DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

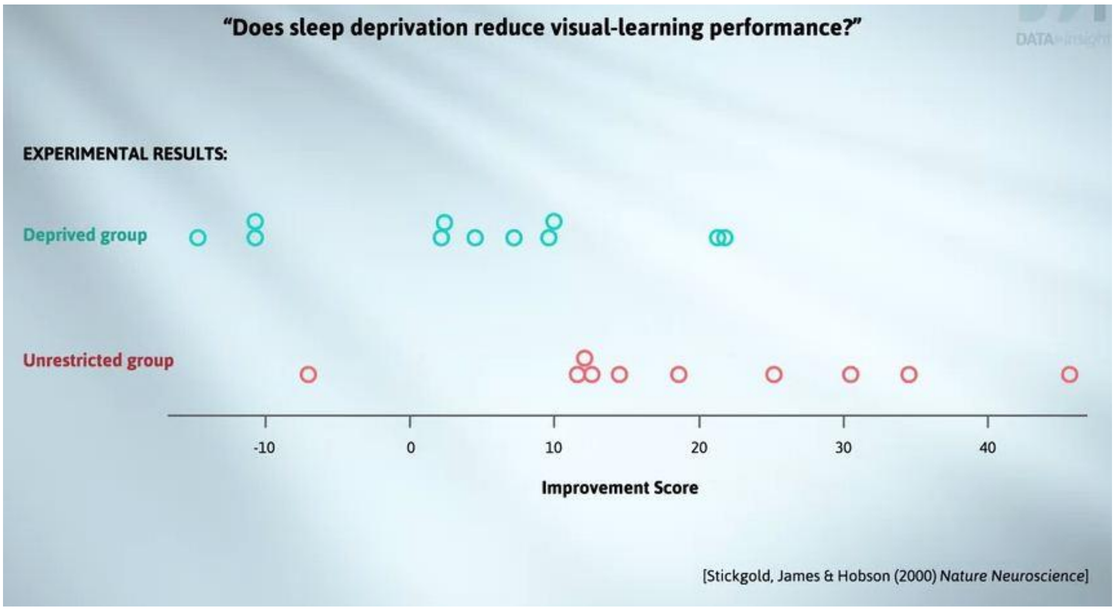
THE UNIVERSITY OF AUCKLAND

WEEK 7

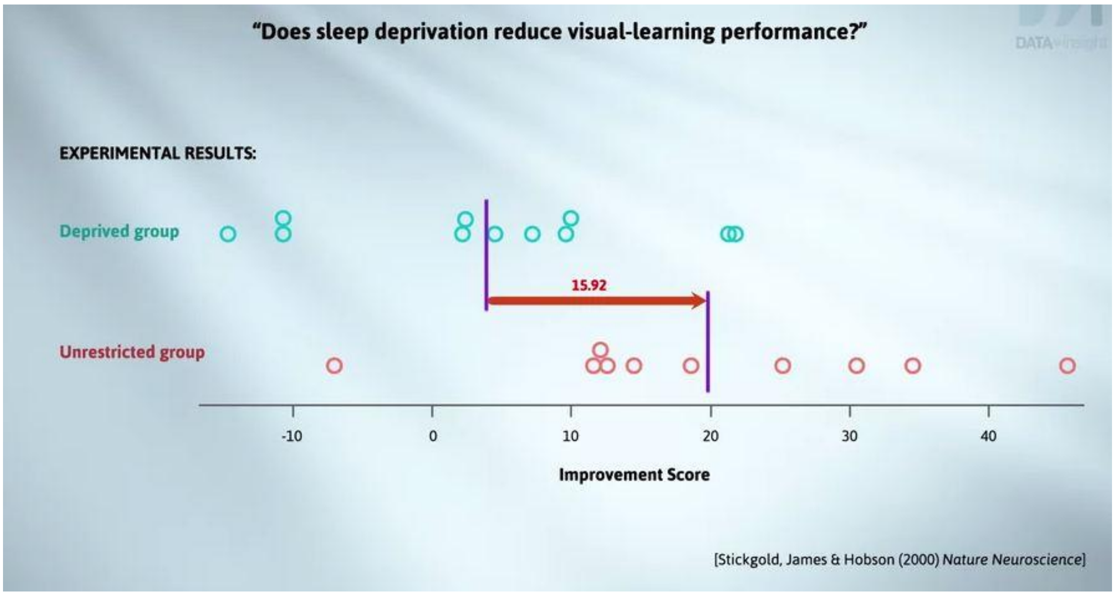
RANDOMISATION TEST by Chris Wild

In the last video, we discussed randomisation variation. Randomisation variation motivates a type of test we do to try to determine whether the effects we've seen in a randomised experiment demonstrate real treatment differences, or whether they could just be "the luck of the draw".

In the last video, we saw how randomly labelling some people as belonging to Group A and others to Group B produces differences in the means of these purely artificial groups. When the group sizes are comparatively small, as in this example, the differences in group means that we observe can be quite large, despite the fact there's no underlying true difference at all.



This data comes from a randomised experiment to investigate the effect of sleep deprivation on performance on a visual learning task. Subjects in the group labelled "deprived" experienced some sleep