Intro to Data

DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D. and Angela Lui, Ph.D.

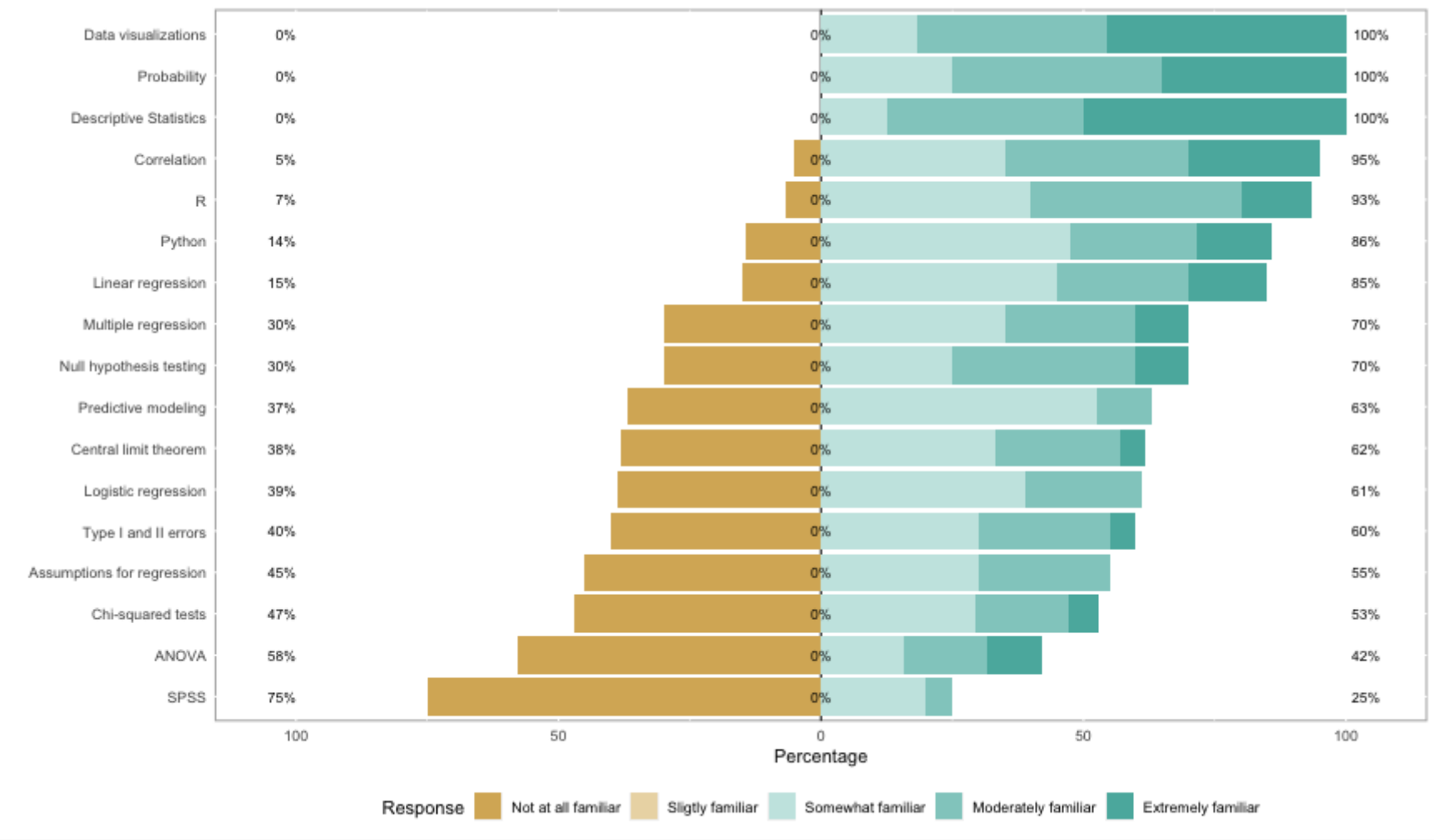
September 1, 2021

Announcements

- DAACS
  - A question for everyone: How well did the results align with your understanding of you as a learner?
    - We have not sent emails yet (an may not need to) to get access to your DAACS results. If you completed it, you are good.
    - DAACS feedback, on the DAACS website, is tailored to your responses. It is worth taking a few minutes to see what your strengths are, as well as areas that might be holding you back.
- Lab 1 and homework 1 is due Sunday.
- Labs - When submitting the labs, you can submit just add your answers to the existing document (i.e. you can leave all the text there so you have all the content together).
- Link to RStudio cheat sheets: https://rstudio.com/resources/cheatsheets/

