

# Mike and Janet's Research Design Course

Michael Drinnan & Janet Wilson

Modernising Scientific Careers

University of Newcastle upon Tyne



## TABLE OF CONTENTS

<b>1 SCIENTIFIC RESEARCH</b> .....	<b>4</b>
What is scientific research? .....	5
Picking the right research question .....	6
The effect of random variability .....	7
Forming your hypothesis .....	8
The different types of variable .....	9
Independent and dependent variables .....	10
Hypotheses and p values .....	11
The p value and the effect - statistical and clinical significance .....	12
<b>2 CHOOSING AND REFINING YOUR RESEARCH QUESTION</b> .....	<b>13</b>
<b>3 THE BASICS OF STUDY DESIGN</b> .....	<b>18</b>
Introduction to study design .....	19
Bias .....	20
Observational studies .....	21
Experimental studies .....	22
A randomised controlled trial (RCT) .....	23
Picking the control .....	24
The crossover .....	25
Explanatory and pragmatic studies .....	26
Randomisation .....	26
Sample size .....	27
Type I errors, type II errors and power calculations .....	28
<b>4 STUDYING THE BEHAVIOUR OF RANDOM OR NOISY VARIABLES</b> .....	<b>29</b>
How do random variables behave? .....	30
How does the normal distribution help? .....	31
Populations and samples .....	32
Sampling bias .....	33
Degrees of freedom .....	34
<b>5 PRESENTATION OF DATA</b> .....	<b>35</b>
Summary or descriptive statistics .....	36
Graphical description of data .....	37
Plotting the results of a clinical study .....	38
Confusing stuff .....	39
Parametric or non-parametric statistics? .....	40
Standard deviation, standard error or confidence interval? .....	41
<b>6 COMPARING BETWEEN GROUPS: THE T-TEST</b> .....	<b>42</b>
The single sample t-test .....	43
The 2-sample t-test .....	44
The paired t-test .....	45
Multiple comparisons and analysis of variance (ANOVA) .....	46
More about analysis of variance .....	47
Non-parametric tests .....	48
Writing up your statistical test .....	49
<b>7 LINKS IN NUMERIC VARIABLES: CORRELATION AND REGRESSION</b> .....	<b>50</b>
Demonstrating a link in numeric data - correlation .....	51
Some examples of correlation .....	52
Quantifying the effect - linear regression .....	52
Non-parametric tests .....	53
<b>8 RELIABILITY, VALIDITY AND AGREEMENT</b> .....	<b>54</b>
Agreement - what's it all about? .....	55
Measuring agreement – categorical stuff, reliability and Kappa .....	57
Measuring validity - evaluating a diagnostic test .....	59
Statistical tests if you want them - the chi-squared test .....	61
Some more about cutoffs – the receiver-operator characteristic .....	62
Measuring agreement – The use and abuse of correlation .....	63
<b>USEFUL RESOURCES</b> .....	<b>64</b>

# Ten things to remember

Write a protocol before starting the study.

Talk to a statistician BEFORE you mess up the study - not afterwards.

**BACK UP YOUR DATA** - computers are cheap and replaceable, but research data are neither cheap nor replaceable.

Plot your data.

Use a computer stats package to do your statistics - don't do them by hand.

The p value indicates the probability that there is no link between the test variables, and the observed pattern (or one even more extreme) arose by chance. It will be improved by: a bigger effect; making more measurements; less random variability between measurements.

Don't confuse statistical significance with clinical importance. It's the effect you're interested in - not the p value. A highly significant p is NOT evidence of a clinically important effect.

You can never be ABSOLUTELY sure a link is real.

The correlation coefficient shows there is a link between two variables - but is NOT a measure of agreement.

**KEEP IT SIMPLE!**

# 1

Scientific research

# 1

## What is scientific research?

---

Everyone has a rough idea what science is about - it's the study of the way nature works. Of course, many doctrines claim to explain how nature works, but you probably don't believe the world is supported by four giant elephants, all standing on the back of an even bigger turtle. What singles out science as being special? One way to think about science is as a rigorous and objective way of studying nature. The scientific method goes something like this:

**1 You observe some interesting phenomenon in nature**

*With few exceptions, children can always learn to speak.*

**2 You form a hypothesis to explain how nature is working**

*Language is an innate ability, conferred at birth by God. (!)*

**3 You conduct a study to test the hypothesis**

*A group of children are separated from their parents at birth, and kept in isolation. If the hypothesis is correct, then they should nevertheless learn to speak. This is known as 'the forbidden experiment'; Holy Roman Emperor Frederick II (allegedly) performed it, with a predictable outcome.*

**4a You reject your hypothesis based on the outcome of the study**

*The children don't learn to speak, so you must reject your hypothesis.  
You must now go back to step 2 with a new hypothesis.*

*OR (IF YOU'RE REALLY LUCKY)*

**4b The study supports or at least, doesn't completely refute your hypothesis**

*Your hypothesis gains some credibility.  
You (or other people) go back to step 3 and test the hypothesis further with a new study.*

**5 After some time, your hypothesis becomes an accepted theory**

*With each successful study, your hypothesis gains credibility.  
Eventually, it becomes a theory - people generally accept that your explanation is correct.*

### But notice...

- No matter how many successful studies, you can never be *completely* sure you have the right theory. Newton's laws of motion were accepted for 300 years, until Einstein showed they don't work at high speeds.
- It might only take one study to discredit your theory, as happened with Newton's laws. Michelson & Morley carried out the experiment in question, but they didn't follow the result to its conclusion. One of Einstein's achievements was to suggest that Newton's laws of motion might be wrong, even though he didn't have the means to test his new hypothesis.
- Nevertheless, Newton's laws are so close to being correct in our everyday world that the difference doesn't matter. It took high-speed aircraft and atomic clocks to show Einstein was right. A simple theory that isn't quite right might be more use than a complex but accurate theory.

### And so...

The scientific method is fairly straightforward but to do successful science, you need to get a few things right:

- There is a lot of nature out there - what are the interesting phenomena that are worth studying?
- How do you form a hypothesis about the thing you've chosen to study that can be tested experimentally?
- How do you design an experiment to test the hypothesis properly?
- How do you interpret the results of the experiment to support or reject your hypothesis?

This course is about understanding how these things work, so with any luck you'll be able to get them right.

## Picking the right research question

---

Janet is going to talk to you about this subject and as you'll appreciate, there's no substitute for experience in knowing what the interesting research questions are. In many cases, you'll need to apply for soft money to fund your study - from the likes of the UK funding bodies (Medical Research Council, etc.) or from the medical charities. Generally, applications will be sent out to two or more reviewers, and the money is awarded to the projects with the best reviews. According to people who know this stuff, the funding bodies normally want to know two things from their reviewers:

### Is this study worth doing?

This is where the experience of the experts comes in - there are always current sexy topics, and you need to be in the know.

### Can these people do it?

Here, you need to demonstrate that:

- You have the skills and resources to complete the study in the time set.
- The proposed study will actually answer the research question you've raised.

The rating of your study will be a product of the answers to the two questions. If you can get a strong YES answer to both these questions, you've got a fighting chance of getting the support for your research.

## Forming your hypothesis

No matter what study you're doing, you will probably be trying to show a link (formally, an *association*, or a *relationship*) between two variables. For most studies, a *study hypothesis* can be formed along similar lines:

*We hypothesise that XXX is linked with YYY*

### This is true whether you're doing epidemiology...

*We hypothesise that oral cancer is linked with cigarette smoking.*

### ... a drug trial ...

*We hypothesise that the prognosis of patients with oral cancer is related to the treatment given.*

### ...or an inter-observer agreement study.

*We hypothesise that Jack's classification of these pressure traces will be related to Jill's.*

You'll struggle to think of an example that can't be phrased as a hypothesis in this way with a bit of thought. The most obvious exception is an observational study where you're just trying to quantify the incidence of a disease, for example.

***A good study will have a single and clear main hypothesis that could be written up as a single article.***

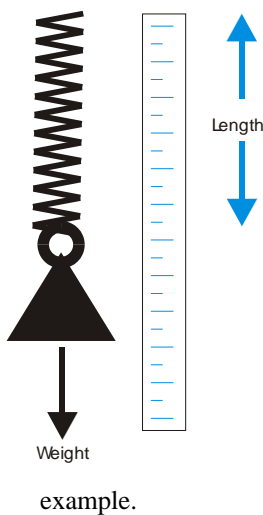
## What happens next?

- Perform your study. For the time being, you assume the null hypothesis is true - there is no link between the variables you're studying.
- Assuming the null hypothesis is true (there is no link), you use a statistical test to calculate how likely it would be that your results arose *by chance alone*.
- If it is highly improbable that the results are due *to chance alone*, you reject the null hypothesis in favour of the alternative study hypothesis.
- Notice you can never be *absolutely* sure the link is real - you can just show that it is highly improbable that it arose by chance.

## The effect of random variability

If you want to do medical research, you can't escape statistics. In fact, some of the prestigious journals, the BMJ for example, insist on statistical tests in any paper you submit to them, and all papers are sent for statistical review. Why?

### A physics experiment

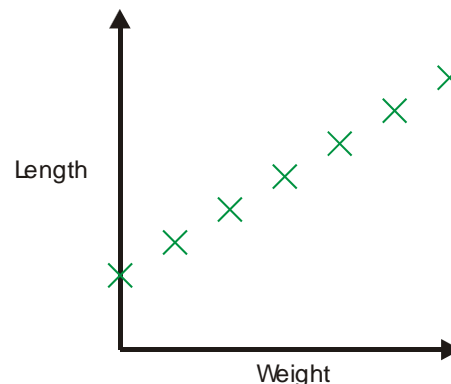


If you ever did physics at school, you probably did this experiment. The idea is to study the link between the weight hung from the end of a spring, and the spring's length.

So - you hang a few different weights from the end of the spring, and measure the length of the spring every time. The results you got might have looked something like this (right):

It's easy to see that there is a link between weight and Hooke's law.

You could even go on and come up with a law for this spring. The length increases by 1 inch for each 1lb



different weights and measure the time. The results looked something

clearly is a link length, which is

come up with a length increases by weight, for

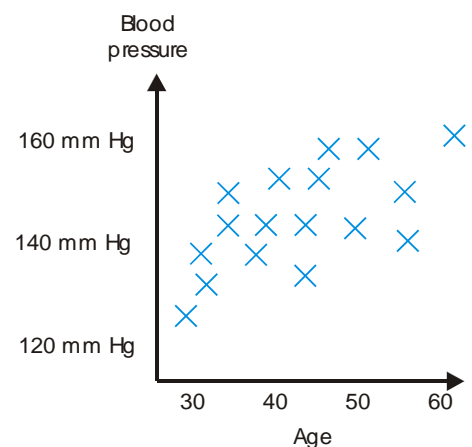
### A medical experiment

This time, you want to study the link between age and systolic blood pressure. So you grab a few people of varying ages, and measure everyone's blood pressure. Your results look like this (right):

#### IS THERE A RELATIONSHIP BETWEEN AGE AND BP?

Well, it looks like there might be. Perhaps older people have a higher blood pressure. But on the other hand, the relationship isn't clear. Perhaps this was just a fluke, and there really isn't any link at all.

This is very typical of medical research. It is VERY RARE to find a clear-cut relationship between the two variables you're investigating. Maybe there is a link, but it is largely masked by the random variability between people, so how can you be sure?



### The need for statistics

A change of subject. If you want to pick randomly between an Italian or Indian restaurant, you might toss a coin. Tossing a coin has entered folklore as a random process, and it pretty much is<sup>♦</sup>. Suppose you eat out with your friend every week for a year. On a given week, nobody can predict whether you'll eat pizza or curry, but there are still things you can predict. For example:

- Over a year, you'd probably expect about the same number of each - 26 Indian, 26 Italian.
- You'd be VERY surprised to eat pizza every week for a year, and might suspect your friend's coin was bent.
- But what if you had 40 Italian, 12 Indian. Is this just chance, or is the coin bent?

Your gut feeling tells you it probably isn't random. Looking at the 'blood pressure' graph, your gut feeling tells you it probably isn't random; there's a link between BP and age. This is where the statistics comes in:

#### Statistics: studying the behaviour of random or noisy variables

If you can understand how random variables behave, you might be able to:

- Sort out which effects are due simply to random variability, and which ones probably aren't.
- Say how big the effect is - the coin lands heads 4 of every 5 tosses, BP rises 10 mm Hg every 10 years.

<sup>♦</sup> By all accounts, a coin is biased - roughly 51% tails - because of the slightly higher weight of the head.

## Forming your hypothesis

As we already said, you will normally be trying to show a link between two variables. Let's take the following study hypothesis:

*We hypothesise that oral cancer is associated with cigarette smoking.*

### The null hypothesis

For every study hypothesis you put forward, there will be a corresponding null hypothesis. That is:

**Study hypothesis:** *We hypothesise that XXX is associated with YYY.*

**Null hypothesis:** *XXX is NOT associated with YYY, and our results reflect random variability alone.*

For example:

**Study hypothesis:** *We hypothesise that oral cancer is associated with cigarette smoking.*

**Null hypothesis:** *Oral cancer is NOT associated with cigarette smoking.*

### A testable hypothesis

When you form a hypothesis, remember that you will have to test the hypothesis in a study. Therefore:

***Your study hypothesis must be testable.***

To make a testable hypothesis, (to show a link between XXX and YYY), *you must be able to measure something about XXX and YYY.* In the *oral cancer* example:

Things you could measure about smoking	Things you could measure about oral cancer
Does your subject smoke? (Y/N)	Does your patient have oral cancer? (Y/N)
How many cigarettes per day are smoked?	The grade of cancer
How long has he/she smoked for?	The location of the cancer
What is the lifetime consumption of tobacco?	How long does the patient survive?

You could look for a link between anything in the left column, and anything in the right column. For example:

- Lifetime consumption of tobacco, and the grade of cancer.
- Number of years smoking, and survival.

You would then probably decide which things:

- You were most interested in, because they fit best with your hypothesis;
- Ought to show the strongest link;
- Were easiest to measure.

### An untestable hypothesis

A poor hypothesis is one where you can't make any meaningful measurements to test the hypothesis:

*We hypothesise that the Daily Telegraph is a better newspaper than The Sun.*

How would you interpret *better*? Cheaper? More readers? A readership higher up in the socio-economic scale? Any of these might be used, depending who performed the study and what they were trying to prove. Here's a testable hypothesis:

*We hypothesise that the average number of syllables per word in the Daily Telegraph is higher than in The Sun.*

### The types of variable

Any of the things that you measure would be termed a *variable* - something that can vary between patients, or with time.

- The answer to the question *do you smoke* is a variable that can only be *yes* or *no*.
- The *number of cigarettes per day* is another type of variable - a number.

Clearly, there are different types of variable, and it's useful to consider them now.



## The different types of variable

---

The names and types of variables are a great cause of confusion. Like in medicine, complicated names for simple everyday ideas.

### Numeric (AKA quantitative) variables

You'll be familiar with numeric variables; a *numeric variable* is something you can naturally put a number to.

#### Continuous numeric variables

The term *continuous* simply means the variable can take any value. Continuous variables can be recorded to any degree of accuracy, limited only by your measuring instrument.

*Height; weight; age; body mass index; temperature; vocal cord frequency.*

#### Discrete numeric variables

A *discrete variable* is one that can only take certain values - it will typically be a count of something. For example:

*Number of patients in study; number of teeth missing; Number of children.*

When you record a continuous variable, you will inevitably round it. For example, you might record temperature to the nearest 0.1 °C, or weight to the nearest 1 kg. In either case, you have converted a continuous variable to a discrete one. As you might expect then, the same statistical methods can often be used for both.

### Categorical (AKA nominal, qualitative) variables

A *categorical variable* is one where the measurement can be placed into one of several categories.

#### Ordered nominal (AKA ranked, ordinal) variables

An *ordered nominal variable* is one that takes a range of values with a natural order. The response to questionnaires is very often an ordered nominal variable.

- *None mild moderate severe*
- *Strongly agree agree not sure disagree strongly disagree.*

In many cases, you will put numbers to the categories (none=0, mild=1, moderate=2, severe=3), and use the same statistical methods here as for numeric variables. However, be careful how the numbers are interpreted. The difference between 120 and 130 cm is the same as 170 to 180 cm. The difference between *none* and *mild* might not be the same as from *moderate* to *severe*.

#### Nominal variables

A *nominal variable* describes something about your patients where no natural order can be applied. For example:

- *Blood group (A/B/AB/O)*
- *eye colour (blue/green/brown)*
- *treatment (aspirin/paracetamol/ibuprofen/placebo).*

There is no natural ranked order to blood group or eye colour. As you might expect, you generally need different statistical methods to handle nominal variables.

#### Dichotomous (binary) variables

A *dichotomous* variable is a simple nominal variable that can take only one of two values. In very many studies you will classify your subjects according to dichotomous variables, as the list below hints:

- *survived* or *died*
- *diseased* or *normal*
- *treatment* or *placebo*
- *pre-op* or *post-op*

Here, you can generally use similar statistical methods as for nominal variables, but often things are made easier when you have only the two categories.

## Independent and dependent variables

---

### Cause and effect

In very many cases, you'll be interested in cause and effect. For example, if you found a link between oral cancer and cigarette smoking, you would suspect that smoking causes the cancer. It would be far-fetched to imagine that the cancer would lead people to start smoking. Likewise, you would anticipate in a drug trial that the drug affects the disease outcome, and not vice versa. Therefore, in many studies there is naturally an *independent* variable and a *dependent* variable.

### Independent variable (AKA category, predictor, explanatory variable, cause, factor)

The *independent variable* is the thing you think is causing the outcome.

### Dependent variable (AKA outcome, effect)

The *dependent variable* is the thing you measure to assess the outcome. It might be blood pressure or survival in a drug trial.

### For example...

We hypothesise that smoking is associated with oral cancer:

*Number of cigarettes per day might be the independent variable.*

*Presence or absence of cancer might be the dependent variable.*

### Experimental studies...

In most experiments, an *independent variable* is something you the experimenter have control over. Very often, you are responsible for recruiting patients of particular types, or for assigning the subject into one of the categories:

*Treatment or placebo; pre-op or post-op; diseased or non-diseased; male or female.*

You then perform an experiment on some or all of the subjects. This would be termed an *experimental or interventional study*, because you actively experiment on the subjects. For example, to study the link between oral cancer and smoking, you would like to do the following experiment:

- Recruit a group of non-smoking subjects.
- Assign half at random to a *smoking* and half to a *non-smoking* group. The *smoking* group are to immediately begin smoking 80 per day. The group (*smoking* or *non-smoking*) is the independent variable.
- Measure outcome, survival in years perhaps, in each group. This is the dependent variable.
- Look for a link between the group (*smoking* or *non-smoking*) and survival.

You assigned the subjects completely at random, so there should have been no overall difference between the groups *before the experiment*. Therefore, if you find a link later, *you can argue convincingly that it is due to their being treated differently*. This would be a *randomised controlled trial (RCT)*. We'll consider it later.

### ...and observational studies

You can't do such an experiment for obvious reasons, but what you can do is observe the smokers already in the community. For example:

- Recruit two groups of subjects, *smoking* and *non-smoking*, according to their existing habits.
- Measure outcome (survival, perhaps) in each group. This is again the dependent variable.
- Look for a link between the group and survival.

This would be termed a *case-control study*, a type of *observational study*. You don't actively intervene in the subjects' lives - you merely observe what happens to them. Unfortunately, you the scientist can't assign people to *smoking* or *non-smoking* groups at random - it's already been decided by their smoking habits. It's therefore difficult to know whether the outcome is really due to smoking, or perhaps to some other *confounding factor*. For instance, perhaps the smokers also do less exercise, and it's the lack of exercise that is the *real* risk factor.

## Hypotheses and p values

As we said, you always start from a *study hypothesis*, normally based on an interesting observation about how nature seems to be working. Let's take another example related to food:

*You and your colleague buy lunch from the local sandwich shop on alternate days. You've noticed something interesting about your colleague's choice:*

- *When you're buying, he normally has the smoked wild salmon in a squid-ink and avocado sauce.*
- *When he's buying, he has cheese.*

### The study hypothesis

You think this isn't simply due to chance, and so you study the phenomenon further. First, you raise your study hypothesis. Remember, in a good hypothesis you must be able to measure the outcome, so:

*I hypothesise that the price of my colleague's lunch is linked to who is buying it.*

### The null hypothesis

But there's always another possibility - the *null hypothesis*:

*There is no link between the price of my colleague's lunch and the person buying it.  
My observations are due to chance alone.*

### Perform the experiment

Now you perform the experiment. For 20 days you record:

- Who bought lunch (the predictor, a binary variable);
- The cost of your colleague's lunch (the outcome, a continuous variable).

