DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

WEEK 7

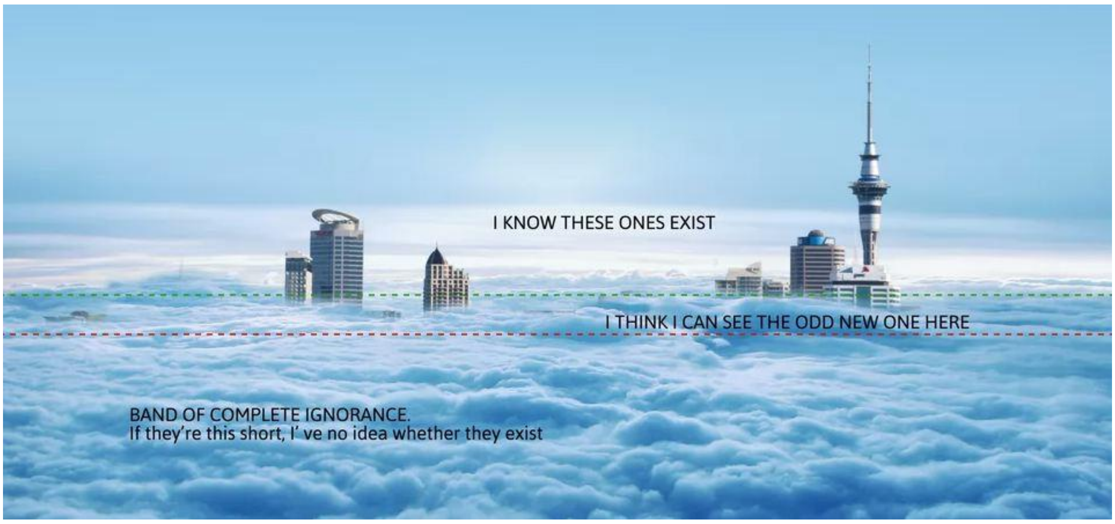
OUT OF THE FOG AND INTO POWER by Chris Wild

Hi. When are statistical tests good at detecting when true treatment differences exist, and when are they bad at it?

We'll be talking about the concept of the power of a statistical test, and also of Type I error, concluding true differences that don't actually exist at all. A warning -- this is quite deep water, ideas that many people have difficulty with. So if you've already found the earlier videos in this week heavy going, you should probably skip this one and go on to Week Eight. You won't be tested on this in the Week Seven test.

So now I'm only addressing those brave souls who are prepared to give this a go. I've mentioned power, but how does the fog come into this?

Here's a picture of the Auckland skyline.


I KNOW THESE ONES EXIST
I THINK I CAN SEE THE ODD NEW ONE HERE
BAND OF COMPLETE IGNORANCE.
If they're this short, I' ve no idea whether they exist

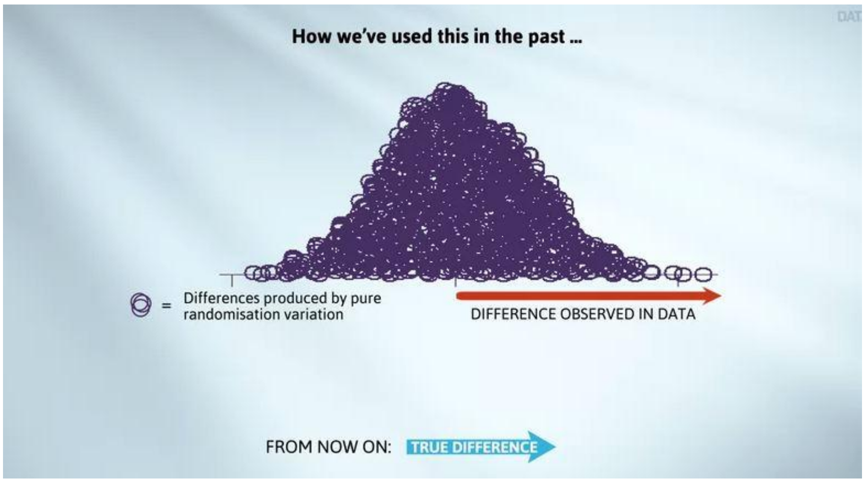
Here's the same picture, but with a thick, ground-hugging sea fog hiding most of the city. If that was our only view, the only buildings we'd know actually existed would be those tall enough to poke up above the fog.

