National
SAMR
Fostering Medical Research

Understanding statistics: a guide for medical students

Author: James Gupta
Communications Officer, NSAMR 2013
Medical Student, University of Leeds

Understanding statistics: a guide for medical students

About This Guide

'Statistics' probably isn't what you had in mind when you decided to pursue a career in medicine, but at least a basic understanding is required to allow you to critically appraise published medical research,

Undoubtedly, medical statistics is a vast, complex field, but fortunately you can get a good grounding by learning a few of the key concepts, which this guide aims to introduce you to.

We're going to use a real research paper as our case study, along with a few made up examples. The study aimed to find out whether the 'Mediterranean' diet was effective at reducing heart attacks and can be found here:

http://www.sciencedirect.com/science/article/pii/S0140673694925801

It may seem like nonsense now, but by the end of this guide, you should have no problem understanding this statement found near the end of the paper:

"With the proportional-hazards model, the risk ratio was 0.24 (0.11-0.55, p<0.001), 0.27 (0.12-0.59, p=0.001), after adjustment (97% CI 0.11-0.65)."

Why Do Clinicians Need Statistics?

Imagine a consultation with a patient. A 55-year old man called John with diabetes who wants to know whether or not he should take an ACE Inhibitor to lower his blood pressure, which has been slightly raised the last few times it was checked.

The hundreds of studies conducted suggest ACE inhibitors are a safe and effective way to lower blood pressure, and therefore reduce the risk of stroke and other diseases, especially in patients with diabetes.

The obvious answer seems to be yes, but consider this: there has never been a study conducted on John, who has never taken this tablet before. John may have a very rare gene or be taking some other combination of tablets that means he has a fatal reaction to ACE inhibitors. On the less-extreme but still important end of the spectrum, they might not lower his blood pressure at all, or they could lower his blood pressure but he still has a stroke.

In short, this is why statistics are vital. There is still a great deal that we do not know in medicine. Genetic medicine is beginning to scratch the surface of our understanding about *why* some people respond better to certain drugs than others, but the era of 'personalised medicine' that could tell us for certain whether or not John would personally benefit from ACE inhibitors is still far away, so for the foreseeable future our best evidence will come from large trials where hundreds or thousands of 55-year old male diabetics were given ACE inhibitors, the results of which will tell us whether or not John will benefit from taking one. Probably.

1) Averages & Standard Deviation (SD)

When looking at the results of clinical trials, we often need to take the average (mean or median) results. A study of a new type of painkiller might claim the following

"On average, pain was reduced by 25% compared with placebo"

This does give you some information – but it doesn't give you the whole story, as shown below:

	Reduction in Pain	
	Trial 1	**Trial 2**
John	27%	1%
Anne	24%	48%
Will	23%	11%
Sarah	23%	60%
Tess	28%	5%
AVERAGE	**25%**	**25%**

The 'average' pain reduction in both trials was 25%, but clearly the results are very different. In Trial 1, we see that there isn't that much variation between each subject, so we can quite accurately say that the drug does indeed reduce pain by 25%. However, we have a lot of variation in Trial 2. Some patients saw hardly any benefit, whilst others saw reductions much greater than 25%.

Standard Deviation (SD) is a way of quantifying these variances. It is usually reported alongside averages to let us know how close (low SD) or far apart (high SD) the data are.

The SD for the average in Trial 1 is 2.1, and the average for Trial 2 is 24.2. This would be reported as '25% ± 2.1' and '25% ± 24.2' respectively.

Averages make a lot of sense when looking at something like pain which can be graded on a continuous scale, but what about 'all or nothing' measurements. A drug may reduce the risk of stroke in at-risk patients from 20% to 5%. This 15% reduction would be fantastic, but it is important to remember that the reality is that the 15% benefit is not shared equally across the patients, and if you give the drug to 100 patients, 5 of them *will* have a stroke and 95 of them *will not.*

Important definition:

The **Standard Deviation (SD)** is a measure of how much variation there is from the average in a set of data

2) Absolute and Relative Differences

Every so often, a story makes its way into the media about a drug that reduces the risk something (heart attacks, cancer, hip fractures etc) by 50%, and there is uproar if the NHS refuses to pay for it. After all, 50% is a huge reduction, so why would we not want to pay for such a miracle drug?

More often than not in situations like this, someone has failed to grasp the difference between absolute and relative differences.

