A Really Simple Guide to Quantitative Data Analysis

Why this guide?

This purpose of this guide is to help university students, staff and researchers understand the basic principles of analysing the typical kinds of quantitative data they may collect or encounter in the course of their learning, teaching or research.

What is statistics?

Statistics is an academic subject that involves presenting, interpreting and reasoning about summary quantities derived from data sets. Common statistical quantities are measures of middle values, such as average (also known as mean), mode and median, and measures of spread, such as range and standard deviation.

There are five main sub-areas of this academic subject:

- **Descriptive statistics** (also known as exploratory data analysis) – this does not involve any decision making
- **Data mining** – a systematic approach to looking for relationships in large data sets that were not anticipated in advance. A classic example is Google Flu Trends[1]. Additionally, **data analytics** uses data mining in the context of decision making within an organisation.
- **Time series analysis** – a systematic approach to analysing time-related events which depend on previous events (such as pulse rates or share prices).
- **Statistical testing** (also known as inferential statistics) – this involves reasoning about statistical quantities derived from a sample from a population, where it is assumed that the events are independent, and making decisions with a certain level of confidence.
- **Probability theory** – this provides the theory that underpins the reasoning in statistical analysis and decision making.

Although statistics is a branch of mathematics, much of its reasoning is very different as it is either qualitative or it involves probability-based decisions rather than exact mathematical proof.

The quantitative research process

This guide focuses on descriptive statistics and statistical testing as these are the common forms of quantitative data analysis required at the university and research level. It is assumed that data is being analysed in the context of a research project involving the following stages:

- Define your aim and research questions
- Carry out a literature review
- For primary data research: establish and conceptual framework and use it to design a data collection instrument to collect your primary data.
- For secondary data research: identify a data source and evaluate its validity and reliability
- Process your data set to make it ready for analysis
- Carry out an exploratory data analysis using descriptive statistics and an informal interpretation
- (Optional: carry out an inferential analysis)
- Report on your findings

University and research level data analysis

The experience of statistics at university and in research is often very different from the way statistics is taught at school. School-level statistics education typically involves summary

[1] See Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014) The parable of Google Flu: traps in big data analysis. *Science*, 343(6176), pp. 1203-1205, available at: http://dx.doi.org/10.1126/science.1248506.

information about contrived problems with simple clean data and one right way to carry out an analysis. University and research-level statistics is often **applied**. This means the data sets tend to be large, complex and messy, with some data missing and other data of questionable validity. Rather than there being one right way to analyse such data sets you need to put forward a plan of analysis that is credible but you should be willing to modify your plan as you go along, depending upon what you find, if necessary carrying out an alternative analysis. This requires an additional skill known as **heuristics** or **metacognition**, which means **being in control of the process**.

What is quantitative data?

Essentially, quantitative data is **factual information involving numbers and categories**. Categories often refer to choices between options, such as your favourite type of food or your opinion in a range from strongly disagree to strongly agree. This leader to three fundamental types of data:

- **Numerical data** (this could be whole numbers or decimals)
- Categories with a natural ordering (such as strongly agree, agree, neutral, disagree, strongly disagree) – this is known as **ordinal** data
- Categories without any agreed ordering (such as protein, dairy, carbohydrate, fruit and vegetables) – this is known as **nominal** data

The best kind of quantitative data in statistical analysis is numerical, followed by ordinal, and lastly nominal. It is important to know what kind of data you are planning to collect or analyse as this will affect your analysis method.

A 12 step approach to quantitative data analysis

Step 1: Start with an aim and research questions

Most research starts with these. Vague investigations are dangerous as they are unfocused and may not be undertaken systematically. There is also a greater risk that you will find something that is just a random event.

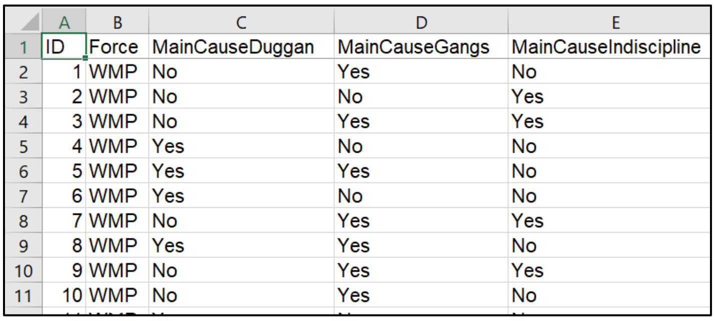
Step 2: Collect data consistent with your aim and research questions

Assuming you have started with a research question you need to think about what data you need to collect in order to investigate this issue. Then there are also the questions: **where** will you get this data, **how** will you approach this process, and **how much** data should you obtain?

