Intro to Data

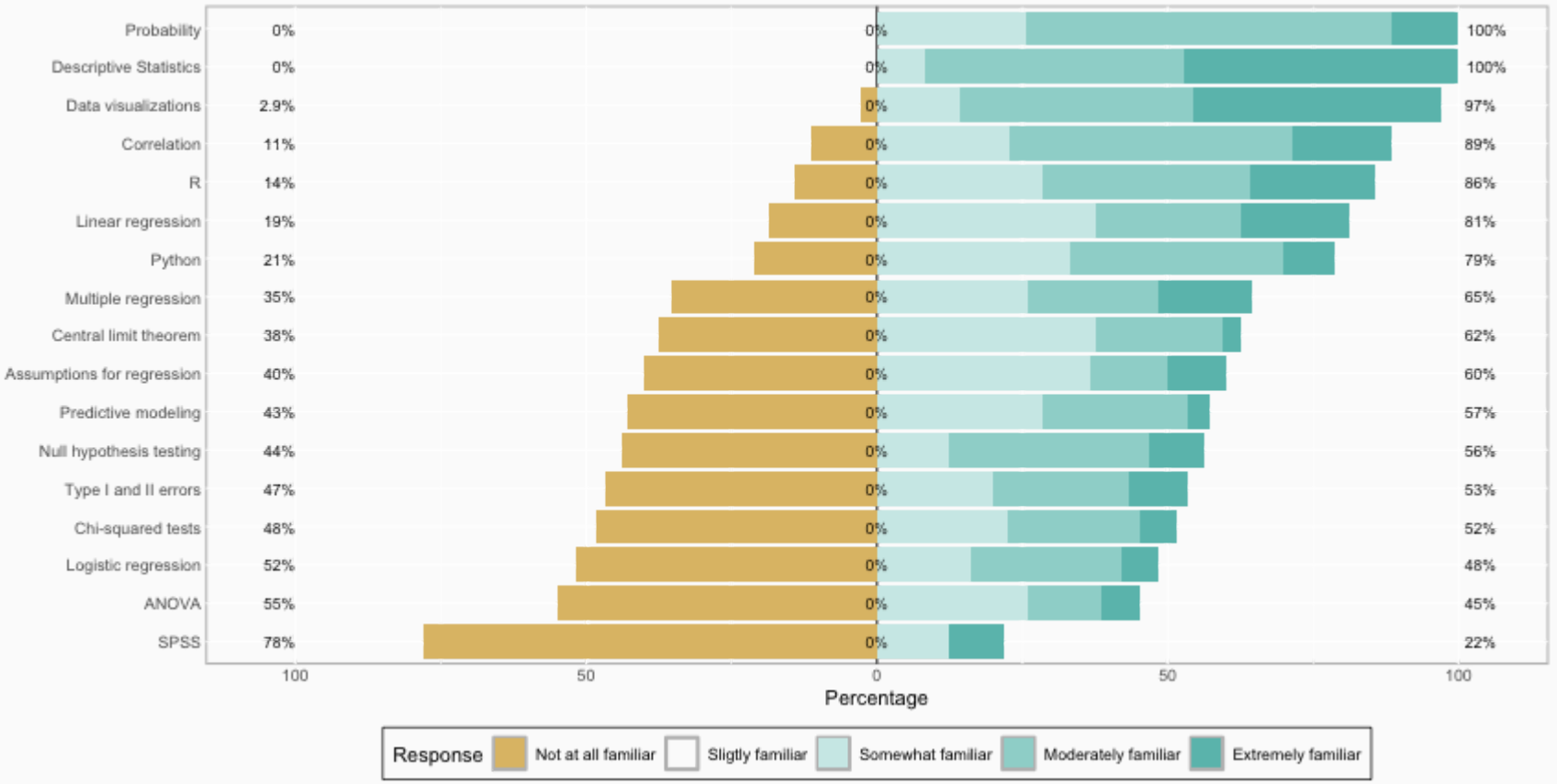
DATA 606 - Statistics & Probability for Data Analytics

