Understanding statistical significance

Inferential statistics are used in both experimental and non-experimental designs. To make sure we interpret our data correctly, we need to consider:

- whether our results could have occurred by chance
- how meaningful our result is in the real world

Could our results have occurred by chance?

As a general rule, researchers expect to infer one of two things from their studies: either that the independent variable has no effect on the dependent variable, or that the two are related.

The notion that there will be no effect is sometimes referred to as the **null hypothesis**. A null (or zero) hypothesis always states that there will be no relationship between the variables being compared.

In contrast, **the research hypothesis** will usually state that a relationship is expected. Researchers will use their current knowledge and previous work to predict what they think this relationship will be.

**Null
Hypothesis**

**Research
Hypothesis**

Research teams rely on statistical tests to analyse their data, to determine whether the null hypothesis or the research hypothesis is more likely. There are many different types of statistical tests, but all of these provide something called a **p-value**. A p-value is a measure of statistical significance. Put simply, it is the likelihood that any relationship observed between the variables is caused by something other than chance.

As the p-value gets lower (i.e. closer to zero), we are more inclined to accept a research hypothesis and to conclude that there is a relationship between our variables. A cut-off of 5% ($p=0.05$) or 1% ($p=0.01$) is conventionally used to indicate statistical significance. This means that any p-value lower than these values is normally accepted as indicating a significant and 'real' result. P-values above the cut-off (of either 0.05 or 0.01) suggest that there is unlikely to be a significant relationship between our variables, and prevent us from rejecting the null hypothesis.

To put this into context, let's look at the EQUIP trial.

The EQUIP trial aimed to answer the following research question:

Do service users treated by a community mental health team trained in user involvement have different outcomes to those treated by teams who have not been exposed to training?

For this question, the null and research hypotheses would be set up as follows:

- **Null hypothesis:** There will be no differences in the outcomes of service users treated by trained and non-trained community mental health teams.
- **Research hypothesis:** Service users treated by a community mental health team trained in user involvement will have different outcomes to service users treated by a community mental health team that has not been exposed to training.

The research question and hypothesis suggest that service users' outcomes are somehow related to whether the community mental health team that treats them is exposed to training. Outcomes included perceived involvement in the care planning process and satisfaction with care. The EQUIP researchers collected data measuring various service user outcomes from a representative sample of service users receiving treatment from community mental health teams exposed, or otherwise, to training. When the researchers analyse these data, we would like to know if the results can be generalised to all service users, not just those who participated in the EQUIP study. This requires the use of a statistical test.

Imagine the researchers analysed the results from this trial and found that they were statistically significant at $p < 0.05$. Therefore, service users treated by a community mental health team that has been exposed to training will have different outcomes to service users treated by a community mental health team that has not been exposed to training. The null hypothesis (i.e. that there are no statistical differences) can be rejected!

A portion of the results reported by the EQUIP trial are provided below.

Outcome	Time	Usual Care			Intervention			Adjusted mean difference	P-value
		Mean	SD	N	Mean	SD	N		
HCCQ (Support for service-user autonomy)	Baseline	5.06	1.66	272	5.27	1.48	332		
	6 Months	4.93	1.78	227	5.01	1.70	269	-0.064	0.654
HADS-D (Depression)	Baseline	9.18	5.30	272	10.05	5.19	332		
	6 Months	8.90	5.81	172	9.82	5.50	208	-0.020	0.963
VSSS (Service satisfaction)	Baseline	3.58	0.62	272	3.53	0.61	332		
	6 Months	3.53	0.80	156	3.52	0.72	191	0.121	0.045

Discussion point 2:

The EQUIP trial aimed to embed shared decision-making in routine community mental health services by delivering a practical and feasible training intervention to improve service-user and carer involvement in the care planning process. They used the Health Care Climate Questionnaire (HCCQ) to measure the extent to which service-users felt supported in their care decisions. They also measured their satisfaction with community mental health services (measured by the Verona Service Satisfaction Scale, VSSS) and their levels of depression (using the Hospital Anxiety and Depression Scale, HADS-D).

Have a look at the p-values in the results table above. Do you think the results of the EQUIP trial were statistically significant? What conclusions should the team have drawn?

Answer: The results of the EQUIP trial showed that there were no statistically significant differences in HCCQ scores between the intervention and usual care groups six months after training. Depression scores were not significantly different between the two groups. The intervention group reported higher satisfaction with services compared to the control group. This difference was very small and only just statistically significant at the 5% level.