deprivation, while those in the group labelled "unrestricted" were able to sleep normally. The outcome variable was an improvement in score on a particular type of test of visual learning, over a three-day period.



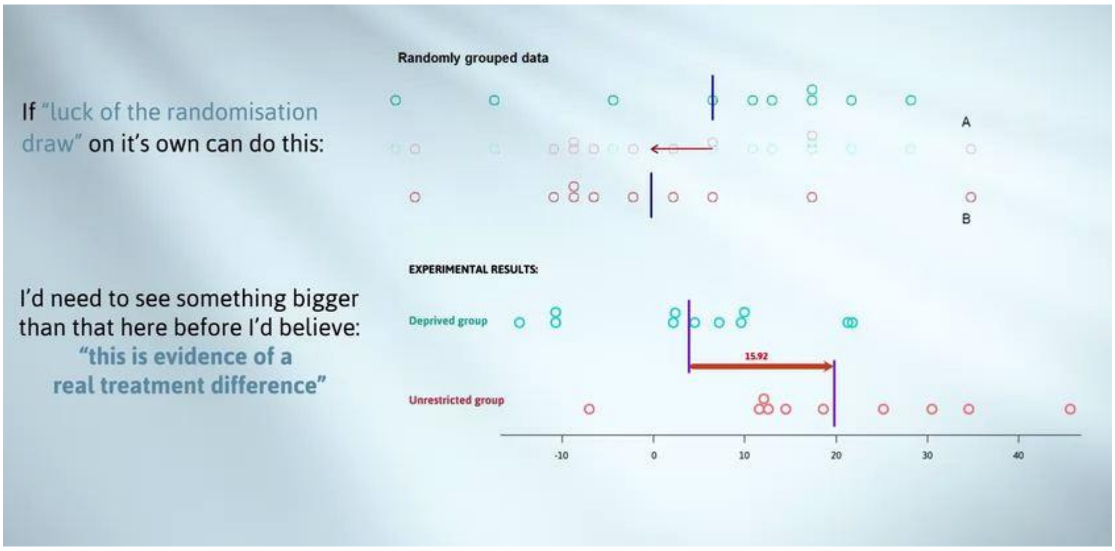
"Does sleep deprivation reduce visual-learning performance?"

EXPERIMENTAL RESULTS:

[Stickgold, James & Hobson (2000) *Nature Neuroscience*]

Here are the mean test scores for each group. And the difference? It looks like the "unrestricted" group is doing quite a bit better, which would imply that sleep deprivation leads to reduced scores. But then we remember this. Could what we're seeing here just be due to this, "the luck of the draw"? Maybe the sleep deprivation is not making any real difference at all. Maybe all we're seeing is one of these pure "luck of the draw" differences from the randomisation.

If randomisation-luck (on its own) can do this, we're going to need to see something bigger in our experimental data, before we'll believe that the experimental data provides evidence of a treatment effect.