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

September 7, 2026

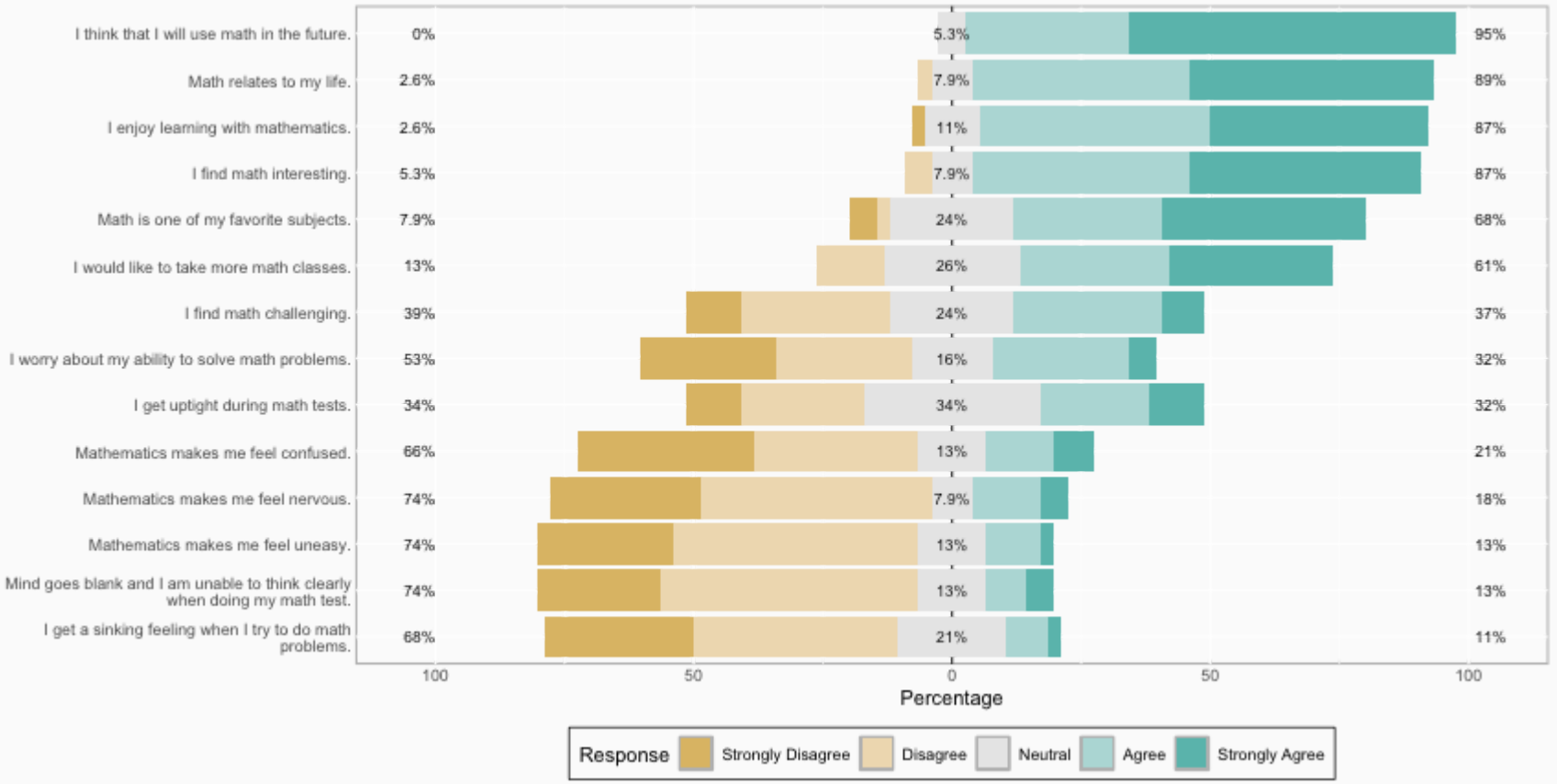
Familiarity with Statistical Topics

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



Math Anxiety Survey Scale

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



Validity and Reliability

- An assessment is **valid** if it measures what it is supposed to measure.
- An assessment is **reliable** if it measures the same thing consistently and reproducibly.
- Trusted assessments must be reliable AND valid.