The EQUIP trial team concluded that their training intervention was not sufficient to embed shared decision-making into routine community mental health services. They concluded that successful intervention was likely to require much greater investment of resources. They suggested that the effects of a staff training intervention may take time to emerge and may only become apparent after six months.

Type 1 and type 2 errors

Sometimes, you hear research teams talk about type 1 and type 2 errors.

A type 1 error occurs when you conclude that two variables are related to one another when in fact they are not (a false positive). A type 2 error occurs when a researcher concludes that two variables are not related when in fact they are (a false negative).

Type 1 and type 2 errors can mislead people and might have serious consequences. Let's think for a moment about a randomised controlled trial testing the effects of a new drug. A type 1 error would mean that we concluded the drug was effective in treating a person's symptoms when in fact it wasn't. At the extreme, this could lead to unnecessary cost in producing and distributing the drug, unnecessary use of the drug (possibly with accompanying side-effects) and a failure to properly treat or control the target illness. A type 2 error would mean that the research team would conclude the drug was not effective, when in fact it could have helped a lot of people.

Useful ways of avoiding these errors and reducing threats to the validity of a research study are to:

- have an adequate sample size (to ensure the study has sufficient power)
- recruit a representative sample (to ensure that the sample providing data resembles the wider patient population)
- follow an appropriate research design (to minimise confounding variables)
- use reliable and valid outcome measures (to ensure that any changes are detected and recorded)

Using inferential statistics to interpret research data: Understanding clinical significance

Statistical significance is not always the same as clinical significance. Even if a result is statistically significant we still need to consider clinical significance, because this tells us how important a study's findings are likely to be in the real world.

Clinical significance, sometimes called practical significance, is our 'so what' question. To decide if a study's findings are meaningful in the real world you need to consider carefully the research questions and the measures used to collect the data. For example, if a study shows that depression scores reduce by one point after treatment, is this a useful change? The result might be statistically significant (i.e. have a p-value less than 0.05) but what does a reduction of one point really mean for patients? Could a reduction of this magnitude noticeably enhance a person's mental health, daily functioning or quality of life? Would a greater reduction in scores normally be required before people noticed an effect?

The minimal clinically important difference is the smallest difference or change in outcomes which patients perceive to be beneficial. In the context of a randomised controlled trial examining the effects of a new treatment, it is the smallest difference that would justify changing patient care.

Decisions about clinical significance are not always easy, and can require detailed knowledge of the area. However, there are some key concepts that are usually quoted in research papers that may help you in this respect:

Effect size: The effect size quantifies the difference between two or more groups. In an RCT, it is a measure of the difference in the outcomes between the experimental and control groups. Effect sizes are based on the mean and standard deviation of the outcome scores in each group, and are often standardised so that differences across several different outcome measures with different units can easily be compared. Effect sizes of 0.2 or below are usually considered small, effect sizes over 0.5 are medium and effect sizes over 0.8 are large.

Odds ratio: An odds ratio is a bit like an effect size in that it also allows the outcomes of an experimental and control group to be compared. However, unlike an effect size, it is often used for categorical data and provides a more relative measure of effect. The odds ratio represents the odds that a particular outcome will occur given a particular context or treatment compared to the odds of the same outcome occurring in the absence of that context or treatment.

If the outcome is the same for both groups (i.e. no difference between them) then the odds ratio will be 1. The odds ratio needs to be more than 1 for the intervention to be considered better than the control. If the odds ratio is less than 1 then it means the control group is better than the intervention.

Confidence interval: Confidence intervals are used to indicate the level of uncertainty around an effect reported in a research study. Earlier we discussed how it was practically and financially impossible for a whole population to take part in a research study. Instead we recruit a 'sample' of people, our research volunteers, who provide us with data that we can use to estimate the range of values that we would be likely to see in a whole population.

Estimating means we are unlikely to measure exactly the right result. If we were to run our study lots of times, we might get a slightly different result each time. Confidence intervals help us to decide how much these different estimates are likely to vary. By having an upper and lower confidence limit, we can show the range of values between which we think the true result lies. A narrow confidence interval suggests our estimate is likely to be fairly accurate, as there is little room for it to vary. A much wider confidence interval suggests that our estimate may be less precise.

Statistical analysis plans: What are they and why do researchers have them?

Before conducting any research, a team should agree on what they would like to investigate and how they plan to go about it. Making a detailed plan before you start can help to make sure that you can answer all of your questions by the end of the study, and that you don't make any mistakes along the way.