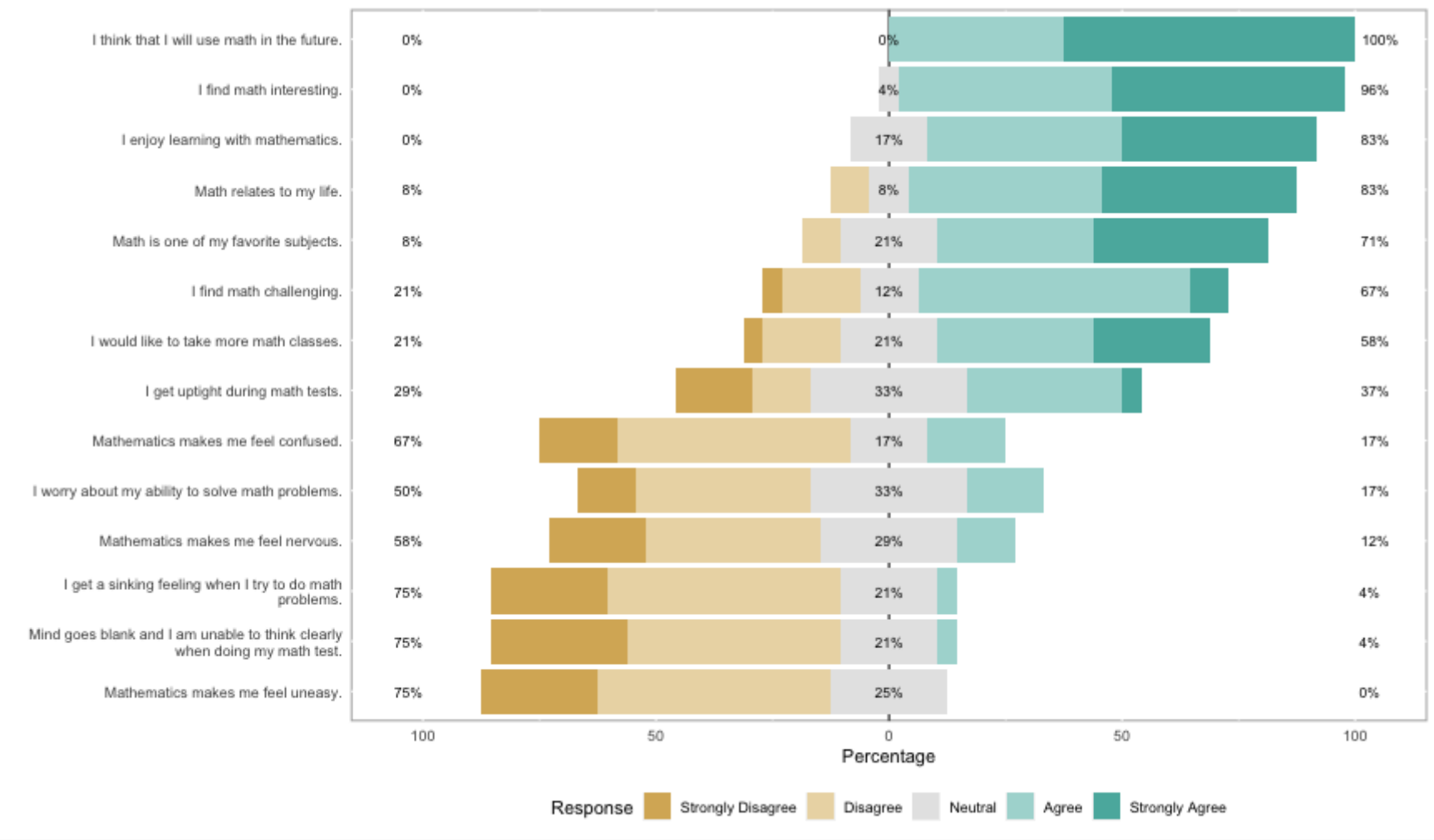
Familiarity with Statistical Topics

```
likert(stats.results) %>% plot(center = 2.5)
```



Math Anxiety Survey Scale

```
likert(mass.results) %>% plot()
```



Sampling vs. Census

A census involves collecting data for the entire population of interest. This is problematic for several reasons, including:

- It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.
- Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
- Taking a census may be more complex than sampling.

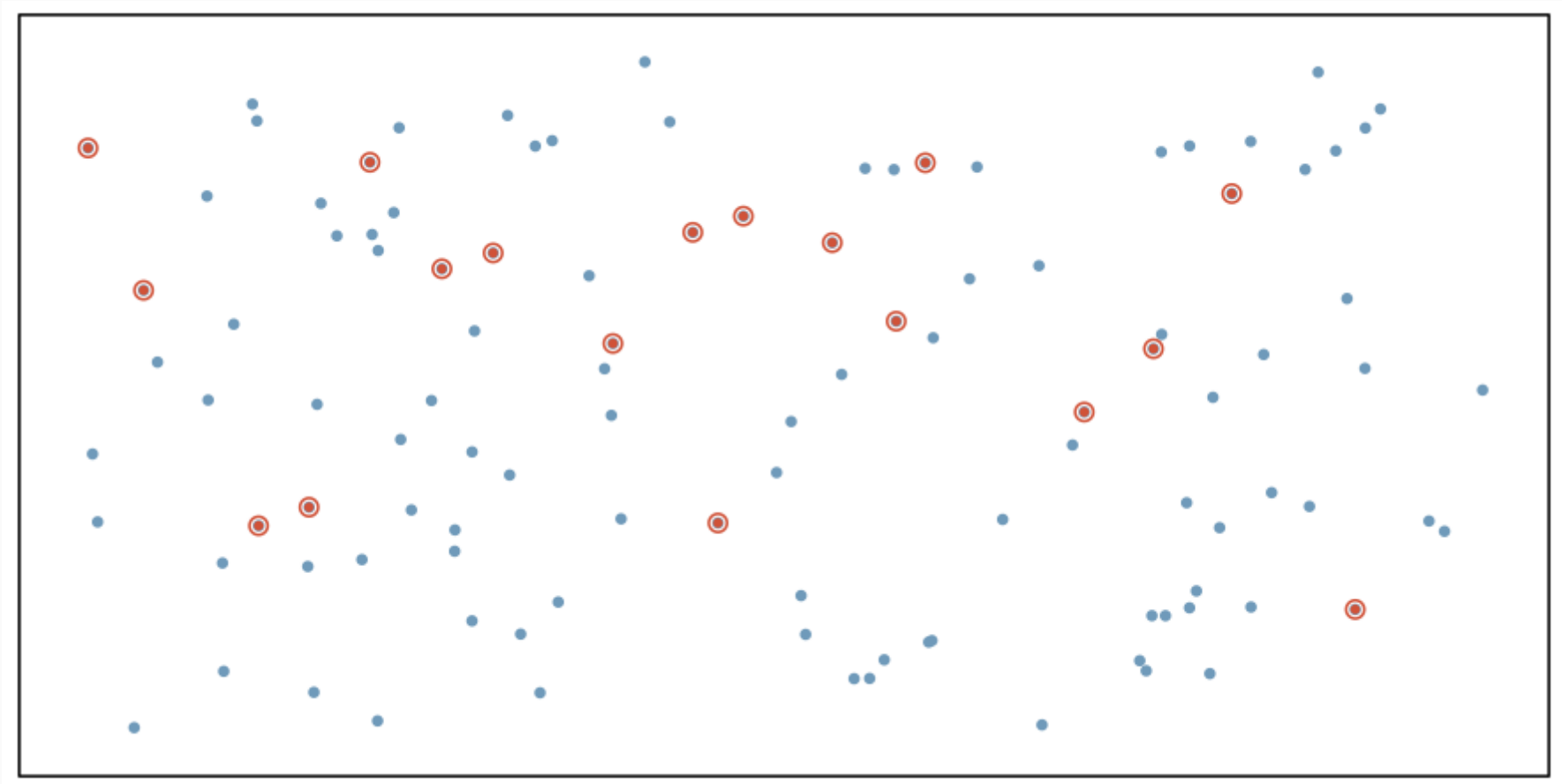
Sampling involves measuring a subset of the population of interest, usually randomly.