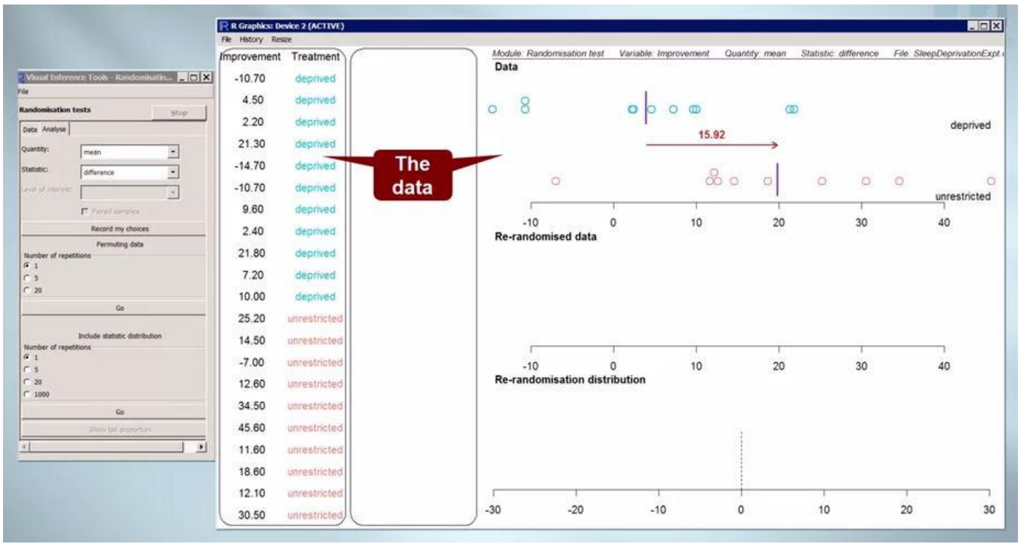
If "luck of the randomisation draw" on it's own can do this:

I'd need to see something bigger than that here before I'd believe: "this is evidence of a real treatment difference"

Here, that sleep deprivation causes reductions in test scores.

Sound like common sense? That's the basic thinking behind a "statistical significance test" or "statistical hypothesis test". In situations where the effect I see in experimental data could quite easily be produced by the randomisation alone, then the experimental evidence doesn't prove anything to me. It could all just be due to chance. I'd need to see something bigger.

So how do we turn this basic idea into the randomisation test? We'll use VIT's randomisation test module to develop the ideas.



R Graphics: Device 2 (ACTIVE)

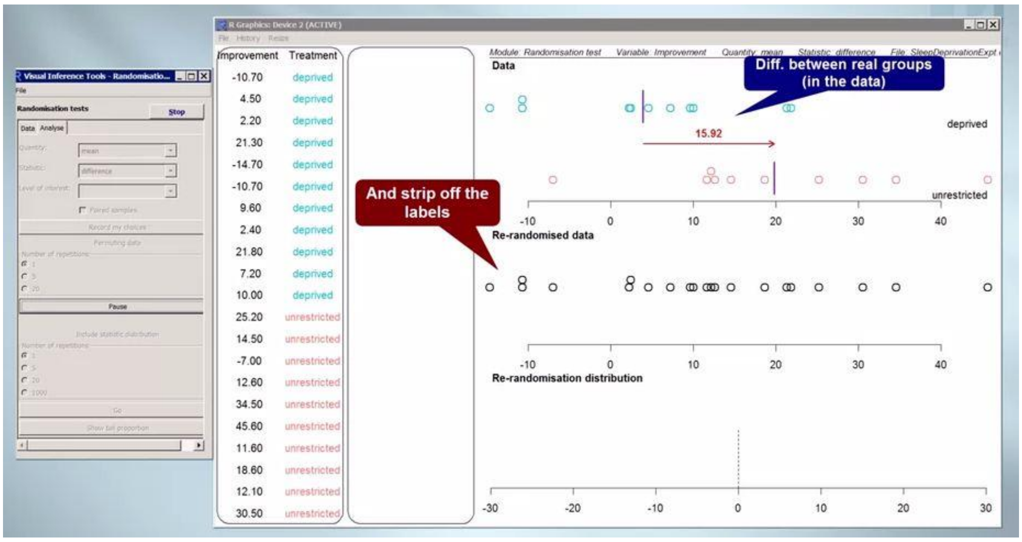
Module: Randomisation test Variable: Improvement Quantity: mean Statistic: difference File: SleepDeprivationExpt