Not reliable | Reliable but not valid | Reliable and valid

More info: Riezler, S., & Hagmann, M. (2021). Validity, Reliability, and Significance Empirical Methods for NLP and Data Science

Validity Evidence

1. Content
2. Response process
3. Internal structure
4. Relations with other variables
5. Consequences

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Case Study

Treating Chronic Fatigue Syndrome

- Objective: Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
- Participant pool: 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- Actual participants: Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Source: Deale et. al. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial*. The American Journal of Psychiatry 154.3 (1997).

Study design

Patients randomly assigned to treatment and control groups, 30 patients in each group:

- **Treatment**: Cognitive behavior therapy -- collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
- **Control**: Relaxation -- No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

	Yes	No	Total
Treatment	19	8	27
Control	5	21	26

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

Understanding the results

Do the data show a "real" difference between the groups?

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- The observed difference between the two groups (70 - 19 = 51%) may be real, or may be due to natural variation.
- Since the difference is quite large, it is more believable that the difference is real.
- We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

These patients had specific characteristics and volunteered to be a part of this study, therefore they may not be representative of all patients with chronic fatigue syndrome. While we cannot immediately generalize the results to all patients, this first study is encouraging. The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.

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.

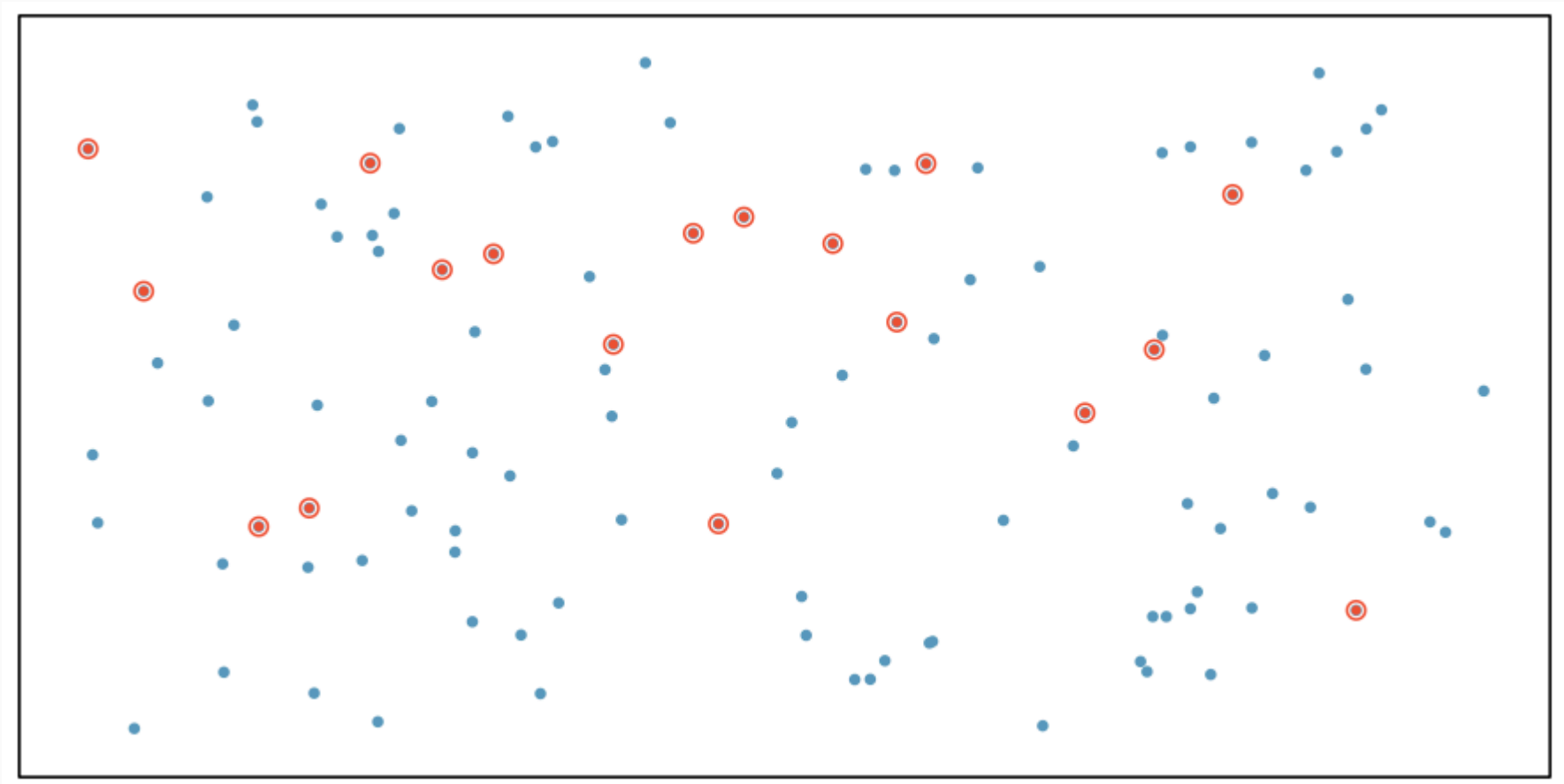
Sampling involves measuring a subset of the population of interest, usually randomly.