- Where – is known as your **sample** – this is assumed to come from a larger **population**
- How – this is your sampling method – is it **random** or **non-random**? Most statistical testing assumes data has been sampled randomly. For a questionnaire you should also consider how to maximize your response rate in order to reduce bias.
- How much – essentially, you should collect **as much data as possible**. It should also be as good quality as possible. There are rules of thumb about what a minimum acceptable amount of data is and a formal process known as sample size calculations. However, both methods suffer from the weakness of not knowing whether there is something that can be found in the first place.

Step 3: Process your data and create a raw data spreadsheet

This step is often overlooked. Data analysis should start with a spreadsheet with types of collected data in the columns and instances in the rows rather than summary statistics derived from raw data. If you are downloading data from an

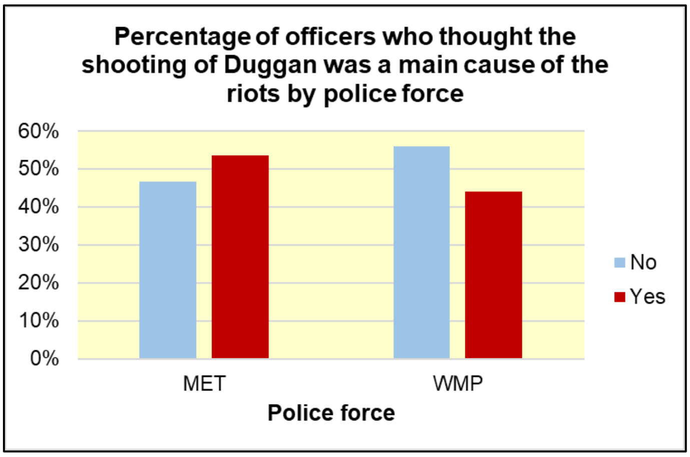


	A	B	C	D	E
1	ID	Force	MainCauseDuggan	MainCauseGangs	MainCauseIndiscipline
2	1	WMP	No	Yes	No
3	2	WMP	No	No	Yes
4	3	WMP	No	Yes	Yes
5	4	WMP	Yes	No	No
6	5	WMP	Yes	Yes	No
7	6	WMP	Yes	No	No
8	7	WMP	No	Yes	Yes
9	8	WMP	Yes	Yes	No
10	9	WMP	No	Yes	Yes
11	10	WMP	No	Yes	No

online questionnaire, it is often quite messy and needs tidying up first.

Step 4: Get a feel for your data with a descriptive analysis

Descriptive analysis involves creating tables, charts and summary statistics from your raw data. This might start with individual types of collected data (known as a **variable**), but it is often more useful to compare one variable against another. Your choice of which variables to compare against each other should be guided by your aim and research questions. Do not do this at random and do not report on everything.



Also, the choice of a table or a chart should be based on what best explains the question being addressed to the reader. Tables can be difficult for your reader to process if they contain too many figures. The shape of the data is often more important that the specific numerical values in interpreting its meaning.