Sampling Bias

- **Non-response**: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- **Voluntary response**: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.
- **Convenience sample**: Individuals who are easily accessible are more likely to be included in the sample.

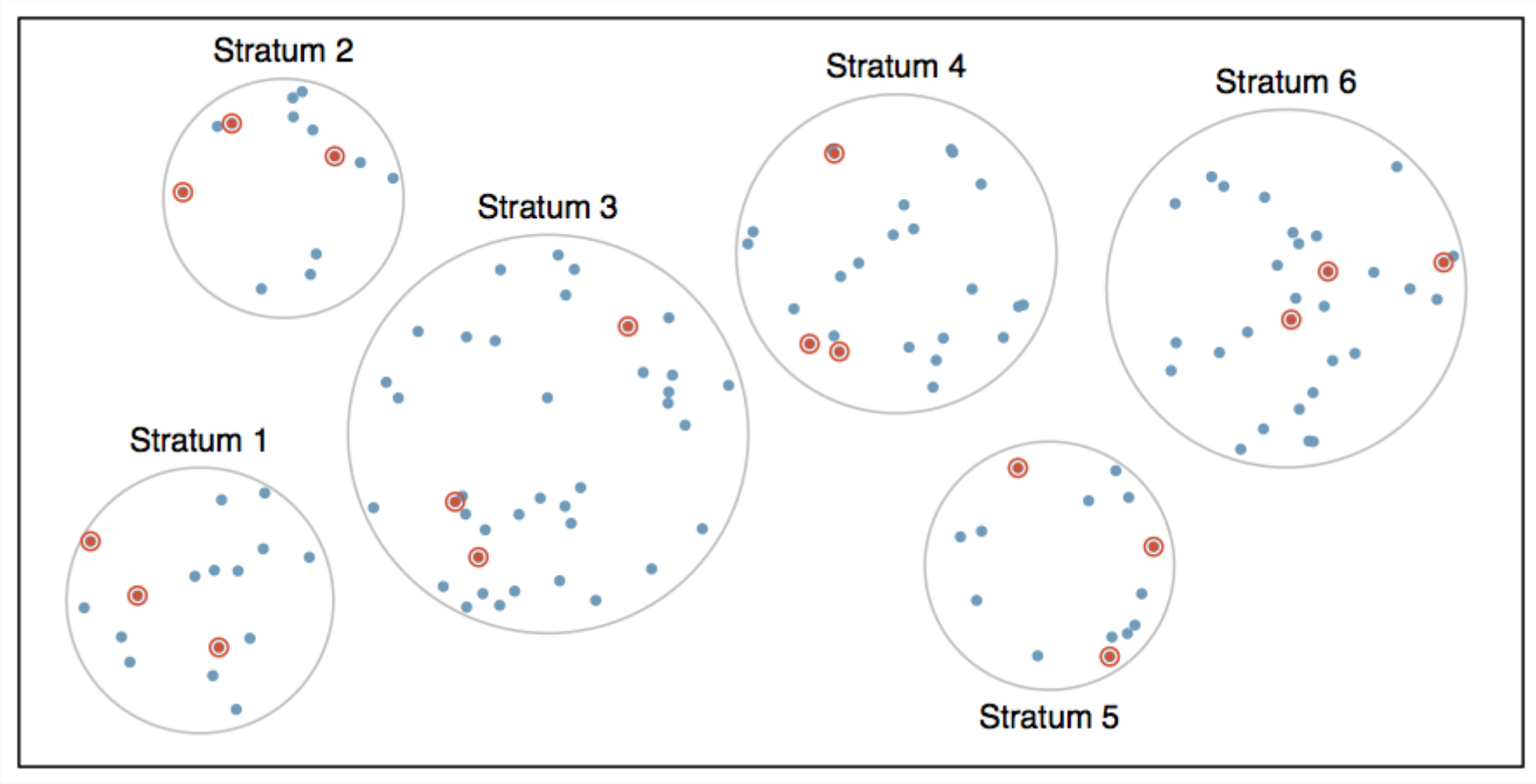
Simple Random Sampling

Randomly select cases from the population, where there is no implied connection between the points that are selected.



