STATISTICS & DATA SCIENCE EDUCATOR

Laying the foundations for the randomisation test

Emma Lehrke
The University of Auckland, New Zealand
Email: e.lehrke@auckland.ac.nz

Overview of lesson

This lesson can be taught in two sessions to Year 13 statistics students or in one or two lectures to stage one statistics students. The lesson introduces four key ideas that underpin the randomisation test: the meaning of unusualness; chance acting alone behaviour; the omnipresence of chance; and evidence and argumentation. Students participate throughout the lesson with a mixture of hands-on and online interactions.

Learning objectives

By the end of the lesson students should be able to use randomisation distributions to:

- recognise what chance acting alone means
- reason with a chance model
- recognise what is required to make a causal claim.

Suggested age range

16+ years

Time required

Two (or one) 50-minute sessions

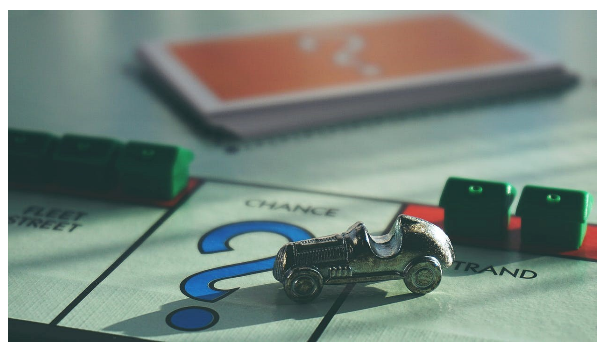
Keywords

randomisation test, experiments, chance

Introduction

My experience with teaching the randomisation test is that students often struggle with some of the key ideas, particularly *chance acting alone*. In the introductory statistics course of 500 students that I teach, the topic on the randomisation test builds the foundation for hypothesis testing, which makes up the majority of the second half of the course. Because of this I believe it is vitally important that students grasp the key concepts of this topic early. Teaching via simulation-based inference has many benefits to student understanding (see Pfannkuch et al., 2013). However, these are only capitalised on when students are able to connect the simulation with what it is trying to simulate – this is the challenge!

Much research in statistics education supports the use of hands-on activities before moving to computer simulations (see Hancock and Rummerfield, 2020). Unfortunately, working in a large-scale lecture environment, hands-on interactions are often time consuming and difficult to manage. Nevertheless, I wanted to incorporate the use of online interactives to speed things up without eliminating the learning that happens through a hands-on process. I also wanted students to use data that they themselves had generated to increase their engagement and understanding of the context.



Lesson outline

The lesson follows four parts, each with an active component and a key message. The four parts build on one another with the fourth being a culmination of all the skills and ideas learned in the earlier three.

The purpose of the first part is to develop students' ability to reason with distributions built from data that would be familiar to them. From this prior knowledge we can then go on to establish the premise that "unusual things happen in the tails." The concept is introduced in the context of height and used to decide if someone is unusually tall, a low threshold task.

The second part of the lesson is to use the distributional reasoning from part one but apply it to reasoning about chance. Students have to decide if there is evidence that someone can taste the difference between pink and white marshmallows. Constant linking between the distribution of what we would expect if they were just guessing, and the language of chance and the real-life activity is crucial at this stage to build up students' reasoning skills. Beginning with an example of a single proportion is a much easier starting point for building up chance reasoning than jumping straight to comparing means. Rossman (2019) gives a good explanation in his blog.

The third part moves from categorical data to numeric. It is designed to demonstrate that even with a no treatment variable (or no effective treatment), there are still likely to be differences between randomly allocated groups. This part sets the stage for the randomisation test as it highlights the need to rule out "chance acting alone" as an explanation for any observed differences.

The final part pulls everything together to perform a randomisation test on the difference between two means. Using a version of Anna Fergusson's statistical modelling framework (see Fergusson, 2017), I aimed to keep the "real world" and the "model world" (which I refer to as the "chance world") separate, as students often struggle to differentiate between what is being shown in the randomisation distribution (a model of chance) and what is shown in the observed data (see Appendix A). I utilise the line-up protocol (Chowdhury et al., 2018) as an informal way to make a decision about whether the outcome of an experiment stands out among simulations of chance acting alone.