Quantity: mean

Statistic: difference

Improvement	Treatment
-10.70	deprived
4.50	deprived
2.20	deprived
21.30	deprived
-14.70	deprived
-10.70	deprived
9.60	deprived
2.40	deprived
21.80	deprived
7.20	deprived
10.00	deprived
25.20	unrestricted
14.50	unrestricted
-7.00	unrestricted
12.60	unrestricted
34.50	unrestricted
45.60	unrestricted
11.60	unrestricted
18.60	unrestricted
12.10	unrestricted
30.50	unrestricted

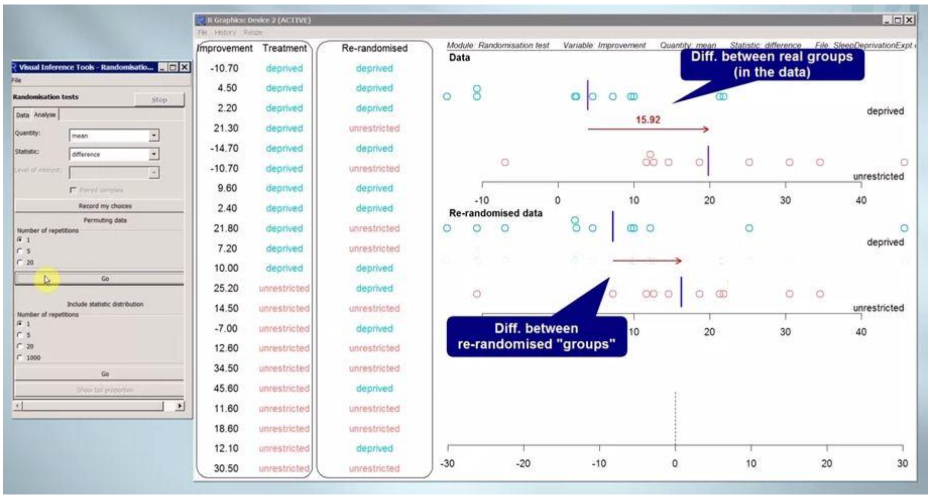
We begin with the real data from the experiment, which involves real treatment groups. We think, maybe sleep deprivation makes no difference, and the difference in means we say is just due to randomisation. Well, if sleep deprivation made no difference, then it wouldn't matter what treatment people got. The results would have been the same. So we'll strip the treatment labels and colours of everyone.



Improvement	Treatment
-10.70	deprived
4.50	deprived
2.20	deprived
21.30	deprived
-14.70	deprived
-10.70	deprived
9.60	deprived
2.40	deprived
21.80	deprived
7.20	deprived
10.00	deprived
25.20	unrestricted
14.50	unrestricted
-7.00	unrestricted
12.60	unrestricted
34.50	unrestricted
45.60	unrestricted
11.60	unrestricted
18.60	unrestricted
12.10	unrestricted
30.50	unrestricted

How big are the differences we'd get from pure random reassignment, just by the luck of the draw?

We'll randomly form new groups. There were 11 people in the sleep deprivation group and 10 in the unrestricted group. So we'll randomly put green "deprived" labels on 11 of them and red "unrestricted" labels on the rest.

