Statistical analysis

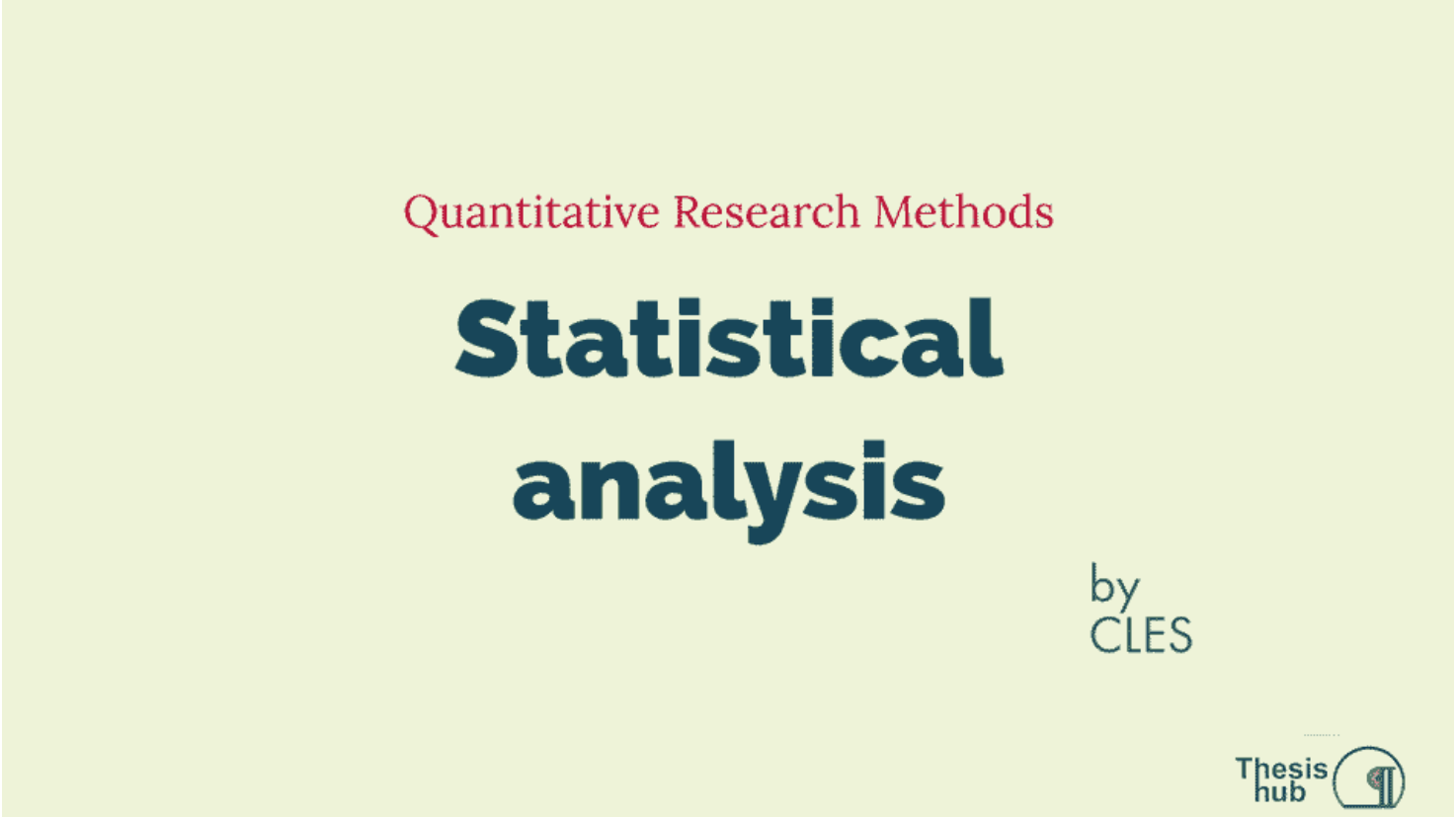

Table of contents

What is the method?

Statistical analysis is a mathematical method of interrogating data.

This is done by looking for relationships between different sets of data. Statistical analysis can be complex, and this following section aims to explain some of the basic considerations, to an audience without an assumed mathematical background. At the end of this section there are a wide variety of links to further reading, which can help you through the process of statistical analysis.

There are two types of statistics:

- Descriptive statistics: numerical summaries of samples (what was observed);
- Inferential statistics: from samples of populations (what could have been or will be observed).