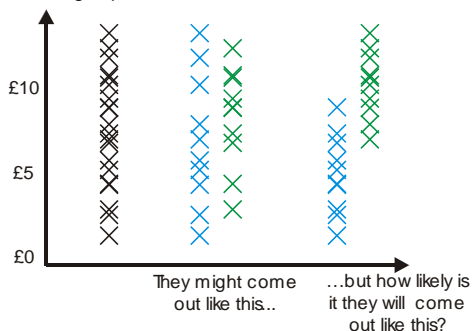
### Plot the data

This should *always* be the first step in analyzing your results - plot the data. Here (*right*), we've plotted a cross for every day, separated according to the person who bought the lunch on that day. Your gut feeling tells you immediately that you were right, but your colleague flatly denies the evidence. You spent over £100 in total, he spent under £50, but he claims these results are due to chance alone.



### Perform a statistical test to show the link

Take these 20 lunches and split them at random into two groups..



Maybe it is due to chance. But what is the probability? One way to figure it out would be like this:

- Write down the price of each lunch on a piece of paper.
- Put all the pieces of paper in a hat.
- At random, pull out 10 into one pile, marked *ME*.
- Put the other 10 in the other pile, marked *COLLEAGUE*.
- Write down the total cost of each pile.

If you repeated this enough times, you might find that only about once every 1,000 times would the total in your pile be £100 or more. What you've just done is simulate what happens by chance alone.

*The probability of your results being due to chance alone is 1 in 1000.  
 $p = 0.001$*

On this basis you would probably reject the null hypothesis and accept the study hypothesis. Your conclusion would be written as follows:

*The price of my colleague's lunch is associated with the person buying it ( $p=0.001$ ).*

***But it MIGHT be chance.  
You can never be ABSOLUTELY sure the link is real.***

*In this case, the statistical test just confirms what you already knew from the graph. It is RARE for a statistical test to contradict flatly the gut feeling you get from looking at the data in graphical form. PLOT YOUR DATA!!!*

## The p value and the effect - statistical and clinical significance

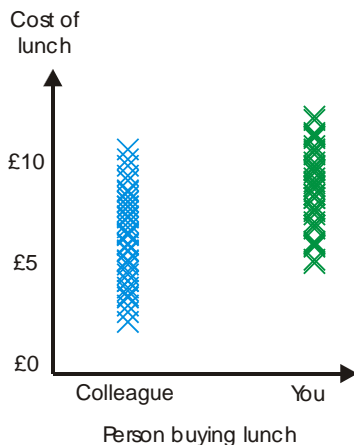
### Statistical significance and the number of measurements

In the last example, you demonstrated to everyone's satisfaction that there is a link between the independent (*the person buying lunch*) and dependent (*the cost of the lunch*) variables. But is the link worth worrying about?

Well in this case, it seems you are paying twice as much as your colleague, and so you might well be a bit upset. But what if things had come out differently? You wait for twenty days, and get these results (*right*).

In this case, your colleague has paid an average of £7, whereas you paid £8. Is there still a link, or not?

Well, it turns out the probability of this happening by chance is about 1 in 10 ( $p=0.1$ ). Not likely but it *could* just be luck, so you stay quiet and keep your records for a further few months.



Now the results look something like this (*left*). Your colleague still paid an average of £7, and you still paid an average of £8. Nevertheless, the chance of this arrangement arising *by chance alone* is now just 1 in 1000 ( $p=0.001$ ).

The *effect* is exactly the same, but the *statistical significance* has increased. The only thing that changed is the number of measurements you made!!!

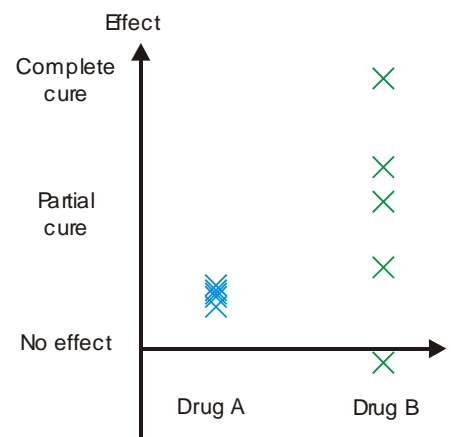
Once again, you've clearly demonstrated that you are paying more than your colleague. This time, maybe you aren't too worried, because the effect only amounts to £1 per day.

***A highly significant p value doesn't tell you that the effect is important!!!***

### Statistical significance and the spread of data

This time, consider a drug trial of two drugs, A and B, for (say) headache. Each drug is tested in 5 patients. Drug A has a small but consistent curative effect, but drug B shows a considerable spread of effect (*right*). Which would you rather have?

Most people would probably pick drug B, because it seems the effect of B is bigger. However, if you perform the statistical tests, the observed effect of drug A is less likely to be due to chance alone. That's not because it has a bigger effect, but because *the effect is more consistent between patients*.



### To summarise...

The thing you are interested in is probably the effect. However, the statistical significance (the p value, the probability that your observations are due to chance alone) is improved by any of the following:

- A bigger effect***
- Making more measurements***
- More consistency (ie. less random variability) between the measurements***

***Statistical significance means ONLY that the link you found is unlikely to be due to chance alone. It is up to you to decide whether the effect is large enough to be of any clinical importance.***

# 2

## Choosing and refining your research question

# 2

## Choosing and refining your research question

Often those new to research complain they don't have any ideas

After a very short time, however it is clear – the problem is too many questions, not too few.

### Some drivers to research question development



**I wonder why's:** this is the best kind of question. If you ask this Q – others will too (just need to make sure they haven't already answered it....)



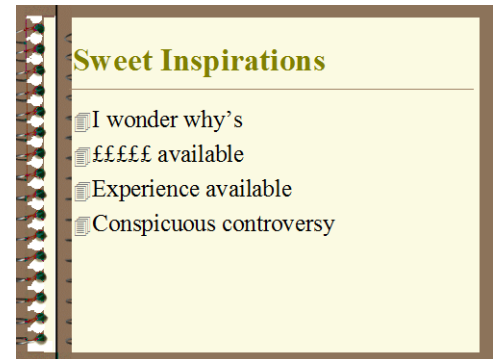
**£££££ available.** Research priorities, like much else, have fashions. NHS research priorities have fashions. If stroke is fashionable and you are into dysphagia – stroke dysphagia is an area to scrutinize.



**Experience available.** In research, credibility is all. If your three most interested with research experience pals are physios, then a study using them is more likely to succeed than one requiring radiographers.



**Conspicuous controversy.** To feed or not to feed. To rest or to exercise. To irradiate or to operate.



### 3 broad categories of research question



**Clinical:** most MSc students on this course will select this category, which in many ways is the hardest to do really well. Human variables. Human frailty (yours and your subjects). Human error. Not for the flaky or the squeamish.



**Science:** test tubes don't talk back: nice work if you can get it.



**Epidemiology:** the world is full of amateur epidemiologists who think they are pros. Easiest by far to make a real pig's ear of without noticing.

### Brass tacks



**Common conditions:** MSc = time limited. No time to await people happening by. Also, less interesting to the masses. Harder to get dissemination outlet.



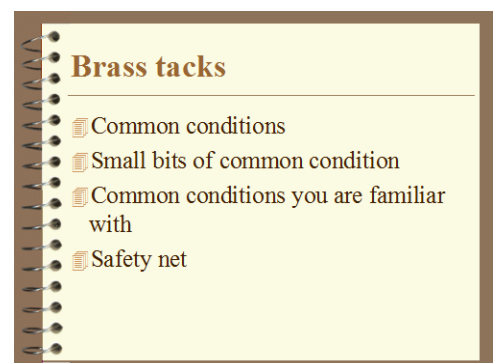
**Small bits of common condition:** common things are commonly researched. Find the gap (in knowledge).



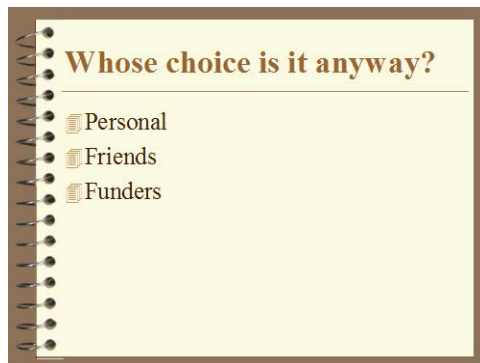
**Common conditions you are familiar with:** you are the expert. Never forget that. New to research maybe, but you know these people.





**Safety net:** always predict the outcome of your study. If you predict a positive result – fine. Positive results are much more likely to be published. (Fact.) But always run the 'what if' scenario too. What if it's negative? What if it's impossible? – What is salvageable? Never EVER go into a research project without a second publication parachute. Even if it's just the review of the subject.




## Whose choice is it anyway?





 **Personal:** helps to be aflame. But, realistically not an absolute requisite.


 **Friends:** if the group is well experienced and are en route, there are many worse ways to start than just jumping aboard.


 **Funders:** as above: there may be calls to submit proposals. Such 'commissioned' research is increasingly common – and why should not the experts influence the course of things, at least where very large sums are involved.


## First Draft


 **On paper:** don't waste too much time drawing project in the air. If it can't be written down, it can't be done. Or, if you can't write it – YOU can't do it.


 **Crib a protocol:** never overlook the obvious. Borrow a protocol of a study. From a friend. From the local R+D department. From a tutor. Close to your area is good, but not essential. At least it puts the heading down:

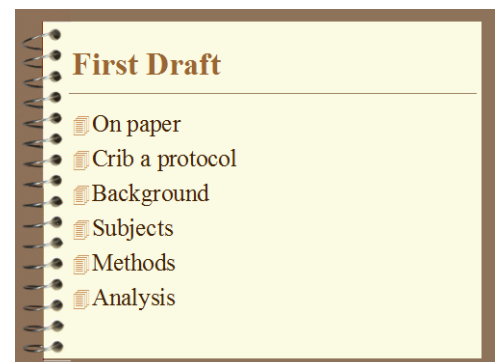
 **Background**

 **Subjects**

 **Methods**


 **Analysis**


 **Dissemination**





## Checklist



 **Feasibility:** nothing fails to succeed like the unfeasible.

 **Timescale:** things take longer than you expect. Getting started takes longer. Recruitment takes longer. Analysis takes longer. Writing up takes longer. Your patients, your coworkers, your librarian, your statistician are also just like you – they have holidays too....

 **Resources:** make sure all is included if you are applying for funding. But don't err on being generous to yourself. Committees respond rather well to good housekeeping.

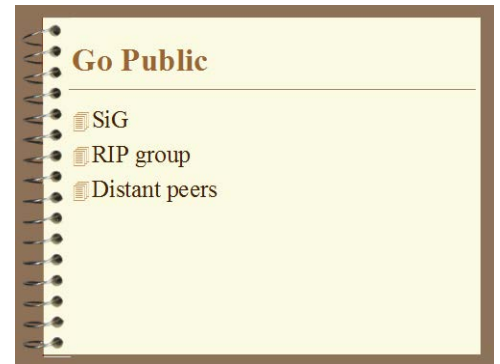
 **Ethics:** for the purposes of this section the key thing to remember is that good science is good ethics. However unintrusive, if your study is rubbish, it's unethical.



## Go Public



**SiG:** pinching research ideas does happen. If you really feel you have a hot potato, you might decide to keep it secret until a meeting presentation with a published abstract that will identify your claim. But –let’s face it, in view of the fact that you will be addressing a very small part of a common question, and perhaps even doing a piece of work commissioned by someone else in the first place....Plus the fact that ideas are NOT the rate limiting step in research and you colleagues will have more than enough to do without your thoughts... **THUS** get as many opportunities as possible to present your work. It may emerge a battered shadow of its former – but better that experience for your design than your results....

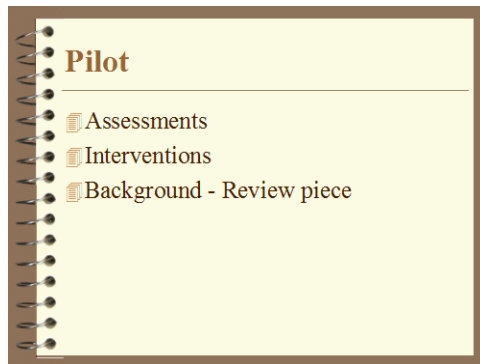


**RIP group:** mutual support by a Res In Progress group is invaluable.



**Distant peers:** if you are a bit lonely, pick up that modem! Even senior peers you have never met may be surprisingly willing to offer advice. Not a detailed critique, necessarily – but this type of activity many regard as part of their general academic responsibility.

## Pilot



**Assessments:** the easiest way to refine a new tool, or to assess how well your subjects will cope with an established one is to pilot it.



**Interventions:** how long will something take? How easily is it documented? What do patients or volunteers think of it? All can be addressed most efficiently by just trying them out.



**Background - Review piece** may be the only publishable part of a study at the end of the day, so don't neglect the pedestrian. Many have regretted postponement of the definitive review until their study is complete – only to turn up a couple of nuggets that would have made life tremendously much easier!!

## Things best avoided by the inexperienced



**Children:** they are surrounded by parents and ethical issues. BUT for the same reason, there is much less research on them. Thus if you do get results, gold dust!!



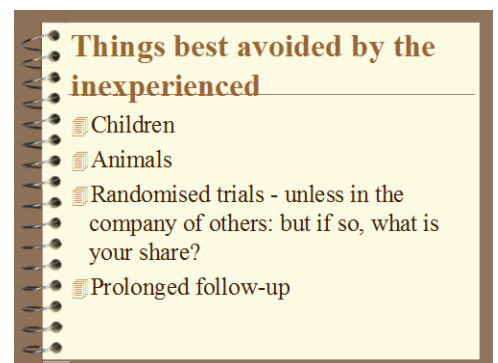
**Animals:** licensing etc is only for the specialist group. Maybe a problem for vegetarians, too.



**Randomised trials** - unless in the company of others: but if so, what is your share?

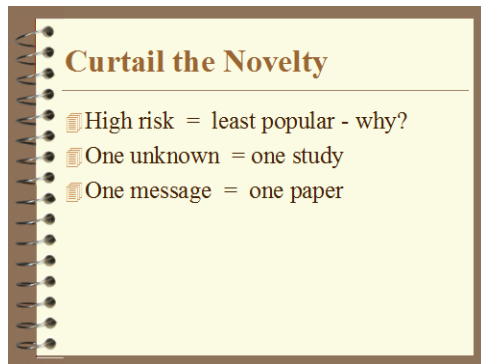


**Prolonged follow-up**





## Curtail the Novelty



**High risk = least popular - why?** We shall discuss this as a group. Some of the reasons are very obvious

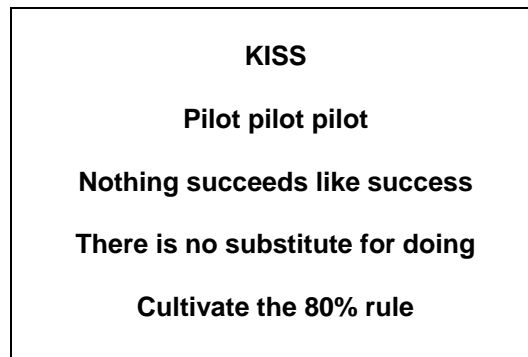


**One unknown = one study**



**One message = one paper**

## Memos



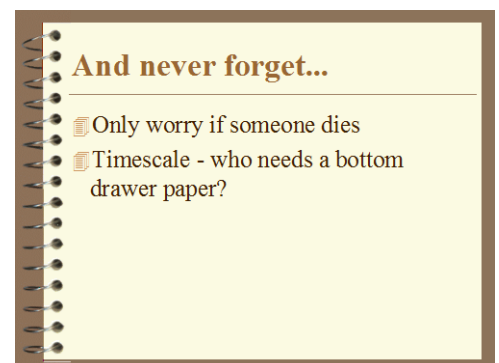
## And never forget...



**Only worry if someone dies.** Research must be ethical, but ethics should not be paralyzing.



**Timescale - who needs a bottom drawer paper?** Even if you are not on an MSc, set yourself a deadline. If you haven't worn that jacket in 3 years.....



# 3

## The basics of study design

# 3

## Introduction to study design

---

Medical research studies are generally long and difficult affairs. Most last at least a year, and many last considerably longer. You would be a bit upset to come to the end of a year's study and find you couldn't answer the question you had posed because of some fundamental flaw in your study design.

### Grant applications and protocols

When you write a grant application for a piece of research, you will inevitably be asked to prepare a research protocol. The protocol is effectively the study design, and will include:

- The research question to be answered;
- The population from which the subjects are to be recruited;
- The number of subjects to be studied;
- The assignment of subjects to sub-groups within the study;
- The experiments or interventions to be made on the subjects;
- The measurements to be made in each subject at each stage;
- The proposed statistical analysis for the data.

Sometimes, some of these things won't be relevant. For example:

- In some studies, all subjects undergo the same treatment; there will be no sub-groups.
- In an observational study, there will be no interventions; you merely watch what happens to the subjects.

However, these points should be made clear from the protocol.

The grant application procedure serves to some extent as a screening process:

- First, it makes you think about your study design;
- Second, if the reviewers are doing their job then they will point out shortcomings in your protocol, and perhaps suggest modifications and improvements;
- Third, if the protocol is truly awful, the application ought to be rejected.

### Write a protocol!

You might someday find you have a research project to conduct, but don't have to write a grant application, perhaps because:

- You're in full-time employment, and you're doing the research as a sideline your clinical job;
- You've found a consultant with a pot of gold and a research interest.

If so, you're in a lucky position. But nevertheless:

***WRITE A PROTOCOL  
and  
INVOLVE YOUR LOCAL STATISTICIAN!***

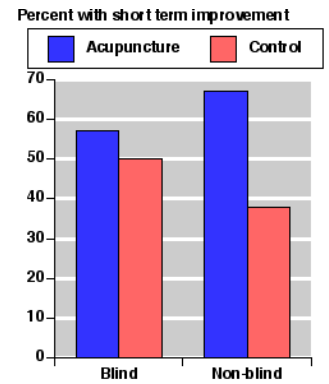
**Statisticians reserve a special place in hell for scientists who run a poor study,  
and who then ask for help in analysing the data.**

In the rest of this section, we'll consider some of the things to think about when you're designing a research study.

## Bias

One of the key criticisms of research studies is their susceptibility to bias. Unfortunately, it is often too late to sort out the problem once you've collected the data, so this is something you have to think about *right from the outset*. Below are some of the main sources of bias in clinical studies. You may not be able to avoid them all, but at least think about them. We're not saying your study has to be perfect, but if you could easily have avoided a situation that later turns out to be a big problem you'll kick yourself.

If you're dubious about bias, the figure (right, copied from *Bandolier*) shows trials of acupuncture with and without blinding (see later). The journal concludes that non-blinded studies over-estimate the true effect by 17%.



### Sampling bias

Sometime early on in the study, you decide the population you wish to study. Thereafter, *every person in the population should have the same chance of being selected*. For example, you may wish to study all subjects with a particular condition. Your clinical head (Professor ...add your own name here...) has agreed to recruit your subjects. Of course, said Professor is sent all the intractable cases while the junior doctors and registrars deal with the others. Your sample is now biased towards difficult patients, who probably won't do so well as a randomly chosen cross section.

### Volunteer bias

Volunteer bias arises when the subjects themselves have the power to include or exclude themselves in the study. A good example would be the 1992 general election; Neil Kinnock (Labour) was ahead by about 20 seats in the exit polls, but John Major won the election with an overall majority of 21 seats. It seems that people were happier to volunteer their opinions if they were a Labour supporter, and effect dubbed *Shy Tory Syndrome* by the pollsters. This can happen in experimental studies. For example, people with headaches that don't respond to other painkillers might be keen to participate in a new trial, but might well do worse than a random sample of subjects. Since subjects usually have the power of veto in a research study, volunteer bias can rarely be discounted completely.

### Allocation bias

Allocation bias arises where patients are assigned to different groups in a manner that is not completely random. For example, the researcher may be inclined to assign difficult patients to the control group, on the grounds that they are less likely to comply with treatment. Proper randomization and blinding should avoid allocation bias.

### Response bias (the placebo effect)

Also known as the placebo effect, a subject's response is biased by the belief they are being treated.

### Assessment bias

Similar to allocation bias; the researcher knows which group the patient is in, and might be influenced in the interpretation of the outcome. This can apply in any study where measurements are open to interpretation. Studies with randomized, blind assessment should avoid this bias. At least, data should be analysed without knowledge of the source.

### Intention to treat

Once a patient has been admitted to the trial of a treatment, the *intention to treat* has been established. From here, the data should be analysed as if the patient received the treatment, *whether or not they really did*. This fits in with the *pragmatic* trial described later. The alternative analysis is *on treatment* ie. subjects are analysed according to whether they actually had the treatment, as described for the *explanatory* trial. This introduces the volunteer bias described earlier. The ones who actually comply and undergo the treatment have somehow selected themselves, perhaps because they are the sicker/not-so-sick/more compliant/older/younger subjects. This may have implications for outcome.

### Lost to follow-up

In a similar vein, patients are sometimes lost to follow-up. They might have died from their condition, or been so disenchanted with their treatment that they refuse to return for follow-up. Alternatively, maybe they suddenly recovered and took a round-the-world cruise. You can imagine that this will introduce some bias.

