

# Introductory Statistics with Excel

Andrew Parnell

University College Dublin



Class 2 - Experimental Design and probability

# Learning outcomes

In this class we will cover

- Experimental design
- Randomised controlled trials
- Confounding
- Observational studies
- An introduction to the normal distribution

# Obtaining data from experiments and observational studies

Sometimes we have the ability to design an **experiment** (or **trial**) that enables us to answer specific questions of interest. In such cases, we might be able to determine the **cause and effect** relationship between variables.

In this class we will meet **Randomised Controlled Trials (RCTs)**, possibly the most important idea in science!

When an experiment is not possible, we might run an **observational study**

# The general problem

Suppose I have a pill which is hypothesised to cure a particular disease. I give the pill to an animal and it recovers from the disease. How do I know whether the pill caused it to recover from the disease? Suppose it had not recovered from the disease. Do I then know that the pill hasn't worked?

# Types of research studies

**Observational studies** occur when researchers simply observe (or collect data via e.g. a survey) the participants about behaviour. For example we may run an observational study in which we record the blood pressure and frequency of milking for a number of cows.

**Experiments** occur when researchers manipulate one or more variables with the possibility of affecting some particular outcome of interest. For example, the researchers might forcefully change the number of milkings and then record their blood pressure

It is only in the second situation that we can really determine whether number of milkings has an effect on blood pressure

# Explanatory and response variables

Most experiments are run because we want to determine the relationship between two or more variables. The 'cause' variable is known as the **Explanatory variable** and the 'effect' variable is known as the **Response variable**. In the previous example, frequency of milking is the explanatory variable and blood pressure is the response variable.

Which are the explanatory/response variables in the following situations?

- **Whether taking aspirin or not** and **cholesterol level**
- **Likelihood of voting for the Green party** and **agreement with statement 'humans are responsible for global warming'**

# Problems in observational studies: confounding variables

## Example 1

A recent study showed that people who drink red wine tend to live longer than those who do not drink red wine. The study concluded that red wine might have beneficial health effects

## Example 2

Another study found that people who wore their shoes to bed tended to wake up with a headache. They concluded that wearing shoes to bed might be detrimental to a person's health

In both these cases a **confounding (or lurking) variable** is present that affects both the response and explanatory variables and casts doubt on the causal link between them

# Confounding variables: the solution

We can remove the influence of confounding variables by conducting an experiment where we randomly assign people to different groups.

- In example 1, we could randomly assign people to one of several **treatment groups**: e.g. no red wine, 1-2 glasses of red wine per day, 3+ glasses of red wine per day. We could then follow up how long they live over several years and determine the health benefits of red wine
- In example 2, we could randomly assign people to the treatment groups: wear shoes to bed, do not wear shoes to bed. We could then record how often the people in the two groups woke up with a headache

This **experimental design** works because we are breaking the confounding variable's control over the response and explanatory variables



# Randomised Controlled Trials

# Designing a good experiment

There are a few things that generally make a good experiment:

- **Randomisation** (randomly allocating subjects to different groups) lessens the effect of any confounding variables. The benefit of randomisation is that each of the treatment groups should be roughly equivalent aside from the explanatory variable
- **Blocking** Grouping variables over which we have control over into categories
- **A control group** which receives no treatment (or a dummy treatment) allows us to compare the true effect on the response variable had no treatment occurred
- **Blinding** in which neither the subject nor the researcher knows which treatment is being applied. This is particularly important in medical studies

# Control groups and placebos

- We use a control group to determine whether the change in response is due to the treatment or not. For example, it is well-known that people recover from coughs and colds without any treatment at all given enough time. A treatment should thus only prove acceptable if it can reduce the amount of time people are sick
- A **placebo** is a dummy treatment that is identical in every way from the normal treatment except that of the active ingredient. For example, a sugar pill of exactly the same colour and shape to the pill being tested for medical reasons is often given to a placebo group

# The remarkable placebo effect

- It is well known that giving someone a placebo will tend to improve the response variable more than giving them no treatment at all
- Remarkably, it has also been shown that green sugar pills are better than red sugar pills for treating anxiety. Similarly, a salt-water placebo injection was found to be more effective than a standard white sugar pill
- Even more strangely, in the 1950s a placebo operation was found to be a more effective treatment for Angina than the real operation it was replacing

The placebo effect is not fully understood in science and remains an active area of research.