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.

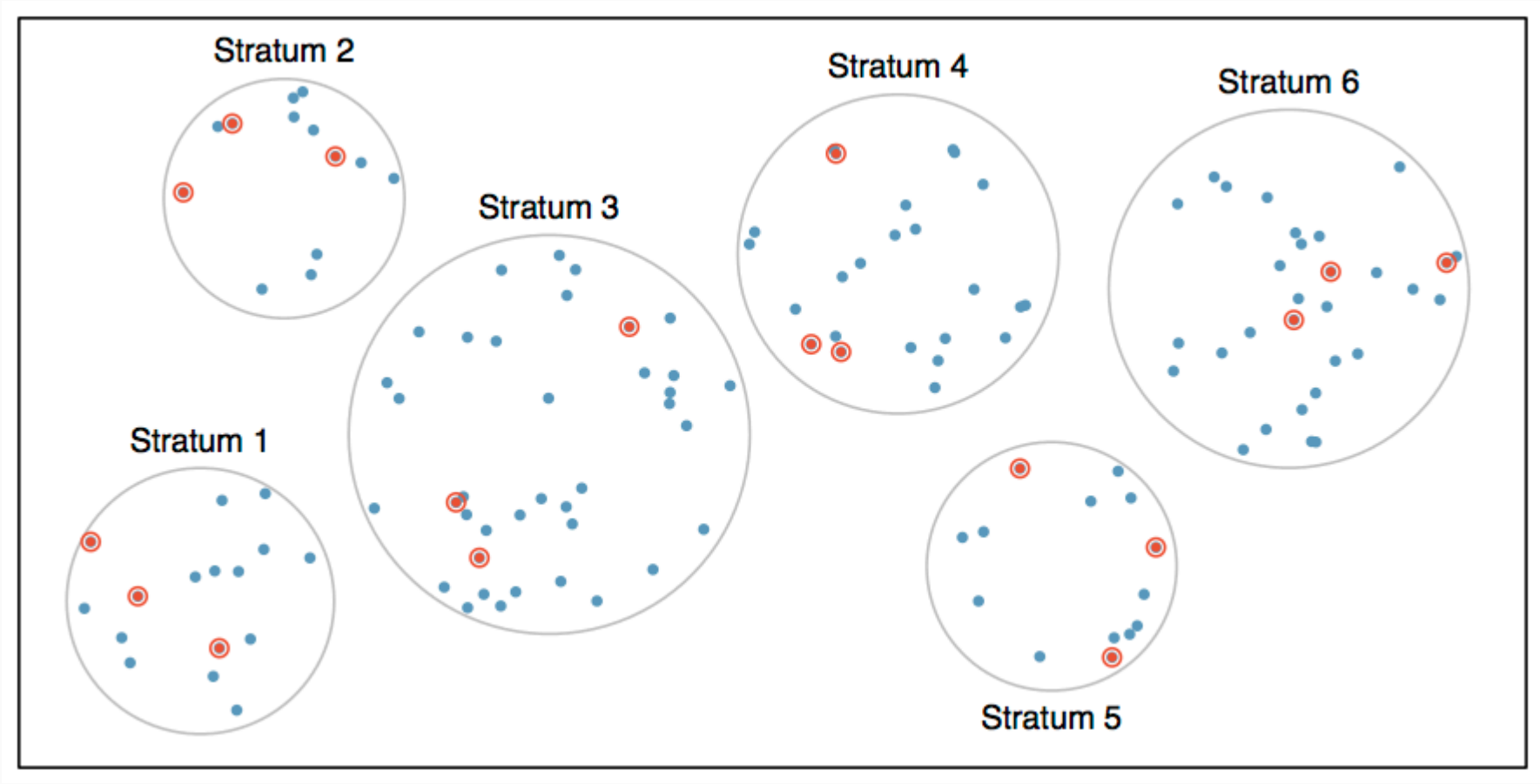
Simple Random Sampling

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



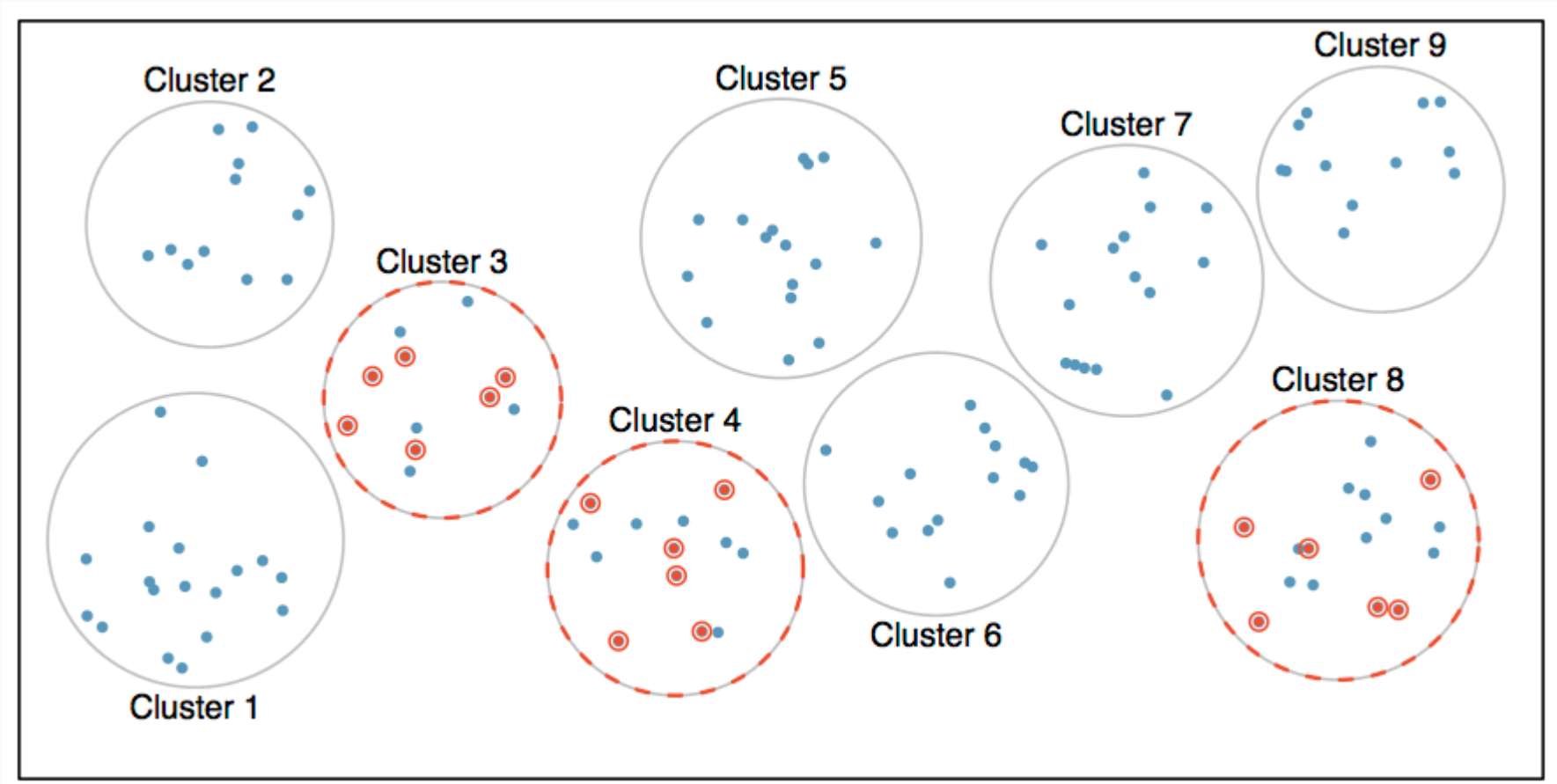