## Observational studies

---

As you already know, a key distinction is between observational and experimental studies. In an observational study, you *simply observe the way things are, without intervening in any way*. It's fair to say that most of the studies you will deal with will be experimental, but observational studies are used in the following circumstances:

### Cross-sectional studies (retrospective)

In a cross-sectional study, you take a cross-section of the population you are interested in (the sample), and measure something about them. Cross-sectional studies are often used in epidemiology. If you wanted to quantify the incidence of undiagnosed reflux in the community, you would use a cross-sectional study. It's important that your sample represents a true cross-section of the population, and that each person in the population has the same chance of being recruited. For example, you might pick names at random from the register of electors. But of course, this would exclude children, the homeless, those recently moved, pending visa status etc.

### Cohort studies (longitudinal, prospective)

There are times when an experimental study could not be justified on ethical grounds. The link between cigarette smoking and heart disease would be a good example. In a cohort study, you might recruit a group of children at birth; they form the cohort. You follow these subjects at intervals through life, keeping track of:

- Their smoking habits;
- The presence and degree of any heart disease.

You might then show a link between smoking and heart disease. The cohort study would be considered:

- Longitudinal, because it studies changes with time;
- Prospective, because the subjects were recruited before you knew what was going to happen to them.

A cohort study is susceptible to *confounding factors*. Perhaps overweight subjects who don't do any exercise are most likely to start smoking, and it is the weight and lack of exercise that are the *real* risk factors for heart disease. This argument was used by the pro-smoking lobby for a long time with great success. The weight and exercise regime are the *confounding factors*.

A cohort study can take a long time to complete, because you have to wait long enough for the subjects to expose themselves to the risk factor (smoking), and then wait to see the heart disease manifest itself. In the example given, you could not obtain meaningful results in less than 40-50 years.

### Case-control studies (retrospective)

The case control study can be used in similar circumstances, but can be completed much more quickly. Take the same example. In a case control study, you would:

- Recruit a group of subjects with heart disease (the cases);
- Recruit a second group of subjects with no heart disease (the controls).

Now, you ask each subject about their smoking habits. If smoking is associated with heart disease, then you would expect more of the *case* group to be smokers than the *control* group. This is a *retrospective* study, because you recruited the subjects after bad things already happened to them; in a sense the patients select themselves, which can introduce bias.

### Matched controls

As described above, your study is still susceptible to confounding factors. Perhaps your *case* group are heavier and less active than the *control* group, and that weight and lack of exercise are the real risk factors. In a case-control study you match your controls with the cases:

- Recruit a group of subjects with heart disease (the cases);
- For each subject with heart disease, recruit a second subject with the same sex, age, height, weight and exercise habits, but *with no heart disease* (the controls).

Very often, matched controls will be recruited from the same GP's surgery as the case, because the GP will have the information on height, weight, age, etc. You can match for all the factors you think might be important, but you might struggle to find a 6'3" tall, 80-year-old lady who runs marathons. The more factors you match, the more difficult it will be to find a suitable control.

## Experimental studies

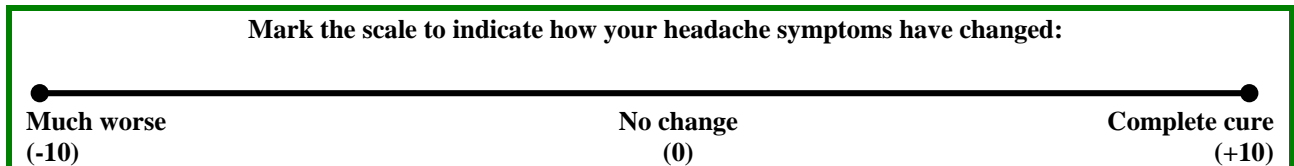
In an experimental study, you do something to your patients that wouldn't normally happen to them, and watch what happens. For example, you might:

- Give them a drug;
- Perform a surgical procedure on them;
- Ask them to swallow boluses of water, yogurt and biscuit.

For the sake of illustration, let's consider a fairly simple and common study - a drug trial. Suppose you want to investigate the efficacy of a new painkiller for headaches. As usual, start with the hypothesis:

*We hypothesise that an improvement in headache is linked with being given the new painkiller.*

This isn't a great hypothesis. How do you measure 'improvement'? Well, you could perhaps measure subjective improvement on a visual-analogue scale like the Newcastle Headache Assessment Scale, HAS:



Now you can write a testable hypothesis:

*We hypothesise that taking the new painkiller is associated with an improvement in HAS score.*

And a null hypothesis:

*We hypothesise that the new painkiller has no effect on HAS score.*

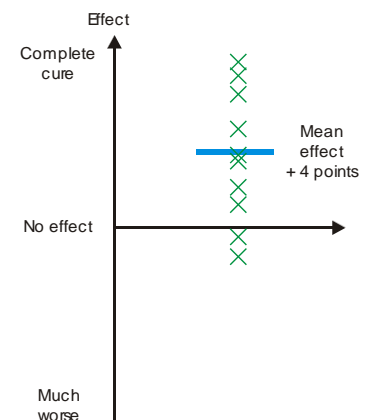
## A naïve study

Your first attempt at a study design might go something like this:

- Recruit 10 subjects with a headache;
- Administer the new painkiller;
- After an hour, the subject completes the headache assessment scale.

Your results for the 10 subjects are shown (right). You observe that:

- The HAS measurements are largely positive. Using a suitable statistical test (*the single-sample t-test, see later*), the chance of the null hypothesis being true, and these results arising by chance alone is 1 in 50 ( $p=0.02$ ). You therefore reject the null hypothesis and conclude that the effect is real.
- The mean effect is +4 points. This seems like a worthwhile and clinically important effect. The new drug is marketed.



## What's wrong with the study?

In this study, you concluded that any outcome better than 'no effect' would vindicate the use of this drug. You have made the assumption that the improvement was due to the drug, but there are other possibilities:

- Headaches don't last for ever. The patient might have improved anyway, even if you hadn't given the drug.
- The improvement wasn't due to the drug itself, but was produced simply because the patient knew he or she was being treated. This is the *placebo effect*.

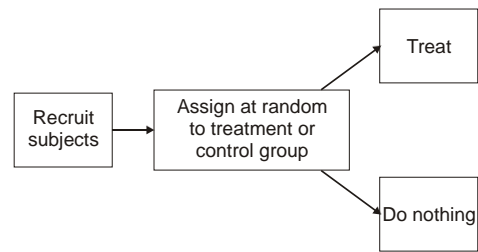
## The randomized controlled trial

We'll now present the randomized controlled trial. This has achieved a certain status as the right way to do things, particularly for drug trials, because it is felt to be the best way to avoid biases in a study design. You might not be doing a drug trial, but the ideas extend to all types of clinical research.

## A randomised controlled trial (RCT)

The general idea is this:

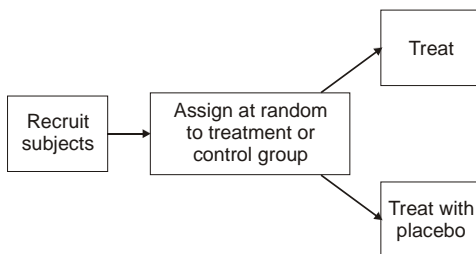
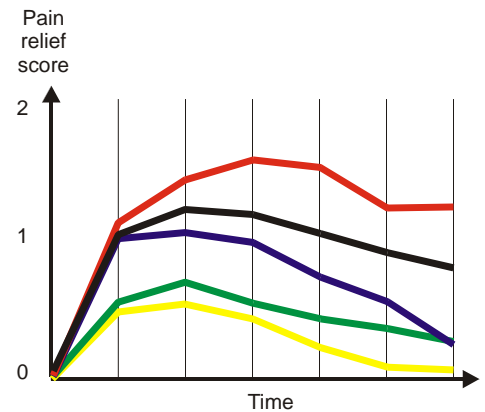
- Recruit a group of subjects;
- Assign subjects at random to the *treatment* or *control* group;
- The *treatment* group are treated, the *control* group aren't;
- Perform a statistical test to see whether outcome is associated with the subject's group (*treatment* or *control*).



If the patients were going to get better anyway, then patients in the *treatment* and *control* groups should *both* get better. This way of doing things is known as the randomised controlled trial (RCT), because the patients should be allocated to the *treatment* or *control* group at random. Only if patients are allocated *at random* can you say for sure that the two groups were the same before treatment, and claim with any conviction that the effect was due to the treatment.

### Placebo control

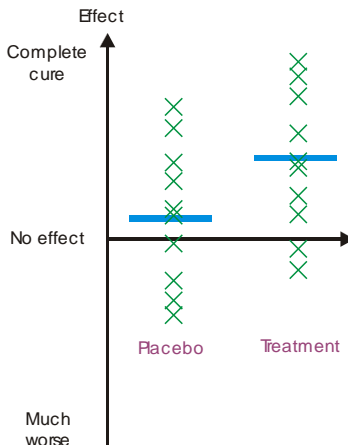
The controlled study (*above*) will deal one of the problems with the naïve study. However, it doesn't deal with the *placebo effect*, because the *control* group are aware they haven't been treated, and so wouldn't be subject to the placebo effect. If you're in any doubt about whether the placebo effect is real, look at the figure (*right*). In 1974, Huskisson *et al* assessed aspirin (*black*) against red, green, blue and yellow placebos. The outcome was measured hourly for 6 hours. The message is clear. When you've got a headache, choose a red tablet.



The solution is a placebo control. A placebo gives the impression of treatment to the subject, but has no therapeutic properties. In a drug trial, it will probably be similar in appearance to the real treatment, but will lack the active ingredient.

The outcome data might look group do slightly better than 'no effect', You can demonstrate this formally *sample t-test, described later*). This is considered:

- Prospective: you allocate the
- Longitudinal: you observe



something like this (*right*). The *placebo* and the *treatment* group do better still. using the proper statistical test (*the 2-* a well-used study design, which would be

subjects to the study groups; what happens with time.

### In summary - the randomised controlled trial

The most important feature of an RCT is this:

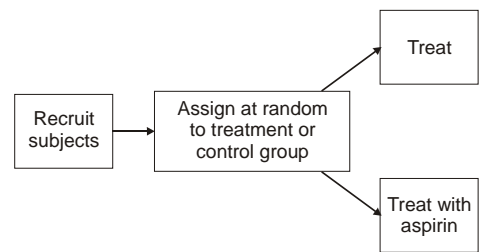
**All subjects should be drawn from the same population, and allocated to study groups at RANDOM!**

RCTs are often used in relation to drug trials, but the guiding principles of random selection and allocation are useful *in many situations*:

- Selecting X-ray images for a study of inter-observer agreement in assessment. You should pick the images at random from all those available, and have the raters assess them in a random order.
- Evaluating a new diagnostic test against the existing gold standard. Here, your control measurement is the old test, which is being evaluated against the new test. If your subjects can't have both tests, they should be allocated to the old or the new test at random. If they can have both tests, give them in a random order.

## Picking the control

If your new drug were first on the market, you'd probably run a placebo-controlled trial. In reality you are in competition with Aspirin, Paracetamol, and all the other headache remedies on the market. In this case, it's no good demonstrating that your new drug is better than placebo; you would want it to be better than the current treatment of choice. In this case, your control group might take aspirin or paracetamol.



### Control in your study

If a control group is appropriate in your study, you can use the same general approach. For example, if you are evaluating a new diagnostic test, then the most appropriate control would be the existing gold standard test.

*Note in passing that diagnostic accuracy might not be the only outcome. For instance, if your new test is fast or has significantly lower risk for the patient, then that might be measured as an outcome too. Screening programmes are often based on tests that have relatively poor diagnostic accuracy.*

## Blinding

Given the placebo effect, it's important that the subject doesn't know whether they are in the *treatment* or *control* group. Even where a true placebo isn't possible (for example, in a trial of a surgical technique), the patient might be given some passive therapy (ultrasound treatment, for example) for placebo. A study where the subject doesn't know their group (*treatment* or *control*) is *single blind*.

*Blinding* is more generally applicable. In particular, it is preferable the researcher recruiting the subjects and/or analysing the outcome data doesn't know which group the subject is in. It is suggested that:

- The researcher might give subtle clues to the subject as to which group they are in;
- More overtly, the researcher might preferentially allocate certain subjects to a particular study groups (*allocation bias*);
- If the outcome data are subject to interpretation, the researcher might be swayed by knowing the patient's group (*assessment bias*).

The situation where the researcher doesn't know the patient's history would be known as a *double blind* trial. It's easy to arrange a double-blind drug trial, where the *treatment* and *control* groups are revealed only after the study is complete. It's difficult to arrange double blind surgery, since one would hope the surgeon at least would know the procedure being carried out. Nevertheless, researchers aspire towards this situation.

***If you have different experimental groups, allocate your subjects in random order.***

### Blinding and assessment bias in your study

There is one key area where blinding can and should be applied in most studies - to counteract assessment bias. Assessment bias typically arises where the worker analyzing the results from the study knows where the data came from. In particular, it's important you don't know anything about other measurements in the same subject. For example, in a study of voice changes with age, knowledge of the subject's age would despite your best intentions *affect your judgement of the voice*.

The only circumstance where assessment bias is not a potential problem would be where the outcome measurement is *completely objective and not open to any interpretation*: for example, *survived* or *died*.

The best solution is to be sure the person analyzing the data doesn't know anything about the source of data.

- This might mean that someone not involved in the study performs the analysis;
- Alternatively, you could arrange for a colleague to put all your data in random order before you make the rating.

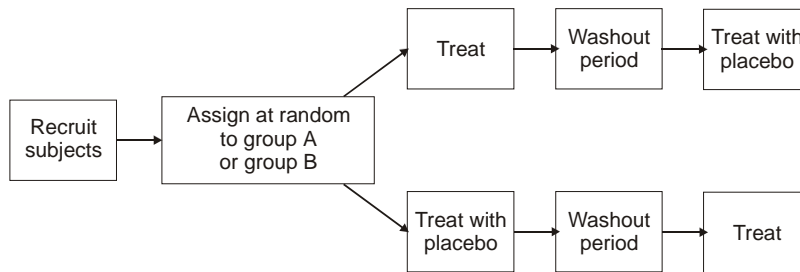
Analysing your data in random order is always a good idea. Otherwise, you might get better with practice, biasing your results towards the later measurements. This is even more true if you are sharing the work in analyzing the data. We know (for example) that there are *huge* differences in assessment of sleep stage. If you always analyse a pre-treatment study, and your colleague always analyses the post-treatment, you might find the patients were suddenly and remarkably cured of their sleep disorder.

***Analyse your results in random order, without knowledge of the source of data.***



## The crossover

As you might imagine, it would be most useful if the control group was identical to the treatment group. You could then be absolutely sure that any effect you found wasn't due to some unexpected difference between the groups. In the discussion of the case-control study, one approach was discussed. For every *treatment* subject, you recruit a *control* subject who is matched for sex, age, weight and any other variables you think might affect the outcome.



In many experimental studies it is possible to make *both measurements in the same subject*. If so, you can use a crossover design. The idea is like this (*left*). Each subject receives the active treatment *and* the placebo, though the order is again random.

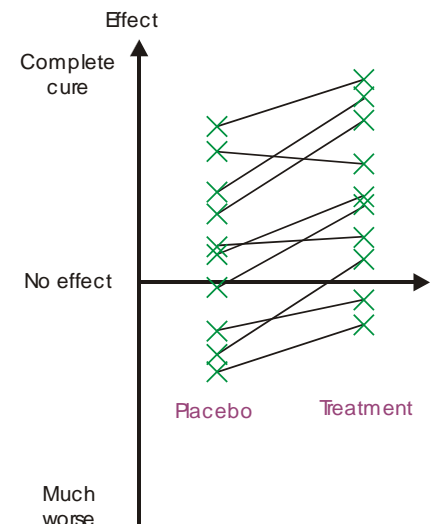
Between the active and placebo treatments there is a washout period, long enough to ensure the effect of the first treatment has worn off before the second treatment starts.

With this design, each subject receives both treatments and acts as their own control. *The treatment and control groups are one and the same.* Any effects can be attributed to the active treatment, because there are no other differences between the groups.

The results might be as shown (*right*). If you ignore the black connecting lines for the minute, there is only a slight overall difference between placebo and treatment, with a lot of overlap. However, the measurements are paired, as shown by the black lines. All-but-one of the subjects do better with treatment than placebo.

If possible, it's best as shown to allocate the subjects at random to group A (*treatment then control*) or group B (*control then treatment*). This will control for the possibility that people get better with time in which case, the later treatment will always look the best.

As you'd expect, the crossover study is generally a more sensitive way to detect effects than the earlier controlled studies. There are statistical tests (*Student's paired t-test*, or *Wilcoxon's test*) that are particularly appropriate for crossover studies.



### Applying the crossover design in your study

The crossover method is useful in any situation where you are making two or more measurements in a single subject. For example, you might be evaluating a new diagnostic test for aspiration, and so it would be appropriate to apply the old and new tests sequentially for comparison. Perhaps the subjects become more relaxed after the first test; if you always apply the old test first, the second test will come out rosier than it ought to.

The solution might be:

- To perform the tests in a random order, as shown earlier;
- In some cases, you might even be able to perform the tests *at the same time*. For example, with some effort you can perform videofluoroscopy at the same time as nasendoscopy. So then, you know you are assessing *the same swallow at the same time*. Any differences from one swallow to the next would disappear.

For a second example, perhaps you're measuring agreement between expert observers in evaluating some diagnostic images. Again, this is a crossover design; each rater is being exposed to all the different images. So make the order random; don't present (for example) all the normal examples first.

***If you're making two (or more) measurements in one subject for comparison - make them in a random order.***

You'll appreciate that in some cases it isn't possible to perform the two measurements in a random order, perhaps when the treatment is irreversible (in surgery, for example). You have to make the control measurement first, then apply the treatment. If so, think about whether you could make two consecutive control measurements. If these were similar, it would provide some evidence that your subjects weren't changing with time.

## Explanatory and pragmatic studies

The experiments described so far might be termed *explanatory* studies, because the investigators try to explain the effect of a treatment in tightly controlled circumstances. However, when many treatments or diagnostic tests become part of clinical practice, their performance is not as good as expected from the earlier trials. Of course, there are a number of differences between administration of a clinical trial and use of the same method in day-to-day clinical practice.

- Patients who are part of a research project are typically given more expert attention than the corresponding patients in the community;
- Patients are aware of being part of a research study, which may have a super-placebo, super-compliance effect;
- Patients are monitored closely for compliance with the treatment;
- Data from patients who don't comply may be eliminated or analysed differently.

A *pragmatic* trial will try to establish the likely effect of a treatment or test *in clinical practice*. The term *intention to treat* has relevance here. Once a patient has been recruited to the study, their data are analysed as if they completed the study, whether or not they actually comply. This type of study is likely to be particularly unkind to treatments with a complicated or time-consuming regime or with unpleasant side-effects.

## Randomisation

Given that just about everything needs to be in a random order, how do you assign your patients to *treatment* and *control* groups? Well, for true randomness you could do worse than tossing a coin, but unfortunately a coin is often just *too* random. You might end up with this situation (*right*). There are 20 subjects but only 5 have ended up in the *treatment* group. You can see it's now not so easy to draw conclusions. In the extreme case where *all* the patients end up in one group, you can't say anything at all. It turns out the best situation is *the same number of subjects in each group*, and you'd like to arrange that if possible.

One way would be like this:

- On ten scraps of paper, write *treatment*;
- On another ten scraps of paper, write *control*;
- Mix the twenty pieces of paper in a hat;
- To randomise the next patient, pull out the next piece of paper, read it and then throw it away.

This works well, except you might still find for example that the *control* group are allocated towards the beginning of the study, but the *treatment* group are allocated nearer the end. This might be a problem if:

- You're not sure from the outset how many patients you will study;
- Things change with time. Perhaps the surgeon becomes more skilled with a new procedure.

In either case, it's better to use a block-randomised design.

### Block randomization

This is best illustrated with an example as before. In a study of swallowing, you think subjects might get better with practice. You're going to randomise 20 patients into two groups *yogurt then water* or *water then yogurt*;

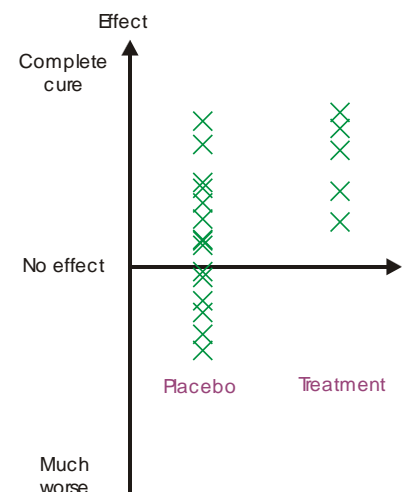
- On two scraps of paper, write *yogurt then water*;
- On another two scraps of paper, write *water then yogurt*;
- Mix the four pieces of paper in a hat;
- To randomise the next patient, pull out the next piece of paper, read it and then put it to one side.