So here are some that I know exist, because they poke up well through the fog. And there's a few more shorter ones I think I can maybe make out. And if they're shorter than that, well, I've no idea whether there's anything there or not.

So where am I going with this? Randomisation variations, sampling and uncontrolled natural variation are like the fog. We can't reliably determine whether a true difference between treatment groups exists unless that true difference is big enough to poke up through the fog.

We will now start developing this basic idea. Earlier, we looked at differences between the centres of artificial groups defined by random labelling. We decided that we could only conclude that a true treatment difference existed if the luck of the randomisation draw never (or at most rarely) produced anything as big as we had seen in our data.

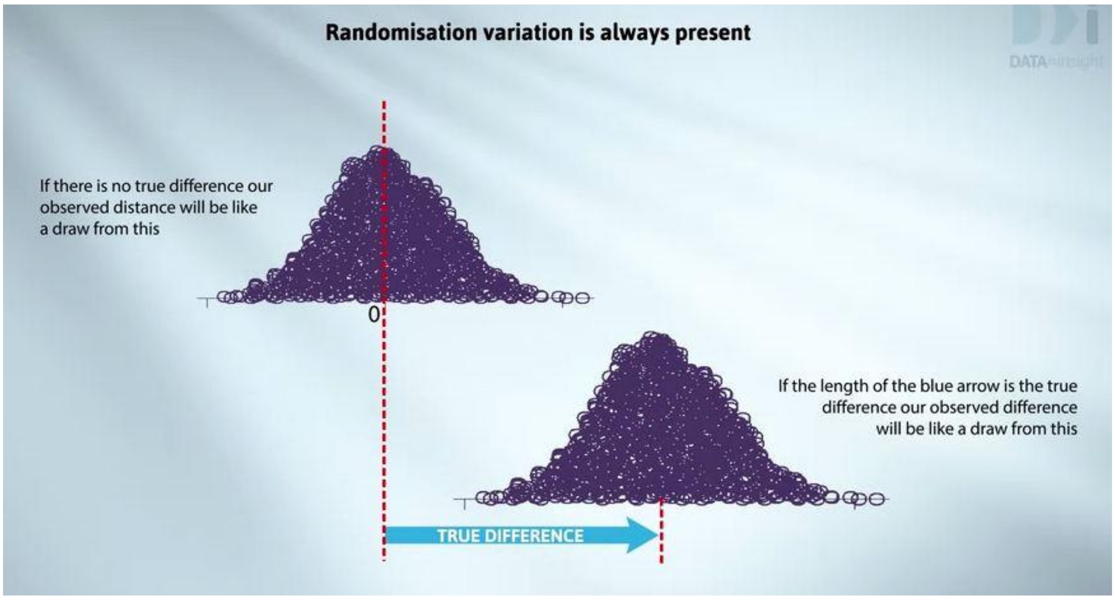
So that's the thinking about taking the treatment difference we see (the one in our data) and thinking about how to reach a conclusion from it.


How we've used this in the past ...