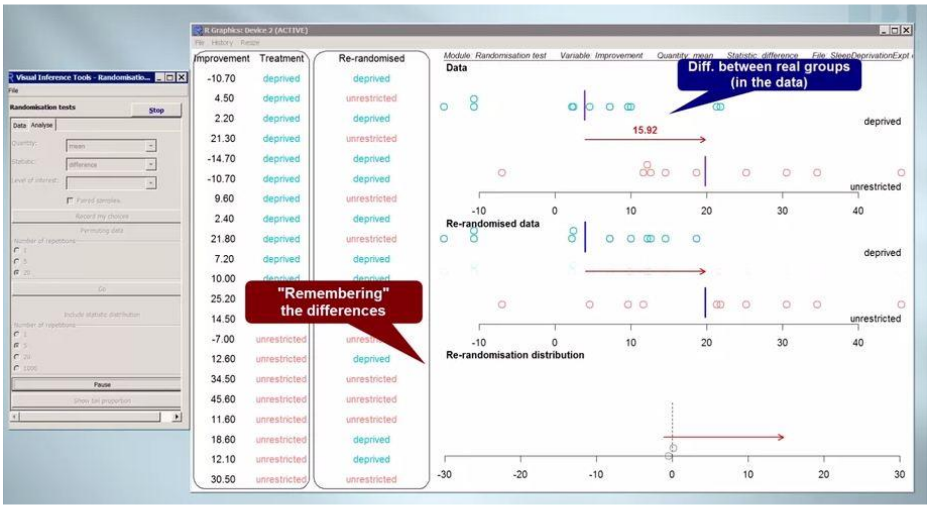


Improvement	Treatment	Re-randomised
-10.70	deprived	deprived
4.50	deprived	deprived
2.20	deprived	deprived
21.30	deprived	unrestricted
-14.70	deprived	deprived
-10.70	deprived	unrestricted
9.60	deprived	deprived
2.40	deprived	deprived
21.80	deprived	unrestricted
7.20	deprived	unrestricted
10.00	deprived	deprived
25.20	unrestricted	deprived
14.50	unrestricted	unrestricted
-7.00	unrestricted	deprived
12.60	unrestricted	unrestricted
34.50	unrestricted	unrestricted
45.60	unrestricted	deprived
11.60	unrestricted	unrestricted
18.60	unrestricted	unrestricted
12.10	unrestricted	deprived
30.50	unrestricted	unrestricted

Now we'll pull them apart into their artificial groups and look at the difference between their means. Let's do this again. Strip the labels off, randomly re-label. Pull them apart into two new randomisation groups. Look at the difference in means.

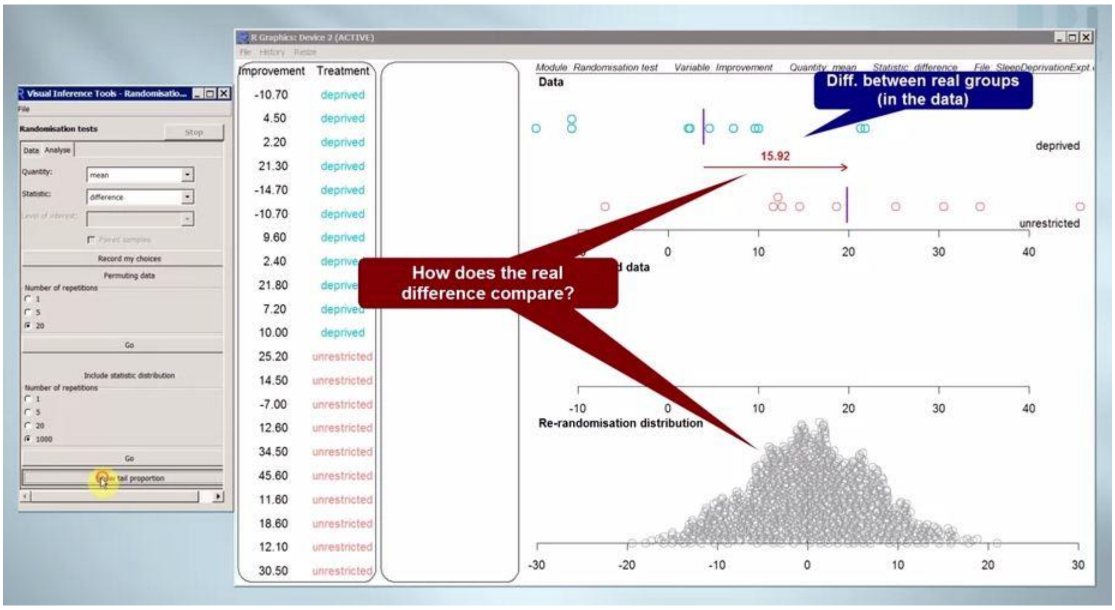
We'll do this 20 times and compare the differences random reassignment is producing to the 15.92 we got in the experiment. I think all of those were smaller. Let's do it again. Most look smaller, but a couple were about as big as the real one from the experiment.