Step 5: Interpret and report on your analysis informally

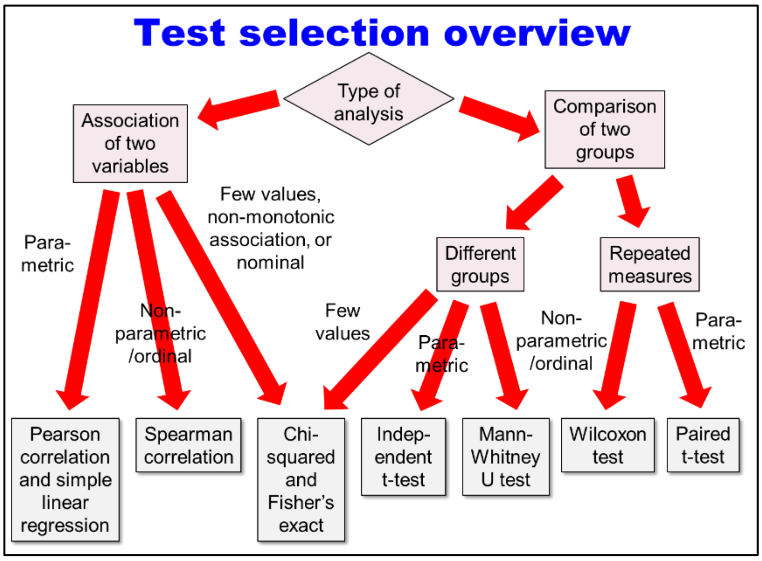
Now you can write a narrative to go with your descriptive statistics. This should seek to answer your research questions by providing an informal interpretation of the meaning of your descriptive statistics. Do not use both a chart and a table to represent the same thing – choose which is best and always write a narrative to go with it. Be careful not to use inappropriate statistical language, such as the word “significant” when you have not conducted any statistical testing.

Descriptive analysis finishes here: the remaining steps relate to statistical testing

Step 6: Decide whether to analyse groups of variables in your data set or just individual variables

Questionnaires, for example, often contain groups of questions on the same thing, known as **scales**. This makes the analysis easier and potentially more accurate as you only have to analyse the values of the scales (which are numeric) rather than the data from individual questions (known as **items**) which make up your scales (which are often ordinal).

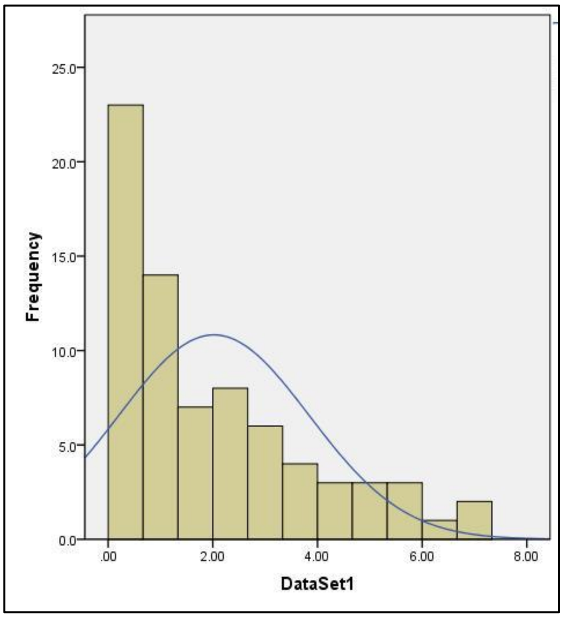
If you choose to use someone else’s questionnaire and wish to use its scales you first need to assess the published literature about it to ensure that its scales are **valid and reliable** (measure what they are supposed to measure accurately). If you have designed your own questionnaire and wish to use the scales you have designed you first need to carry out a **reliability analysis**, but be prepared to remove about half of the items you created. There is also an in between option where you use part of someone else’s questionnaire or modify it, but this is beyond the scope this guide.



There are many myths about reliability analysis. Please see my additional guide for more information.

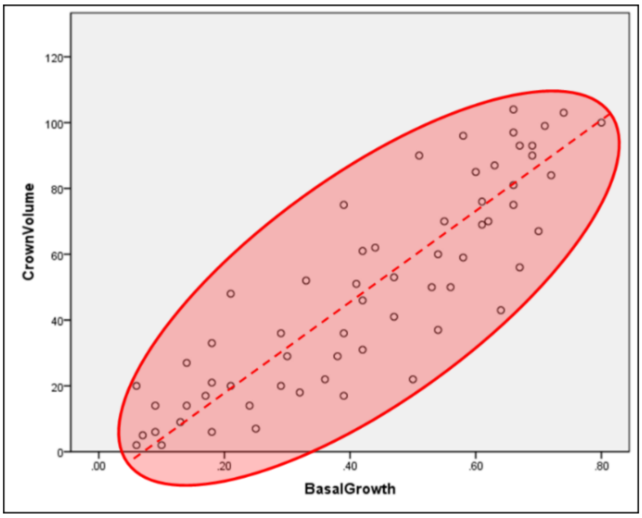
Step 7: Understand your statistical design

There are two main things statistical tests do: investigate differences between groups and explore relationships between variables (known as an **association** or a **correlation**). There is also the issue of whether the same subjects are being measured several times or whether different subjects are being measured. Finally, there are two main kinds of test known as **parametric** and **nonparametric**. Parametric tests are generally more sensitive but they have assumptions that you first need to check before you can run them. The chart above shows a decision tree for choosing simple tests.