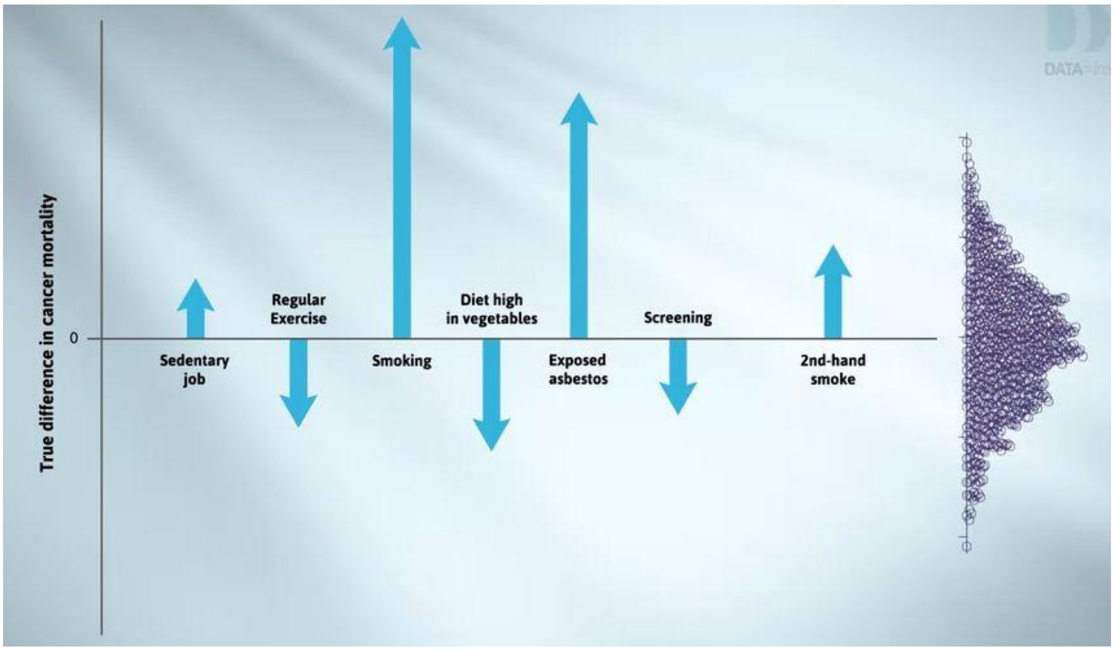
I now want to switch and start thinking about the difference we can't see, the underlying true difference. To take an example of an unknown true treatment difference from the last video. One would be the difference between the percentage cocaine-free if everyone was treated with Desipramine and the percentage cocaine-free if everyone was treated with Lithium.

We're never going to be able to reliably detect the existence of true differences that are smaller than the differences produced by the luck of the randomisation draw. It's as though randomisation variation has produced a fog around zero that makes it impossible to see true differences if they are in that region.

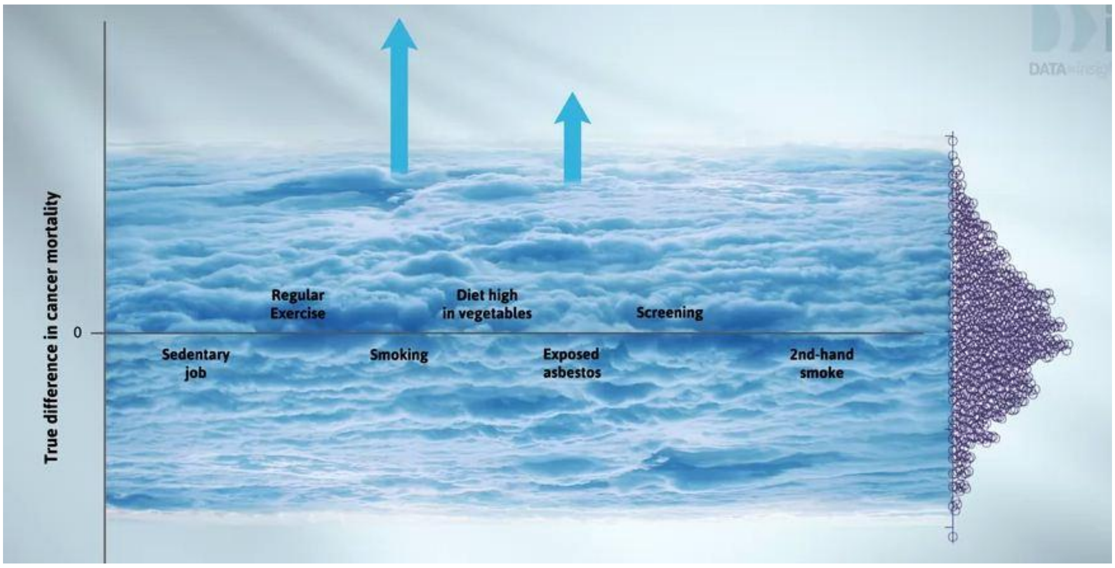
Randomisation variation and other forms of random variation don't go away just because there's a true difference. Instead of variation centred at zero, we then have variation centred around that true difference.