Now let's do it a large number of times and keep track of all the results. We keep a track of the lengths of the arrows, just as we did with the randomisation variation module.



Improvement	Treatment	Re-randomised
-10.70	deprived	deprived
4.50	deprived	unrestricted
2.20	deprived	deprived
21.30	deprived	unrestricted
-14.70	deprived	deprived
-10.70	deprived	deprived
9.60	deprived	unrestricted
2.40	deprived	deprived
21.80	deprived	unrestricted
7.20	deprived	deprived
10.00	deprived	deprived
25.20	[illegible]	[illegible]
14.50	[illegible]	[illegible]
-7.00	unrestricted	unrestricted
12.60	unrestricted	deprived
34.50	unrestricted	unrestricted
45.60	unrestricted	unrestricted
11.60	unrestricted	unrestricted
18.60	unrestricted	deprived
12.10	unrestricted	deprived
30.50	unrestricted	unrestricted

Now let's do it 1,000 times.



Quantity: mean
Statistic: difference
Number of repetitions: ○ 1 ○ 5 ● 20
Include statistic distribution
Number of repetitions: ○ 1 ○ 5 ○ 20 ● 1000

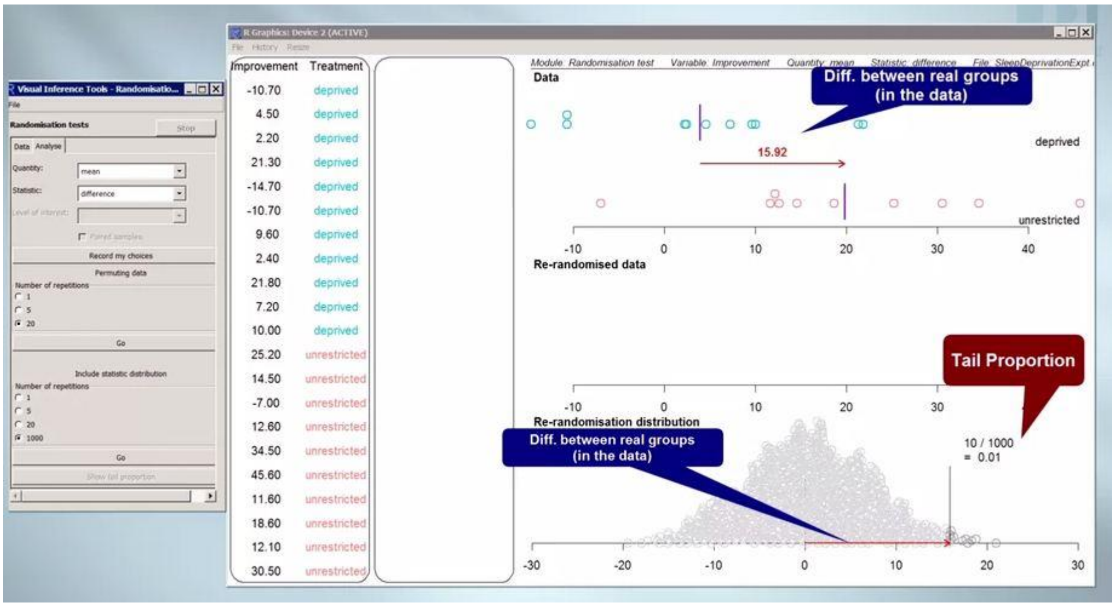
Improvement	Treatment
-10.70	deprived
4.50	deprived
2.20	deprived
21.30	deprived
-14.70	deprived
-10.70	deprived
9.60	deprived
2.40	deprived
21.80	deprived
7.20	deprived
10.00	deprived
25.20	unrestricted
14.50	unrestricted
-7.00	unrestricted
12.60	unrestricted
34.50	unrestricted
45.60	unrestricted
11.60	unrestricted
18.60	unrestricted
12.10	unrestricted
30.50	unrestricted

Visual Inference Tools - Randomisatio...
Randomisation tests
Stop
Data Analyse
Record my choices
Permuting data
Go
Go
Show tail proportion
R Graphics: Device 2 (ACTIVE)
Module Randomisation test Variable Improvement Quantity mean Statistic difference File SleepDeprivationExpt
Data
Diff. between real groups (in the data)
15.92
deprived
unrestricted
How does the real difference compare?
Re-randomisation distribution
-30 -20 -10 0 10 20 30

Okay. The dot plot has recorded all the differences we've got from 1,000 re-randomizations.