It is important to understand which type of statistics you are working with before embarking on analysis.

When should it be used?

The general idea of statistical analysis is to summarise and analyse data so that it is useful and can inform decision-making. You would analyse descriptive statistics if you wanted to summarise some data into a shorter form, where as, you would use inferential statistical analysis when you were trying to understand a relationship and either generalise or predict based on this understanding. Statistical analysis, through a range of statistical tests, can give

us a way to quantify the confidence we can have in our inferences or conclusions.

Statistical analysis should only be used where there is a clear understanding of the reasons for doing so. The use of statistical tests (as detailed above) will provide you with valuable findings if you know how to interpret the results and use them to inform your research.

What do I need to consider?

Variables

A variable is any measured characteristic or attribute that differs for different subjects. Quantitative variables are measured on an ordinal, interval, or ratio scale, whereas qualitative variables are measured on a nominal scale (note in SPSS the Interval and Ratio levels are grouped together and called scale). There are a range of variables that need to be understood, dependent/independent, controlled/continuous/discrete in the application of statistical tests. The independent variable answers the question “What do I change?”, the dependent variable answers the question “What do I observe?” and the controlled variable answers the question “What do I keep the same?”. A variable which can have any numerical value is called a continuous variable (e.g. time). A variable which can only have whole numbers (integers) is called a discrete variable (e.g. the number of people in a group). It is important to understand the variable you have for analysis of data in statistical packages such as SPSS.

Inference

If working with inferential statistics you need a sound understanding of your population (the set of individuals, items, or data, also called universe) and your sample (a subset of elements taken from a population). See the section on quantitative surveys for further discussion on populations and samples. We make inferences (conclusions) about a population from a sample taken from it, therefore it is important that population and sampling is well understood, as any error will influence your inferences (conclusions). In some situations we can examine the entire population, then there is no inference from a sample.

Confidence & Significance

- **The confidence interval** is an interval estimate of a population parameter, this is the plus-or-minus figure reported in, for example, newspaper or television opinion poll results. If you use a confidence interval of 4 for example, and 54% percent of your sample picks one answer, you can be “sure” that if you had asked the question of the entire relevant population, between 50% and 58% would have picked that answer (plus or minus 4). There are three factors that determine the size of the confidence interval for a given confidence level. These are: sample size, percentage and population size (see below).
- **The confidence level** tells you how sure you can be that this inference is correct. Most social science researchers use the 95% confidence **level**, which means you can be 95% certain; while the 99% confidence level means you can be 99% certain. When you apply the confidence level and the confidence **interval** together, you could say that you are 95% sure that between 50% and 58% would have picked that answer.

In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. In statistics, “significant” means probably true, and not ‘important’. The findings of your research may be proved to be ‘true’ but this does not necessarily mean that the findings are ‘important’. In social science, results with a 95% confidence level are accepted as significant.

Factors that affect the confidence interval

The confidence interval is affected by three factors. These are the sample size, percentage and population size.

Sample Size

The larger your sample, the more confident you can be that their answers truly reflect the population. The relationship between the confidence interval and sample size is not linear. An example can be found below:

	Survey 1	Survey 2
Sample	1,000	2,000
Population	20,000	20,000
% of respondents answering 'yes' to a specific question	50%	50%
Confidence Interval	+/-3.02	+/-2.08

Percentage

The confidence interval is also determined by the percentage of the sample that provides the same answer. The confidence interval increases the closer the percentage is to 50%. In survey 1 (above) the confidence interval for a value of 50% is 3.02. This confidence interval would fall to 0.6 if the survey returned a value of 99% or 1%.

It is important that the survey sample size is considered for statistics where 50% of the population answer both ‘yes’ and ‘no’ as this is when the confidence level is broadest and so provides the general level of accuracy for a sample.

Population Size

The population size refers to the number of people within a group that have a similar characteristic. This could be the total number of people living in a town or the number of people with a more specific attribute such as suffering from a disability or residents from a specific ethnic group. Population size is of greatest importance when the population is relatively small and is known.