Randomisation variation is always present

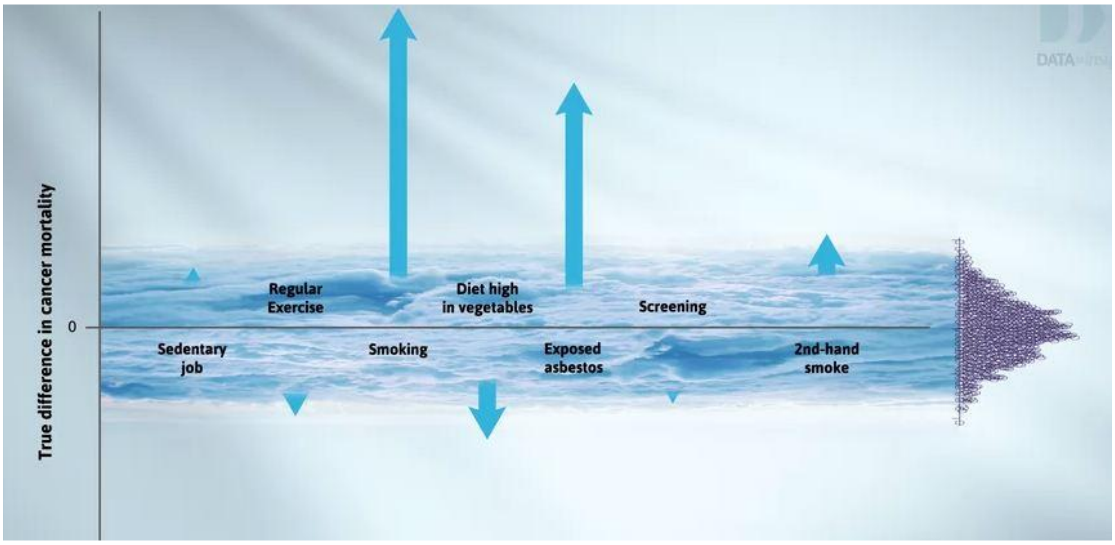
Consider this scenario -- the true differences in mortality rates due to having or not having the depicted characteristics.



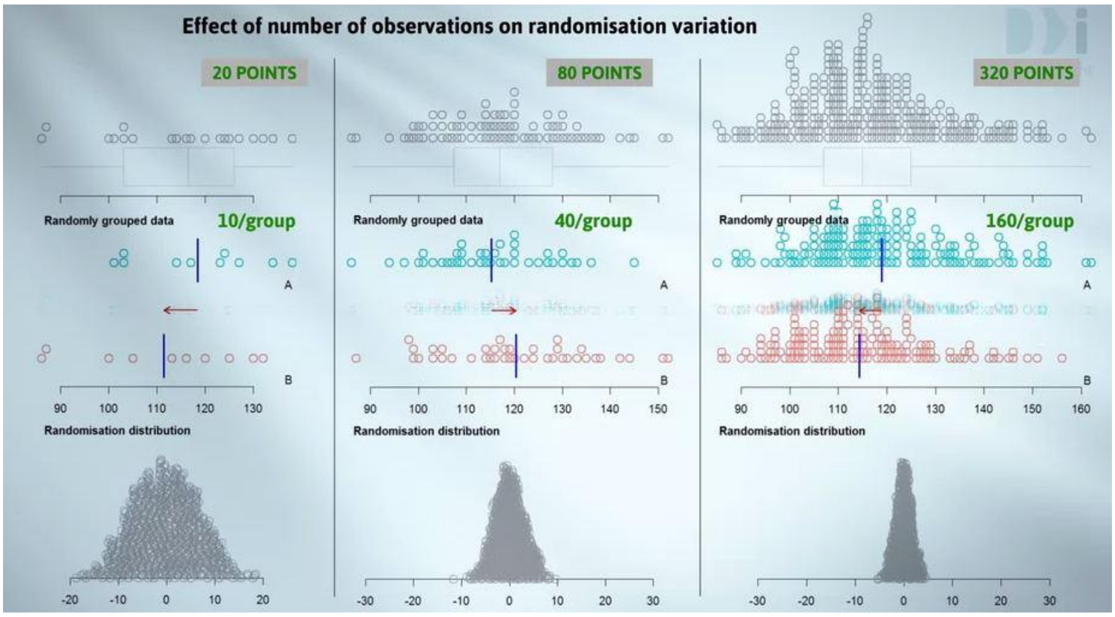
Suppose that randomised experiments were run for each and all had the same levels of random variation. (The randomisation variation and differences is shown off to the right.)