For the lesson the students are given a handout that they write on (see Appendix A, which has been filled in with expected student responses in italics).

Session one

1. What is unusual?

Using a picture of a tall woman, Esra McGoldrick of the Tall Ferns, without other objects to use as a reference to estimate her height, I asked students to write down *how tall they thought she was and whether or not they thought she was unusually tall*. As we discussed some of the students' responses, we realised that it was difficult to know if she was unusually tall or not without any recognisable objects in the picture to use as a reference, and also without really knowing what I meant by "unusual." We discussed what unusual might mean. Some ideas that came up were "taller than most people" and "taller than average plus a bit."

I then showed students the distribution of heights for female New Zealanders (see Appendix B), mentioning how we could see the mean height in the middle of the distribution and how as the height of the distribution got lower and lower there were less and less females at that height. I then showed where Esra McGoldrick's height (188cm) sat on the distribution. Because her height was out in the right tail of the distribution, we reasoned that she was unusually tall.

Next I commented that Esra is a basketball player and asked, *Is Esra unusually tall for a basketball player?* I showed the distribution of NZ Tall Ferns heights (see Appendix B) and moved the marker for Esra in towards the middle of the distribution. At this point students informed me that she was not unusually tall for a basketball player. Hence, I highlighted that if an observation is in the tails then it is unusual **for that distribution**.

Key message 1: To determine if something is unusual or not we need some point of reference. If it's in the tail, then it's unusual ***for that distribution***.

> Part one of the lesson was teacher led. If I had more time or a smaller class then I would give students a longer time to discuss their ideas. In spite of the limitations, I think most students had a good understanding of the key message by the end.

2. Can you taste the colour of marshmallows?

> Instead of using marshmallows, other sweets could be used (e.g., red and brown M&M's). Eating eight M&M's is a more reasonable task than eating eight marshmallows. Regardless of your choice of foodstuff you should always give a health and safety briefing before using edibles in the classroom.

My starter question for the second activity was, *Who thinks that pink and white marshmallows taste different?* There was a pretty even show of hands, probably more students with their hands raised than not.

On the projector I showed eight marshmallows. I asked, *Imagine someone was tasting these eight marshmallows blindfolded. How many out of eight would we expect them to get correct if they really could taste the difference?* A brave student responded with "eight." Following up I asked, *But would they have to get all eight for us to think that they really could taste the difference? What about seven? What about six?*

Then I asked, *What if they could not actually taste the difference, how many would we expect them to get correct out of eight now?* A student, picking up on the ideas from before, told me, "Four, but also maybe not, maybe only 3 or maybe 5."

The students now participated in building up a chance model – a distribution of how many out of eight we would expect to get correct if they could not taste a difference and were only guessing. I got all the students to write down the order that they thought the eight marshmallows would come up in (e.g., WPPWWPWP). Once students had written down their guesses I showed a random order on the projector and had the students record how many of their guesses were a match with the order on the board. Then I asked students to raise their hands for their number as I went through "Zero, One, ..., Eight." As the students were raising their hands I drew a dot plot on the projector of approximately how many hands were going up.

> The hand raising activity, after the students had guessed, was really effective in a large lecture theatre with approximately 500 students. It was easy to see more hands towards the middle (3, 4, 5) and only a few at the extremes. It was also really good to see when there was a hand up at 1 or at 8. I got the student's name and used them as an example later on by saying, "Remember Paula, she got 8 correct when she was just guessing."

Next I asked, *Instead of getting lots of students to do this guessing, how could I generate this information?*

> In a smaller class write up all the student guesses on the board as a dot plot and then ask *How could we get more data?*

We discussed how the data we were recording was not really about marshmallows; we were only recording if the guess was right or wrong. Since, for each guess, there were only two options (pink or white) and they were chosen randomly, the chance of being correct was 50%. Therefore collecting data on right/wrong guesses would be the same as collecting data on flipping a coin where heads is right and tails is wrong. Hence, I used the Rossman/Chance applet (see materials) and demonstrated flipping eight coins and counting how many heads appeared. I linked this to the eight marshmallows and guessing the colour correctly. I repeated the simulation of flipping eight coins 1000 times to build up the distribution of what guessing would look like.