So how does the real one from the experiment compare with what we got by forming artificial groups randomly?

Here, the real group difference (15.92) is laid on top of the ones we got by randomising, the ones we got just by the luck of the draw. Well, it's clear that randomisation alone hardly ever gives us anything as big as 15.92. We'll ask for the tail proportion to see how rare this is.



Quantity: mean
Statistic: difference
Number of repetitions: ○ 1 ○ 5 ● 20
Include statistic distribution
Number of repetitions: ○ 1 ○ 5 ○ 20 ● 1000

Improvement	Treatment
-10.70	deprived
4.50	deprived
2.20	deprived
21.30	deprived
-14.70	deprived
-10.70	deprived
9.60	deprived
2.40	deprived
21.80	deprived
7.20	deprived
10.00	deprived
25.20	unrestricted
14.50	unrestricted
-7.00	unrestricted
12.60	unrestricted
34.50	unrestricted
45.60	unrestricted
11.60	unrestricted
18.60	unrestricted
12.10	unrestricted
30.50	unrestricted

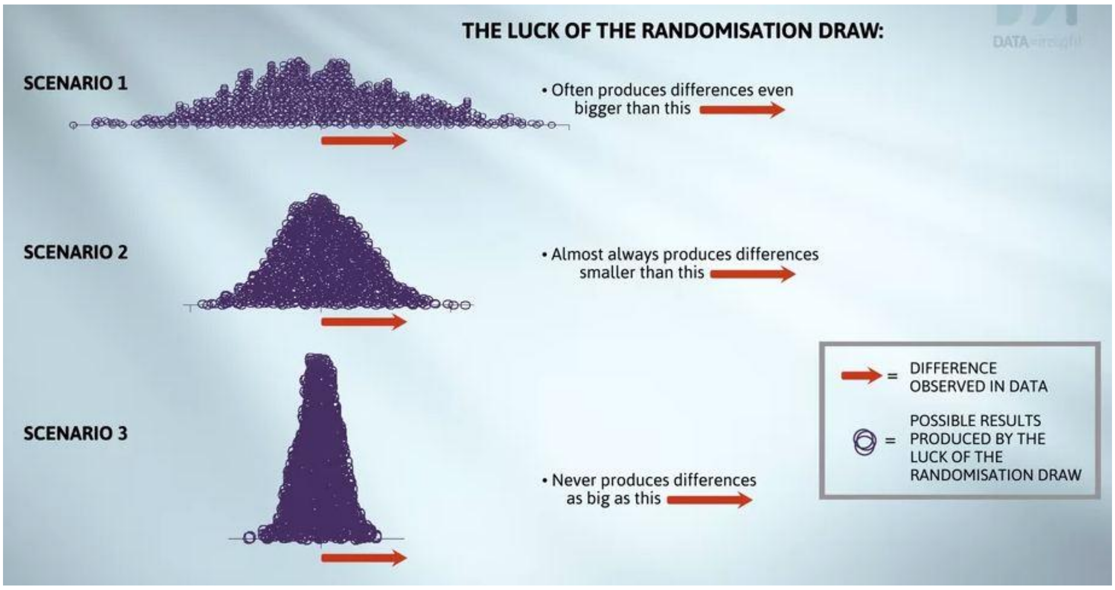
10 / 1000 = 0.01

Visual Inference Tools - Randomisatio...
Randomisation tests
Stop
Data Analyse
Record my choices
Permuting data
Go
Go
Show tail proportion
R Graphics: Device 2 (ACTIVE)
Module Randomisation test Variable Improvement Quantity mean Statistic difference File SleepDeprivationExpt
Data
Diff. between real groups (in the data)
15.92
deprived
unrestricted
-10 0 10 20 30 40
Re-randomised data
Tail Proportion
Re-randomisation distribution
Diff. between real groups (in the data)
-30 -20 -10 0 10 20 30

It tells us that the luck of the draw gave us 15.92 or bigger, 10 times out of 1,000. That's 1% of the time. Put another way, random reassignment almost always gave us values smaller than this. 99% of the time they were smaller.