When it comes to analysing data, we follow the same procedure: before we look at the data we make a detailed plan of everything we would look at and what statistical tests we will use to do this. It is important that this plan, sometimes called a statistical analysis plan (or SAP), is done before we look at the data. This helps to make sure that our decisions aren't influenced by what we can find.

SAPs detail all the different things that will be done to the data in order to get the results to answer our research questions. This usually includes: how we plan to summarise the data, what checks will be carried out to make sure there have been no mistakes when collecting the data, defining all the hypotheses we wish to test and exactly what statistical methods will be used to test these hypotheses. The plan also allows us to describe what might need to be done if parts of the study haven't gone to plan. For example, what will we do if some participants do not take part in the treatment or intervention they have been allocated to? What will we do if some participants do not come for their follow-up meeting and have missing data? What will we do if we look at the characteristics of the intervention and control group and find that one group is much older than the other, or has more women?

Defining all these steps ahead of time means that we can do the statistical analysis quickly at the end of the study. It also gives us a good audit trail. If someone queries what we've done, we can show them the exact steps that we planned to do and the steps that we followed. If they were to carry out the same steps, they should get the same results.

Factors influencing the choice of a statistical test

Various tests are used in research designs, the choice of which is normally made by the team's statistician. The factors influencing the choice of statistical test will depend to a large degree on the specific test; each test has assumptions that should be met before the test is used. General factors that influence choice of statistical test include:

- The research design
- Sample size and sampling method
- Number and nature of the independent and dependent variables
- The spread and pattern of people's responses to an outcome measure

Quantitative data analyses allow researchers to make sense of numerical data gathered from research. Descriptive statistics are used to organise and describe data numerically often through showing measures of central tendency (e.g. mean, median, mode) or data spread (e.g. standard deviation, range). Inferential statistics allow researchers to draw conclusions from their data by testing hypotheses using statistical tests and by calculating estimates of clinical significance such as effect sizes or odds ratios. The SAP provides a clear and transparent roadmap for data analysis, deciding which tests and levels of significance are going to be used in advance. This ensures that there is no bias in the analysis and that the statistical results of a research study can be trusted.

PPI stories from EQUIP

Next Andrew describes his experiences of being involved in quantitative data analysis.

ANDREW'S STORY

We delivered the EQUIP training intervention to 350 professionals from 18 community mental health teams in our trial. Every time we delivered training we used a questionnaire, called the Training Acceptability Rating Scale (or TARS), to measure the acceptability and perceived impact of our work.

We collated all of the questionnaires together and analysed people's responses. We reported the percentage of people who responded and used descriptive statistics to summarise their scores. We calculated the median and range of scores for each question on the TARS.

Our results suggested that our training course was acceptable to health professionals and had had a positive impact on their attitudes, knowledge and skills. It was good to get this feedback from course attendees. I have since co-authored a paper that discusses our findings (Grundy et al, 2017).

REFLECTIVE EXERCISE

- **What is the difference between descriptive and inferential statistics and when might you use each?**
- **What is a power calculation and how might researchers use one?**
- **Define type 1 and type 2 errors.**
- **Consider three factors that might impact on your choice of statistical test.**

REFERENCES AND FURTHER READING

Best, J. (2001) *Damned lies and statistics: untangling numbers from the media, politicians and activists*. Berkeley: University of California Press.

Best, J. (2004) *More damned lies and statistics: how numbers confuse public issues*. Berkeley: University of California Press.

Dancey, C.P., Reidy, J. and Rowe, R. (2012) *Statistics for the health sciences: a non-mathematical introduction*. London: Sage.

Field, A.P. (2013) *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll*. London: Sage.

Grundy, A.C., Bee, P., Meade, O., Callaghan, P., Beatty, S., Ollevent, N. and Lovell, K. (2016) Bringing meaning to user involvement in mental health care planning: a qualitative exploration of service user perspectives. *Journal of Psychiatric and Mental Health Nursing* 23:12-21.

Haslam, S. A. and McGarty, C. (2014) *Research methods and statistics in psychology* (2nd edition.). London: Sage.

Huff, D. (1954) *How to lie with statistics*. New York: Norton & Company.

Kanji, G.K. (2006) *100 statistical tests*. London: Sage.

Levitt, D.S. and Dubner, S.J. (2005) *Freakonomics: a rogue economist explores the hidden side of everything*. New York: William Morrow and Company.

Salsburg, D. (2001) *The lady tasting tea: how statistics revolutionized science in the twentieth century*. New York: W.H. Freeman.