It's blowing out a dense band of fog across the middle of the picture. The consequence is that we cannot see most of the true differences. Our ignorance is total. We can't tell whether they are positive, negative, or zero (don't exist at all).

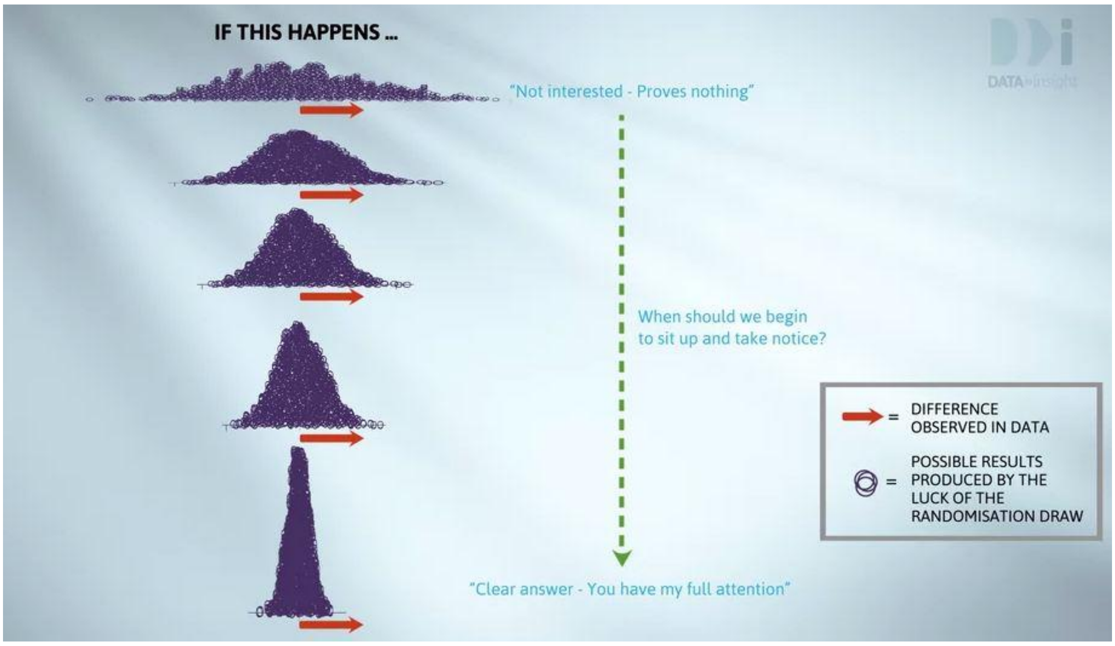


It would be really great if we could do this so that we can see better and detect the existence and directions of more effect. And the key to doing that?



