6.03 Introduction to Statistics

The word “statistics” is commonly used in 3 main ways:

1. Factual information involving numbers: a better word for this is data
2. Quantities which have been derived from sample data, e.g. the mean (or average) of a data set, better called a summary statistic.
3. An academic subject which involves reasoning about statistical quantities

The study of statistics falls into two main categories:

- **Descriptive statistics**
- **Inferential statistics** or **statistical testing**

Doing statistics well involves statistical thinking:

- Understanding how data was obtained, or deciding how to obtain data in the best way
- Presenting data
- Identifying features of potential interest based on research questions
- Selecting and carrying out tests and interpreting the findings
- Drawing conclusions

What is descriptive statistics?

Descriptive statistics is the process of describing and summarising a collection of data. Everyone should be familiar with the basics of descriptive statistics, because most academic subjects need to collect or understand data of some kind.

Data types

There are three fundamental types of data in statistics:

Nominal – data categories that do not have a natural order, e.g. gender, eye colour

Ordinal – data categories which have a natural order but are not numerical, e.g. Likert response scales (strongly disagree – strongly agree)

Scale – data ordered against a constant scale, e.g. date, temperature, length, weight

Advice: It is important to be clear what type your data is, because different types of data require different statistical methods.

Techniques of descriptive statistics

Tables

Present your results clearly in rows and columns. Make sure everything adds up. Label rows and columns clearly and unambiguously. This is easier said than done! Ask a friend if they can understand your table of data.

Summary calculations

For nominal and ordinal data, this involves finding the frequencies of the various categories, and perhaps converting to percentages. (For example, if 28 out of 50 people were dissatisfied with their treatment then that equates to 56%.)

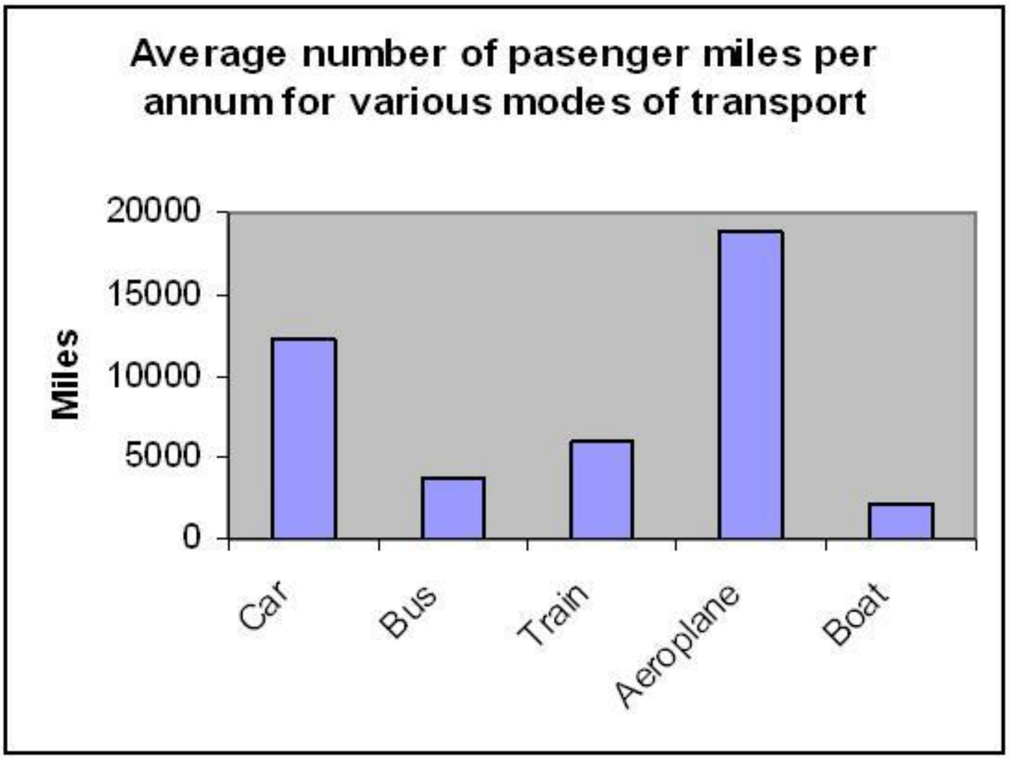
For scale data, summary calculations could include working out **measures of middle value** (averages like the mean, mode and median) and **measures of spread** (standard deviation and range).