You've now randomized the first four patients, but *two will be in each group*. To complete the randomisation, put the four pieces of paper back in the hat and repeat the process until you have twenty patients. Now:

- After every multiple of four patients, there will be equal numbers of both groups.
- At any time, the numbers of patients in each group can be no more than two different.

### Randomisation at outcome

When you come to analyse your data, it's best if possible to use a random order. You might ask a colleague to number the traces and keep the key secret. If you've got two groups, it's again better to use a block-randomisation scheme to make sure you don't get all the of one group early on when you're still a bit rusty.

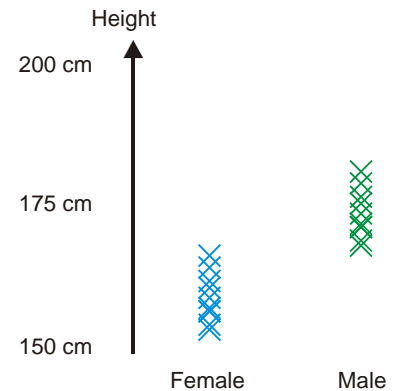


## Sample size

Deciding on a sample size is a difficult problem. Let's take another simple example. Suppose you have the following hypothesis:

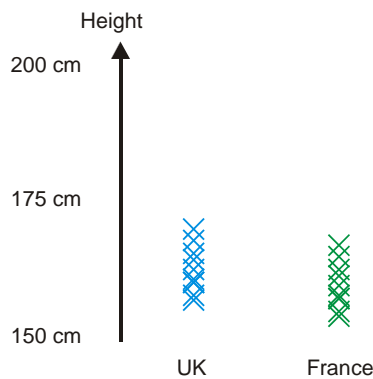
*We hypothesise that a subject's height is linked to their gender*

You do the obvious experiment, measuring 10 randomly selected men and 10 women, and here are your results (*right*). Using the proper statistical test (*two-sample t-test*), you show unambiguously that there is a relationship between gender and height.



With your success, you raise a new hypothesis:

*We hypothesise that a woman's height is linked to her country of origin*



Once again, you study ten subjects but this time (*left*), there's only the merest hint of a link. The solution is to study more subjects. As you saw earlier, your estimates of mean height become more precise as you study more subjects.

But here's a dilemma: how would you know how many subjects to study in order to detect the effect? After all, 10 would have been enough for the first study, but you might need 100 or more for the second study.

Well, you can work it out, but you need to know two things:

- how big the overall effect is;
- the variability between patients, perhaps the standard deviation of height – *see later*.

So generally, you need more patients when:

- The effect is small as in the *UK v France* example above;
- There's a lot of random variability between patients, a high standard deviation between measurements.

But here's the next problem - how do you know the effect or the variability, when you didn't do the study yet?

- Preferably, you perform a small pilot study to determine the likely effect and its variability.
- Guess, using the available literature.
- Failing this, pick the smallest effect that would be of clinical importance. If the effect is too small to be of clinical importance, you don't really care if you don't spot it.

This whole procedure – figuring out how many subjects to study - goes by the name of a power calculation.

## Type I errors, type II errors and power calculations

---

### Type I errors

If you remember:

*Assume the null hypothesis is true - that there is no link in your study.  
The p value is the probability that the effect you actually observe, however big or small, is due to chance alone.*

Suppose you will reject the null hypothesis and conclude you really have found a link on the basis of  $p = 0.05$ . To put it another way, there is a one-in-twenty chance that the result was just a fluke, and there really is no link. You have conducted the study where the observed link really *was* due to chance. This is termed a *type I error*.

### Type II errors and the power calculation

Of course, you could make the opposite error; you might fail to find a link that really is there, simply because you didn't study enough subjects. This is termed a *type II error*. So ... before starting a large-scale study, you should perform a power calculation.

*Statistical power is the probability of demonstrating that your hypothesis is true, assuming that it really IS true.*

In a study with poor statistical power, there is a danger that you *don't* reject the null hypothesis because you can't be sure enough that the observed link is real. It would be wasteful and possibly unethical to conduct a low-powered study where you had little real chance of spotting the effect you were looking for.

Statistical power increases with sample size. Typically, you would like a statistical power of at least 80%, and you can then work back to determine how many patients must be studied. As a rule, power calculations are difficult and best left to statisticians; the exact method depends on which statistical tests you will be applying to the data.

However, here is a useful web site...

<http://www.stat.uiowa.edu/~rlenth/Power/index.html>

...and here are some simple examples to give you an idea of the numbers involved:

*If the mean effect is the same size as the variability between subjects, you need just 7 subjects*

*If the mean effect is half the size of the variability between subjects, you need 18 subjects*

*If the mean effect is one fifth the size of the variability between subjects, you need 99 subjects.*

*If the mean effect is one tenth the size of the variability between subjects, you need 387 subjects.*

There isn't much you can do about the effect size – this is what you are trying to measure. However, notice that the number of subjects rises **DRAMATICALLY** as you increase the variability.

In this context, *variability* means any effect that is not consistent from one measurement to the next. This is something you **DO** have control over:

- Variability **BETWEEN** patients. This is frequently very large, and in some cases can snooker the study completely. The best solution is to use patients as their own controls in some sort of crossover design. This then completely eliminates the effect of variability between patients.
- Variability **WITHIN** patients. If you measure the same thing on the same patient on two occasions, you are unlikely to get the same answer. Think of blood pressure measurements. Wherever possible, you should study patients under the same environmental conditions, at the same time of day, etc. Try to think of all the things that will affect your measurements.

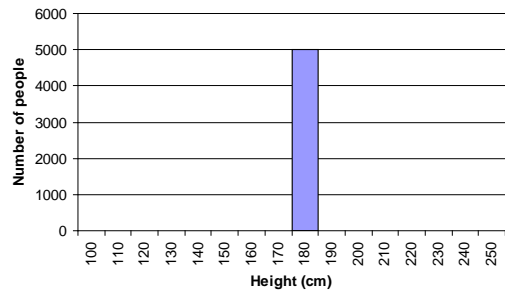
# 4

Studying the behaviour of random or  
noisy variables

# 4

## How do random variables behave?

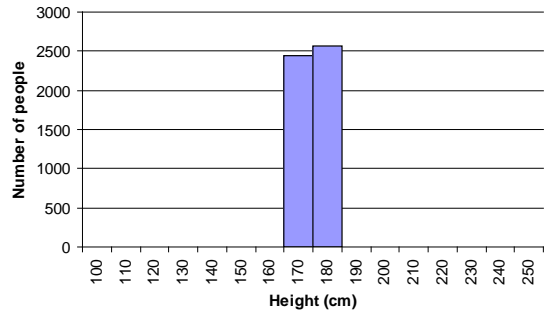
To get a bit further we now need to study how random variables behave, so let's consider one of the best-known examples of variability, peoples' height. Suppose for a moment that people were produced like boxes on a production line, and everyone comes out the same. If you measured 5000 heights, you could plot the results as a *histogram* like this (*right*). There are 5000 people with exactly the same height, and nobody else.



Of course, heights aren't really like this. As you probably know, height is largely dictated by genetic factors, so let's try to simulate that (*I used MS Excel to perform the simulations*).

### A simple person with 1 gene

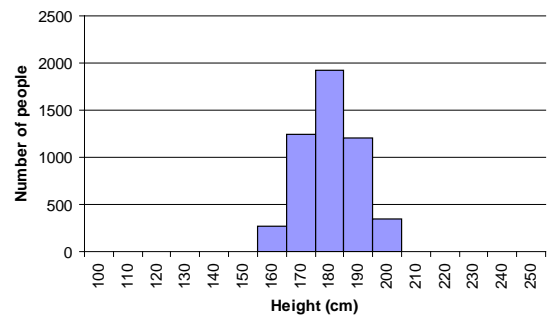
First, here's a very simple simulation with just one gene. The effect of the gene is this. If it's absent, it decreases that person's height a little bit. Here's what happens (*right*).



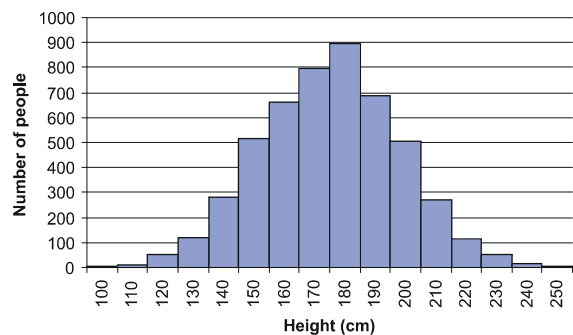
We've now got a group with two heights, depending on whether the gene was present or not. The numbers of people with each height are roughly equal, as you'd expect.

### If people had a few more genes

You can now introduce a couple of extra genes, each of which can either increase or decrease the height a bit. There are still 5000 people in total but as you'd expect, the overall variety of heights increases. However, the distribution is not even. Most of the people lie towards the centre, with relatively few towards the edges.



The bottom figure again simulates a population of 5000 people, but this time with 20 genes. There are about nine hundred people with a height of 180 cm, but less than one hundred with a height of 120 cm.



It's clear what's going on. Since each gene acts randomly, their effects tend on average to cancel each other out. But, for a person to have a height of 120 cm, a big majority of genes must be acting together to reduce the height, and that seems unlikely. It's like tossing 20 coins and getting 18 heads.

It's no surprise, then, that people *really do* show a distribution of heights like this.

In fact, many many things in nature show this distribution with some spread about a central *mean* value. It can be explained very simply as the combined effect of lots of individual factors each having a random effect on the outcome.

As you'll probably be aware, this is called a *Normal* or *Gaussian* distribution.

***If you add together enough independent random values, the result will obey the normal distribution.***

## How does the normal distribution help?

### Parameters of the normal distribution

A normal distribution can always be represented by just two parameters:

- The mean value;
- The standard deviation (SD) about the mean value.

As you'll know, the mean is just the average of all the measurements. In the example (*right*) the mean is about 175 cm.

The standard deviation is a bit more complicated. Your stats package will calculate it for you, but here's how to do it:

- For each measurement in turn, find the difference between it and the overall mean.
- Square all these 'difference' values (5000 'difference' values, in the example).
- Find the mean of all the 5000 squared values.

The number you've got now is the *variance*.

- Find the square root of the variance.

This is the *standard deviation*. In the example, the standard deviation is about 22.5 cm.

Why bother explaining this? Well, as you can see from the rules, the standard deviation is related somehow to the distance of your measurements from the mean:

- If the measurements are generally close to the mean, the standard deviation will be small.
- If the measurements are generally well spread around the mean, the standard deviation will be large.

So the standard deviation gives some indication of the spread about the overall mean value.

### What use is this?

It turns out that the *mean* and the *standard deviation* are the *only* things you need to know about a normal distribution. The normal distribution for mean = 175 cm and SD = 22.5 cm is superimposed (*right*). Now, you can calculate or read off the graph that there ought to be around 45 people with a height of 120 cm. This agrees quite well with the actual result.

You can make the following generalisations:

- Two thirds of all the measurements will lie within  $\pm 1$  standard deviation of the mean value.
- 95% of all the measurements will lie within  $\pm 2$  standard deviations of the mean value.
- 99.9% of all the measurements will lie within  $\pm 3$  standard deviations of the mean value.

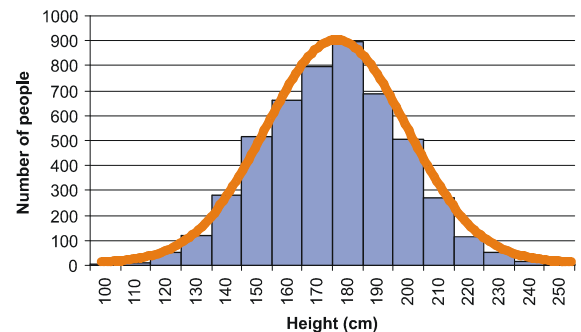
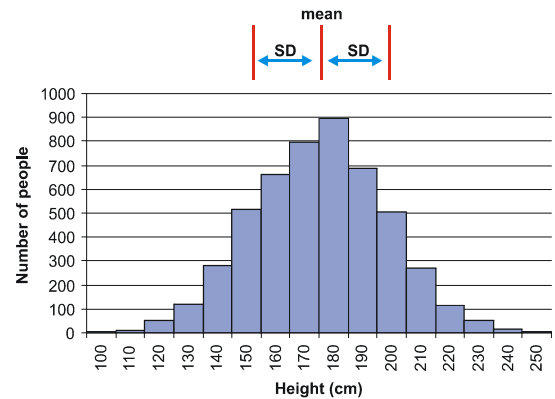
### A simple statistical test

Suppose a person turns up in your clinic with a height of 120 cm. You observe other developmental problems, and wonder whether their height is within normal limits, but all you know is the mean and SD of the population's height. From the normal distribution, you can determine that about 7 in 1000 people would have a height of 120 cm or less. In other words, the probability of this height being due to chance alone is 7 in 1000 ( $p = 0.007$ ). This supports your suspicion of developmental difficulties.

### Standard deviations and the normal range

BEWARE - there are 60 million people in Britain, so 420,000 of them *will* be below 120 cm *purely due to chance alone*.

***Having a statistically unlikely height (or any other clinical measurement) does not itself constitute abnormality!***

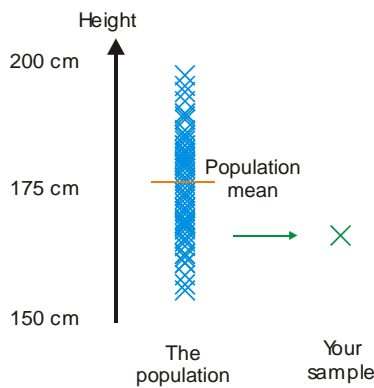




## Populations and samples

When you talk about the mean and standard deviation, you would like to talk about the *population*. For example, we said earlier that the mean and SD of the population's heights are 175 and 22.5 cm respectively. Of course, it's not reasonable to go out and measure everyone's height, and so you would need to make a sample. You will go and recruit perhaps 100 people and measure their heights; this is your *sample*. You hope your sample is representative of the population as a whole, but there are a couple of things you have to get right:

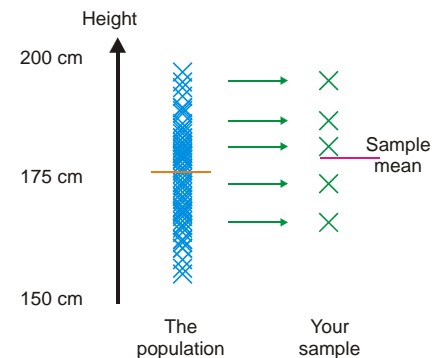
### The number of subjects in the sample



Take the extreme case where you take a sample of just a single person (*left*). There's a good chance your one sample will be quite different to the true population mean.

When you sample more subjects (*right*), the errors will tend to cancel out, and your sample mean will be closer to the true population mean.

Just as earlier, a study with more measurements is better.



### The standard error in the mean (SEM)

The bigger your sample, the more representative it will be of the population, and the more *precise* your estimate of the true population mean. There is a statistic called the *standard error in the mean (SEM)*. You calculate the SEM like this:

- Calculate the standard deviation of all the measurements in your sample.
- Divide it by the square-root of the number of measurements in the sample.

The number you get is the SEM, and tells you something about how *precise* your sample mean is:

- There is a 2 in 3 probability that the *true population mean* is within  $\pm 1$  SEM of the calculated sample mean.
- There is a 95% probability that the *true population mean* is within  $\pm 2$  SEMs of the sample mean.

### The confidence interval (CI)

The latter statistic is so well-used that it is often called the 95% confidence interval (CI). The importance will become clear later, because it is becoming the accepted way of reporting the outcome of statistical tests.

For example, at one time you might have read or written this in a report: *Following administration of the drug, there was a reduction in blood pressure ( $p < 0.05$ ).*

What you really mean by  $p < 0.05$  is this: *There was a reduction in blood pressure. The chance that this was due to random variation, and NOT to the effect of the drug, is less than 5%.*

***BUT REMEMBER*** – by studying lots of subjects, you can spot even a very tiny effect. There is no mention here of the size of the drug's effect, which is what you are really interested in.

***Statistical significance means ONLY that the link you found is unlikely to be due to chance alone. It is up to you to decide whether the effect is large enough to be of any clinical importance.***

Now, you would be encouraged to do it like this: *Following administration of the drug, the mean reduction in blood pressure was 20 mm Hg (95% confidence interval 15 to 25 mm Hg).*

What you really mean is this: *From my sample, the best estimate of the true effect of the drug is 20 mm Hg. If I repeated the experiment, I would probably get a slightly different answer. However, I am 95% certain that the true effect is in the range 15 to 25 mm Hg.*

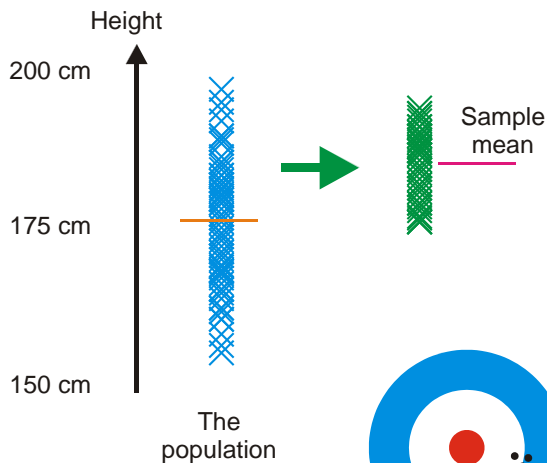


## Sampling bias

The enemy of good sampling is sampling bias. There is one fundamental rule to unbiased sampling:

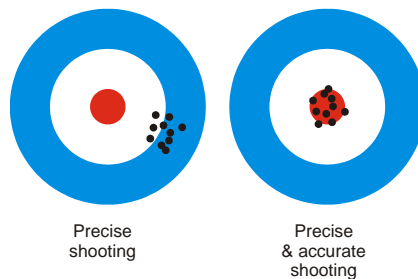
***Everybody in the population being studied should have exactly the same chance of being included in the sample.***

You want to know the height of the adult population, so one Tuesday you sample 100 adults at random from the streets of Newcastle. Therefore, *you have excluded everybody who isn't on the streets of Newcastle on a Tuesday*. The people excluded will tend to be:



- Older or housebound people;
- People who don't live in Newcastle;
- People at work on Tuesday.

So you end up with a biased sample of youngish, unemployed North-Eastern people. As a result, the people in the sample might tend to be taller than the population as a whole. With 100 measurements you end up with a *precise* estimate of height (the confidence interval is small), but it isn't an *accurate* estimate because it misses the target.



This whole idea of *reliability and validity*. A reliable instrument will always come up with the same answer, but it is only valid if it comes up with the right answer. A stopped clock is reliable but not valid.

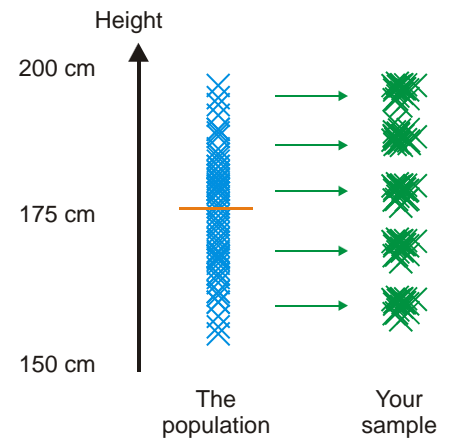
precision and accuracy comes up again later

## Degrees of freedom

Back to the same problem of measuring heights. This time, you decide you need 100 heights to get a good estimate of the population mean, but you leave things a bit late and the streets are looking empty. So instead, you decide to recruit five people, but to measure them twenty times each.

This seems to be cheating but after all you've got 100 measurements, and your stats package still tells you the CI is very small. So surely, the job is done. Once again, you only have to plot the data to see the problem (*right*).

Sure you've got 100 measurements, but they're not *independent*. They are clustered into 5 groups. In fact, the only reason there is any variability *within subjects* is that there's a bit of measurement error that shows up when you measure a person repeatedly.



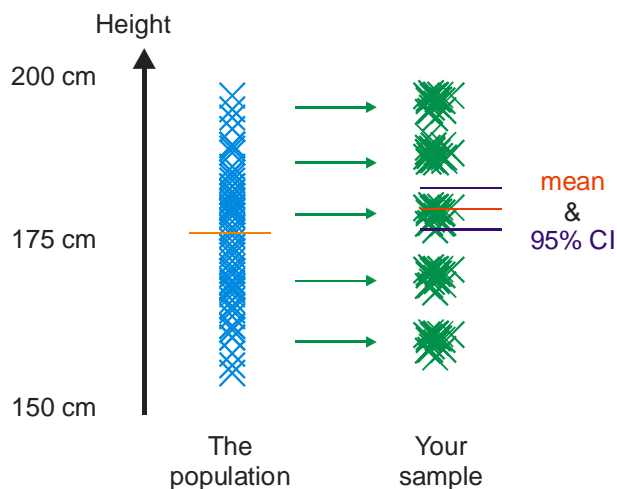
### What is a degree of freedom?