Referring students back to the first activity with Esra, I then asked, *Now, how many would someone have to get correct for you to think that they weren't just guessing?* Again, seven or eight came up. I showed the tail of the distribution and made sure to link back to the idea of unusualness from the first activity and stated, *It would be unusual for someone who was just guessing (for that distribution) to get seven or eight correct.*

I asked for a volunteer, Jasmine, who really liked marshmallows to come to the front of the lecture theatre. She was blindfolded and given the marshmallows one by one and asked to guess the colour. I recorded how many were correct. After Jasmine returned to her seat, I asked, *So what can we conclude?* Jasmine got three out of eight correct, so we concluded that we had no evidence that she wasn't just guessing. I made sure to point out that Jasmine wasn't just guessing, she was really trying, and we saw how hard she was trying! But we didn't know for sure and from just looking at the data we didn't have enough evidence to say that she wasn't just guessing. We definitely didn't have enough evidence to say that she could taste the different colours.

Key message 2: To determine if chance acting alone is a reasonable explanation for what we observe, we need to see what chance acting alone would look like: We need a model of chance acting alone for comparison.

A second volunteer gave the student the marshmallows and that left me free to record whether they were right or wrong. It was quite a big ask to eat eight marshmallows, and after the first couple it went more slowly. I still thought it was worth it though to run the activity, as it gave the students a conclusion to what they were doing and gave it a purpose. The logic at the end, that we could not say she was guessing but we did not have enough evidence to say that she wasn't, is a difficult one for students, as it is the idea that we cannot accept the null hypothesis, rather we can only look for evidence against it. Rossman (2008) writes that assessing the strength of evidence against a claim becomes more difficult when the null model becomes less intuitive and the situation becomes more abstract. Therefore, always linking back to something that actually happened in class may help to build student reasoning. I often referred back to how Jasmine was really trying to determine the correct colour of the marshmallows when teaching about what you can and cannot conclude in my other lectures.

3. Random allocation and estimating age from photos - Is there a photo effect?

For the third activity I used the *Random re-direct tool* (Fergusson, 2020a) and two Google forms to randomly allocate all the students to either Group A or Group B. Each student then had to estimate the age of the person in the photo in the Google form.

Alternative method: Instead of using the *Random re-direct tool* and the Google forms you can do this experiment by hand. You would need to print copies of the photos, half with "Group A" written on them, the other half with "Group B." Ask students to write down their estimate of the age of the person and then collect them in. The downside to doing this by hand is that it will take a while to collect in the data, however it is an opportunity to draw the dot plots onto the board by hand and demonstrate calculating the medians and the difference between them.

I told the students they had been randomly allocated to either Group A or Group B (it also showed "Group A" or "Group B" at the top of the form they filled in). I showed students a dot plot and boxplot of their age-estimate data using a link to *iNZight Lite* (see https://lite.docker.stat.auckland.ac.nz/) and we could see that the age estimates for Group A were, on average, slightly higher than for Group B. I then showed the students what Group A and Group B were shown, which was the same photo! I asked the students, *Was this an experiment?* Even though there was random allocation, there was no treatment and I wanted students to recognise this. I asked students, *If there was no difference between what Group A and Group B did, why were their estimates different?* Consequently, there was a discussion that brought up ideas such as, "just chance" and "it's random."

Key message 3: Even when there is no treatment effect, even when chance is acting alone, there will probably still be differences between groups.

This part went relatively quickly and I think students were perfectly happy to agree that there could be a difference just by chance. Linking this idea to the next part was more challenging.

To finish off Session One of the lesson, I used the *Random re-direct tool* (see note above for by-hand method) again but this time Group A and Group B saw different photos.

Session two

Is there a photo effect? [continued]

To prepare for Session Two, I used *R* (version 3.6.3) and the *nullabor* package (Wickham et al., 2020) to generate a line-up (see Appendix C for *R* code) of 20 dot plots (Figure 1), where 19 plots were from the "non-experiment" with the same photo and one plot was from the experiment with two different photos. Note that I used only the age-estimate data gathered from the first seven students from each group to keep the task and visual images simpler.

Alternative method: As an alternative to using *R*, you could take screenshots of the middle panel of *VIT Online* (see https://www.stat.auckland.ac.nz/~wild/VITonline/) to show individual re-randomisations and mix these up with a screenshot of the top panel showing the experiment data.