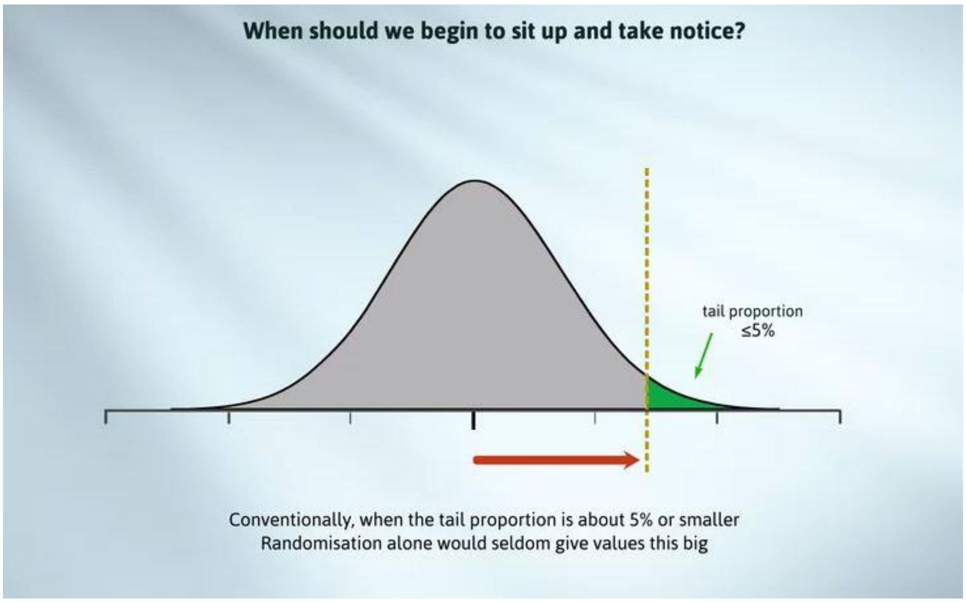
Increasing the size of the study. (But of course, that costs a lot of money, and it's not always practically feasible.)

We're now getting close to the idea of power, which is a key idea for designing studies that can reliably detect the existence of important differences.
Let's go back to this.


IF THIS HAPPENS ...
"Not interested - Proves nothing"
When should we begin to sit up and take notice?
"Clear answer - You have my full attention"
= DIFFERENCE OBSERVED IN DATA
= POSSIBLE RESULTS PRODUCED BY THE LUCK OF THE RANDOMISATION DRAW
DATA insight

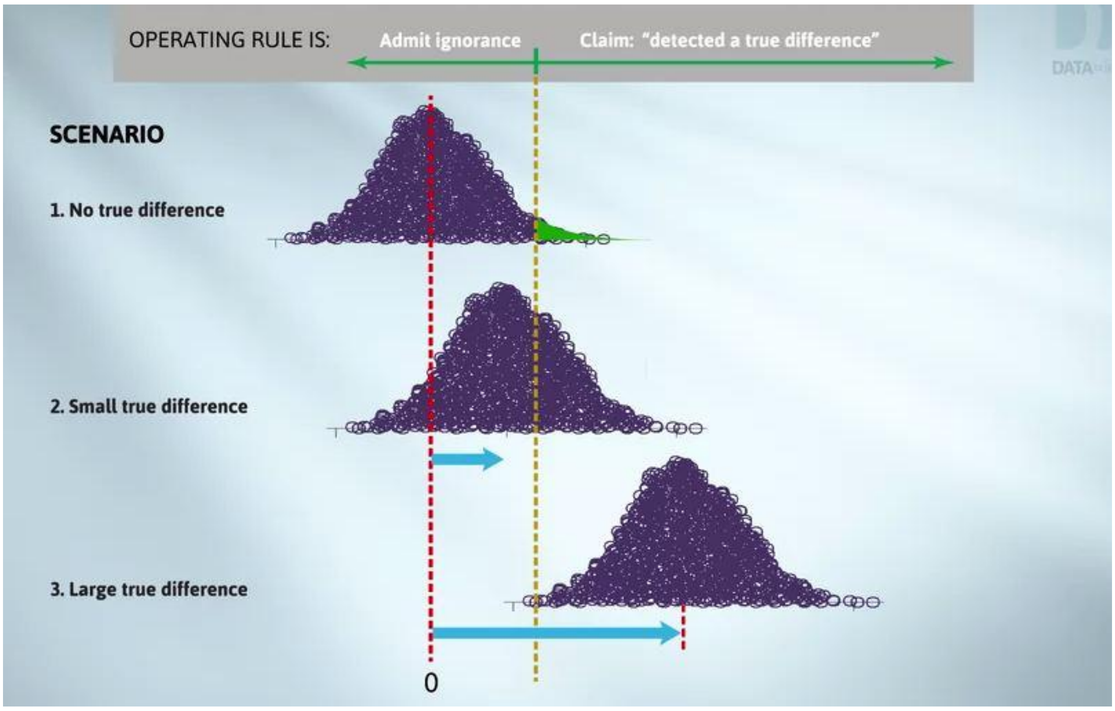
Remember that the length of the red arrows is the difference we observed in our data. Possible degrees of randomisation variation are shown, decreasing as we go down. The question is, "When do we sit up and take notice? "