*Degrees of freedom* give endless trouble for even well-experienced researchers. It's very difficult to give an accurate definition, but here's an attempt.

***The number of degrees of freedom in a sample is the number of independent measurements in that sample.***

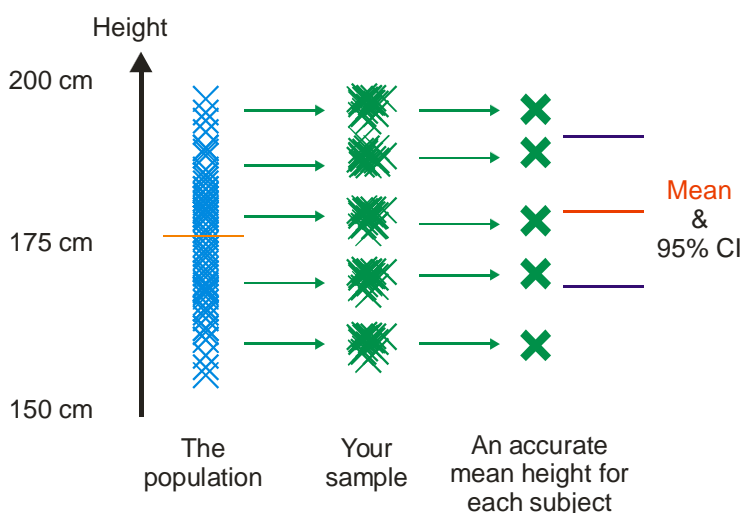
In the example above, you made 100 measurements, *but there are only 5 degrees of freedom*. That's because each person produced 20 measurements that were very closely related to each other and not at all independent.

### How should you analyse the data?



It would be tempting to type the 100 measurements into your stats package, and ask for the mean and 95% CI. Remember that the CI becomes smaller as you make more measurements. The stats package would assume that the 100 measurements are all independent, and give you a very small CI (*left*). You have a *precise* estimate of the mean, but it isn't *accurate* - it missed the target.

In reality, you have just five peoples' heights. However, you've measured each person's height 20 times; there's no reason why you shouldn't take the average of the 20 measurements, to give you a very accurate estimation of those five heights.



Notice that when you do things the proper way (*right*), the standard error in the mean is larger. These results are telling you that your estimate isn't so precise, and the true population mean might actually be quite a bit different from the sample mean you just calculated.

This mistake is *extremely common*. One day, you will read a paper where the authors draw seemingly improbable conclusions from a small number of subjects. Look closer and you might find lots of repeated measurements that have been analysed as though they all come from different subjects in the way we've described.

# 5

Presentation of data

# 5

## Summary or descriptive statistics

---

Every time you write a paper, you will need to provide some descriptive statistics. For example, suppose you're conducting a study of bladder control in normal young people, normal elderly and post-stroke elderly.

You might typically summarise the following parameters:

- Sex;
- Age;
- Body mass index;
- Distribution of subjects into *normal*, *normal elderly* and *post-stroke*.

What's the right way to go about it? Below are the *most common* ways of expressing these things:

### Gender (a binary or dichotomous variable)

The easy one first.

- 20 M, 20 F

Says everything about the gender of your subjects<sup>☹</sup>.

### Age (a continuous numeric variable)

Whereas age is continuous, in practice it is almost always recorded to the nearest year. However precisely the age is recorded, the approach is the same:

- $60.7 \pm 10.3$  years (mean  $\pm$  standard deviation);

is perfectly acceptable. When quoting summary statistics, it's quite acceptable to use one further decimal place than the original values were recorded to, though the value of knowing the mean age to the nearest month could be disputed.

It is also common and useful to add the overall range:

- range 24 to 80 years.

### Body-mass index (a continuous numeric variable)

The same as for age:

- $25.1 \pm 5.3$  kg m<sup>-2</sup> (range 18.7 to 34.4 kg m<sup>-2</sup>)

On this occasion, BMI is quoted to the nearest 0.1 kg m<sup>-2</sup>, because two decimal places would be needlessly precise.

### Distribution of subjects (a categorical variable)

As for gender, you can say everything about the distribution of subjects in a single sentence:

- 10 young, 10 elderly, 20 post-stroke.

### However...

Age, gender and BMI are used simply to describe your subjects. However, your hypothesis might be (say):

*We hypothesise that frequency of incontinence is linked to group (young normal, elderly normal or post-stroke)*

It seems that the grouping of subjects is fundamental to your hypothesis; it's the predictor, the independent variable. The audience would probably like to see how well-matched the groups were, and so you might summarise your subjects separately for each group, probably in a table like this:

	Young normal	Elderly normal	Post-stroke
Number and gender	10 (5M, 5F)	10 (6M, 4F)	20 (9M, 11F)
Age (years)	$31.0 \pm 8.2$ (24 to 40)	$63.3 \pm 11.2$ (52 to 79)	$67.5 \pm 9.7$ (59 to 80)
Body-mass index (kg m <sup>-2</sup> )	$22.1 \pm 4.1$ (18.7 to 25.4)	$25.7 \pm 5.3$ (20.4 to 32.1)	$27.1 \pm 5.3$ (22.9 to 34.4)

---

<sup>☹</sup> Not quite. A colleague was once involved in a long-term study where a subject had a gender re-assignment part-way through the study. How would you deal with that?

## Graphical description of data

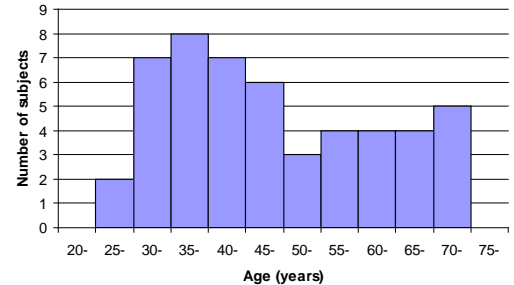
On some occasions, a parameter in your study might be so important that you would like to give more than just summary statistics. On these occasions, a graphical summary might be appropriate. Thinking about the earlier examples, you might do it something like this:

### Gender (a dichotomous variable)

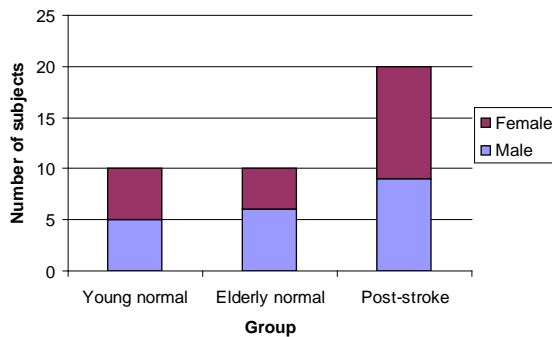
Probably not worth a figure, because it won't convey any more information than the basic values.

### Age (a continuous numeric variable)

If you were studying the link of age with (say) frequency of incontinence, you would probably want to say more about the ages of your subjects. A *histogram* would be an appropriate way of representing your data. Notice that a histogram is used to summarise *continuous numeric data*. Each bar represents a *range* of ages. For example, the first bar (labelled 25-) indicates two subjects in the range 25-29 years. The bars should all be next to each other, indicating that age is a continuous variable; a gap would imply no subjects in that age range.



### Subject groups (a categorical variable)

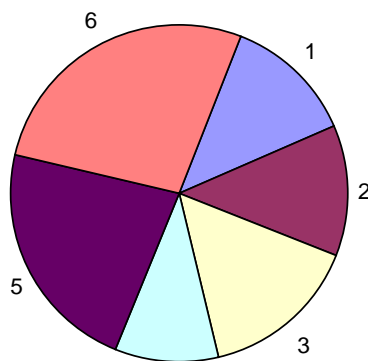


Contrast the histogram with a bar chart (*left*). As shown, the bar chart is used to summarise data by category. In this case, the bars are separate, indicating that the three categories are completely separate.

As shown, it's also possible to divide the bars by (in this case) numbers of male and female subjects.

You might caution against good at judging the different

In the example smallest? In fact,



### Pie charts

summarise the same data using a pie chart. We'd using pie charts because the eye isn't particularly angles, and it's difficult to see at a glance how categories match up.

(*right*), which is the biggest sector? The the data are the same as in the bar chart above.

## Plotting the results of a clinical study

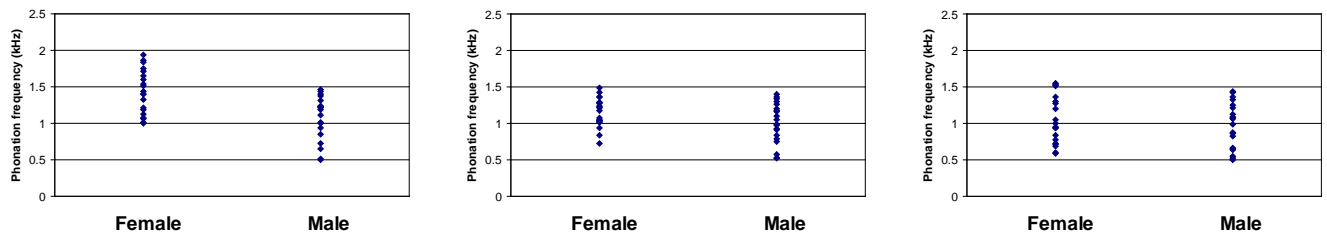
So far, we've considered the presentation of summary data, possibly demographic data for a group of subjects. This might easily be in the *Methods* section of your paper. You might also use such methods in the *Results* section of an observational study. However, in most clinical studies, you will be trying to demonstrate a link between two variables. As we said earlier:

***Plotting the data should be the first step in your analysis.***

By tradition, the most useful type of plot is the *scatter plot* (also known as *X-Y plot* or *dot plot*). The rules for the scatter plot are:

- Measure the independent variable (the predictor, the factor, the group, the cause) on the X (horizontal) axis;
- Measure the dependent variable (the effect, the outcome) on the Y (vertical) axis.
- Plot each measurement in the appropriate place according to the values.

Let's look at an example. You are studying the relationship between gender and the pitch of voice – phonation frequency. So plot the predictor (gender) on the X axis, and plot the outcome (phonation frequency) on the Y axis. Here are three sets of results you might get. Following your gut feeling, in which cases is there a real link?



If you were like me, you'd be quite convinced by the first example. In the second one, you'd not be sure. In the third example, you'd probably want to continue with the experiment and record lots more data.

And you'd be right. You can do the statistics formally using the 2-sample t-test. In the first example, the probability of this arrangement arising by chance alone (if males and females are really the same) would be about 1 in 1600 ( $p = 0.0006$ ). In the second example, it would be 1 in 16 ( $p=0.06$ ). In the third example, 2 in 5 ( $p=0.4$ ).

If the probability is lower than 1 in 20 ( $p=0.05$ ), you would probably conclude the link was real. This means that *your gut feeling was right*. The first example would be convincing to even a skeptical reader. The second example ( $p=0.06$ ) is equivocal and you would probably stay with the null hypothesis of no relationship. The third example shows really no evidence at all of a relationship.

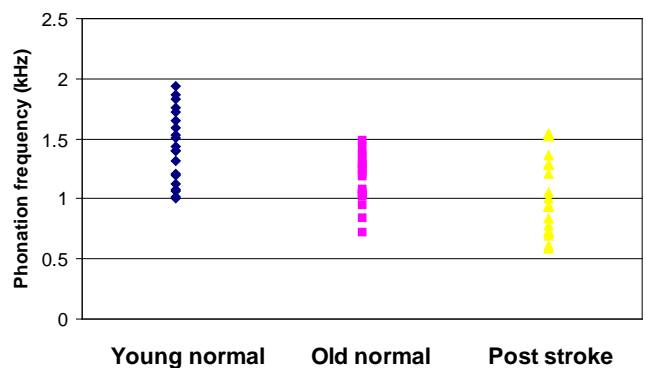
The moral of this is this:

***Gut feeling is useful - plot the data.***

### Plotting more than two categories

Using the same idea, you can handle lots of categories. Here are three categories presented using the same idea in fetching shades of pink and yellow.

Once again, there's a statistical test that can be used naturally to analyse data arranged this way. It's an extension of the t-test called *analysis of variance* (ANOVA). It will be mentioned later.



## Confusing stuff

Before diving into statistical tests in the next section, some stuff that causes difficulties.

### Handling data where both variables are numeric

In very many cases, your data will be arranged like the data just described. That is, you will have subjects in categories that might be:

- Treatment *or* placebo;
- Young normal *or* old normal *or* post-stroke;
- Male *or* female.

And so forth. Then, you measure something from each subject:

- Improvement in clinical condition;
- Incontinence frequency.

and try to demonstrate a link between this, and the subject's group. That's why the various incarnations of the t-test and ANOVA are probably the most-used statistical tests; they are the appropriate tests for data arranged this way.

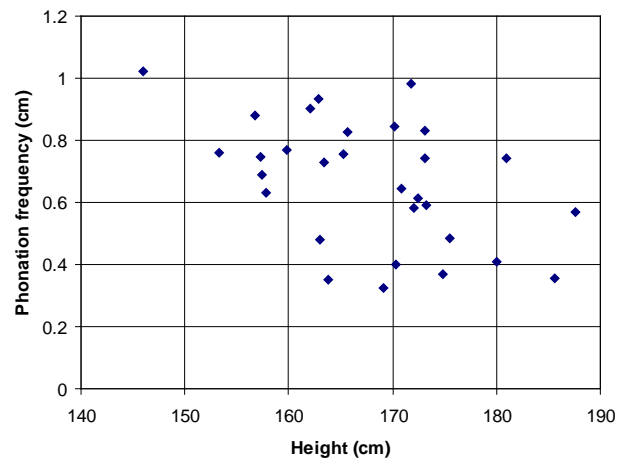
However, there are lots of studies where the independent (predictor) variable is continuous. For example;

- The relationship between quantity of drug administered and improvement in condition;
- The relationship between age and the frequency of incontinence episodes.

Nevertheless, you can plot the data in exactly the same way (*right*). Here, we've again got the independent variable (height) on the X axis, with the outcome (phonation frequency) on the Y axis. The only difference is that this time, the predictor (height) is a continuous variable.

There's another statistical test (correlation) to examine whether the relationship between height and frequency is likely to be due to chance alone. Correlation will be discussed later.

In this example,  $p=0.006$ ; the probability of observing this distribution due to chance alone is about 1 in 160. You'd probably accept that the relationship was real, but it's not absolutely convincing. Your gut feeling would probably agree.



### But beware...

In exactly the same way as the t-test, correlation can be used to demonstrate that there is a relationship or link between two variables. In fact, correlation and the t-test are intimately linked. What correlation does *not* do is show how well the two variables agree, but it is often abused for this purpose. This is *wrong*, and will be discussed later in some depth.

### Footnote

If you look through the *BMJ* book *Statistics at Square One*, all but one of the figures are covered by the few types we've discussed here.\* The reason is simple; the vast majority of studies follow this same general format with an outcome variable, and a predictive factor that you believe might be associated with the outcome.

\* The exception shows survival data from an epidemiological study - and it's not so different.

## Parametric or non-parametric statistics?

The mean and standard deviation are termed *parametric statistics*, because they are the two important parameters that describe a normal distribution. Many of the best-known statistical tests (*Student's t-test*; *analysis of variance*; *correlation*) are based on the properties of the normal distribution, and are termed *parametric tests*.

However, not all variables follow a normal distribution, and you might be advised by your local statistician to adopt *non-parametric* statistics. There are a whole set of non-parametric statistical tests, some of which are described later. These tend to be based on the rank order of the values, rather than the values themselves.

There is LOTS of debate about whether this is the right thing to do or not. An eminent local statistician is broadly against non-parametric statistics, on the following grounds:

- There is a widely-held belief that non-parametric tests are better for small groups. This is simply not true. With some non-parametric tests it is actually impossible to obtain a positive result in small group sizes.
- Though they are based on the normal distribution, parametric tests such as the *t-test* are reasonably robust in the face of non-normally distributed data.
- If you have highly non-normal data, you can use a transform (for example, a logarithmic transform) that will often sort out the problem. The general idea is that you simply take the logarithm of each value before performing the test, but you should take advice.
- There are only a few fairly basic non-parametric tests, non-parametric equivalents to t-tests and ANOVA. There are no tests for many more complex analysis problems. It will look odd if you suddenly jump to parametric tests part-way through your analysis simply because there are no suitable non-parametric tests.

### Some stuff on non-parametrics

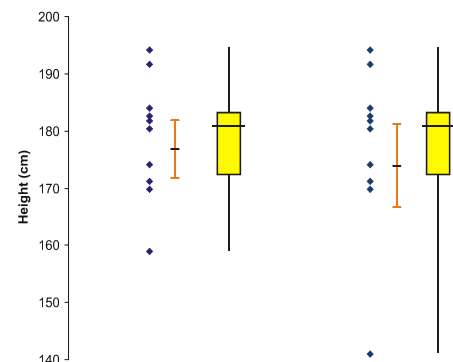
Nevertheless, you might find yourself using non-parametric statistics, or at least reading papers where other authors have used them. If you do want to summarise data using non-parametric descriptive statistics, your stats package should produce them for you, but here are the rules:

- First, arrange all the measurements in order, smallest to largest;
- Number the measurements, from smallest (#1) to largest (#100, say);
- The *median* is the value half-way down the list (or half-way between the two middle values, #50 and #51);
- The *lower quartile* is the value just one quarter way down the list (#25);
- The *upper quartile* is the value three-quarters way down the list (#75).

In the example (*right*), the data are shown without, and then with, an outlying measurement. The outlier affects both the mean and standard deviation of the data.

The *box and whisker plot* is often used to show non-normal data:

- The central horizontal line indicates the median;
- The yellow box marks the upper and lower quartiles, indicating the *inter-quartile range*;
- The vertical whiskers indicate the overall range of the data, or possibly some other measure of overall spread eg. they contain 95% of all the measurements. Check to be sure.
- Outlying points are shown separately. These are the ones that aren't included by the whiskers, if any.



Since the median in the figure is closer to the upper quartile than the lower, this indicates that the distribution of heights is skewed. Notice that the median and the inter-quartile range are not affected by the outlier, indicating that the overall distribution of the data hasn't changed. However, the extended whisker indicates the presence of the outlier. In other cases, this might be shown as a separate point.

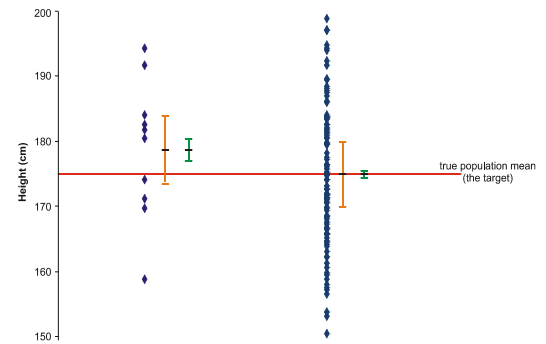


## Standard deviation, standard error or confidence interval?

It's quite common to show error bars on graphical data. The question is: *should you present the standard deviation (SD), the standard error in the mean (SEM), or the 95% confidence interval (CI)?*

It's tempting to use the SEM, because it's always smaller than the SD or the 95% CI, and so makes your results look better. The rules are these:

- If you want to describe the population as a whole, then use the SD.
- If you specifically want to quantify the mean value of something, then use the SEM or the 95% CI.



In the example (*right*) the SDs are shown in orange, with the SEMs in green. Notice that the SD stays (more or less) same as the number of subjects increases, but the SEM gets smaller. This reflects the fact that the mean is becoming a more accurate estimate of the true population mean.

- Remember that the 95% CI is just about double the SEM – so it doesn't matter hugely which is used, so long as you are clear in your report. You don't need both, but I'd probably favour the 95% CI.

In your paper, you might say:

*The heights of the men studied had a mean of 175 cm, with an SD of 10 cm.*

Here, you are describing the study population. You wouldn't expect to see the distribution of heights to change just because you studied more subjects.

*In our results, the mean height of the swimmers (175cm, 95% CI 173 to 177 cm) was significantly different to that of the runners (168 cm, 95% CI 165 to 171 cm) ( $p < 0.01$ ).*

Here, you are demonstrating the grounds on which you made your statistical judgement. The confidence interval is appropriate, because more subjects lets you quantify the mean more accurately and therefore be more confident in your results.

Notice again that we have quoted the size of effect in our results ie. the difference in height between the two groups is 168 to 175 cm, or about 7 cm. Remember...

***Statistical significance means ONLY that the link you found is unlikely to be due to chance alone. It is up to you to decide whether the effect is large enough to be of any clinical importance.***

# 6

Comparing between groups: the t-test

# 6

## The single sample t-test

Once again, let's invent a hypothetical study. You think that people are getting taller with time.