Since randomisation variation almost always give differences smaller than we got from this experiment, I'd be prepared to rule out the luck of the draw as the plausible explanation and conclude that, in this experiment, sleep deprivation really did lead to a reduction in visual learning test scores.

Let's expand on this logic. We want to compare the real difference in the data with all the differences generated by re-randomizing.



We'll look at several scenarios of what might happen. Here, the luck of the randomisation draw quite often produces differences even bigger than the one from our data.

In the second one, re-randomization hardly ever produces a difference this big. They're almost always smaller.

And in this last one, re-randomization never produces a difference this big.

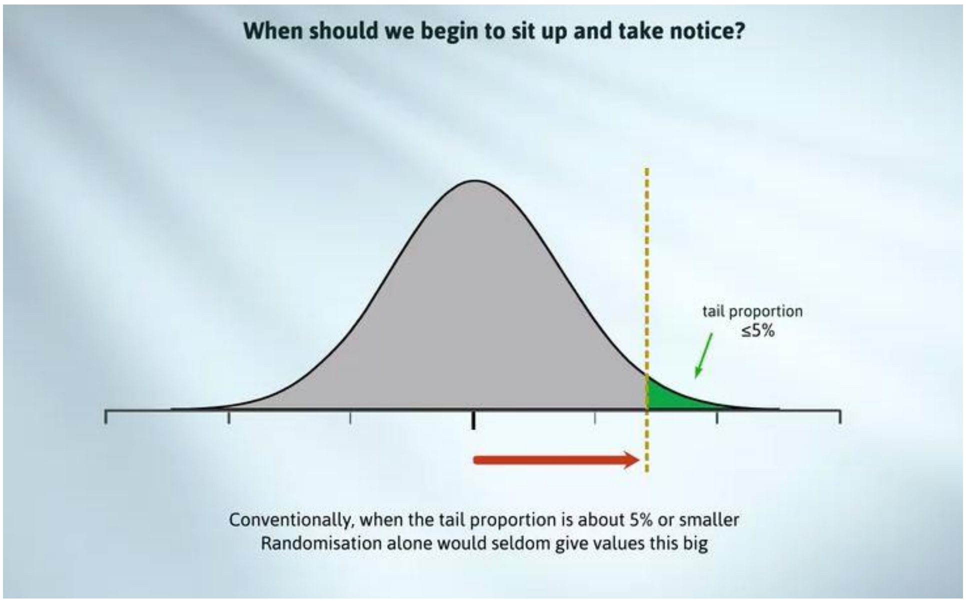
What can we conclude from that?

In this top one, the luck of the randomisation draw alone can easily make differences even bigger than this. There's no evidence that there's a real difference between the groups. At the bottom, the luck of the randomisation draw can't make something that big. We have conclusive evidence that there's a real difference.

But what about the middle one? Re-randomization hardly ever makes differences that big. So surely we have some sort of evidence of a true difference.

Let's think about intermediate cases. At the top, we're simply not interested. The data proves nothing.

At the bottom, there's a clear result. The researchers'd have our full attention.
But in between, at what point will we start to get interested, to realise start to sit up and taking notice?



When should we begin to sit up and take notice?

Conventionally, when the tail proportion is about 5% or smaller
Randomisation alone would seldom give values this big

Conventionally, people start concluding something real has probably happened if the tail proportion is about 5% or smaller. In this case, randomisation alone would seldom give values this big. 19 times out of 20 it would be smaller.
There are a number of important subtleties about interpreting tail proportions, (formally they're called "p-values"). There are also some really bad consequences of a rigid application of the 5% cut-off as a gold standard for having proved something. We'll address these in articles to follow.

That brings us to the end of this video. In the next video, we'll apply these ideas to percentages and extend to multiple groups.