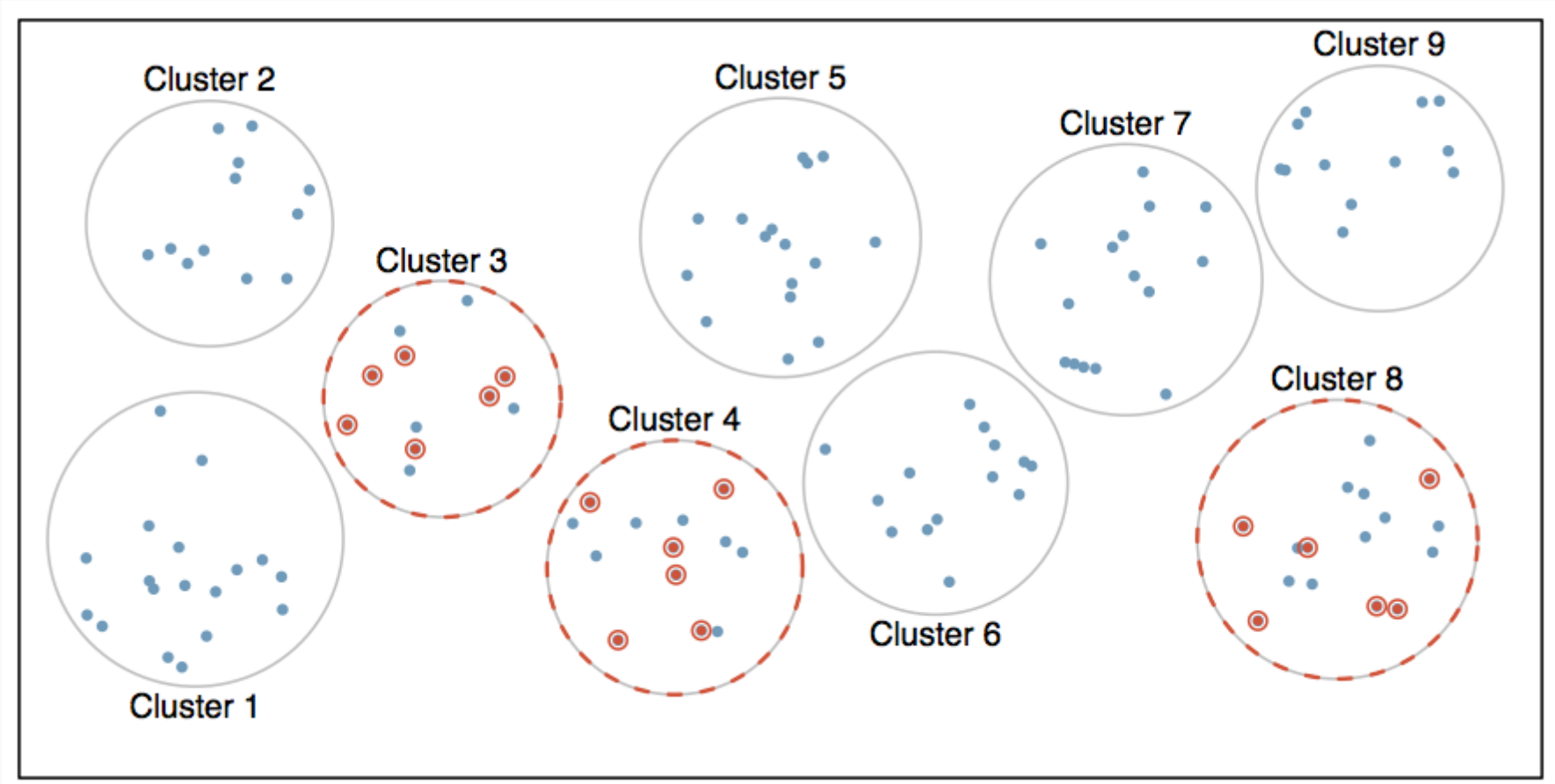
Stratified Sampling

Strata are made up of similar observations. We take a simple random sample from each stratum.



Cluster Sampling

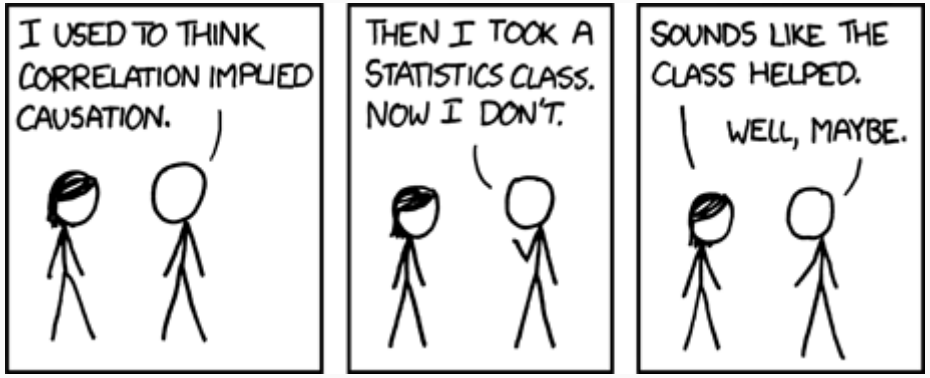
Clusters are usually not made up of homogeneous observations so we take random samples from random samples of clusters.



Observational Studies vs. Experiments

- **Observational study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.

- **Experiment**: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.