I reminded the students that at the end of Session One that they had been randomly allocated to Group A and Group B to estimate the age of the person in the photo. I then showed them that the photos for Group A and Group B were different, one of me in good lighting and the other not so good. We discussed how this time it was an experiment and what the units, treatment variable and response variable were. I then asked students to draw dot plots of what they might expect the data to look like. To make the task easier, I asked them to assume there were 14 volunteers and to draw seven dots for Group A and seven for Group B.

Figure 1 was shown to the students and I explained that 19 of the plots were from the "non-experiment" where all students saw the same photo and one of the plots was from the experiment with two different photos. I asked students, *Which one do you think is the plot from the experiment?* I wanted students to be thinking about the dot plots they had just drawn and what they expected the data to look like.

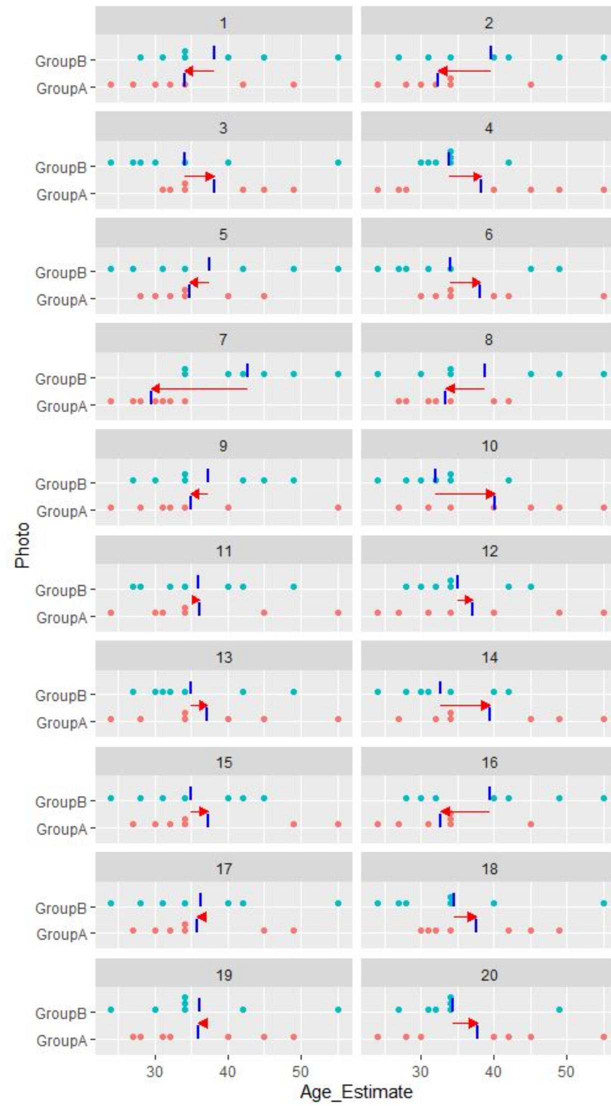


Figure 1: Line-up of plots from *R*, 19 showing chance alone and one showing the observed data from the photo experiment of age estimates gathered from the first seven students in each group

I got students to put their hands up for which plot they thought was from our experiment and when nearly all the hands went up for plot 7 we had a winner! The takeaway from this line-up of 20 dot plots was that if the plot of the experiment stood out, then it was different from all the "chance alone" plots and so we had evidence of a treatment effect. I did not go too deeply into this at the time but I did point out that it looked different from what "chance alone" looked like. In terms of "chance alone", plot 7 was unusual.

Jumping from the Google forms to the *R* output is technically challenging, if doing the lesson in one session. It is easier to split the lesson over two sessions to get the line-up ready. Students were reluctant to draw what they expected the data to look like. I didn't know if this was due to them not being clear about the instruction or just an unwillingness to put pen to paper, or an anxiety at getting it wrong. With encouragement they did eventually draw some dot plots. At the end of the line-up activity students wrote a summary of what it meant to easily spot the real plot (see Appendix A). From walking around the lecture theatre, I could read what they were writing and there was a mixture of good and not so good summaries. In a small class situation these summaries can be shared and discussed.