Stratified Sampling

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



Cluster Sampling

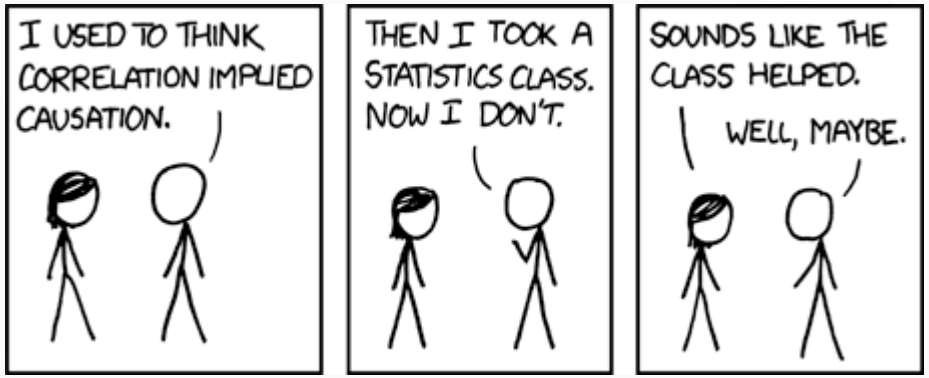
Clusters are usually not made up of homogeneous observations so