> ***Examples***
>
> *Confidence A survey of 1,000 households has been completed, in a town of 20,000 households. 54% of households felt that crime had the largest impact on their quality of life. Using a 95% confidence level a confidence interval of 3.01 can be assumed. So you can say that between 51% and 57% of the town’ss population feel the crime has the largest impact on quality of life.*
>
> *Significance A survey is distributed to all 20,000 households in a town, there are 1,000 responses to the survey, equal to a 5% response. In accepting an interval level of 3, the sample size needed for significant results at the 95% confidence level is 1013, therefore the response rate is just short of significance at the 95% level.*

The significance of change over time in survey findings

In measuring the confidence interval of survey data when survey results are compared over time, it is important to understand if, for example, economic activity has changed over time or if the change in results is caused by survey error. To understand whether actual change has taken place, this requires the confidence interval of the difference between the two means to be tested (see further reading for a link to a web tool for measuring the confidence interval between two means).

> ***Example***
>
> *Survey 1 finds that economic activity stands at 49% using a sample of 1,000 residents. Another sample is selected one year later. Survey 2 finds that 51% of residents are economically active. In this case the 95% confidence interval is from -0.05 to 0.03 meaning that we cannot be sure whether the economic activity rates have actually increased or whether this is a result of survey error. This is because the 95% confidence interval has values which are either side of zero.*
>
> *If economic activity increases to 55%, the 95% confidence interval is from -0.09 to -0.01 meaning we can be 95% confident that economic activity has actually increased.*
>
> *Considerations: Both surveys must be based on a sample that is representative of the population. The sample used in survey 2 also needs to be independent from the sample used in survey 1.*

Cross-tabulation

Cross-tabluation is about taking two variables and tabulating the results of one variable against the other variable. This can be done quite simply in data analysis tools such as Microsoft Excel or SPSS. A crosstabulation gives you a basic picture of how two variables inter-relate, so for example you may have a question in your survey about employment, by running a cross tabulation of the survey data obtained for this question against that of age or gender for example (or both), would give you a table showing the employment status of both males and females, broken down by the age ranges you coded in your survey. This can provide quite powerful levels of information and is a useful way of testing the relationships between variables.

Statistical tests

For more complex statistical analysis there are a range of statistical tests that can be applied to your data. To select the right test, you need to ask yourself two questions:

1. What kind of data have you collected?
2. What variables are you looking to establish a relationship between?

Choosing the right test to compare measurements can be a tricky one, as you must choose between two families of tests: parametric and non- parametric:

- Parametric tests – include Mean, Standard Deviation, t test, analysis of variance (ANOVA), Pearson correlation, regression (linear and non linear);
- Non-parametric tests – include Median, interquartile range, Spearman correlation, Wilcoxon test, Mann-Whitney test, Kruskal-Wallis test, Friedman test.

Choosing the right test

Choosing between these two families of tests can be difficult. The following section outlines some of the basic rules for deciding which family of tests suits your data.

- You should choose a parametric test if your data is sampled from a population that follows a normal distribution (or Gaussian distribution). The normal distribution is a pattern for the distribution of a set of data, which follows a bell shaped curve. This means that the data has less of a tendency to produce unusually extreme values, compared to some other distributions.
- You should choose a non-parametric test if the population clearly does not follow a normal distribution.

Where values may be “off the scale,” that is, too high or too low to measure, a non-parametric test can assign values too low or too high to measure.

What do these tests tell you?

Parametric tests

Mean – The mean is more commonly called the average, however this is incorrect if “mean” is taken in the specific sense of “arithmetic mean” as there are different types of averages: the mean, median, and mode.

Standard Deviation – The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data, which may have the same mean but a different range.

t test – The t-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups.

Analysis of variance (ANOVA) – This is used to test hypotheses about differences between two or more means as in the t-test, however when there are more than two means, analysis of variance can be used to test differences for significance without increasing the error rate (Type I).

Pearson correlation – This is a common measure of the correlation between two variables. A correlation of +1 means that there is a perfect positive linear relationship between variables. A correlation of -1 means that there is a perfect negative linear relationship between variables.

Regression (linear and non linear) – A technique used for the modelling and analysis of numerical data. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modelling of causal relationships.

Non-parametric tests

Median – The median is the middle of a distribution: half the scores are above the median and half are below the median. The median is less sensitive to extreme scores than the mean and this makes it a better measure than the mean for highly skewed distributions. The median income is usually more informative than the mean income for example.

Interquartile range – The interquartile range (IQR) is the distance between the 75th percentile and the 25th percentile. The IQR is essentially the range of the middle 50% of the data. Because it uses the middle 50%, the IQR is not affected by outliers or extreme values.