Here's a quick summary of the results from the Mediterranean diet study:

	Normal Diet	**Mediterranean Diet**
Total people	303	302
Heart attacks	17	5

5 out of 302 people on the Mediterranean diet had heart attacks, compared to 17 out of 303 on the normal diet.

Normal diet: 17/303 had heart attacks = 5.61%

Mediterranean diet: 5/302 had heart attacks: 1.67%

1.67 / 5.61 = 29.5%

So here, we can see that the 'treatment' appears to have reduced heart attacks by almost 30%, which sounds great, but can be a bit misleading. A heart attack is actually a fairly rare event. Even on the normal diet, only 5.61% of people had a heart attack. The relative reduction between the 'normal' and 'Mediterranean' groups is 29.5%, but the 'absolute' reduction (calculated as 5.61% - 1.67%) is just 3.95%.

This is not to be sniffed at – the authors have shown that we can reduce heart attacks by almost 4% without any drugs or surgery, but it paints a very different picture to the more dramatic sounding '30% reduction'. It is important to consider both.

Important definition:

The **Relative Risk Reduction (RRR)** describes the effect of an intervention in relative terms between two groups

The **Absolute Risk Reduction (ARR)** describes the effect of an intervention in absolute terms, so is typically far lower than the RRR as it takes this into account

Understanding the difference between the two is **vital**!

3) Risk Ratios

Risk ratios are simply another way of describing the difference between two groups. Mathematically, it is calculated as:

Risk Ratio (RR) = risk in treated group / risk in control group

Where the risk of a given group is simply the number of people who experience an event (i.e. a heart attack) divided by the total number of people in that group – simple!

A RR of 1 would mean that risk in both groups is the same.

A RR >1 would mean that risk in the treated group is greater.

A RR < 1 would mean that risk in the treated group is lower.

So an RR of 0.5 means that the treated group has half the risk of the non-treated group.

Always bear in mind that 'risk' in this case is not always bad! Imagine a trial comparing the rate of successful pregnancies in a new IVF procedure compared with the standard procedure, the results may look like this:

New IVF: 10 women treated, 6 became pregnant

Old IVF: 10 women treated, 3 became pregnant

The 'relative risk' of the new IVF treatment is 1.5 – whilst it may seem counterintuitive to see a 'good' outcome described as a 'risk', that's just the way it is!

Important definition:
The Risk Ratio (RR) compares the 'risk' of a particular event in the treated group compared to that of the control group. The 'event' can be good or bad, and an RR < 1 means that the risk is lower with treatment

P Values and Confidence Intervals

These cause a lot of confusion, but they really are quite simple. If you flip a coin 10 times, you would expect to get 5 heads and 5 tails. However, it's entirely possible that you will get all heads, all tails or anything in-between. The same concept applies to experiments, so if you take a trial of 10 people and they all get better, how do you know this wasn't just a chance result?

We're going to move away from the Mediterranean study for a bit and instead focus on a made-up drug called GlucoVax, which aims to reduce blood sugar levels in diabetic patients. So when we're looking at a statement such as:

"Patients on GlucoVax had blood sugar levels 35% lower than patients receiving placebo"

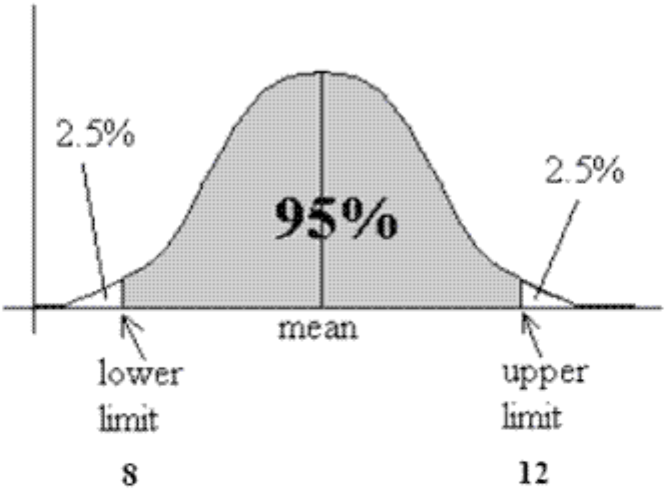
How can we know it is representative of the 'true' result, and not just a chance occurrence? This is where p-values come in: they tell us how likely a result is to be 'true'. In a real paper, the result would be reported like this:

"Patients on GlucoVax had blood sugar levels 35% lower than patients receiving placebo (p=0.02)".