we take random samples from random samples of clusters.



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.

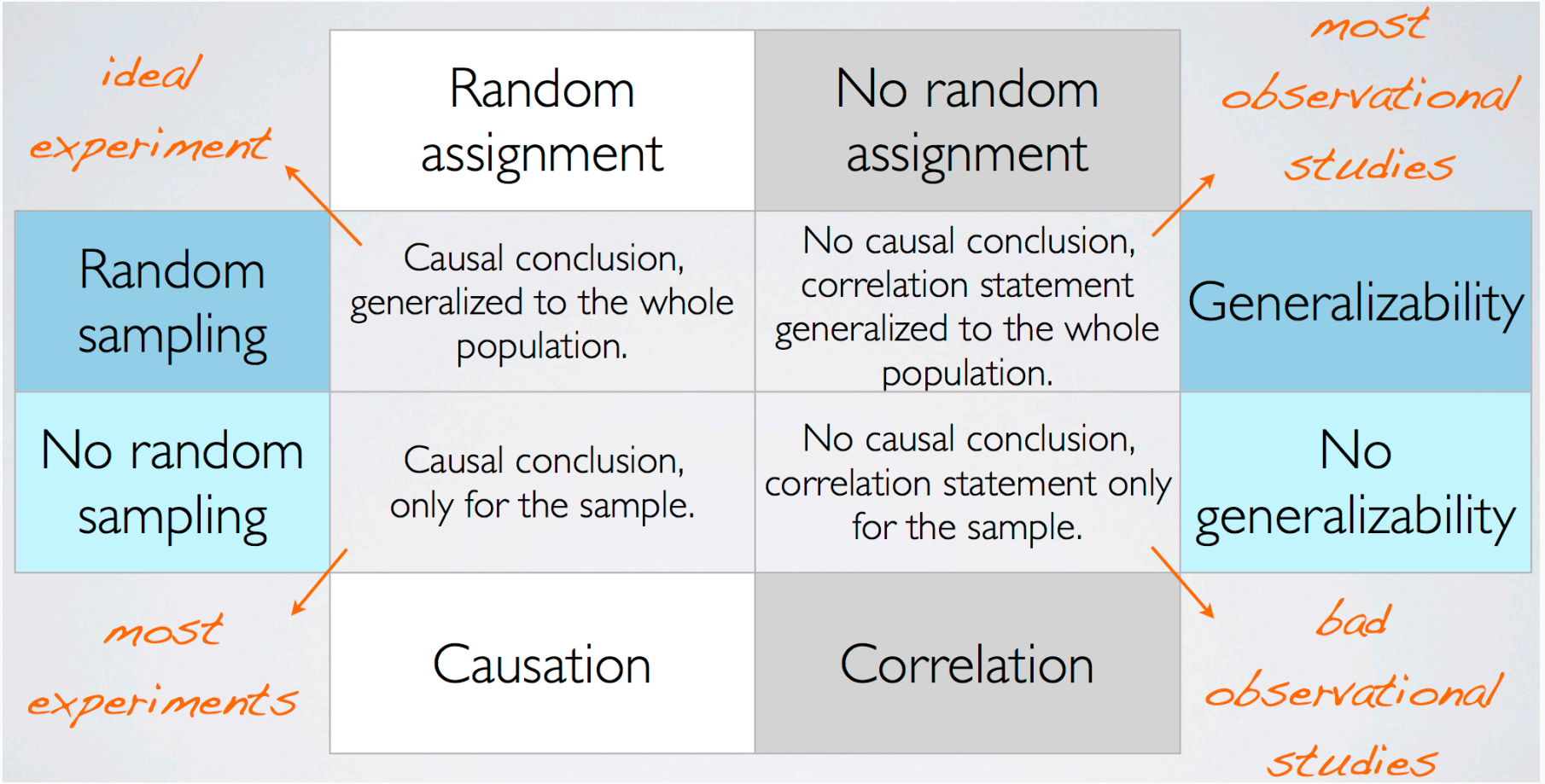
Difference between blocking and explanatory variables