Graphs

A picture speaks a thousand words, and a well-produced graph can illustrate the key features of a data set. Nominal and ordinal data are best presented using bar charts or pie charts. Numerical data can be presented using histograms, line graphs, box plots or scatter graphs.

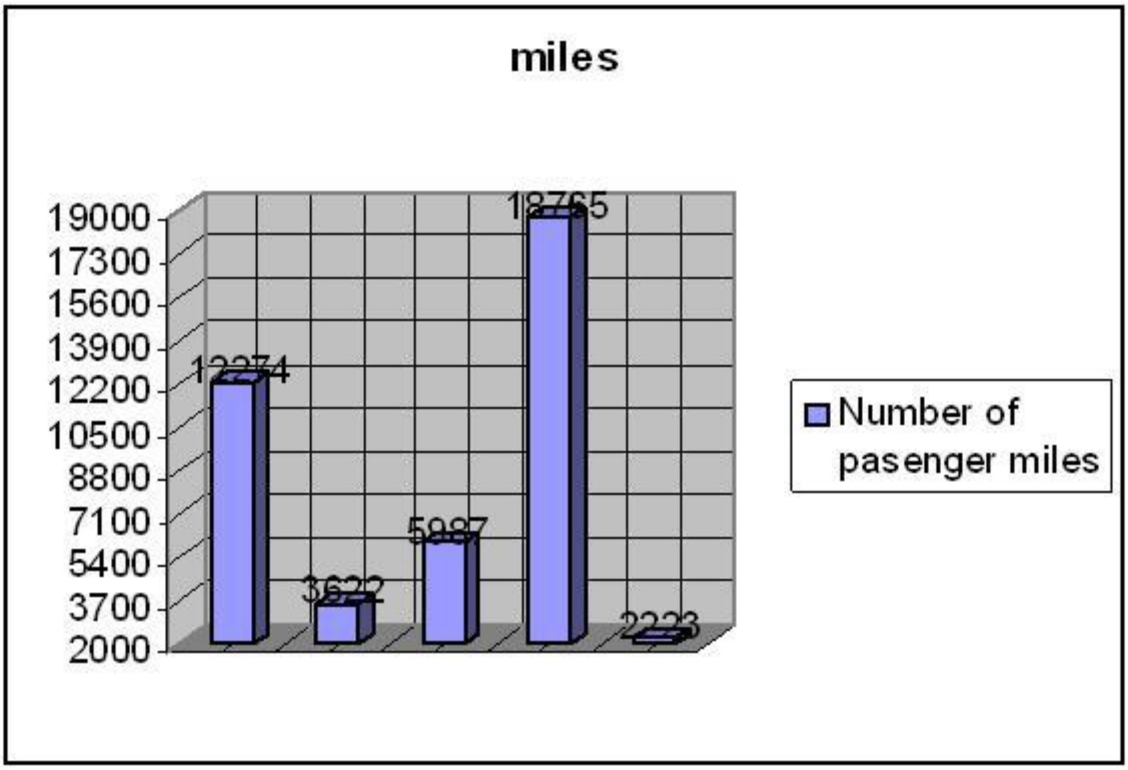
Advice: Computer software packages, such as Excel and SPSS, have many pre-defined graphs with quirky features (e.g. Excel creates a legend with a single data series). Don't just accept what they automatically produce for you: decide yourself in advance how you want your graph to look and use the software as a tool to produce what you want. Also, don't get carried away with all the fancy options – simple clarity is best (e.g. only use two dimensional graphs for academic purposes). See below for an example.