Spearman correlation – Spearman’s Rank Correlation is a technique used to test the direction and strength of the relationship between two variables. In other words, it’s a device to show whether any one set of numbers has an effect on another set of numbers.

Wilcoxon test – The Wilcoxon test compares two paired groups of data. It calculates the differences between each set of pairs, and analyses the list of differences.

Mann-Whitney test – The Mann-Whitney test is a non-paramet- ric test for assessing whether two samples of observations come from the same distribution, testing the null hypothesis that the probability of an observation from one population exceeds the probability of an observation in a second population.

Kruskal-Wallis test – A non-parametric method for testing equality of population medians among groups, using a one-way analysis of variance by ranks.

Friedman test – The Friedman test is a nonparametric test that compares three or more paired groups.

What is the output?

The output of statistical analysis will depend on the statistical test you apply to your data, a detailed understanding of the test is required to be able to interpret the results. The output will most probably be further tables of data, with a number of things being reported. It is important to understand the information you need from a table of results, as you may only require a single figure, but be presented with a range of information which may be confusing if you are new to statistical analysis.

How should it be analysed?

Microsoft Excel

Microsoft Excel includes a collection of statistical functions, within the add-on Data Analysis ToolPak. Excel can analyse descriptive statistics at a simple level and when used effectively, can be very useful in the exploratory analysis of data, cross tabulations (pivot charts), viewing data in graphs to detect errors, unusual values, trends and patterns and summarising data with means and standard deviations. However, Excel is of very limited use in the formal statistical analysis of data unless your experimental design is very simple. The Analysis ToolPak is also no easier to use than more formal statistical packages, however there are plenty of guides and tutorials to be found on the internet.

Formal Statistical Packages (SPSS, SAS, Stata)

Inferential statistics are more often analysed in specialist statistical packages such as SPSS which provide greater functionality compared to Excel. The package used by the researcher often depends on which

package the researcher is familiar with and has access to. ‘ttese formal statistical packages can summarise data (e.g. frequencies), determine whether there are significant differences between groups (e.g. t-tests, analysis of variance) and examine relationships among variables (e.g. correlation, multiple regression). Further, these packages can produce charts, graphs and tables from the results of the analysis.

Formal Statistical Packages

	Pros	Cons
Microsoft Excel	Commonly used and widely available	It is not possible to see a record of the analysis you have previously conducted
	Easy to use for basic data analysis	Statistical analysis is only possible if data is sorted or in blocks
	Easy to import information from other packages.	Limited by space - MS Excel has a size limitation of 256 columns and
	Creating and amending charts is simple	over 65,500 rows meaning it has limited capacity for analysing larger datasets
Formal	Widely used.	Expensive to purchase.

Statistical Packages (SPSS, SAS, Stata)	More recent versions are more user friendly than earlier versions (menus to select rather than having to use syntax)	Need to buy add-ons to get full functionality
	Allows a wider range of statistics test to be conducted compared to Excel	Output isn't user friendly for beginners
	Easy to analyse survey / questionnaire responses	Charts are poor quality and difficult to amend - need to copy information into Excel
	File size is only dependent on your computers capacity	
	Survey data can be given assigned labels	
	It is easy to analyse sub groups of a large dataset	

Further reading

Guide to Good Statistical Practice – 'ttis resource is based at the Statistical Services Centre, University of Reading and consists of a series of guides on good statistical practice, intended primarily for research and support staff in development projects. Guides can be downloaded in HTML and PDF on subjects such as Data Management and Analysis. Links include: training courses and workshops; consultancy; resources (such as publications, software, external links)

Introduction to Central Tendency, David Lane – A useful guide explaining some basics to Statistical Analysis.

Statsoft Electronic Textbook,

http://www.statsoft.com/textbook/stathome.html – 'this Electronic Statistics Textbook offers training in the understanding and application of statistics

Simple Interactive Statistical Analysis,

http://home.clara.net/sisa/ – SISA allows you to do statistical analysis directly on the Internet. User friendly guides are available for statistical procedures.

Excel For Statistical Data Analysis,

Raynald's SPSS Tools – A website offering tools and tips for users of SPSS software, the site offers an archive of 400+ sample SPSS syntax, scripts and macros classified by purpose, as well as an FAQ, tips, tutorials and a Newbie's Corner. It invites contributions from other SPSS users to create a shared, open-source resource.

Choosing the correct Statistical Test

Confidence interval between two means – the following link provides a tool for measuring the confidence interval between two means

---

October 27, 2016