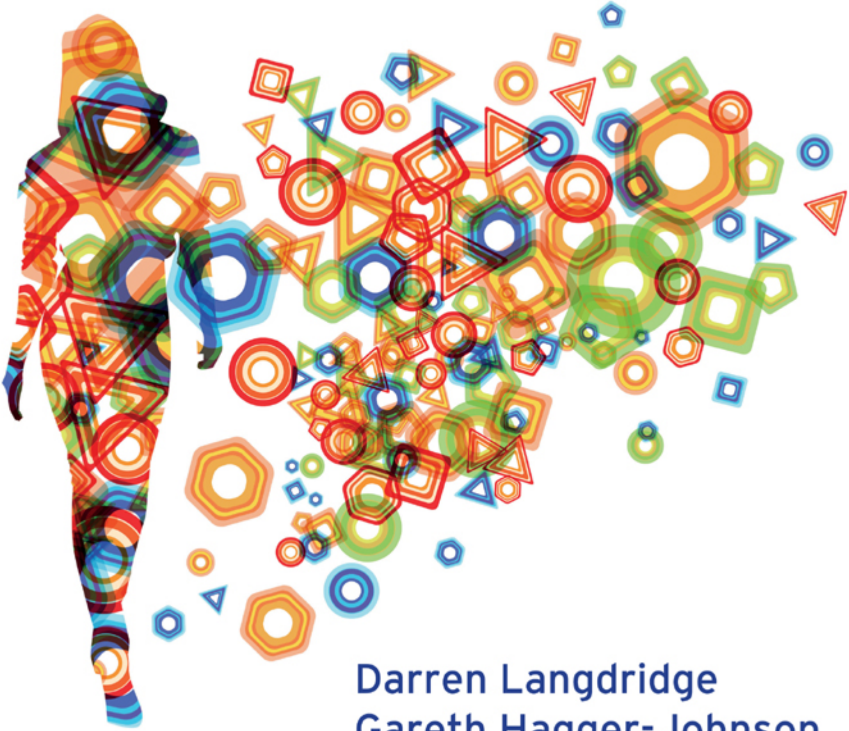
Introduction to Research Methods and Data Analysis in Psychology

Third Edition



Darren Langdridge
Gareth Hagger-Johnson

ALWAYS LEARNING

PEARSON

Introduction to Research Methods and Data Analysis in Psychology

8 Fundamentals of statistics

- This chapter begins by addressing mathematics anxiety and then introduces some of the basic principles of statistics.
- It describes the difference between descriptive and inferential statistics.
- The chapter covers measures of central tendency and dispersion.
- It then moves on to cover probability, the normal distribution and *z*-scores.
- Finally, this chapter will introduce you to Type I and Type II errors, statistical power, and differences between parametric and non-parametric tests.

INTRODUCTION

We think it is probably worth spending a little time addressing any mathematics anxiety that you may be feeling before we move on to explore the topic of statistics in detail. While it is undoubtedly true that the topic of statistics is one of the least popular elements of a psychology degree, it is an essential element and one that is within your capabilities. If you are that rare individual who loves maths then you can probably skip this section and move on to Section 8.1. However, if you have unwelcome memories of learning mathematics at school looming into consciousness as we approach the thorny issue of statistics, then stick with us.

First, some research on mathematics anxiety. There is considerable evidence that mathematics performance is significantly affected by anxiety (Ashcraft & Faust, 1994). Anxiety about your mathematics ability will produce poorer performance at mathematics tasks. And this is one of the major problems for any teacher of statistics. Early negative experiences of mathematics (often at school) can leave a student with a lack of confidence that makes any future learning of statistics more difficult than it need be (Monteil *et al*., 1996). If your confidence can be increased then learning will become easier and your mathematics ability will

▶

improve by leaps and bounds. So, we need to increase your confidence, and this can only happen if you engage with statistics with enthusiasm and an open mind. Belief in your own ability is the key to success!

Now let us explore your experience and feelings about mathematics (and hence statistics) a little more (see Box 8.1). We hope that if we *work together* you will find that statistics really are quite straightforward. We will endeavour to keep our presentation of statistical information clear and simple so that you can try to read and engage with the material. We will try to use real-world examples wherever possible and you will try to apply these statistical ideas to your own experience wherever possible. And finally, we guarantee that if you put in the effort to understand this material you will be able to conquer any fear you may have. You may never find statistics to be your favourite aspect of psychology, but we are sure that you will be able to learn enough to tackle this essential aspect of the discipline with confidence.