A p-value of 0.02 means that there is a 2% chance that this is just a random occurrence. In science, a p-value lower than 0.05 (a less than 5% chance) is considered 'statistically significant'.

Confidence Intervals (CI's) serve a similar purpose to p-values, but before we jump in to CI's, let's talk about Normal Distribution.

Many biological variables such as height, IQ and even the size of your red blood cells are 'normally' distributed. This means that, whilst they do vary from person-to-person, there is a 'modal' (peak) value, and the vast majority of the population isn't too-far from this.



A CI gives us a range of values that we can be fairly confident that the 'true' value lies within. Technically, any interval could be used but as with p-values, there is an accepted convention whereby the 95% CI is used, therefore giving a range of values we can be 95% certain the 'true' value lies within. It is reported like this:

"Patients on GlucoVax had blood sugar levels 35% lower than patients receiving placebo (95% CI 30-40%)".

We can't really say that "GlucoVax reduces blood sugar levels by 35%", especially on such a small trial. However, we can say that the reduction is, on average 'probably' no smaller than 30% and no greater than 40%.

On balance, a result such as this would probably make us inclined to prescribe GlucoVax. When looking at CI's, one of the most obvious things to see is whether it crosses the 'no effect' threshold, for example:

"Patients on GlucoVax had blood sugar levels 10% lower than patients receiving placebo (95% CI -15-20%)".

In this case, the lower limit is -15%, so we can't confidently exclude the possibility that it actually raises glucose levels!

Now let's go back to the Mediterranean study and evaluate their claims, the authors state that:

"The risk ratio of cardiac death was 0.19 (95% CI 0.06-0.65, p < 0.002), a reduction of 81%."

Knowing what you now know about p-values and CI's, does this seem reasonable to you? We have a 81% (relative) reduction, the p-value is well within the range for 'statistical significance'. The Confidence Interval is intriguing – whilst the risk ratio was 0.19, the 95% confidence range is given as 0.06-0.65. So, whilst the results of this particular trial gave a risk ratio of 0.19, the 'true' result could lie anywhere between the confidence interval. In the worst-case scenario, therefore, we have a risk ratio of 0.65 (a 35% reduction in risk), but in the best-case, we could see a risk reduction of 94%!

Important definition:
In pragmatic terms, the **p-value** tells you the probability that these results occurred by chance.

The **95% Confidence Interval (95% CI)** gives you a range of values we can be 95% certain that the 'true' value lies within

5) Statistical Tests

There are hundreds of statistical tests available, ranging from the relatively simple ones which could be calculated by anybody competent with maths, to the fairly complex ones understood only by professional statisticians, to the overwhelmingly complicated, multi-factorial ones which can only be performed by a computer.

Fortunately for us, it isn't necessary to learn all the ins and outs of statistical tests, and in reality there are only a handful that are used on a regular basis. What we do need is a basic appreciation of how the commonly used tests work, when they should be used and in which circumstances they are unreliable.

When deciding on a statistical test, the main consideration will be what type of data you're looking to analyse. Let's say we were investigating pain levels where patients were asked to rate their current pain on the following scale:

1) No pain
2) Mild pain
3) Moderate pain
4) Severe pain

There is a clear rank order here – a score of 1 indicates a lower pain score than a score of 3. If we recorded an individual's pain levels every day for a week, we could calculate the mean pain score and the result (say for example, 2.5) would be meaningful – the patient on average reported pain levels between mild-moderate.

However, whilst we have established a certain value in these scores, they're not *really* numbers in the same way that Weight (kg) would be.

	Pain (7 day average)	**Weight**
John	2.0	76kg
Adam	4.0	152kg

There are a few things we can say about these data:

- Adam weighs more than John
- Adam weighs twice as much as John
- Adam was in more pain than John

However, what we can't say is 'Adam was in twice as much pain as John'. Whilst we know that Severe pain is worse than Mild pain, we don't know that Severe pain is specifically 'twice as bad' as Mild pain – there is no reason to assume that these variables exist on a linear scale.