Source: [XKCD 552 http://xkcd.com/552/](http://xkcd.com/552/)

Correlation does not imply causation!

Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.

2. **Randomize**: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.

3. **Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.

4. **Block**: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

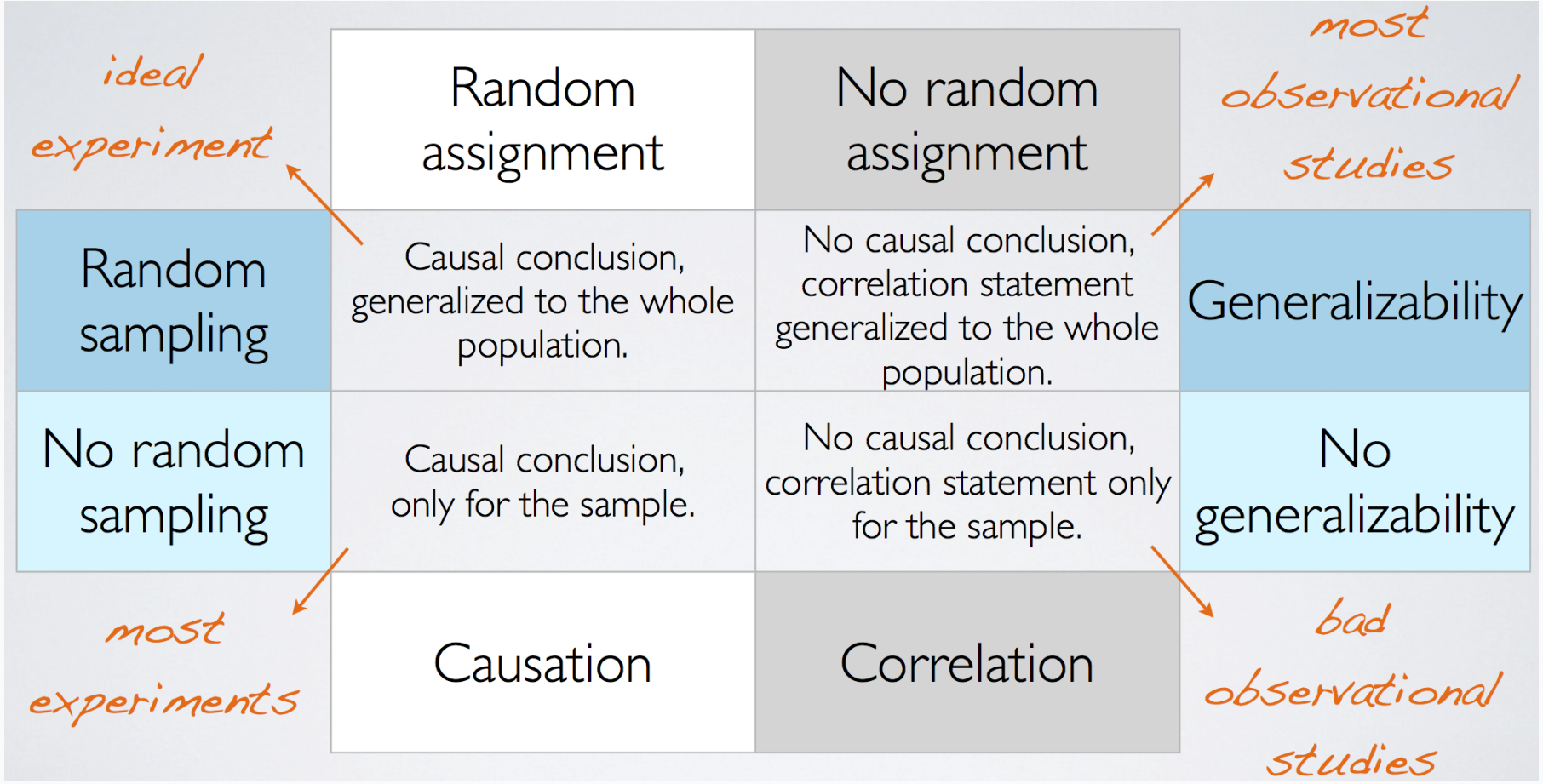
Difference between blocking and explanatory variables