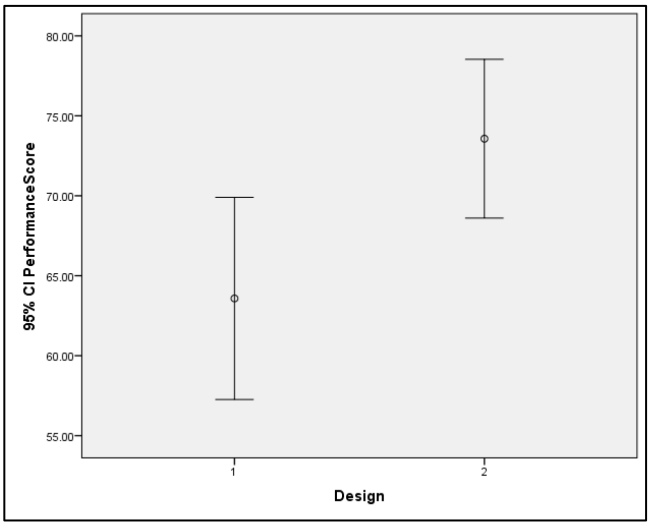


Step 8: Generate advanced level descriptive statistics and check test assumptions

The assumptions of most parametric tests are for the data to be normally distributed. This can be checked by producing a histogram with a fitted normal curve.



There are also tests of normality, such as the Shapiro-Wilk test. Other assumptions are: the equality of variances for an independent samples t-test, which can be assessed using a Levene's test; and an elliptical distribution shape of a scatter plot for linear correlation and regression, which can be assessed qualitatively.



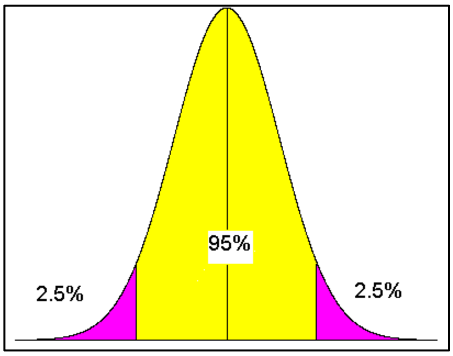
Confidence intervals are a useful advanced descriptive statistic that bridges the gap between exploring data and a statistical test. These are often displayed on an error bar chart.

Step 9: Understand the null hypothesis statistical testing process

Whilst it is often criticised[2], the null hypothesis statistical testing process provides a clear approach to making a decision about a comparison of groups or variables. Imagine that you are a judge in a courtroom and your data is on trial. The assumption that you data is innocent is known as the **null hypothesis**. This usually refers to there being no difference between two groups or no relationship between two variables. Your job is to evaluate to decide whether there is sufficient evidence to convict your data of having a difference or a relationship **beyond reasonable doubt**, or to acquit your data. The beyond reasonable doubt level is usually set at **95% confidence**. The evidence often comes in two forms – a **statistic value** which represents the event that occurred in your sample and an associated probability value (known as a **significance value**) that measures how

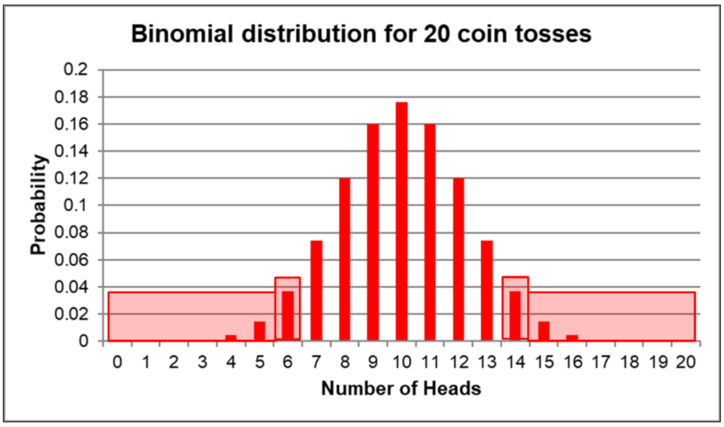
[2] For example, see Häggström, O. (2017) The need for nuance in the null hypothesis significance testing debate. *Educational and Psychological Measurement*, 77(4), pp. 616-630, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5991794/, and Halsey, L. G. (2019) The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15(5), available at: http://dx.doi.org/10.1098/rsbl.2019.0174.