The difference is important to appreciate as it determines whether you should use a parametric or a non-parametric test. In a nutshell, a parametric test focuses on the absolute differences between your data. They are more likely to give a favourable result demonstrating statistical significance, but they are only valid if your data follows a normal distribution and if the scale is meaningful.

The result of a parametric test would not be reliable when evaluating our pain scores above, so a non-parametric test which focuses on the rank order of the variables rather than the difference between them, should be used.

Another consideration when choosing a statistical test is whether the data are 'paired' or not. Consider two studies:

Study 1: Compare blood glucose measurements taken from diabetic patients before and after a meal

Study 2: Compare Blood glucose measurements taken from diabetics and non-diabetics

On the face of it, these studies have a lot in common: they are both comparing two measurements of blood glucose. Blood glucose is a *real* number (a reading of 8.0 is 2x greater than a reading of 4.0) and normally distributed, so we can use a parametric test.

The difference, however, is that Study 1 is comparing two observations from the same sample – there is a meaningful relationship between the two measurements (they are 'paired'), whereas Study 2 is comparing two observations from different samples and is therefore said to be 'unpaired'.

Four of the most commonly used statistical tests are summarised below:

	Parametric	**Non Parametric**
Paired	Paired t-test	Wilcoxon matched-pairs test
Unpaired	Unpaired t-test	Mann-Whitney U test

Important definition:

Parametric data follows a normal pattern of distribution, and applies to many (but not all) biological phenomena

Data are said to be **'paired'** if it consists of two or more measurements taken from the same subject

6) Proportional Hazards and Risk Adjustment

Let's go back to the statement we were trying to understand:

"With the proportional-hazards model, the risk ratio was 0.24 (0.11-0.55, p<0.001), 0.27 (0.12-0.59, p=0.001) after adjustment (97% CI 0.11-0.65)."

Hopefully it makes a lot more sense to you now. There are still a couple of things we haven't covered: what's a proportional hazards model, and why have they 'adjusted' the results?

An in-depth explanation of proportional hazards is something well beyond the scope of this guide, but at a basic level these are functions widely used in clinical papers. They take into account the factor being tested (in this case, the Mediterranean diet) and also other potential factors (i.e. age, weight, cholesterol levels etc).

In this study, the two groups were matched as closely as possible, but Table 6 shows some differences between the factors such as Age, Weight, Triglyceride count etc. In this case 'adjustment' of the results refers to the authors performing a calculation that gives the results 'if both groups were exactly equal at the beginning of the study'. For obvious reasons, the reliability of these calculations can be disputed but in this case, the adjusted values are not much different from the non-adjusted, and they seem proportional to the level of difference between the two groups. Furthermore, for the adjusted results they have used a harsher test (97% CI instead of 95%) to assess its significance.

7) Putting It All Together

"With the proportional-hazards model, the risk ratio was 0.24 (95% CI 0.11-0.55, p<0.001), 0.27 (0.12-0.59, p=0.001) after adjustment (97% CI 0.11-0.65)."

So, is a Mediterranean diet good for your heart? The non-adjusted and adjusted results are pretty similar, so let's just focus on one of them to keep things simple – I'll choose the non-adjusted result.

The risk ratio is 0.24, so in this study people taking a Mediterranean diet experienced a quarter of the amount of cardiovascular events as people with a 'normal' diet. The 95% CI ranges from 0.11 to 0.55 so we can conclude that a Mediterranean diet probably does reduce the risk of a heart attack by at least about 50% (0.55, upper

limit), and quite possibly more.
We have a p-value <0.001, so it is very unlikely that these results have come about by chance.

The authors conclude that "An alpha-linolenic acid-rich Mediterranean diet seems to be more efficient than presently used diets in the secondary prevention of coronary events and death" – do you think the stats back up this claim?

8) Number Needed to Treat (NNT)

One final measurement I felt important to include is the Number Needed to Treat (NNT). It isn't reported on a lot of research papers which is unfortunate because it's a very simple and effective way of expressing the impact of a treatment.

Earlier we discussed the differences between absolute and relative risk – how a 50% reduction in a 1-in-100 event is only actually a 0.5% reduction in absolute terms. We also mentioned 'all or nothing' events, and the fact that for events such as heart attacks, describing a % reduction can be slightly misleading, as the fact is that for any given patient, they either 100% do or 100% do not have a heart attack.