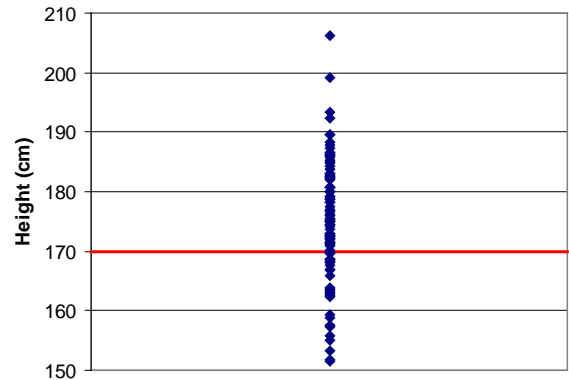
*We hypothesise that the height of adult men is linked to the year in which they were born.*

Unfortunately, you missed the boat in making the measurements, but know (from a big historical study) that the average height of men born 1800 was 170 cm. So, you'd like to compare a group of modern men to see how they measure up. Here are the results from 100 men (*right*). You can now use the single sample t-test. The process with your stats package will be something like this:

- Arrange all the measurements in a column;
- Select the *single sample t-test*;
- Pick the column containing the data;
- Enter '170' as the value for comparison.

The package will say something like this:

- $N = 100$ ,  $Mean = 175.06$
- $Standard\ deviation = 10.17$ ,  $SEM = 1.02$
- $t = 5.0$ ,  $p = 0.0000024$
- $95\% CI = 173.02\ to\ 177.10$



### Interpreting the t-test

Here's what this all means:

- The number of values in the test is 100. The mean and standard deviation were described earlier.
- The SEM (standard error in the mean) was also described earlier. Remember this gives some indication of how precise your estimate of the true population mean is. With the data given you can interpret it this way:
  - The probability the estimate is within 1 SEM (1.02 cm) of the true mean is 2 in 3;
  - The probability the estimate is within 2 SEMs (2.04 cm) of the true mean is 19 in 20 (95%);
  - The probability the estimate is within 3 SEMs (3.06 cm) of the true mean is 999 in 1000 (99.9%).
- The t statistic (5.0 in this case) indicates that the calculated mean of 175.06 cm is actually 5 SEMs away from 170 cm.
- The p value is the probability that you get your observed results, assuming there is *no link* between height and birth year. The stats package has determined that  $p = 0.000\ 002\ 4$ . Pretty unlikely.
- Finally, the *95% confidence interval* is given. The probability the estimate is within 2 SEMs (2.04 cm) of the true mean is 19 in 20 (95%). You can be 95% confident that the true population mean is within 2.04 cm either way of the calculated mean of 175.06 cm. This leads to the 95% CI of 173.02 to 177.10 cm.

### Interpreting the p value

For those interested, here's another hand-waving interpretation of the p value:

- Fill a big hat with a billion (or more) random numbers. The overall mean should be 170, and the overall standard deviation should be 10.17.
- Pick 100 numbers from the hat at random, and calculate their average.

The p value indicates the probability that the average of the 100 balls will be 175 or higher. It's about once in every 400,000 repeats, so don't try it.

### One or two sided?

In some packages, you can select a *one-sided* test, which is slightly more sensitive than the two-sided one you should be using. The one-sided test does not allow for the possibility that change could be in *either direction*; the mean height could be either side of 170 cm.

*You should rarely if ever use a one-sided test.*

## The 2-sample t-test

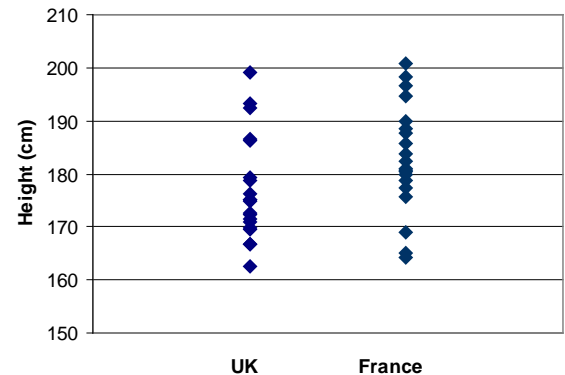
The single-sample t-test as described in the last page isn't so common. That's because in a well-designed study, you will aspire to have some control data, for the reasons we've already covered. In the same vein, but much more common, is the *two-sample t-test*. The two-sample t-test goes naturally with data arranged as shown (*right*). In this example, you are looking for a link between height and the country of origin.

The operation of your stats package will be something like this:

- Arrange all the *UK* measurements in a column;
- Arrange the *France* measurements in a second column;
- Select the *two sample t-test*;
- Pick the columns containing the data.

Your data might look something like this:

UK	France
173	182
186	201
175	164
...	...



Some stats packages do things a bit differently. You need all the measurements in the same column. A second column is used to identify *UK* and *France* data, like this (*right*):

In the example, '1' is used to indicate a measurement from the UK, and '2' is used to indicate a measurement from France. It's a bit more fiddly doing things this way, but can be useful when you come to do more complex analyses.

Height	Country
173	1
186	1
175	1
182	2
201	2
164	2

### Interpreting the two-sample t-test

The output will look very similar to that for the 1-sample t-test. Here's the first part of the output from SPSS. This is just the descriptive stuff, split for the UK data (country 1) and the France data (country 2).

	Country	N	Mean	Std. Deviation	Std. Error Mean
Height	1.00	20	176.9951	9.7908	2.1893
	2.00	20	182.9806	10.2224	2.2858

Here's the second part:

		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% CI of the Difference		
										Lower	Upper
Height	Equal variances assumed	.014	.907	-1.891	38	.066	-5.9855	3.1651	-12.3929	.4219	
	Equal variances not assumed			-1.891	37.929	.066	-5.9855	3.1651	-12.3933	.4223	

### Things to note:

- There are two versions of two-sample t-test, slightly different. Since both sets of data (*UK* and *France*) have approximately the same variance (the variance is just the SD squared), they give almost identical results.
- The F-test can be used to decide whether the variances are *really* equal. This gives you a hint as to which test to believe; a low p value means they probably *aren't* equal. The p value of 0.907 means you can use either test.
- The t statistic (-1.891) is shown, along with the number of degrees of freedom. Don't even ask how this is calculated.
- The Sig (significance) column is what we've been calling the p value.
- The mean difference and the rest of the values are as for the single sample t-test. They relate to the *difference* between the UK and France data. The mean difference is -5.99 cm, with a 95% CI of -12.4 to +0.42 cm.

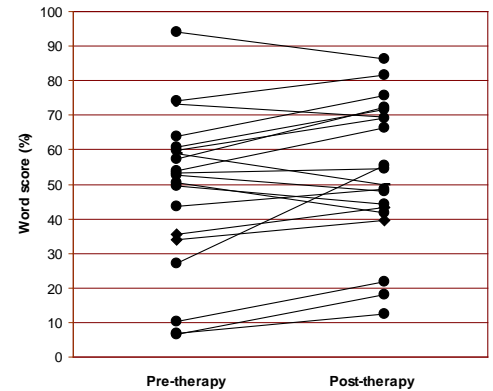
## The paired t-test

The paired t-test is appropriate for use in any situation where the measurements are paired in some way. In particular this includes controlled studies where:

- Subjects act as their own controls, such as a crossover study;
- The controls are matched *per patient* with the treatment group.<sup>β</sup>

In the example (*right*), you are investigating a bunch of patients post stroke. In the *pre-therapy* measurement, you evaluate their ability to speak with some kind of word score.

In *post-therapy*, you measure the same subjects again after 3 months of speech therapy. The question is: did the therapy have any effect?



In this case, you clearly have paired data, and so the paired t-test is appropriate. You might arrange the data in two columns as for the two-sample t-test. However, it's possible you'll do it as shown (*left*).

WordScore	Subject	Study
15	1	1
25	1	2
27	2	1
29	2	2
73	3	1
62	3	2

This table shows the data for the first three patients.

- The measurements are in the first column;
- Column 2 indicates which subject the measurement is from;
- Column 3 indicates whether this is the pre-therapy (1) or post-therapy (2) measurement.

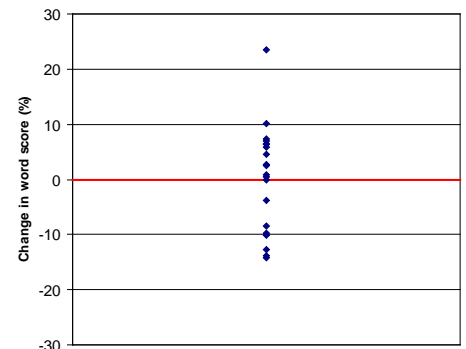
This allows the stats package to pair up the measurements correctly.

## Interpreting the paired t-test

Interpretation of the paired t-test is very similar to that for the single-sample t-test, and for good reason because they are essentially the same test. If your stats package doesn't do a paired t-test (for example, Minitab doesn't), you can easily do one like this:

- For each subject, subtract the 'pre-therapy' value from the 'post-therapy' value. This gives you a list of differences.
- Compare these differences to 0, using the single-sample t-test.

The figure (*right*) shows the differences. If the therapy isn't having much success, then you'd expect the differences to lie either side of the red line, and the mean difference to be about zero.



In this case, the mean difference from zero is +5.2 points, and the SEM is 2.1 points. Therefore, your stats package will tell you that:

- The chance of this distribution arising by chance alone is 1 in 50 ( $p=0.02$ );
- The mean effect is +5.2 points;
- The 95% confidence intervals are  $\pm 2$  SEMs from the mean, or from 1.0 points to 9.4 points.

Of course, it's now up to you to decide whether or not this is a worthwhile improvement.

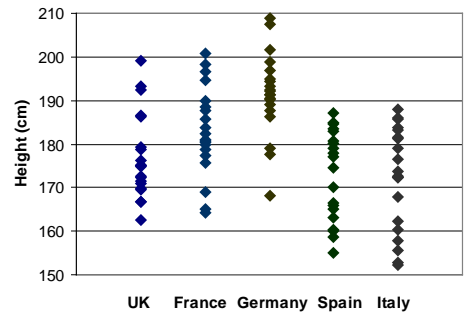
## Footnote...

This example is a bit like a crossover trial (*see earlier*), but shows you why randomisation of the treatment order is important. Much as you have detected an effect, *you can't say for sure that it was due to the therapy*. That's because the patients might have got better anyway. The solutions might be:

- Set up a control group who didn't get any therapy. Would this be ethical?
- Study each patient twice at 3-monthly intervals before the therapy began. Would this be ethical?

## Multiple comparisons and analysis of variance (ANOVA)

Back to the 'comparing heights' example. It's possible you just want to compare heights from two countries, but more likely you would want to compare three or more. Your data might look like this (*right*). How do you perform the statistical tests to show the link between height and country of origin?



### Multiple testing

One way would be to use the two-sample t-test exactly as before. The trouble is, you've now got five countries, and the t-test can only be used to compare two sets of data. You would have to perform the following tests:

*UK v France*      *UK v Germany*      *UK v Spain*      *UK v Italy*      *France v Germany*  
*France v Spain*      *France v Italy*      *Germany v Spain*      *Germany v Italy*      *Spain v Italy*

There are two problems here:

- It's tedious.
- Typically, you might accept that a link is real if the probability of it occurring by chance alone is less than 1 in 20 ( $p < 0.05$ ). However, you are performing ten separate comparisons. There is more than a one-in-three chance that one of the comparisons will produce such a result *by chance alone*.

### Fishing for links

Let's tackle the second problem first. It's not uncommon that the workers will measure a large number of parameters, and look for a link between every combination of two parameters. This is particularly true for correlation analysis, but the same principle applies to any statistical tests. There's a saying in statistics:

*If you torture for long enough, then the data will eventually confess.*

And sure enough, the workers might well discover some bizarre link. In one paper on sleep, the researchers went fishing for links between daytime sleepiness, demographic factors, and 24 separate indicators of overnight sleep quality. They found a link between daytime sleepiness and the subject's level of postgraduate education (!), and then spent some time debating where the relationship might come from.

### The Bonferroni correction

Bonferroni says something like this:

*If you make  $N$  comparisons, then make your criterion for believing a result  $N$  times more severe.*

In the earlier example, you were making 10 separate comparisons, and so you must make your criterion 10 times more severe. Instead of taking 1-in-20 as the acceptable value, you would take 1-in-200 ( $p < 0.005$ ).

### Analysis of variance (ANOVA)

Generally, analysis of variance (ANOVA) is the right way to tackle these problems. You'll arrange your data as for the two-sample t-test (*right*), except this time there are more countries.

The output from the stats package will also be very similar to that for the two-sample t-test. In fact, the two-sample t-test is just a special case of ANOVA. If you performed ANOVA with just two countries, *you would get exactly the same results*.

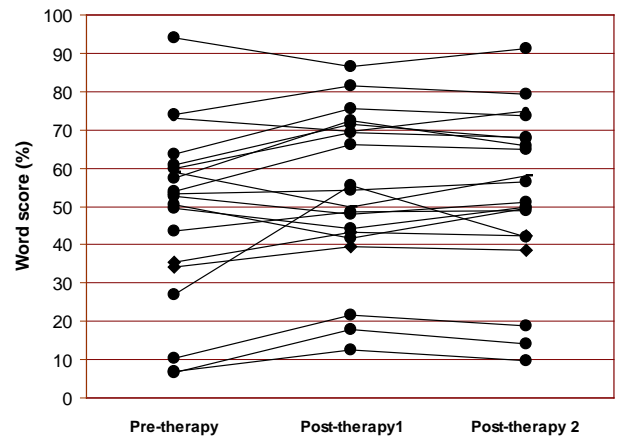
Height	Country
173	1
186	1
175	1
168	2
192	2
174	2
182	3
201	3
164	3

## More about analysis of variance

### Repeated measures ANOVA

The comparison above would be called a *one-way ANOVA*, because each height measurement was classified only by country. However, you might well have a situation (*right*), where the data are paired (or tripled, in this case). It's the same example as for the paired t-test, but this time you're making an extra measurement after 6 months of therapy.

WordScore	Subject	Study
15	1	1
25	1	2
27	1	3
29	2	1
73	2	2
62	2	3
...	...	...



The *repeated-measures ANOVA* is the generalization of the paired t-test when you have more than two measurements in each subject (these are the repeats). As for the paired t-test, you will classify each patient by subject number (1..20) and by study (1, 2 or 3).

### In summary...

- ANOVA is a generalized form of the t-test that can deal with more than two categories of data.
- One-way ANOVA is a general form of the two-sample t-test.
- Repeated measures ANOVA is a general form of the paired t-test.

If you have a complex experimental design (for example, subjects might be classified by *gender* and *young/old age* and *diseased/normal* and *pre/post-therapy*) you can also do three, four, five, one-hundred way ANOVA.

### Cautions about ANOVA

Beware. When you use two- or more way ANOVA, there are all kinds of traps to fall into. In the example above, you have a repeated measures design. The same measurement is made three times on the same 20 patients. There are two factors: patient (1..20) and measurement (pre/post1/post2). However, the measurements are not necessarily independent because there were three measurements on each patient. Particular problems arise when some of the measurements are closer-related than others.

Alternatively, perhaps you have 60 different patients but each patient belongs to one of three groups (first factor) and each patient has one of four treatments (second factor). These measurements may very well be independent, and under some circumstances will need to be analysed differently than the first example.

There are other situations where you have multi-level repeated measures, for example you might make a series of measurements on a given day, and then further series of measurements at monthly intervals. If you have data of this complexity, it's best to get good advice on the proper analysis.

## Non-parametric tests

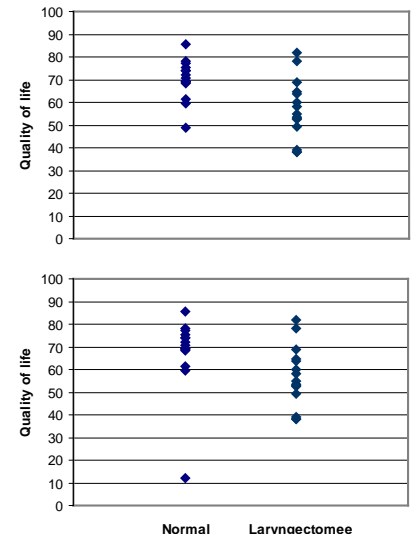
Before diving into this section, read the earlier warnings about non-parametric statistics.

In this study, you want to show a link between laryngectomy and a reduced quality of life, as measured using a validated questionnaire. So you recruit 15 laryngectomees and 15 normal controls, and measure their quality of life.

Here are your results (*right*). Your gut feeling says you were right, and the two-sample t-test says this link would only arise by chance with a probability of  $p=0.0035$ , and so you're convinced.

But suppose your data had come out like this (*right*). Exactly the same, *except* the one normal subject who now clearly has the weight of the world on his shoulders. Your gut feeling probably hasn't changed because this *outlier* is clearly not representative of the group. However, the t-test now says  $p=0.06$ . This one outlying measurement has converted your study from a convincing *yes* to a probable *no*.

The problem goes back to the way the t-test works. This one outlier has reduced the mean and *greatly* increased the SD of the measurements in the *normal* group, to the extent that the t-test is no longer sure the two groups are different.



## Non-parametric tests

The three non-parametric tests appropriate for discussion here are Wilcoxon's sign-rank test, the Mann-Whitney U test and the Kruskal-Wallis test. These are all based on *rank order*; the actual measurement is *not important*, just its rank order in relation to the other measurements. The stats package will do the ranking for you, but the process is something like this:

- Arrange all the measurements in order, smallest to largest;
- Number the measurements from 1 (smallest) to (in our case) 30 (largest);
- If two measurements are exactly the same, they will share a number.

### The Mann-Whitney U test

This is a non-parametric replacement for the two-sample t-test used above. In SPSS, you need to arrange the data *exactly as for the two-sample t-test*, but simply choose the alternative test.

- For the first set of results, Mann-Whitney says  $p = 0.007$ .
- For the second set of results, Mann-Whitney says  $p = 0.009$ .

These values can be interpreted exactly as for the t-test - the probability that your result (or one more extreme) would have been observed, *if there was no link between patient group and quality of life*. Notice that:

- In the first set of data, Mann-Whitney is not as sensitive as the t-test. When the data follow an (approximately) normal distribution, the t-test is generally better.
- However, Mann-Whitney gives a very similar result for the second set of data. When the data are non-normal (particularly, with outliers), Mann-Whitney is better.

There are formal ways to test your data for normality, but like all statistical tests they become more sensitive as you increase the sample size. For large samples, they tend to indicate your data are not normal even in cases where the t-test would be absolutely fine. The moral is: **plot your data first!!!**

### The Wilcoxon signed-rank and Kruskal-Wallis tests

These tests work in a very similar way to Mann-Whitney:

- Wilcoxon is a direct replacement for the paired t-test;
- Kruskal-Wallis is a direct replacement for one-way ANOVA.

In SPSS, you arrange your data *exactly as for the corresponding t or ANOVA test*, then simply pick the alternative non-parametric test from the menu. Much as this might not be considered good etiquette, it's easy to perform both tests to see how things work for your own data.



## Writing up your statistical test

A bit of re-iteration here. When you come to write up your results in a paper, the traditional way has been something like this:

*Apnoea duration in the normal group was significantly different to that in the post-stroke group ( $p < 0.05$ ).*

There are some shortcomings in this style of report:

- Statistical significance shows *only* that the effect is probably not due to chance. It says nothing about how clinically important the effect is.
- You get no impression of how big the effect is.
- The style  $p < 0.05$  is a throwback to the days of statistical tables, where the exact probability couldn't be computed. It's crazy to suggest that  $p = 0.049$  is dramatically different to  $p = 0.051$ . These days, you should be using a computer that can give the exact value, so quote it and let the readers judge the evidence for themselves.

You would be better using something like this:

*Mean apnoea duration in the normal group was 0.45 s (95% CI 0.31 to 0.59 s), but in the post-stroke group was 0.63 s (95% CI 0.52 to 0.74 s).*

*The mean difference in apnoea duration between the groups was 0.18 s (t-test,  $p = 0.023$ , 95% CI 0.11 to 0.25 s).*

Or perhaps, you might summarise the data in a table:

	Normal group (n=10)		Post-stroke group (n=10)		Difference (normal to post-stroke)		
	Mean	95% CI	Mean	95% CI	Mean	95% CI	P ( two-sample t-test)
Apnoea duration	0.45 s	0.31 to 0.59 s	0.63 s	0.52 to 0.74 s	0.18 s	0.11 to 0.25 s	0.023
...	...	...	...	...	...	...	...

The important points are:

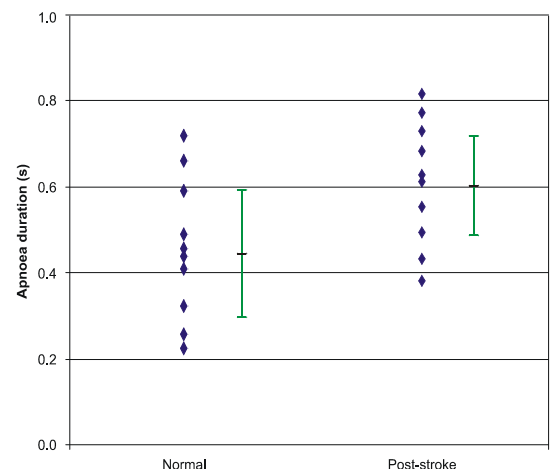
- Give the means for the two groups and the difference between them, so you get an impression of the effect (about 0.18 s) in relation to the individual measurements.
- The 95% confidence intervals are given. These let the reader judge how confident you are that your mean values are accurate. These values will be produced by your stats package.
- There's nothing wrong with p values; just give the exact value and discuss it later. Let the readers judge for themselves how to interpret your data. Again, the p value will be produced by your stats package.
- You can also include the numbers of subjects in each group, though this should of course be given elsewhere in the paper, and cite the test being used.