Nice graph



Centre for Academic Success
Birmingham City University
0121 331 7685 success@bcu.ac.uk
bcu.ac.uk/cas

Not so nice graph



Common bad practice in descriptive statistics

1. Working with data that is incomplete or biased in some way
2. Presenting irrelevant summary calculations. Computer software will work out everything for you, but it is for you to be selective.
3. Producing meaningless or misleading graphs. Choose sensible and consistent scales, and ensure everything is clearly labelled.

What is inferential statistics?

Inferential statistics deal with analysing **samples** from a **population** and drawing conclusions from them. This is routinely done by researchers. Obviously, a sample can never be totally representative of the whole population, so procedures are devised to take **sampling erro**r into account. Try to ensure as much as possible that your data set is **unbiased**, i.e. it is a **random sample**.

Techniques of inferential statistics

The vast majority of research questions are to do with establishing a difference or identifying a relationship. This is your **research hypothesis**. A **difference** could measure the effect of two treatments, for example, does the yield of a certain crop depend on the type of soil it is planted in? A **relationship** looks at links between two or more variables, for example, does personal esteem increase as earned income increases?

Statistical testing normally involves trying to show that the null hypothesis (the logical opposite result to your research hypothesis) is very unlikely. This is like trying to convict someone of a crime **beyond reasonable doubt**. The normal minimal threshold, or **significance level**, for rejecting the null hypothesis (equivalent to making a conviction) is **95%**.

For example, if you are trying to test for whether a coin is biased your null hypothesis would be that it is unbiased. You would then measure the likelihood of your experimental result (such as 20 coin tosses) and against the probability distribution for this experiment (which would be the binomial distribution here) and only reject the null hypothesis if the result (e.g. 18 heads and 2 tails) was within one of the **tails** of the distribution curve corresponding to a probability of less than 5% (or 0.05, this example is indeed significant at 95%).

Common pitfalls in inferential statistics

"I have collected my data, but I don't know what to do with it."

This comment will have your tutors/supervisors tearing their hair out! **Never** allow yourself to get to this stage. Without planning your strategy in advance, there is a real danger that you have wasted time collecting data that is inappropriate for further analysis. We suggest that you:

1. Attend the theory classes and learn about some standard statistical tests (e.g. t-test, chisquared test, ANOVA, etc.). This will be your "toolbox" for future reference.

2. Only set yourself a research question if you know you have the tools (i.e. an appropriate statistical test) to answer it and you can access data of the type and quantity required.

3. Never use the word "*significant*" in your write-up unless you have carried out a statistical test to show that your results actually are significant (as described above).

Suggested reading

Clarke, G. M. and Cooke, D. (2004) *A Basic Course in Statistics*. 5th edn. 519.5/Cla.

Coolican, H. (2014) *Research Methods and Statistics in Psychology*. 6th edn. e-book. (5th edn. also available from 150.72/Coo. This is a good reference for everyone, not just psychologists!)

Crawshaw, J. (2001) *A Concise Course in Advanced Level Statistics: With worked examples.* 519.5076/Cra.

Gonick, L. and Smith, W. (1993) *A Cartoon Guide to Statistics*. 519.50207/Gon.

Kapadia, R. and Andersson, G. (1987) *Statistics Explained: Basic concepts and methods*. 519.5/Kap.

Phillips, J. L. (1992) *How to Think about Statistics*. 519.5/Phi.

Rowntree, D. (1981) *Statistics without Tears*. 519.5/Row.

Rugg, G. (2007) *Using Statistics: A gentle introduction*. 001.4/Rug and e-book.

Salkind, N. J. (2004) *Statistics for People who (think they) Hate Statistics*. 519.5/Sal.

The Open University (2008) *Exploring data: graphs and numerical summaries*. Available at: http://openlearn.open.ac.uk/course/view.php?id=1710.

Wood, M. (2003) *Making Sense of Statistics: A non-mathematical approach*. 519.5/Woo and e-book.