As the name implies, calculating the NNT tells you how many people you would have to treat in order to prevent one 'event'. Calculating the NNT is very intuitive: it's the inverse of the absolute risk reduction (1 / ARR). If your absolute risk reduction is 0.2 (20%), it makes sense that you would need to treat 5 people to reduce one event.

So how many people would we need to 'prescribe' a Mediterranean diet to in order to prevent a cardiovascular event?

Normal diet: 33/303 had events = 10.89%

Mediterranean diet: 8/302 had events: 2.64%

2.64 / 10.89 = 24.24%

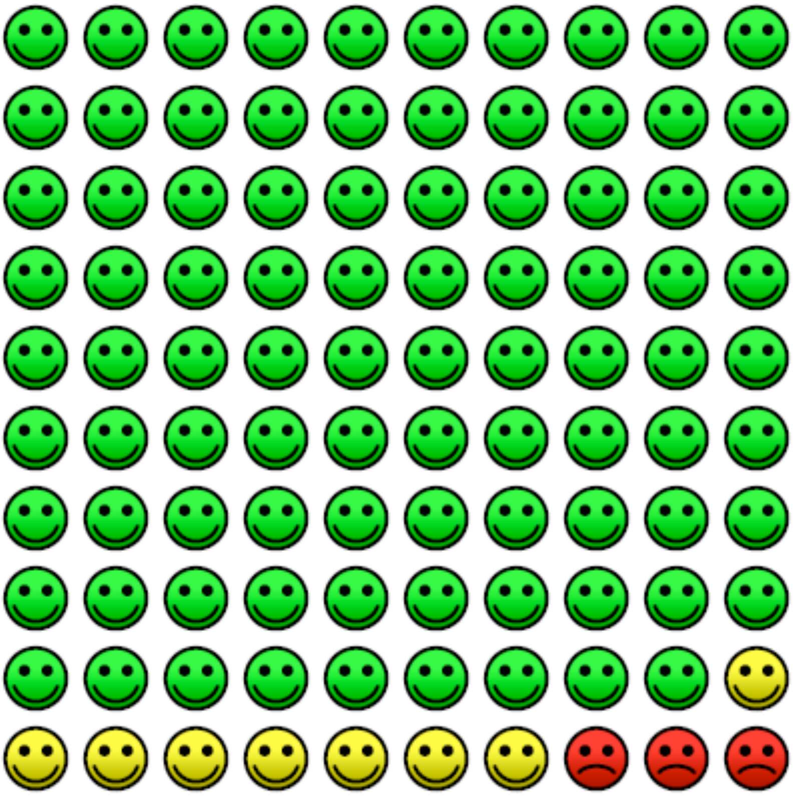
Relative risk reduction = 24.25% (0.24)

ARR = 8.25% (0.0825)

NNT = 12

So, if 12 people who were at risk of a heart attack followed the Mediterranean diet, we would prevent one heart attack.

From this, you can also create the fantastically visual and highly intuitive 'Cates' Plot', as shown below:



We have 100 faces here, each a ‘patient’ in the Mediterranean study. The green faces represent people who *would not* have a heart attack, regardless of whether they kept to a normal diet or the Mediterranean one. The red faces are people who *would* have had a heart attack, regardless of whether they had the Mediterranean diet or not.

The yellow faces are what we’re really interested in – these are people who would have had a heart attack on a normal diet, but would be ‘saved’ by the Mediterranean diet.

This is a great way to show both relative and absolute risk in a meaningful way, and if you’re ever struggling to get to grips with the statistics presented in a paper, one of these diagrams with the NNT is a good place to go.

Important definition:
The **Number Needed to Treat (NNT)** is the inverse of the Absolute Risk Reduction, and tells you how many people you would need to treat with a particular intervention in order to prevent one event.

Sugested external resources

I hope you’ve found this guide helpful! It really is just a crash course in statistics though, if you want a more comprehensive introduction you should read the excellent book ‘How to Read a Paper’, by Dr. Trisha Greenhalgh.

There is a website, www.thennt.com which publishes highly accessible appraisals of various papers to assess whether or not a treatment is effective which I’d also recommend.

Otherwise, the best way to learn is just to read through papers which you will do over the course of your career anyway. The CASP tool is a fantastic guideline for critically analysing papers in a structured way (http://www.casp-uk.net/).

Finally, if you want to generate your own Cates plots you can do so here: http://www.nntonline.net/visualrx/.