- Factors are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

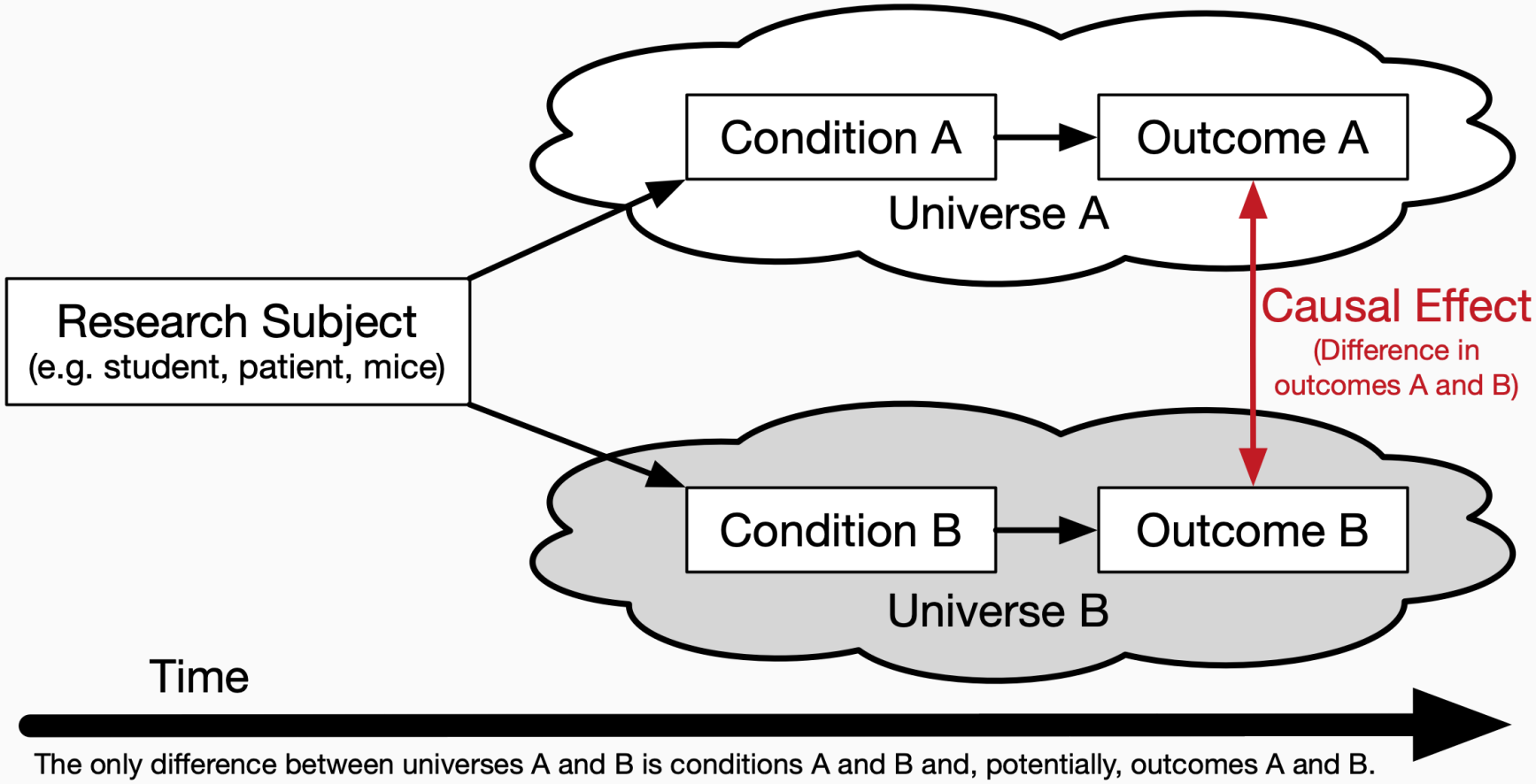
More experimental design terminology...

- **Placebo**: fake treatment, often used as the control group for medical studies
- **Placebo effect**: experimental units showing improvement simply because they believe they are receiving a special treatment
- **Blinding**: when experimental units do not know whether they are in the control or treatment group
- **Double-blind**: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