- 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.

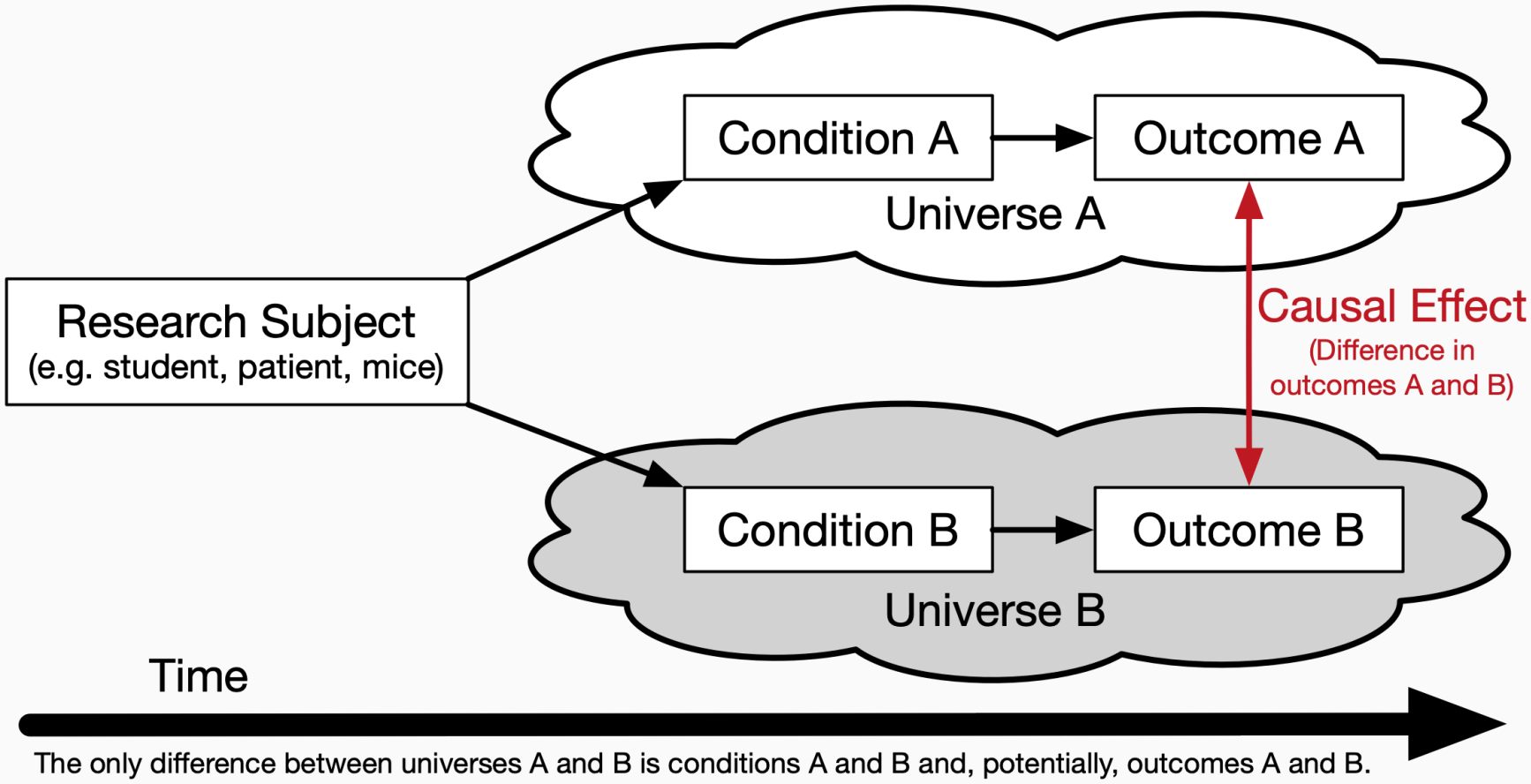
More experimental design terminology...