# Blinding

Suppose you are running an experiment into a potential cure for a serious disease. For each person who arrives in your office you must allocate them to either the proper treatment or a placebo. A very sick looking person arrives into your office. Do you allocate them to the treatment or placebo?

You are taking part in a trial to see whether paracetamol or ibuprofen is better for reducing headaches. You have been allocated to the aspirin group but you really don't like aspirin and far prefer taking ibuprofen

In both these situations it is beneficial for the experimenter and the subjects not to know whether they are receiving the treatment or placebo. If only the experimenter knows which treatment you are taking it is known as **single blinding**. If neither the experimenter nor the subject know it is called **double blinding**

# Randomised controlled trials: the gold standard

- 1 Take a **large sample** of the population of interest. A large sample will be more representative of the population.
- 2 **Randomly allocate** each person in each block to either receiving the **treatment** or a **placebo** (or the current best treatment). The randomisation will lessen the effect of confounding variables. The placebo will allow the measurement to focus only on the effect of the treatment and not the action of being treated
- 3 Make sure that neither the experimenters nor the subjects know to which group they have been allocated. This **double blinding** will stop subjects/experimenters biasing the response to the treatment
- 4 Record your response variable(s) of interest and report the **difference** in the response variable(s) between the treatment group and the placebo group

The above is remarkably simple and incredibly powerful yet almost completely unknown to the general public

## Some other useful things to know

- It is often the case that RCTs use volunteers. Those who volunteer early in the study might be fitter or more optimistic. It is therefore best to randomise the order to which they are allocated to treatment/placebo
- If the subjects are undergoing more than one treatment (e.g. two different drugs) it is usually best to randomise the order of the treatments, to leave a gap between their administration, or to give them both simultaneously
- In some situations (e.g. physiotherapy) it is very hard to find good placebos. Finding a good placebo requires imagination and creativity. See for example the studies in Acupuncture

# Extending RCTs

- If we know in advance that there will be a strong confounding variable, for example age or income, we might divide up our sample into **blocks** representing different groups of this variable. Within each group we would then randomly allocate to treatment or placebo
- A specific example of this is a **matched pair** design whereby subjects are paired together based on equality of some important variable. We then randomly allocate one to placebo and one to response
- If we have the resources we may make more than one measurement on each subject. This will enable us to see whether each subject responds similarly to the same treatment or not. These are known as **replicated** or **repeated measures** experimental designs



# The Hawthorne and other effects

In addition to the placebo effect, a number of other factors have been observed to affect the results of an experimental study

- The **Hawthorne effect** occurs when a response variable is changed simply by the manor of being in an experiment. This was first noticed in the 1920s when workers in a factory improved their output when any change was made to their working conditions
- The **Experimenter effect** occurs when the experimenter has a vested interest in the outcome of the experiment and so is likely to (unconsciously) bias the results. This effect can be lessened when the experiment is double-blinded
- The **Nocebo** effect occurs when subjects believe the treatment will actively harm them. In such cases the subjects tend to report increased side effects and reduced effectiveness of the treatment

# When would you **not** use an RCT?

The main issue with RCTs is that they are usually expensive, artificial, and occasionally unethical. Consider the following examples:

- We want to know whether breast-feeding or formula-feeding is most beneficial for babies' development. A standard RCT would involve telling mothers to breast-feed or formula-feed based on their allocation to treatment group
- A new drug is available to relieve pain in the very last stages of people dying from a serious disease. The drug seems to be safe and works in various animal trials. There are no pre-existing treatments for comparison purposes
- We want to learn whether people who are left-handed are more likely to suffer from post-traumatic stress disorder. The proportion of people who are left-handed is approximately 5% and the proportion suffering from post-traumatic stress disorder is approximately 0.1%

In such cases we might use an **observational study** instead

# Observational studies

# Designing good observational studies

There are two main types of observational study:

**Retrospective studies** where information is obtained from the past. For example, subjects may be asked to recall their past exercising behaviour

**Prospective studies** where information is obtained by following subjects over a period of time. This approach is generally better as people tend to forget or mis-report their past behaviour

A third category (which may be part of either a retrospective or prospective study) is that of a **case-control** study, whereby pairs of subjects with similar explanatory variables are matched between categories of the response variable.