4. Transitioning to the randomisation test

Alternative method: Previously I have done this exercise by hand with students tearing the labels off paper print outs of the data and shuffling them by hand. When doing it by hand it took a very long time in a large lecture theatre and some students still didn't get time to do it properly. Doing it by hand also caused more stress than it should have, even though the students calculated the difference between the medians rather than the means. Moving to an online tool has a similar user experience without all the drama. I think the "feel" of tearing off the labels that you get with paper is lost, which is a shame, but other than that I found the experience similar and I think that the students learnt as much from it as before. If you are teaching in a smaller classroom you might still prefer to do this part as a hands-on activity, in which case I would recommend working with medians rather than means.

The re-randomised differences between the means collected from the students, via the *Shuffle Tool* were then explored. Using *iNZight Lite* (see https://lite.docker.stat.auckland.ac.nz/) I showed a plot of their re-randomised differences between the means (Figure 2). We discussed how it made sense that it was centred around zero, as that was what we would expect if there was no effect of the Group A or Group B photos. We also looked at how spread out the plot was, showing up other differences between the means that we could expect just by chance. The students were particularly surprised by the re-randomised differences at the edges of the plot and commented that it was surprising to get differences that large/small just by chance. I agreed with their comments but added that while differences that big did happen by chance, they didn't come up very often.

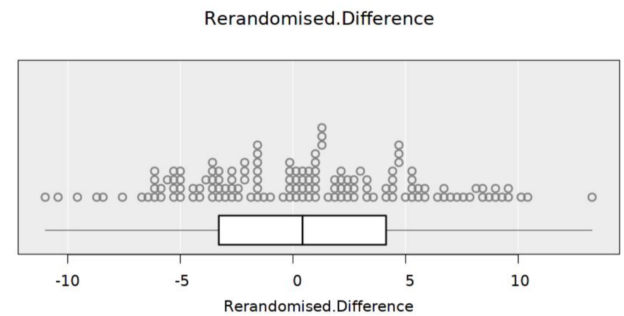


Figure 2: *iNZight* dot plot of the student generated re-randomised differences between the means of age estimates from the *Shuffle Tool*

Referring back to what we had done with the marshmallows, I asked, "*How big a difference between the means would we need to see before we believed that there was a photo effect?"* We discussed how for the marshmallows we wanted our person guessing to be unusual for the distribution of just guessing. We wanted the experiment result to be unusual for the distribution of chance acting alone. We looked to the right tail of the distribution and estimated that a difference between the means of at least about 9 or 10 would be needed.

I then showed students how *VIT Online* (see https://www.stat.auckland.ac.nz/~wild/VITonline/) runs a randomisation test. We could see the data from our experiment in the top panel, showing a difference between our observed means of 13.3 years. I used the demonstration features of *VIT Online* to show the process of the randomisation test. I explained that what was being shown in the middle panel was each re-randomisation, which was the same as what happened for each of them when using the *Shuffle Tool*. I explained how the size of the difference between the means from each re-randomisation is then recorded in the bottom panel as a dot on the randomisation distribution. In the randomisation distribution (Figure 3) we could see again, a difference between the means of around 10 would be needed to get into the tail of the chance distribution. I then clicked the tail proportion button to show students where the difference between the means (13.3) from our experiment sat. Students then completed their worksheet (Appendix A) filling in what it meant that the observed difference between means sat in the tail of the distribution. Students (teacher-led) also wrote a conclusion to the experiment.

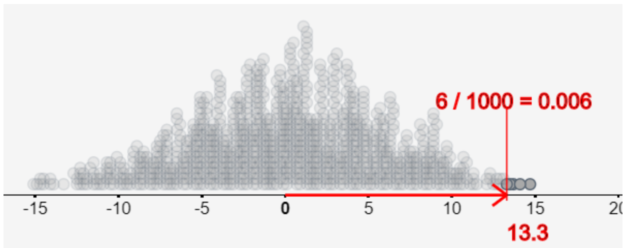


Figure 3: *VIT Online* output of randomisation distribution with tail proportion for photo experiment

Key message 4: If what we have observed in our well designed and well executed experiment would be unusual when chance is acting alone, then we have evidence that there was a treatment effect (i.e., we have evidence against chance acting alone).

The fourth activity is teacher led. As this lesson is only the introduction to the randomisation test topic, I teach future lessons with more examples where students have an opportunity to use *VIT Online* themselves and formulate their own descriptions.