We want to reach for some sort of operating rule here, based on the tail proportion. We'll claim that we've detected a true difference if the tail area is small enough (say, no larger than 5%). If that's the case, the difference is said to be statistically significant.


When should we begin to sit up and take notice?
tail proportion ≤5%
Conventionally, when the tail proportion is about 5% or smaller
Randomisation alone would seldom give values this big

So our operating rule is: Claim that we've detected a true difference if the observed difference from our data is to the right of the brown dashed line.

This has consequences. We don't know what the true state of the world is.



If Scenario One holds (so that in truth there's no difference at all) the difference we get from our data is like a random draw from the top dotplot. There's a small chance that just from the luck of the randomisation draw, our observed difference will fall to the right of the brown dashed line, and we will make a false claim that there's a true difference. Statisticians call this "making a Type I error", and take comfort in the fact that it rarely happens. It's the cost of doing business in this way.

Perhaps the true state of the world has the true difference given by the blue arrow in Scenario Two. Our observed difference would then be like a random draw from the second dotplot. There's a high chance that we would not end up claiming there was a true difference. If it was really important to know that there was a treatment difference and if, in reality, it was this big, then this would not be a satisfactory experiment to run.

If, on the other hand, the state of the world was like Scenario Three, we would basically always end up detecting a true difference. If the real difference was this big, the experiment would reliably tell us the existence and direction of a treatment effect.

The power of a study is the probability that it would detect an effect exists when the true effect was of a nominated size. It's like this: A helicopter pilot might say, "If a building's 100 metres high or higher, I want to know that there's a building there. If it's shorter than that, I'm OK with not knowing."

So as part of designing a study, we have to specify a minimum effect size at which we'd want to know that a true effect exists. We then ensure that the study uses enough experimental subjects so that if the true state of the world was our specified effect size or larger, we would almost certainly conclude that a treatment difference existed.

Before leaving this image, I should point out a weakness in our buildings and fog analogy. With a building that pokes above fog, we have some information about the height of a building, as well as knowing that it's there. Significance testing in small tail areas (or small P-values) are all about evidence that a real effect exists. They give no information at all about the size or practical importance of an effect. For that, we need confidence intervals.

So the steps are:

- Specify a minimum effect size you want to be able to detect.
- Specify a Type I error rate and the power you'd like to have.
- Calculate the number of experimental units that you'd need to achieve that level of power.

We don't take significant effects very seriously if we know they come from a study that was under-powered for detecting effects of a believable size. I stress again that if a difference is statistically significant, we're pretty sure a real difference exists.

Statistical significance says nothing at all about the size or the practical importance of the difference. To estimate its size, we need a confidence interval.

Well, thanks for hanging in there. These are very hard ideas to get your head around. I promise that Week Eight will be much gentler, but also very useful.