### Once again...

At risk of becoming boring, we'd like to see you plot your data (*right*). From the figure, you get immediate impressions:

- The numbers of subjects studied are relatively small.
- There's quite a lot of spread, giving considerable overlap between the two groups;
- Nevertheless, the *post-stroke* group have a mean apnoea duration that's about 25% longer than the *normal* group.
- The difference probably isn't due to chance alone.

This just backs up what the statistical data in the table tell you.



# 7

Links in numeric variables: correlation  
and regression

# 7

---

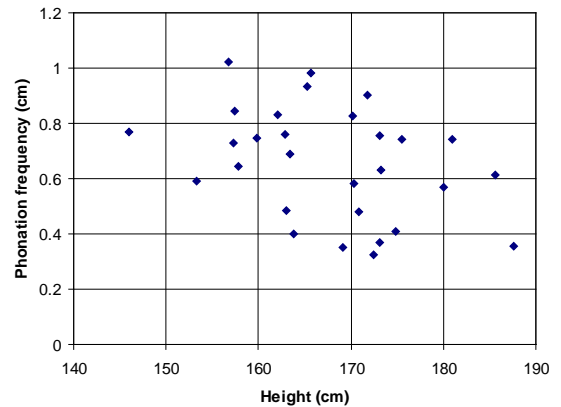
## Demonstrating a link in numeric data - correlation

Here are some results from a 'height versus phonation frequency' experiment (*right*). You think there's a link between height and phonation frequency, but would like the statistical test to show it formally. That test is correlation.<sup>ψ</sup>

### Performing the correlation

In your stats package, you'll set up your data something like this:

Height	Frequency
156	0.88
145	1.02
159	0.77
...	...



And select the *correlation* option from the menu. You'll need to select the two columns. The values are clearly paired, so both columns must contain the same number of measurements.

Many stats packages let you pick 3 or more variables, and produce a *cross correlation table*, where everything is correlated with everything else. This is fishing for links, a recipe for torturing the data. As we said earlier, *be very careful what you read into multiple comparisons*.

### Interpreting the correlation

For the data shown, the stats package will say something like:

- $r = -0.48$
- $p = 0.007$

The  $r$  value is loosely equivalent to the  $t$  statistic. It's a test statistic that quantifies the agreement between the two variables. The  $r$  value has the following properties:

- It is always in the range -1 to +1;
- A positive  $r$  value means that the variables tend to increase and decrease together
- A negative  $r$  value means that as one variable increases, the other tends to decrease (*as above*)
- A value of +1 indicates a straight-line relationship between the two variables with a positive slope.
- A value of -1 indicates a straight-line relationship between the two variables with a negative slope.
- A value of 0 indicates no relationship whatsoever between the variables.
- Correlation shows *only that two variables are related!!!* It is *completely insensitive* to the actual numerical values of the two variables. You could multiply all the heights by 100, or add 50 to each frequency - *the correlation coefficient would be exactly the same*.

Next, the  $p$  value. This is interpreted exactly as for every other  $p$  value as the probability that this value of  $r$  (or one more extreme) would have arisen by chance alone. For our example, you could interpret it like this:

- Write down all the heights on scraps of paper, and put them in one hat;
- Write down all the frequencies on scraps of paper, and put them in a second hat;
- at random, pick a height and a frequency, your first pair of measurements;
- repeat until you have all the pairs of measurements;
- now work out the correlation coefficient for the pairs of measurements you just created.

The  $p$  value is the probability that your 'random chance' correlation coefficient is 0.48 or higher. Since  $p=0.007$ , you'd expect it in 7 out of every 1000 trials. Once again, don't try it.

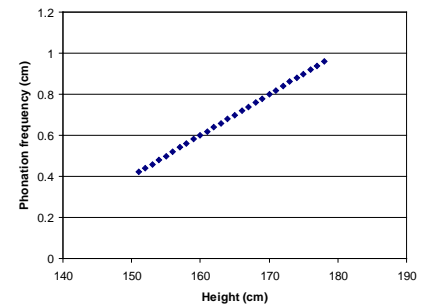
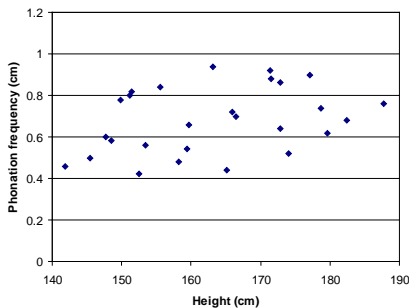
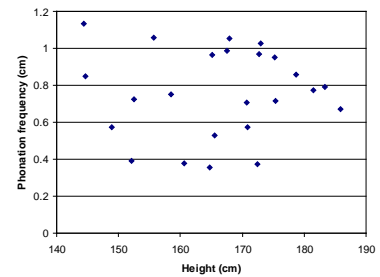
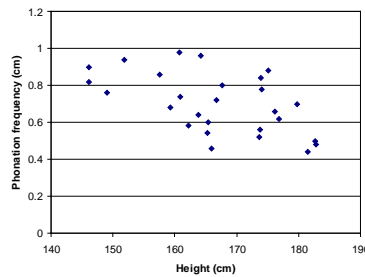
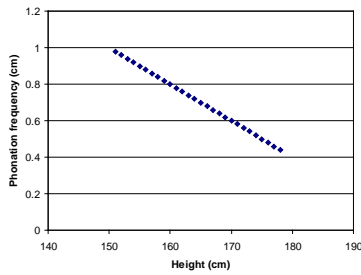
***Correlation shows only that two variables are related!!!  
It is completely insensitive to the actual numerical values of the two variables.***

<sup>ψ</sup> The correct name is Pearson's product moment correlation coefficient, sometimes called 'the Pearson', 'the correlation coefficient', or 'the  $r$  value'.

## Some examples of correlation

Here are examples of data with a range of correlation coefficients:

- (top row)  $r = -1.0$ ,  $r = -0.62$ ,  $r = -0.21$ .
- (bottom row)  $r = +0.50$  and  $r = +1.0$ .



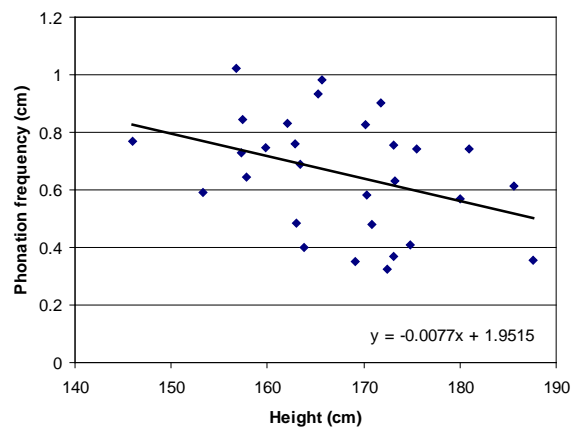
## Quantifying the effect - linear regression

Once you've established the relationship is real, you will want to say something about the *effect*. In this case, you want to know how much the height affects the phonation frequency.

Superimposed on the graph (right) is the *regression line*. Clearly, you couldn't draw a straight line that went through all the points. The stats package has calculated the regression line such that the errors in phonation frequency are reduced to the absolute minimum.

Underneath is printed the regression equation:

$$y = -0.0077x + 1.9515$$



This equation simply represents the line on the graph. Remember that:

- The x axis represents height in cm (the independent or predictor variable);
- The y axis represents frequency in kHz (the dependent or outcome variable).

You could interpret the equation like this:

- At a height of 0 cm, you would expect a phonation frequency of 1.9515 kHz.
- For every 1cm *increase* in height thereafter, the phonation frequency will *decrease* by 0.0077 kHz.

So this agrees in principle with the correlation coefficient; as height increases, frequency decreases.

If you knew a person's height, you could now use the regression equation to predict their phonation frequency. Since there's lots of variability it probably wouldn't be exactly right but on the evidence of this study, *this would be your best possible estimate of their phonation frequency*.

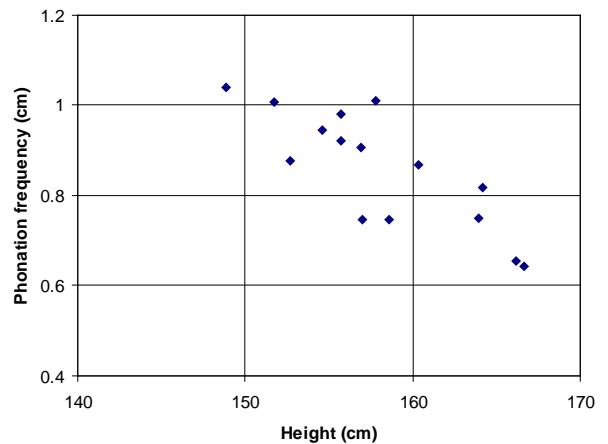
## Non-parametric tests

Just as for the t-test, there is a non-parametric version of the correlation coefficient.

Here (*right*) are some results you might have obtained from the 'height versus phonation frequency' experiment. There's clearly a strong relationship here, and for these data:

$$r = -0.81, \quad p = 0.000138$$

So the statistical test agrees with you.

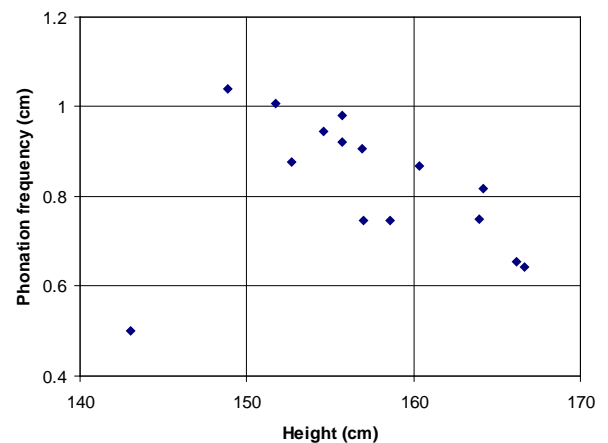


### Correlation and outliers

Now, introduce just a single outlier (*bottom*):

$$r = -0.2, \quad p = 0.47$$

All of a sudden, the story isn't so convincing. This is quite a common problem with correlation; it is VERY sensitive to outliers.



### Rank correlation

There is a non-parametric test called *Spearman's Rank Correlation Coefficient*. Your stats package will probably do it; Minitab and Excel don't, but it's quite simple:

- Rank the height measurements in order, smallest = 1, biggest = 15;
- Rank the frequency measurements in order, smallest = 1, biggest = 15;
- Calculate a standard correlation coefficient by replacing each measurement *with its rank number*, rather than on the measurements themselves.

So:

- For the first set of data, Spearman's rank correlation coefficient ( $r_s$ ) is -0.79 ( $p < 0.001$ ).
- For the second set of data, Spearman's rank correlation coefficient ( $r_s$ ) is -0.53 ( $p = 0.043$ ).

## Parametric or non-parametric?

We've already covered this, but for completeness, here are the issues again.

- There is a wider range of parametric tests available.
- Parametric tests are better when the data are normally distributed, and will also work well with small deviations from normality.
- Non-parametric tests are generally more robust in the face of outliers.
- There is a general feeling that non-parametric tests work better with small sample sizes. This isn't true!

Most statisticians would probably stick with parametric tests, perhaps using a logarithmic transform if necessary. So, once again:

***Check with the statistician.***

# 8

## Reliability, validity and agreement

# 8

---

## Agreement - what's it all about?

This is a relatively new addition to the manual, and we've added it because so many people have needed it in the past. Reliability and validity are both well-used when describing the performance of a clinical test. The terms are often used interchangeably, and in fact both are measures of agreement. In some cases, you can use the same statistics to describe both, which just adds to the confusion. So here's the thing:

### Reliability

Reliability is to do with repeatability. A reliable test will give the same answer every time you use it. If you want to measure the reliability of a test, you need to apply the same test on two or more occasions.

**Reliability: agreement across repeated applications of the same test.**

For any test, the hope is that the outcome of the test reflects *only* the clinical status of the patient. Ideally, the test could be performed by a different person, using a different instrument on a different occasion and they would get the same answer. In practice, that isn't the case – there are lots of sources of error:

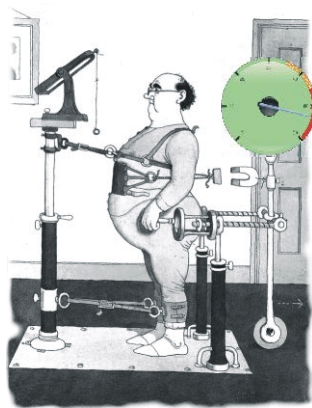
### Instrumental errors

Measuring instruments are not entirely reliable. To assess instrumental errors, you will ideally make the measurement on a so-called phantom, an inanimate object with known properties that are absolutely constant. For example, you could use a rock to assess the errors in a set of weighing scales. Presumably, this led to the stone being adopted as a unit of weight. If you must use live subjects, you would want to make repeated measurements on the same subject, closely spaced in time to be sure the subject themselves hadn't changed.

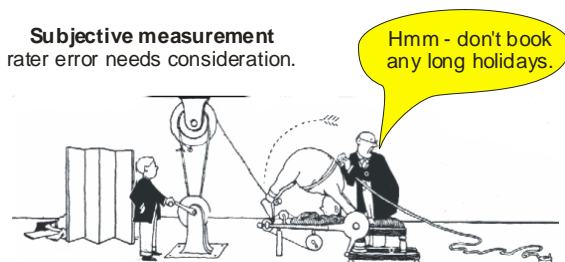
You might split instrumental errors down further. For *within instrument* errors, you are interested in the variation from one measurement to the next on a given set of weighing scales. These errors might be due to friction in the mechanism, or perhaps exactly where on the weighing pan you put the rock. For *between instrument* errors, you would study the agreement between two different sets of scales. For example, one set might be reading low relative to the other because it has been calibrated incorrectly.

### Rater errors

This is all fine for an objective measurement, where the test is not open to interpretation by the user. For example, a set of digital weighing scales is reasonably objective. Things start to get more complicated for old-fashioned scales with a moving dial. The weight you get depends what angle you read the scales from (or how optimistic you're feeling), but it's still hard to affect the measurement by more than a kilogramme or two. You probably wouldn't be worried that the person reading the scales was having a big effect on the measurement.



**Objective measurement**  
rater error is probably unimportant.



**Subjective measurement**  
rater error needs consideration.

Many tests are at the opposite end of the scale. For example, assessment of radiological images is highly subjective. The clinician doing the rating is actually part of the measuring instrument.

This being the case, you can measure *within-rater* (*intra-rater*) and *between rater* (*inter-rater*) agreement, just as you can for *within-instrument* and *between-instrument* agreement. There's no difference because the rater is just part of the instrument.

### Within-subject errors

Even with an instrument that gives good repeatability on a phantom, it is likely that repeated measurements on a live subject will not be identical. Weight varies according to hydration status, and blood pressure from minute to minute.

## Test-retest errors

If you're keen you can separate out all these different sources of error. But unless you're very keen, you are probably just interested in the overall effect on your own experiment. The easiest way is to measure test-retest agreement, ie. you make the measurement now, and you make the same measurement again in a little while. It's likely though not essential that you will use the same measuring instrument and the same rater on both occasions. In this case, you will be estimating the combined effect of *within-instrument*, *within-rater* and *within-subject* errors.

## Validity

Reliability is *agreement across repeated applications of the same test*. Our set of scales can be reliable by measuring the same thing repeatably – but are the measurements valid? Let's say we're trying to diagnose reflux. I can give you a highly reliable diagnosis of every patient you meet: *they have reflux*. The next time I assess the same patient, I give the same answer. It's reliable, but of very little value.



So ... reliability is no use without validity. Less obviously though, you can't have validity without reliability. If a test isn't reliable (the results aren't consistent from one measurement to the next), then some of the time it must be wrong.

**Validity: agreement of a test with the right answer.**

Validity is always in the eye of the beholder. A test that is valid for detecting cancer is unlikely to be much use in measuring voice quality. To measure validity, you need to measure agreement with some kind of standard that tells you the 'right' answer. Notice the scare quote marks around 'right' - this is where the problems start.

### Criterion validity

The best way is if you have an indisputably correct answer. A good example might be the diagnosis of cancer – presumably a patient either has the disease, or doesn't. That is the correct answer. If you were developing a new screening test for cancer, you would want to validate it against the known diagnosis (from biopsy, say). That's criterion validity – agreement with a criterion that gives the known correct answer. Of course if you dig a little deeper you realise that occasionally a biopsy gets it wrong – but not that often, so it's a reasonably good standard.

### Construct validity

Sometimes, there is no right answer. Suppose you were developing a new questionnaire to measure reflux. It comes up with a number from 0 (normal) to 10 (very bad). Is it valid?

Well, in this case you there isn't a right answer. It isn't going to give exactly the same answer as an instrumental investigation but you could make some predictions:

- It will probably have some relationship with other methods to measure reflux such as pH-metry.
- If a person's reflux changes, it probably should measure the change. For example, you might apply the questionnaire before and after prescribing a course of PPI; it ought to show a change in response to the therapy.

This is construct validity – showing that your new test responds to the things you predict it should respond to.



## Measuring agreement – categorical stuff, reliability and Kappa

Reliability is agreement across repeated applications of the same test. So – how should you measure reliability? Let's take an easy case, a clinical diagnostic test for aspiration. The answer from the test is 'yes' or 'no': we're not going to allow a 'not sure'. Two experiments come to mind:

**Test-retest agreement:** Perform the test now, and repeat it in an hour. Measure agreement between the two tests.

**Inter-rater agreement:** You and your colleague observe the same test, and privately form your own view on the results. Measure agreement between your judgements.

It doesn't matter which experiment you perform; you can measure agreement the same way.

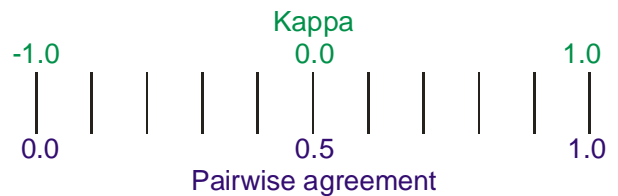
### Pairwise agreement

The easiest thing is just measure how often the pair of tests agreed. This is pairwise agreement. If you perform the test on 20 patients and agree on 14, then the pairwise agreement is 70%. You can do exactly the same thing for three or four repeats of the test, or for three or four raters, just measuring agreement between each pair of raters in turn. It's perfectly okay to quote this as your measure of agreement.

### The Kappa statistic

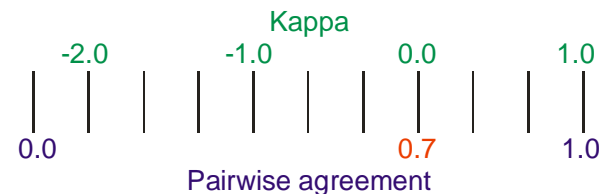
Pairwise agreement makes no allowance for raters sometimes agreeing by chance. In our simple example, two raters who were just guessing would expect to agree about half the time. The Kappa statistic is an attempt to correct for the expected chance agreement. Lots has been written about Kappa – some love it, others hate it – but here are the facts that no-one will dispute...

Kappa uses pairwise agreement, but simply changes the measurement scale. Work out what pairwise agreement you would expect by chance, and call that *zero agreement*. In our simple case above, you would re-label 50% agreement as *zero agreement*, as shown in the picture. And that's it. Think of it like changing from Fahrenheit to Centigrade: the actual numbers change, but higher is still hotter (or better agreement).



### The Kappa statistic and bias

The problem arises with this *expected chance agreement* business – how do you work it out? Well if there are two choices then surely it's just 50%? But then, I can make myself look like a great rater just by guessing 'yes' all the time; then I agree with myself all the time. Kappa sorts this out by taking account of bias. If the ratings are biased, they will tend to agree more often, and the *expected chance agreement* is higher. You might end up with something like this (*right*), where the expected chance agreement is 70% (0.7).



Even this isn't as simple as it sounds. There are versions of Kappa described by Fleiss and Cohen, and they measure *expected chance agreement* differently. But maybe the ratings are biased towards 'yes' because all the patients really *are* aspirators. In that case, it seems a bit unfair to put the agreement down to bias.

***Kappa assumes that any bias is not real, and is the fault of the raters.  
It gives a misleadingly low agreement when the patients or whatever being rated are truly biased.***

If you are going to use Kappa, make sure the data being rated are not heavily biased towards one outcome or the other.