Adaptations

- For the activity where there is no treatment (the photos are the same) teachers could allow students to explore the data in *VIT Online*. In the re-randomisation panel the students would be able to see other examples of data from two groups where any difference in means is due to chance alone. Another way of exploring the difference in means between two groups where there is no treatment is to use the *VIT Online Randomisation Variation* link. Data from a distribution (e.g., weights) can be entered and then students can observe the data being randomly allocated into two groups and the consequent randomisation distribution of the difference between the means under chance alone.
- In a small classroom setting some of these activities could be modified to pen and paper tasks. For example, instead of using the *Shuffle Tool*, students could put the data on paper and do the shuffling by hand (see Alan Rossman's (2019) blog under relevant reading for his penguin activity).
- For the activity where students share the difference in means they got from the *Shuffle Tool*, the students could share their results in groups and then have a group member come up and plot their difference in means onto a dot plot on the board.
- I had the whole class do the photo experiment but to work with smaller amounts of data I have only used data from the first 20 or so students. You might, instead, choose to run the experiment with 10 or 15 volunteers from the class. In this case you could get them to write their age estimate onto a post-it note and then use the post-it notes to create a dot plot on the board.
- Using live data in class does run the risk of technical problems. I was lucky in that I had data from the previous semester that I could use as a back-up in case things didn't work. It might pay to prepare some back-up data if you run these activities.
- I have presented the lesson as taking two sessions. However, due to time restrictions, I did the lesson in one lecture. It was a push to get through all four activities in one lecture and if time allows I would recommend splitting the activities over two sessions, which will give more time for class discussions.

Teacher notes

The reference to "in the tails" is used as a visual judgement for whether or not a value is considered unusual or not for that distribution. This is done to purposefully avoid using a "cut-off" value such as $p < 0.1$ as students can quickly get caught up in using the value of the tail proportion and no longer look at the distribution and use the visual cues that it provides. This is also in line with current discussions around the ubiquitous and potentially inappropriate nature of $p < 0.05$ (see Wasserstein et al., 2019). For all of the examples used in these activities the value in question is either clearly in the tail end of the distribution or obviously towards the middle so it is not challenging for students to make a visual determination.

In later lectures, I do introduce criteria for determining if there is sufficient evidence against chance acting alone but I do not think it wise to jump to this right at the start. If questions do arise around what we might conclude if, for example, a person guessed six marshmallows correctly I would discuss how there is a grey area there and it is difficult to determine. Playing it safe we would say we do not have enough evidence that it wasn't just chance alone and that in future we will learn about some methods to help us make decisions when our value is in that grey area. Some students may already be familiar with using criteria (such as $p < 0.1$) from learning about the randomisation test in secondary school. Those students are still able to complete the tasks as they are presented and their patience is rewarded when they see familiar criteria in later lessons.

Materials required

R (or use alternative method with VIT)

Random re-direct tool (or use alternative hands-on method)

Shuffle tool (or use alternative hands-on method)

Rossman/Chance coin flip simulation

Marshmallows or other appropriate confectionary

Copyright information

Acknowledgements

The development of this lesson has been guided and supported by many conversations with Anna Fergusson, including re-use or modifications of her teaching materials and use of the digital tools she has developed.

References

Chowdhury, N. R., Cook, D., Hofmann, H., & Majumder, M. (2018). Measuring lineup difficulty by matching distance metrics with subject choices in crowd-sourced data. *Journal of Computational and Graphical Statistics*, *27*(1), 132–145. doi: 10.1080/10618600.2017.1356323

Fergusson, A. (2017). Informally testing the fit of a probability distribution model (unpublished master's thesis). University of Auckland, New Zealand. https://researchspace.auckland.ac.nz/handle/2292/36909

Fergusson, A. (2020a). Random re-direct tool [computer software]. Retrieved from https://teaching.statistics-is-awesome.org/tools/random-redirector/

Fergusson, A. (2020b). Shuffle tool [computer software]. Retrieved from https://annafergusson.online/LSinteractives/examples.html

Hancock, S. A., & Rummerfield, W. (2020). Simulation methods for teaching sampling distributions: Should hands-on activities precede the computer? *Journal of Statistics Education*, *28*(1), 9–17. https://doi.org/10.1080/10691898.2020.1720551

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. (2013). "Bootstrapping" students' understanding of statistical inference. Teachings and Learning Research Initiative. Retrieved from http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report_0.pdf

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19.