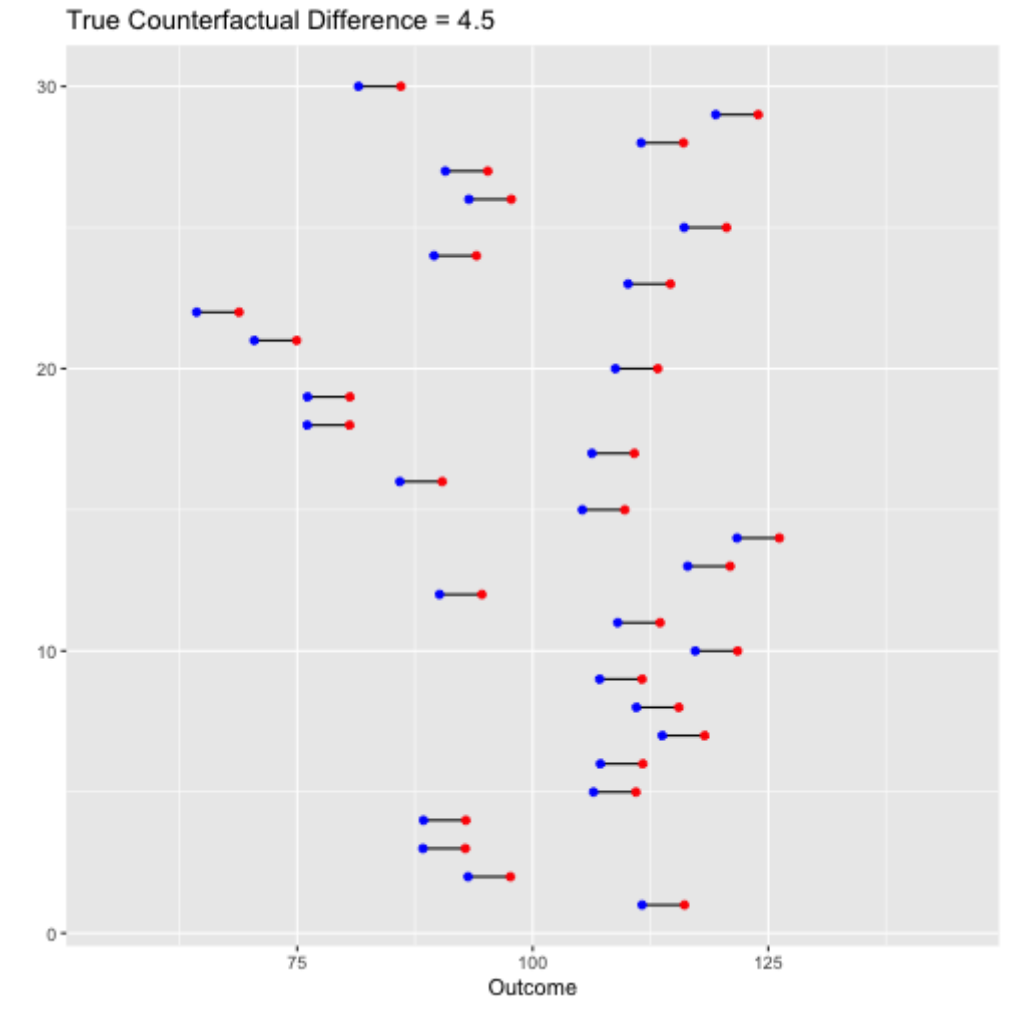
Random assignment vs. random sampling



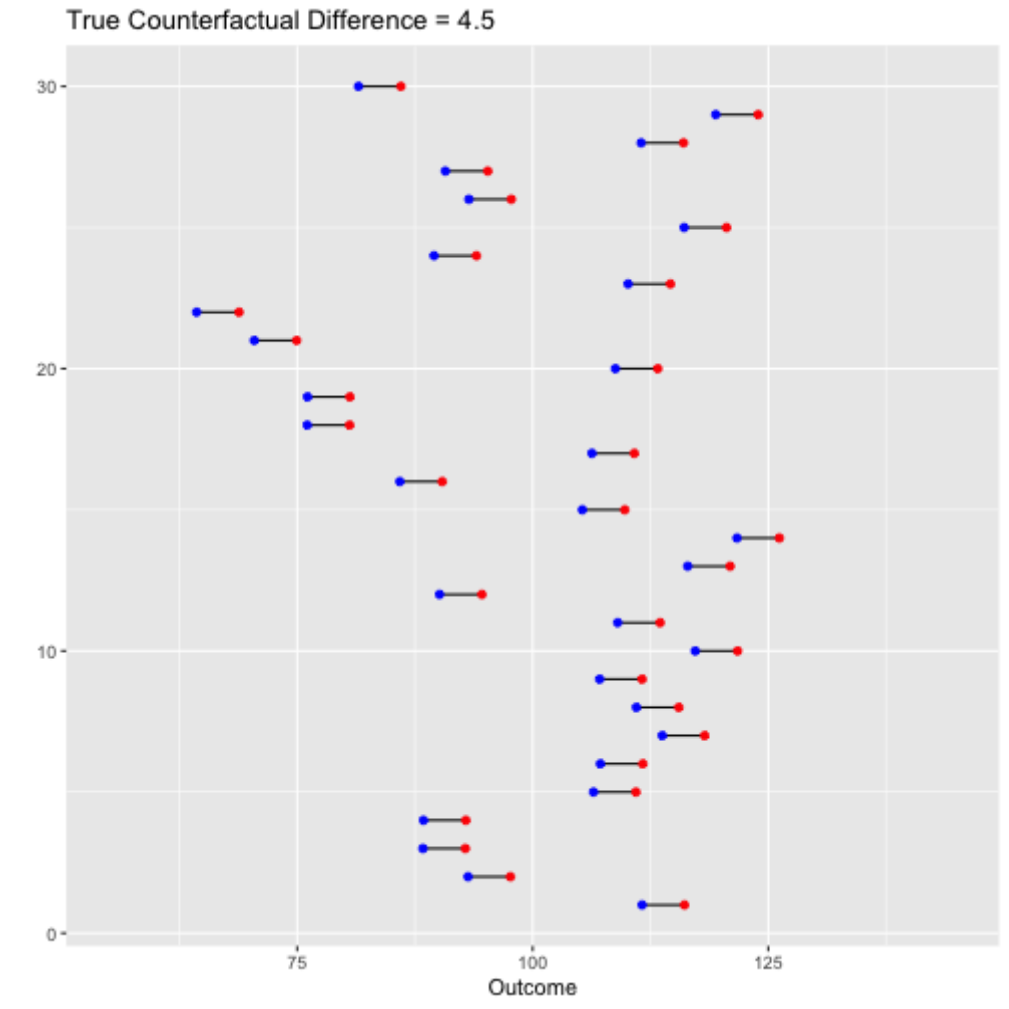
Causality

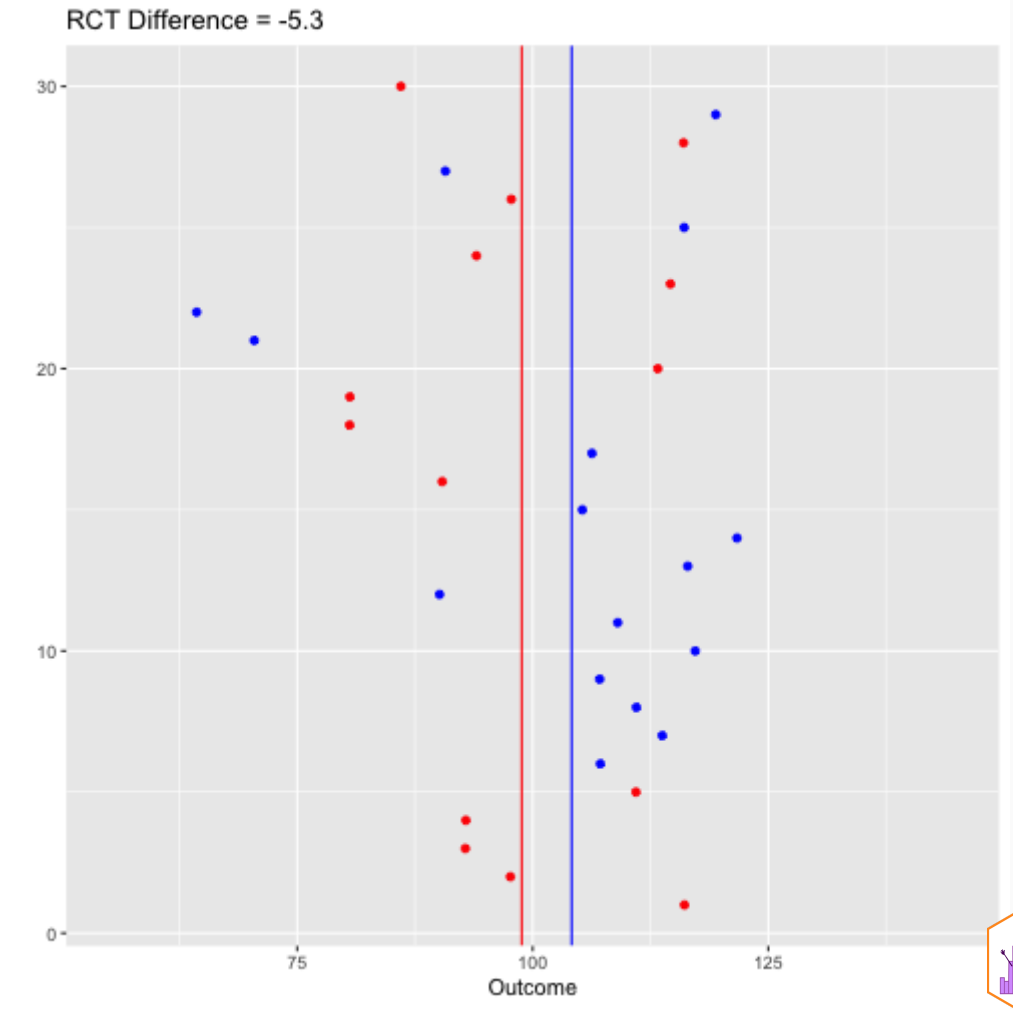

Condition A
Outcome A
Universe A
Causal Effect
(Difference in outcomes A and B)
Research Subject
(e.g. student, patient, mice)
Condition B
Outcome B
Universe B
Time
The only difference between universes A and B is conditions A and B and, potentially, outcomes A and B.

Randomized Control Trials



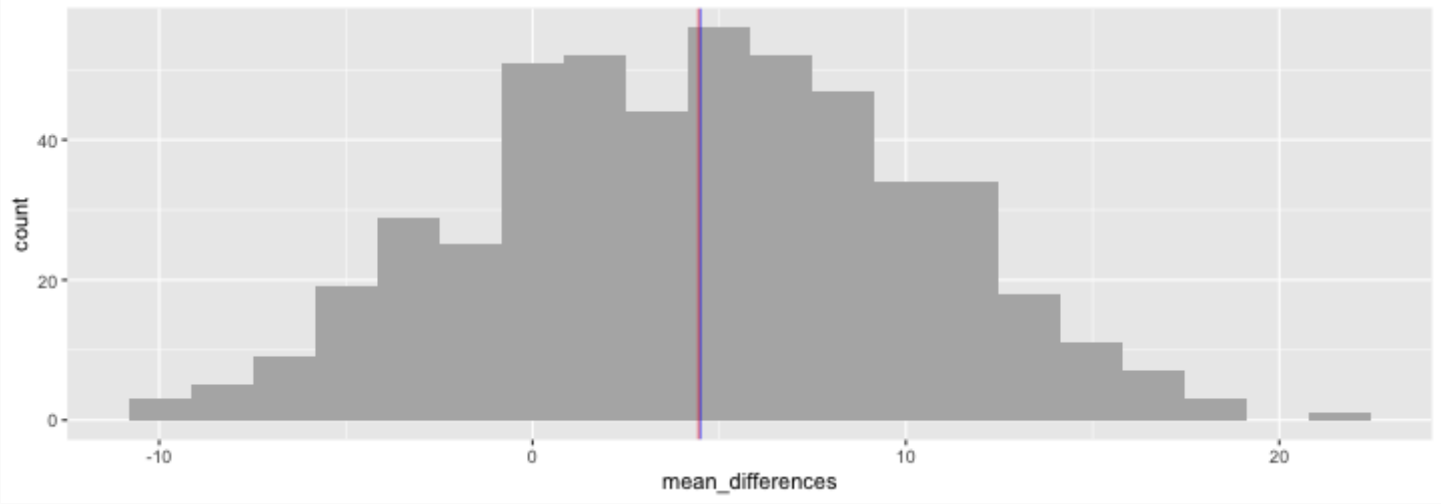
Randomized Control Trials





What if we take a lot of random samples?

```
mean_differences <- numeric(500)
for(i in 1:length(mean_differences)) {
    thedata$RCT_Assignment <- sample(c('placebo', 'treatment'), nrow(thedata), replace = TRUE)
    thedata$RCT_Value <- as.numeric(apply(thedata, 1,
                    FUN = function(x) { return(x[x['RCT_Assignment']]) }))
    tab.out <- describeBy(thedata$RCT_Value, group = thedata$RCT_Assignment, mat = TRUE, skew = FALSE)
    mean_differences[i] <- diff(tab.out$mean)
}
ggplot() + geom_histogram(aes(x = mean_differences), bins = 20, fill = 'grey70') +
    geom_vline(xintercept = mean(mean_differences), color = 'red', alpha = 0.5) +
    geom_vline(xintercept = pop.sd * pop.es, color = 'blue', alpha = 0.5)
```



Complete the one minute paper: https://forms.gle/ENFqTnDB5fJDw3kx9

1. What was the most important thing you learned during this class?

2. What important question remains unanswered for you?