## A bit more oddness...

An unexpected corollary of this effect is seen if you work out the intra-rater agreement separately for each of the raters in a study. It's possible that rater one might actually agree with themselves more frequently than rater two, but get a lower Kappa score because they are more biased in their ratings. Look at the table...

		Patient number																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
<b>Rater 1</b>	1 <sup>st</sup> rating	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
	2 <sup>nd</sup> rating	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
<b>Rater 2</b>	1 <sup>st</sup> rating	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓
	2 <sup>nd</sup> rating	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

Rater 1 and rater 2 both rated the same 20 patients on two separate occasions. They both disagreed on just one occasion (patient 20). They both had the same pairwise agreement (95%). But rater 2 has a higher Kappa score, simply because he or she was less biased; the expected chance agreement for rater 1 was 50%, but for rater 2 was 53%.

## Weighted agreement

Pairwise agreement and Kappa are most appropriate for categorical things: *yes* or *no*, *normal* or *abnormal*. But many questionnaires and scales have four categories (0=normal, 3=severe). Either two ratings agree or they don't, but this doesn't seem quite fair. If the two ratings are 2 and 3, then this seems like better agreement than 0 and 3.

In this case, it seems reasonable to award some credit for partial agreement. If we award an agreement score of 3 for two identical ratings, we might award a score of 2 when the ratings are one point apart, or a score of 1 when the ratings are two points apart. This is weighted pairwise agreement. You can go on to calculate weighted Kappa exactly as before, but using the weighted pairwise agreement.

But by using Kappa like this, you're getting on dodgy ground. In the extreme, you could have a rating scale with 100 categories and Kappa wouldn't be the right choice. Some people will argue that you should use the intra-class correlation coefficient instead, but this isn't necessarily correct. First, the two give almost the same answer; the ICC is often used to estimate Kappa. But more importantly, correlation isn't the right answer either. Read the later page on correlation.

## Remember...

This is quite high-level stuff. We've put it in here because it's something that lots of people come across, but don't worry if you don't understand it. All will become clear when you actually try to use it.

## Measuring validity - evaluating a diagnostic test

We'll talk about validity by example. You are investigating the swallows of normal elderly, and of patients who are known aspirators. Let's suppose you suspect a link between aspiration and apnoea duration.

### Picking a gold standard

If you wanted to measure the validity of your clinical diagnostic test, you first need to decide on the standard you're going to measure against. Let's suppose you pick VF as your criterion standard, so you're measuring *criterion validity*.

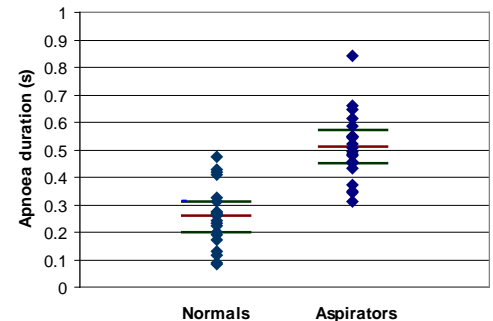
### Performing the study

You recruit 40 patients of whom 20 aspirate according to the VF gold standard. You need to compare the apnoea duration of the *normal* and *aspirator* groups. Using the two sample t-test,  $p = 0.000\ 000\ 065$ , and so the probability of this outcome by chance alone is tiny.

The *effect* is measured as follows:

- The mean apnoea duration for normals is 0.26 s.
- The mean apnoea duration for aspirators is 0.51 s.

This, along with the 95% confidence intervals, is shown on the figure. Subjects with a longer apnoea duration are more at risk of aspiration. { Except that it's more likely to be the other way round, but we'll ignore that! }



Given results this convincing, you might reasonably think to use the apnoea duration as a non-invasive clinical test to detect aspiration without exposing the subject to X-rays.

### Setting the test cutoff

To use the apnoea duration as a diagnostic test, you need to pick a cutoff value of apnoea duration.

- Below this cutoff level, you will describe the measurement as *normal*.
- Above this cutoff level, you will describe the measurement as *abnormal*;

There will, of course, be a trade-off:

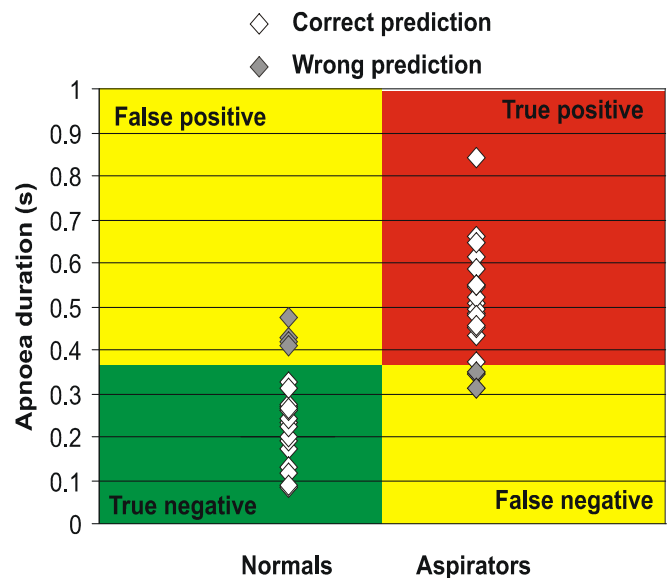
- If you pick a low cutoff level, you will detect most of the *aspirators*, but probably detect some of the *normals* as well. These would be called *false positives*.
- If you pick a high cutoff level, you will have few false positives, but probably miss some of the true aspirators. These would be called *false negatives*.

As a first attempt, you might well pick 0.37. It's half way between the two mean values, so might be a reasonable trade-off. The figure (*right*) shows what happens if you pick that cutoff.

- Subjects in the red and green sections were predicted correctly.
- Subjects in the yellow sections were predicted wrongly.

By just counting the numbers of dots in each section, you can create a contingency table:

	Normal on video	Aspirator on video
Predicted aspirator	4	17
Predicted normal	16	3



## Validity, pairwise agreement and diagnostic accuracy

As a first attempt, you can use very similar statistics as before; *pairwise agreement* between the new test and the VF standard. This would be called *diagnostic accuracy* ie. the overall proportion of times that the new test gets the correct answer. In our case, this was 16 out of 20 normals and 17 out of 20 aspirators. In total, the new test was correct on 33 out of 40 occasions, so the diagnostic accuracy is 83%. You can correct for chance agreement just like for Kappa. The expected chance agreement is about 0.5 (50%), so our corrected agreement by treating chance agreement as zero is 0.66. This isn't called Kappa, but it's the same idea.

## Sensitivity and specificity

In practice, you can do a bit more because you know the right answers from your VF standard. Each patient can be classified according to the following table:

		What the gold standard said	
		Negative	Positive
What the new test said	Positive	FALSE POSITIVE	TRUE POSITIVE
	Negative	TRUE NEGATIVE	FALSE NEGATIVE

### Sensitivity

*The proportion of people with the disease that are correctly detected.*

In our case (previous page), we detected 17 out of the 20 aspirators, or 85%. Notice that to calculate sensitivity we only need to study people who definitely do have the disease and see how many we spot with the new test. People who don't have the disease do not enter into the calculation.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

	✓

### Specificity

*The proportion of non-diseased people that are correctly identified as such.*

We correctly identified 16 out of the 20 normals, or 80%. To calculate specificity, we only need study people who definitely don't have the disease and see how many we reject with the new test.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

✓	

## Positive and negative predictive values

Sensitivity and specificity are important, but don't take account of the population split, the proportions of normal versus diseased. Suppose you have a screening test for ALD (a rare disease, see *Lorenzo's oil*) that is correct 90% of the time, ie. sensitivity and specificity are both 90%. What does a positive test mean? There's 1 in 20,000 people who have the disease, but a 1 in 10 chance the test is wrong. Which is more likely?

Clinically more appropriate statistics are predictive values. This is what matters to your patient; if they have a positive (or negative) test, how likely is it that they really have (or don't have) the disease? In our ALD example, a positive test means you have about a 1 in 2,000 chance of having the disease. Yes really. You don't believe it, do you?

### Positive predictive value

*The proportion of positive tests that really have the disease.*

17 out of the 21 indicated positives were actually positive, so PPV = 81%.

$$\text{PPV} = \frac{TP}{TP + FP}$$

	✓

### Negative predictive value

*And likewise...The proportion of negative tests that really don't have the disease.*

16 out of the 19 indicated negatives were actually negative, so NPV = 84%.

$$\text{NPV} = \frac{TN}{TN + FN}$$

✓	

## Statistical tests if you want them - the chi-squared test

You've reduced the numeric measurement (*apnoea duration*) to a binary categorical variable (*predicted aspirator* or *predicted normal*), and given some summary statistics sensitivity, specificity, etc. that tell you how well the test works. Often you won't need any more, but if you only have a few patients you might want to show that the effect you see isn't just a fluke.

The chi-squared (pronounced kye squared) test is appropriate for analysing such data. The table (right) gives some clue how the test works:

	Observed N	Expected N	Residual
1.00	4	10.0	-6.0
2.00	17	10.0	7.0
3.00	16	10.0	6.0
4.00	3	10.0	-7.0
Total	40		

- Assume that the subjects ought to be distributed evenly amongst the four quadrants, 10 in each\*;
- Work out the residual - the difference between the actual and expected distribution of subjects;
- Work out how likely the actual pattern is by chance alone.

Here's the rest of the output, very like the output for the other tests. The chi-squared statistic is loosely equivalent to the *r* value in that it's a statistic calculated directly from the data. The sig (or p) value can be calculated from chi-squared if you know the number of degrees of freedom. In this case, the value of 0.001 means this arrangement of patients among the four quadrants is very unlikely.

Chi-square	17.00
Df	3
Asymp.sig	.001

More commonly, you will want to compare one test with another to show which is best. A slight alteration of the chi-squared test can be used.

### Some other things about chi-squared

- Outliers aren't a problem for chi-squared, because the data are categorical and not numeric;
- Since the data are not numeric, and chi-squared doesn't rely on properties of the normal distribution it is a *non-parametric* test.

### Fisher's exact test

Chi-squared can be adjusted for small numbers of measurements, but it is not recommended for use when any of the expected counts are less than 5. You can read the expected counts from the SPSS output above. In this case, you should use Fisher's exact test. Fisher's exact test works for contingency table data just as for chi-squared, but works out the *exact* probability of the observed distribution by laboriously going through all the different possible combinations.

*NOTE: for chi-squared, the actual observed count might be zero. It's the expected count that is important in deciding whether or not you have enough data to believe the results.*

### But beware...

What was described here was a retrospective analysis; the new test was developed, and then evaluated *on the same data*. It would be better to do a *prospective* trial of the new test with fresh patients. The performance would inevitably be poorer, because it is very unlikely that the best cutoff for one set of subjects would be best for a second set.

---

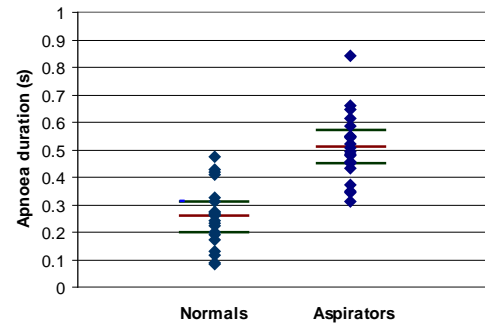
\* It isn't always this easy to work out expected values. Our numbers came out like this because there were equal numbers of normals and aspirators. In reality, go to a stats book to find out the proper way to work out the expected values.

## Some more about cutoffs – the receiver-operator characteristic

Back to the original data. We picked a cutoff of 0.37 because it seemed about right, but you could put the cutoff anywhere. In our example, a lower cutoff would improve sensitivity with more true positives so we would detect more aspirators. This is at the expense of specificity because with more false positives we falsely diagnose more normals as aspirators.

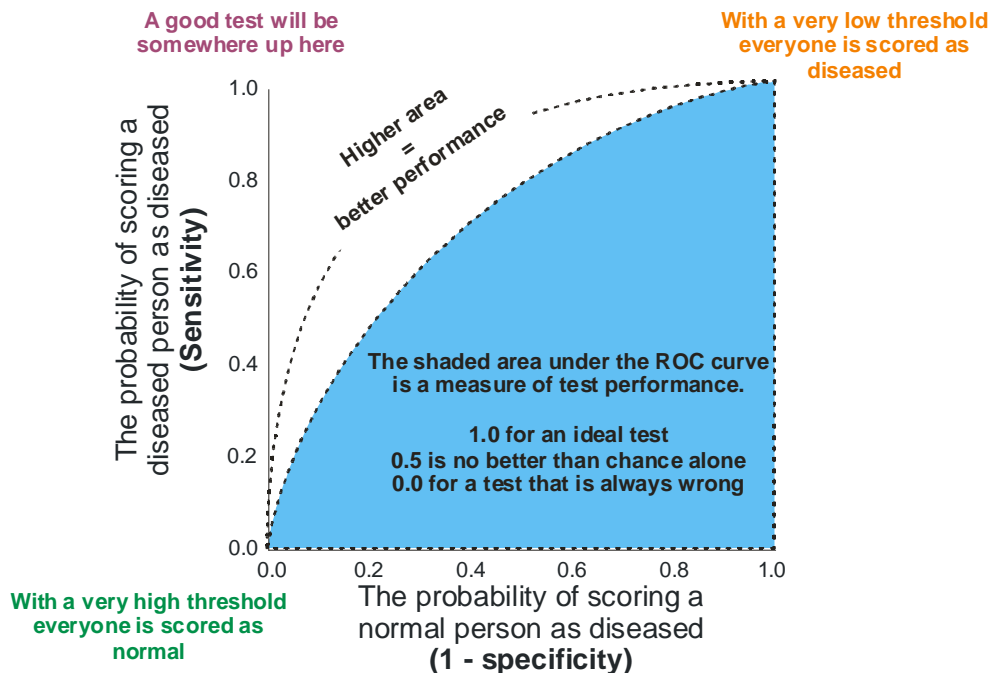
Clearly, there is a trade-off to be had. The cutoff we choose to use will depend on what the test is for. If this was a screening test for cancer, it is essential we don't miss any positives. We would want a very high sensitivity, at the expense of specificity. False positives scare the patient and are a drain on healthcare, but false negatives are potentially fatal.

If this test was being used to recommend radical surgery, things change. We don't want to operate on anyone unless we're absolutely sure they have the disease, so we would require high specificity.



## Receiver-operator characteristics

The receiver-operator characteristic is a graphical means of showing this trade-off between sensitivity and specificity.



- Suppose we use a very high cutoff (0.9, say) to diagnose aspiration. We'll score everybody as normal, whatever their true status. That gives us a point on the bottom left of the graph.
- If we used a very low cutoff (less than 0.1), everybody gets scored as aspirators. That gives us a point on the top right of the graph.
- For a good test, you want to score a diseased person as diseased, but not to score a normal person as diseased. This would (ideally) give us a point at the top-left of the graph. The closer to the top left, the better.

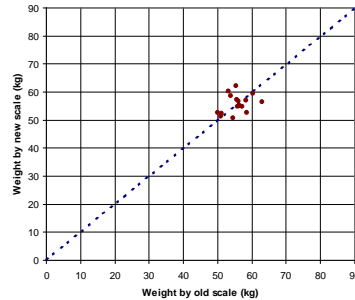
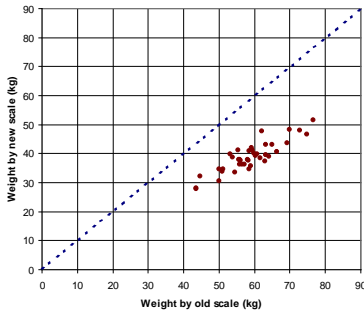
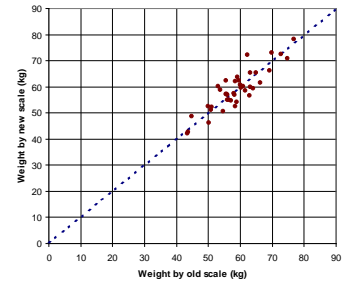
It turns out that the area under this curve, the blue shaded area in the picture, is an absolute measure of how well the test can separate the two patient groups. It is independent of the particular cutoff you choose to use.\*

\* If you really want to know, the area under the ROC curve is *the probability that a randomly chosen aspirator will have a higher apnoea duration than a randomly chosen normal person*. With a value of 1, this would mean that any aspirator has a higher duration than any normal – so the test is very useful. With a value of 0.5, it's 50-50 – this means the aspirators and normals are all mixed up, and the test has no diagnostic value.

## Measuring agreement – The use and abuse of correlation

**Correlation shows only that two variables are linked!!!  
It is completely insensitive to the actual numerical values of the two variables.**

Correlation is often used to measure agreement between two different instruments that are meant to be measuring the same thing. Suppose you just bought a new set of weighing scales. You want to check they read the same as the old ones, and so you recruit 40 of your best friends and weigh them using both sets of scales. Here are your results (*right*). If the two scales agree, the points should lie on the blue line of equality. The correlation coefficient is 0.88, and you conclude there is good agreement between the two sets of weighing scales. But is there? Look at the two figures below:



In the first example, there's a clear problem. The new scales are systematically under-measuring the weight relative to the old scales, *but the correlation coefficient is still 0.88*. That's because *the correlation coefficient is insensitive to the numbers involved*. There is still a relationship between the measurements, and so the correlation coefficient is the same.

In the second example we've used the same weighing scales and the same measurements, except this time plotted only 15 subjects' data, excluding some with very high and low weights. But now,  $r$  is only 0.33. However, it would be ridiculous to say the scales have suddenly gotten worse because the data still lie close to the blue line of equality. Nevertheless, if you imagine the graph axes were taken away, there is now no clear linear relationship between the values.

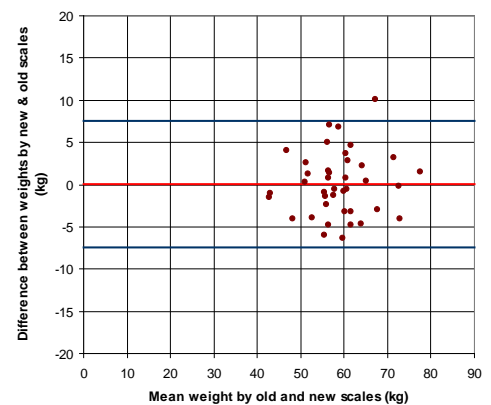
**A correlation coefficient improves when the overall range of the data is increased.**

The proper way to analyse the data was described at length by Bland & Altman, and is shown (*right*). You should:

- Calculate the difference between each pair of measurements. This represents the error, though you can't say which pair of scales is causing the error.
- Calculate the mean of each pair of measurements. This represents your best estimate of the true weight.
- For each subject, plot the difference (y-axis) against the mean of the two measurements (x-axis).

The appropriate statistics are:

- The mean difference between the two measurements (0.05 kg), which reflects any systematic difference between the scales.
  - The standard deviation of the difference between the two measurements (3.75 kg), which reflects the overall random error.
- NOTE: you can't say which set of scales is the inaccurate one.*



You can then mark the mean difference (in red), and the limits of agreement (in blue, 2 standard deviations either side of the mean) on the graph, as shown. In an ideal world, all the points would lie on the X-axis where there is no difference between the measurements. Deviations from the ideal can be interpreted as follows:

- For 95% of all measurements, the difference will lie within the limits of agreement.
- If the mean difference (red line) is not zero, then one of the tests may be systematically over-estimating with respect to the other.
- You could calculate a standard error of the mean, as described earlier, to better evaluate whether the discrepancy is statistically significant. In effect, you would be conducting a paired t-test between the two sets of scales.



# Useful resources

## Books we like

- **Statistics at square one.** *TDV Swinscow and MJ Campbell.* BMJ Books.  
A very simple, but extremely useful book. Covers a lot of the same things we do. Now in its tenth edition, so it can't be doing too badly.
- **An introduction to medical statistics.** *Martin Bland.* Oxford Medical Publications.  
A classic of its time. More comprehensive than the BMJ book, but more maths too.

## Papers we like

*Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995; 274(8): 645-51.*

Contains guidelines on good practice if you're developing a diagnostic test, but is also of general interest.

*Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476): 307-10, 1986.*

Another classic of its time - on measuring agreement. Read it now.

*Begg CB, Cho MK, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA. 1996; 276: 637-639.*

If you're doing a controlled trial, you should read this paper. Some journals insist you follow the guidelines.

## Web sites we like

- **Statsoft**  
A very comprehensive on-line statistics handbook, with some cute animated pictures:

<http://www.statsoft.com/textbook/stathome.html>

You used to be able to download the whole thing, but at last check that had changed. If you have trouble, contact Michael D for your own copy.

- **Power calculations**

<http://www.stat.uiowa.edu/~rlenth/Power/index.html>

This web site is very useful for research design. You can work out sample sizes for a range of experimental designs.

- **Bandolier**

<http://www.jr2.ox.ac.uk/bandolier/>

An on-line journal for evidence-based medicine. Very evangelical, but worth a look.