Rossman, A. (2019, September 23). #12 Simulation-based inference, part 1. Retrieved May 20, 2020, from https://askgoodquestions.blog/2019/09/23/12-simulation-based-inference-part-1/

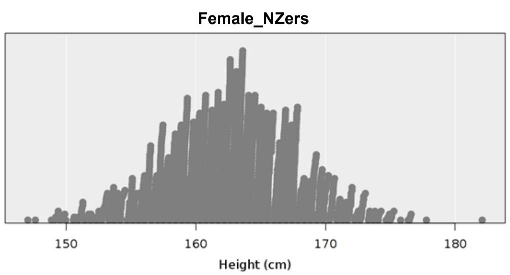
Wasserstein, R., Schirm, A. & Lazar, N. (2019) Moving to a world beyond "p<0.05", *The American Statistician, 73*(sup1), 1–19, doi: 10.1080/00031305.2019.1583913

Wickham, H., Chowdhury, N., Cook, D., & Hofmann, H. (2020). nullabor: Tools for graphical inference. *R* package version 0.3.9.

Appendix A

Randomisation Test Handout

Question: What is unusual?
Summarise what you learned from looking at this distribution of heights:



Key message 1:

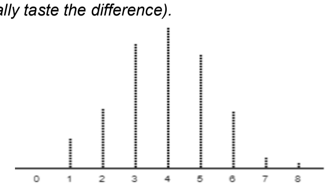
To determine if something is unusual or not we need some point of reference.
If it's in the tail, then it's unusual for that distribution

Question: Can you taste the colour of marshmallows?

How many out of 8 would you expect to guess correctly (assuming that you cannot taste a difference)?

4 (ish)

Draw the distribution of how many correct we would expect if <u>chance was acting alone</u> (if someone was only guessing and couldn't really taste the difference).



Our volunteer ______________ got ____ correct out of 8.
Based on this we would conclude:

Answers here depend on what happened in the lecture

Key message 2:

To determine if chance acting alone is a reasonable explanation for what we observe, we need to see what chance acting alone would look like:
We need a model of chance acting alone for comparison

Question: Is there a photo effect?

Why was the first activity *not* an experiment?
There was no treatment - both groups saw the same photo. Random allocation alone does not make an experiment.

Even though there was no treatment, when we looked at the data the two groups were not identical.
Key message 3:
Even when there is no treatment effect, even when chance is acting alone, there will probably still be differences between groups.

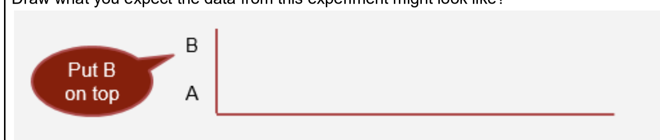
Question: Is there a photo effect?
What made the second activity an experiment?
There was random allocation to treatments.

Units: The 14 volunteers

Treatment variable: The photo (PhotoA, PhotoB)

Response variable: Age estimate

Draw what you expect the data from this experiment might look like?

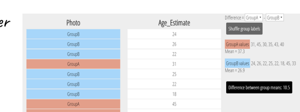


Any drawing of two dotplots/boxplots showing group A as being generally lower than group B

Could you identify our real data in the line-up? Yes / No

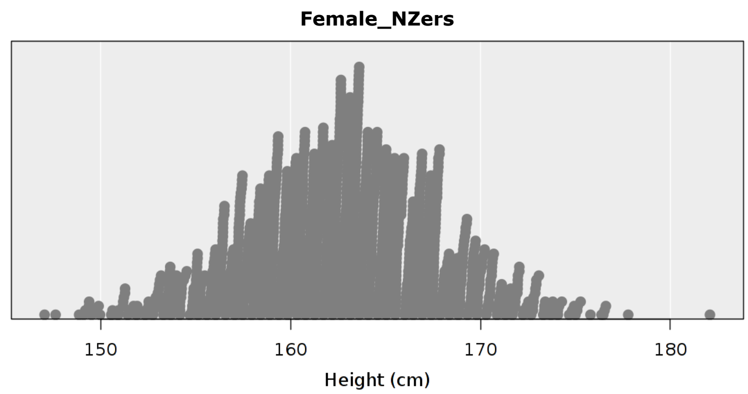
What does this indicate?
If you could then it indicates that our observed results are quite different from what you would expect if chance is acting alone - giving us evidence of a treatment effect.
If you couldn't then it indicates that our observed results are not very different from what you would expect if chance is acting alone - this doesn't give us any evidence of a treatment effect as we have no evidence that it wasn't just chance.