likely or unlikely your event was. If a significance value is **less than 0.05 you reject the null hypothesis**. For example, if you toss a coin 20 times and your throw 6 heads (your statistic value), the probability of this event is about 0.037 but its signifiance value is 0.115 as it is calculated by summing the probabilities of events with less heads (i.e. from 0 to 5 heads) and also the opposite side of the distribution (i.e. from 14 to 20 heads). So getting 6 heads out of 20 coin tosses is not a significant event and you would conclude that there is insufficient evidence to decide that your coin is biased.

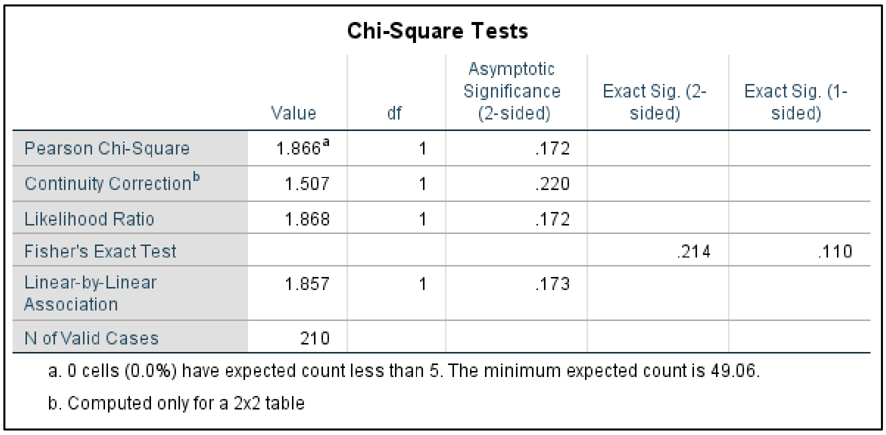


Step 10: Run and interpret an appropriate test

Statistical software such as Excel or SPSS is often used to run statistical tests. The output from these tests requires interpretation.



For example, the table on the right is the output from SPSS for a Chi-squared test of whether there is an association between a cause of rioting and the police force using. The number to interpret is the Asymptotic significance (2-sided) of the Pearson Chi-Square row (0.172). However, the Exact Sig. (2-sided) of the Fisher's exact test (0.214) can also be interpreted. As both of these values is above the 0.05 threshold, we would conclude that there is insufficient evidence of an association. See my additional study guides for more information on running different tests.



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.866[a]	1	.172		
Continuity Correction[b]	1.507	1	.220		
Likelihood Ratio	1.868	1	.172		
Fisher's Exact Test				.214	.110
Linear-by-Linear Association	1.857	1	.173		
N of Valid Cases	210				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 49.06.

b. Computed only for a 2x2 table

Step 11: Report on your results

Results need to be reported on after they are interpreted. This requires quoting relevant probability values, comparing them with the significance threshold in order to make a decision about a null hypothesis and referring this decision back to your research question. It is usually not appropriate to copy and paste software output into your findings but this can be provided in an appendix. You may also need to compare your findings with other people's findings in the literature and discuss any differences or implications.

Step 12: Be prepared to re-analyse your data using metacognition

As already mentioned, in applied statistics, data sets are complex and messy, and there are many ways they could be analysed. In view of this, you should consider whether to run additional analyses to investigate your research questions further. However, be aware that every time you run a statistical test you are introducing the possibility of a false positive result (known as a **Type I error**). If you decide to run several tests, you may wish to increase your confidence threshold, for example from 95% to 99%, and look for correspondingly lower significance values, for example less than 0.01 instead of less than 0.05.