Box 8.1

Activity box **Exploring statistics anxiety**

- Spend some time thinking about your experiences of being taught mathematics at school.
  - What was your worst moment?
  - What was your best moment?
  - What do you think was the main difference between the best and worst moments?
- Do you think your memories of being taught mathematics have influenced your feelings about mathematics in general?
  - For instance, do you now approach all maths topics with some fear and trepidation?
  - Do you avoid engaging with mathematics at all (in daily life and in education)?
- Do you have a strategy for dealing with your mathematics anxiety?
  - For instance, is your strategy to avoid maths completely?
  - If you avoid maths, how will that help you when you must complete a course in statistics to gain a degree in psychology?
- Try to think through alternative strategies for dealing with your mathematics anxiety and instigate a plan to carry out an alternative strategy that you feel comfortable with and that will enable you to maximise your success at statistics. Remember, the key to success is commitment to learn. If you put in the effort you will succeed.

8.1 Descriptive and inferential statistics

Quantitative research is designed to generate numbers, and this is the reason we need both **descriptive** and **inferential statistics**. The first thing we need to do with our data is to summarise some basic features of the numbers that make up our set of data. Just re-presenting the numbers generated in our study is rarely enough. We need to do some work on the data so that we can understand what they mean in more detail. **Descriptive statistics** are usually the first stage in any quantitative analysis, and concern the range of techniques that enable us to summarise our data ('descriptive statistics' are sometimes called 'summary statistics' for this reason). The most fundamental and important descriptive statistics are those concerned with identifying the central tendency (or typical value such as the average) of a set of numbers, and those concerned with how the remaining numbers are distributed (how much they vary) around this central (or typical) value. Both of these issues are explored in detail below.

Inferential statistics are those ways of exploring data that enable us to move beyond a simple description of individual variables to making inferences about the *likelihood* of our findings occurring. We are particularly concerned with identifying whether our findings (relationships or differences between one or more variables) are more or less likely to have occurred than by chance alone. Or, what are the *probabilities* (likelihood) that the relationships/differences occurring between variables within our data set are **significant** (and therefore of interest to us in the social sciences), and not just the result of 'meaningless' chance relationships/differences that occur with all variables?

8.2 Measures of central tendency and dispersion

The arithmetic mean

This is certainly the most important measure of central tendency and also the most familiar. In fact the **arithmetic mean** (often shortened simply to 'the mean') is what is called the 'average' in everyday usage. Very simply, the mean is the result of adding up all the numbers in a set and dividing by the total number of values. So if six people took 6, 8, 10, 15, 20 and 7 seconds to solve a simple maths problem (a series of additions and subtractions), we would need to add up all six numbers and divide by six as there are six values in this set of numbers (as below):

$$\frac{(6 + 8 + 10 + 15 + 20 + 7)}{6} = \frac{66}{6} = 11 \text{ seconds}$$

So, the mean of this set of numbers is 11 seconds. This gives us some idea of the central tendency among this set of numbers. The mathematical symbol used to represent the mean is: $\overline{X}$. This symbol is called 'X bar' (the 'X' of this is pretty obvious and ‘bar’ is what we call the line over the top of a symbol). The mathematical formula for calculating the mean is as follows:

$$\overline{X} = \frac{\sum X}{N}$$

Do not panic when you first see mathematical equations, because we will explain what the symbols mean. Here, you have already calculated this equation. The formula says that we should simply (1) add up all the values, and (2) divide by the total number of values in the set of numbers. The sigma symbol ($\sum$) means ‘add up all the values’ and the values are represented by ‘x’. The total number of values is represented by N.

So, in order to calculate $\overline{X}$(the mean), the formula tells us to add up the numbers in our set of numbers (this is the part on top of the line):

$$\sum X$$

The sigma symbol ($\sum$) tells us to add up whatever follows it, and in this case that is the symbol X, which stands for all the numbers in our set of numbers. So, the top half of our formula tells us to add up all the numbers in our set. The bottom half of the formula has the italicised capital N symbol. This symbol represents the number of values in a set of scores (six in this case). And that is all there is to the formula. We replace specific numbers with symbols so that we can provide general instructions for different sets of numbers. You just need to follow the rules of the formula, insert your numbers in the appropriate part of the formula, add, subtract, multiply or divide as appropriate, and that is all there is to it.