Describe what is happening when you use this tool and why we used it:
The tool is randomly re-labelling each age estimate as groupA or groupB.
If there is no treatment effect (if chance is acting alone) then it doesn't matter which group the person was in they would still have given the same Age Estimate. By randomly re-labelling we are simulating differences that the random allocation process (alone) would produce - we are simulating chance acting alone. Each difference that shows up in the black box is produced by

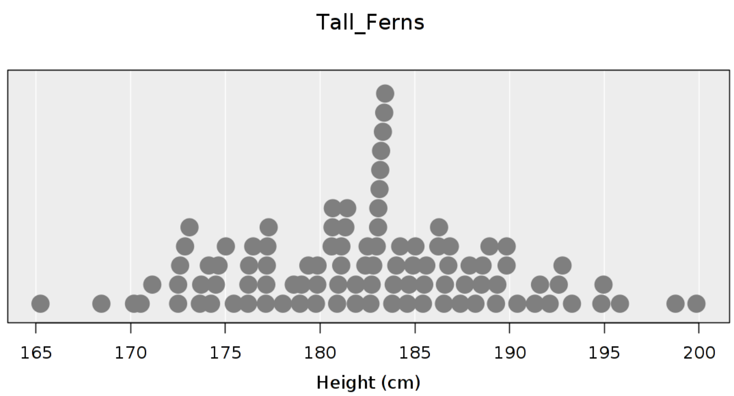


Appendix B

Dot plots were generated using *iNZight Lite* and data simulated from a Normal distribution based on the means and standard deviations stated below. Statistics sourced from https://ourworldindata.org/human-height for female New Zealanders and https://nz.basketball/national-teams/tall-ferns/roster/ for the Tall Ferns.



The heights of f**emale New Zealanders** are approximately normally distributed with a mean of 163.6cm and a standard deviation of 5.3cm.



The heights of **NZ Tall Ferns players** are approximately normally distributed with a mean of 182.4cm and a standard deviation of 6.04cm.

Appendix C

R code used to produce the line-up plots.

```
library(nullabor)
library(tidyverse)

data <- read.csv("data.csv")

lineup <- lineup(null_permute("Age_Estimate"), data)

dotplots <- ggplot() +
  geom_dotplot(data = lineup,
               aes(x = Photo,
                   y = Age_Estimate,
                   fill = Photo,
                   color = Photo),
               binaxis = "y",
               binwidth = 1,
               dotsize = 1.2,
               stackdir = "up",
               show.legend = FALSE) +
 coord_flip() +
 facet_wrap(~ .sample)

# groups into .sample and A and B
data_sum <- group_by(lineup, .sample, Photo)

# add a column that calculates the mean of each A or B group for each .sample
group_means <- summarise(data_sum,
                         meanAge = mean(Age_Estimate,
                                        na.rm = T))

means <- spread(data = group_means,
                key = Photo,
                value = meanAge)

plot_Amean <- geom_spoke(data = means,
                         aes(angle = 0,
                             x = 1,
                             y = GroupA),
                         radius = 0.5,
                         size = 0.8,
                         colour = "blue")
plot_Bmean <- geom_spoke(data = means,
                         aes(angle = 0,
                             x = 2,
                             y = GroupB),
                         radius = 0.5,
                         size = 0.8,
                         colour = "blue")
```

```
plot_means <- dotplots + plot_Amean + plot_Bmean

plot_arrows <- plot_means +
  geom_segment(data = means,
               aes(x = 1.6,
                   xend = 1.6,
                   y = GroupA,
                   yend = GroupB),
               colour = "red",
               show.legend = FALSE,
               arrow = arrow(length = unit(0.20,"cm"),
                             ends = "first",
                             type = "closed"))

plot_arrows

# to get the actual one
attr(lineup, "pos")
```