# Advantages of observational/case-control studies

- Case-control studies do not suffer from the **unethical** aspect of RCTs. The subjects are not required to change their behaviour based on an experimenter's instructions
- They are extremely **efficient**. Consider trying to determine whether pet birds might cause lung cancer. The RCT version would involve randomly giving people pet birds (or placebo birds!) and waiting to see whether the owner's contract lung cancer. A case control study would identify a large group of people with lung cancer and another group who have not and then ask whether or not each group owned a pet bird
- They reduce the chance of **confounders**. Like a matched-pair experimental study, confounders are reduced because we can match subjects based on important variables

# Not just for medicine!

RCTs and observational studies were developed with medical studies in mind. **Since we discovered RCTs and the gold standard our life expectancies have doubled!**

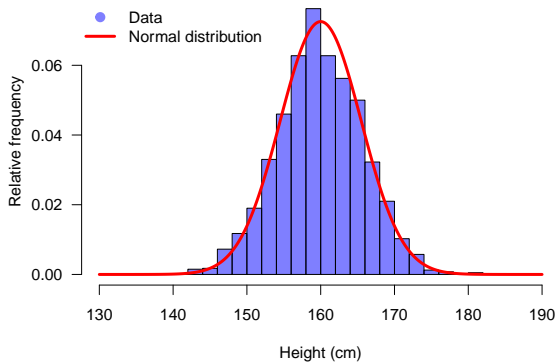
However, their use is not just restricted to the medical field. Consider the following examples

- How might we design an experiment to test whether studying more subjects at primary level leads to increased achievement in later years?
- How might we design an experiment to test whether prison or rehabilitation might improve re-offending rates of drug-related crimes?

# The normal distribution

# The normal distribution

Often, when looking at histograms of data we see the same shapes re-appearing:



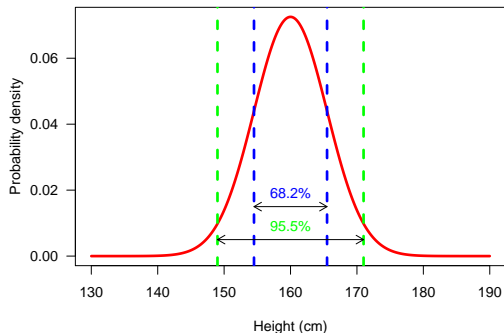
This shape (peaked in the middle, decaying out evenly on both sides) is known as a **bell-shaped curve**. One type of these is the **normal distribution**; possibly the most important and useful probability distribution in statistics



# The normal distribution and the empirical rule

The normal distribution is very useful. If we are happy to assume that our data come from a normal distribution then we can say that:

- 68.2% of the observations will lie within 1 standard deviation of the mean
- 95.5% of the observations will lie within 2 standard deviations of the mean



We can use these facts to create predictions with uncertainty about future observations we might find.

# The normal distribution and z-scores

- If we take each observation and **standardise** it by first subtracting the mean and then dividing by the standard deviation, we obtain a **z-score**
- A **z-score** is useful as a relative measure of an observation in comparison to the others. The standardised observations will have mean 0 and standard deviation 1
- If the data come from a normal distribution, then 68.2% of the **z-scores** will lie between -1 and 1, as we know from the results on the previous slide
- Furthermore, if a particular **z-score** is, e.g. 1.25, then we know that this value is 1.25 standard deviations away from the mean
- We will use this trick later on when we cover **hypothesis testing**

# Class 2 summary

- Remember the key elements of an experiment: **randomisation**, **blocking**, **control groups**, **blinding**
- The bigger the **sample size** the more precise your results will be
- If you can't do randomisation go for an **observational study**, but be wary of **confounding**
- The **normal distribution** and the **empirical rule** allow us to make population statements from statistics