We will try to keep formulae to a minimum in this book. These days there is little need for calculating statistics by hand (or with a calculator). Most people who work with statistics (including many psychologists) will have access to computer programs that do the mathematical work for them. However, it can often be useful to *know* how a statistical test works even if you will rarely have to carry it out by hand. There is value in performing a statistical test by hand at least once, in order to become familiar with how it works. Understanding some of the basic principles can be very helpful in getting a good general overview of what we are doing when we carry out statistical tests. Without some level of understanding of the mathematics involved in statistical tests there is a danger of you falling foul of the GIGO fallacy – Garbage In Garbage Out. A computer program will do as you tell it, and that is the case even if you are telling it to carry out tests that are totally inappropriate for your data. Computers only help with the mechanics – you need to take charge of the important task of deciding what test is needed in what circumstance and why, and then carrying out the vital task of interpreting what the results mean.

Advantages	**Disadvantages**
■ Mean is the basis for many powerful statistical tests ■ It is the most sensitive of the measures of central tendency covered here ■ It takes an exactly central position on an interval and continuous scale	■ Its sensitivity can also be a disadvantage as it is susceptible to extreme values (outliers)

We just want to explain why the sensitivity of the mean can be a disadvantage if we have extreme values (or **outliers**) among our data set. Imagine a seventh person completes our mathematics task. This person, like our previous six participants, is an undergraduate student. But this individual was recovering from a very heavy night on the town and was hung over. We did not know this, but it clearly affected their performance on the mathematics task (as they could not concentrate) and they took 186 seconds to complete the task. If we include this score in our set of scores the mean becomes:

$$\frac{(66 + 186)}{7} = \frac{252}{7} = 36 \text{ seconds}$$

Thirty-six seconds is clearly not the best indicator of what is a typical value in our set of scores. After all, six of the scores are well below this value and only one is above it (and then this value is much higher). This is a danger when calculating a mean and a good reason why you should investigate the cause of extreme values in your data set (or use another measure of central tendency such as the median described below) should they exist. Extreme values in one direction (above or below the mean score) will distort the mean (although extreme values simultaneously above and below may cancel each other out) and reduce its utility as a measure of central tendency.

The median

The **median** (**Med**) is another measure of central tendency that does not suffer problems (like the mean) when there are extreme values in a set of data. The median is simply the central value of a set of numbers. With an odd number of values this is very easy. Using the data for the mathematics task above (seven items), the median is the fourth value along once the data have been put in numerical order:

$$6, 7, 8, 10, 15, 20, 186$$

So, in this case the median of our set of numbers is *10*. However, if there is an even number of values (e.g. without the 7th person added), we need to take the mean of the two central values (in this case 8 and 10) in order to calculate the median value for our set of numbers as follows:

$$6, 7, 8, 10, 15, 20$$

$$\text{So, in this case the median is } \frac{8 + 10}{2} = 9$$

The procedure for calculating the median value is as follows:

1 Find the median position or location (where N is the number of scores in our set of data):

$$\frac{N + 1}{2}$$

2 If *N* is odd, then this will be a whole number:

$$\frac{7 + 1}{2} = 4$$

So, we know that in this circumstance (when the data are in numerical order) the median will be the fourth value along.

3 If *N* is even, then the value is midway between two of the values in the set when numerically ordered:

$$\frac{6 + 1}{2} = 3.5$$

This tells us that in this circumstance the median is midway between the third and fourth values in the set of numbers (when they are in numerical order). Here we take the mean of the third and fourth numbers in our set, and this gives us the median.

If there are a large number of values (or there are ties – equal values where the median falls) then the previous approach can be tedious and tricky. There is a formula that can be used instead that does not require us to order our set of numbers to calculate the median. However, we can honestly say that in all the years we have been carrying out the quantitative analysis of data we have never had to use this formula. A couple of clicks of the mouse will instruct a computer program to do this mathematical work for you. Quite frankly we cannot think of any reason for you needing to know about this arcane formula.

Advantages	**Disadvantages**
■ Easier to calculate than the mean (with small groups and no ties) ■ Unaffected by extreme values in one direction, therefore better with skewed data than the mean ■ Can be obtained when extreme values are unknown	■ Does not take into account the exact values of each item and is therefore less sensitive than the mean ■ If values are few, can be unrepresentative

Finally, we want to explain why the median may be unrepresentative when the number of values in a data set are few. Imagine a set of data as follows:

$$6, 9, 11, 233, 234$$

In this case the median is 11, which does not tell us much about the central tendency of our set of data. We need to be careful in situations like these and avoid use of the median if at all possible.