- **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

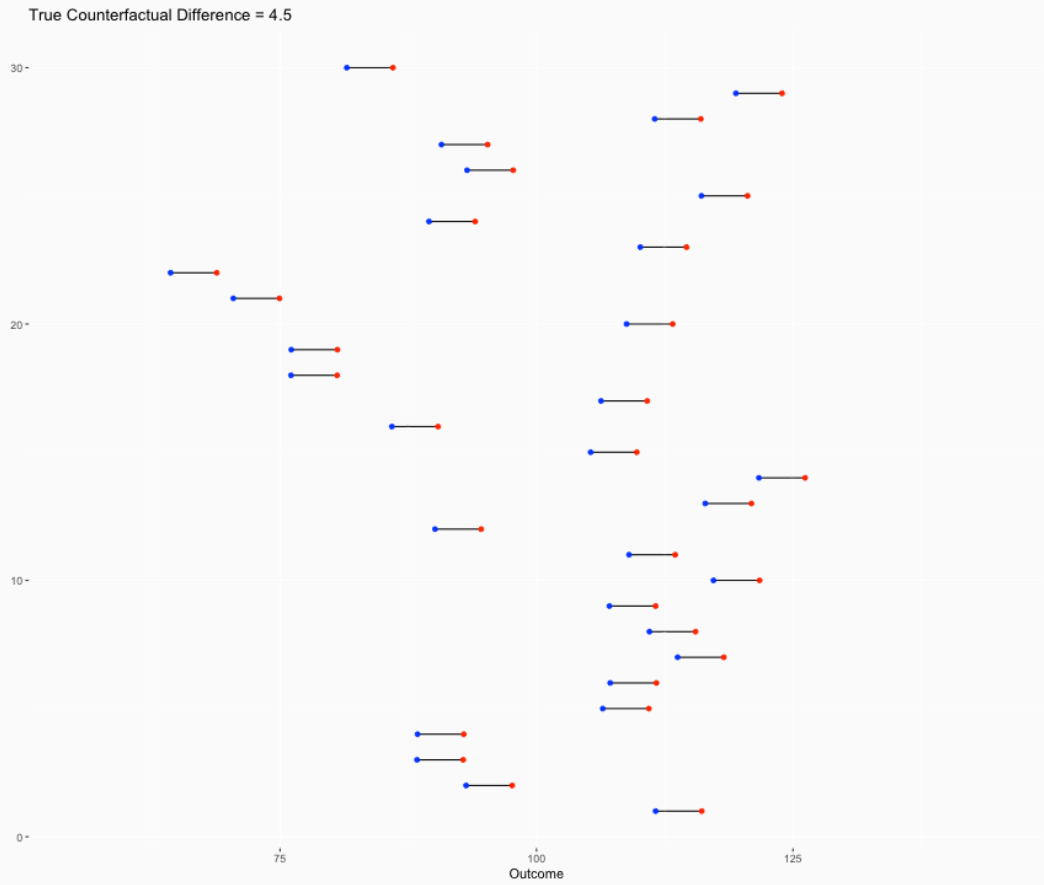
Random assignment vs. random sampling



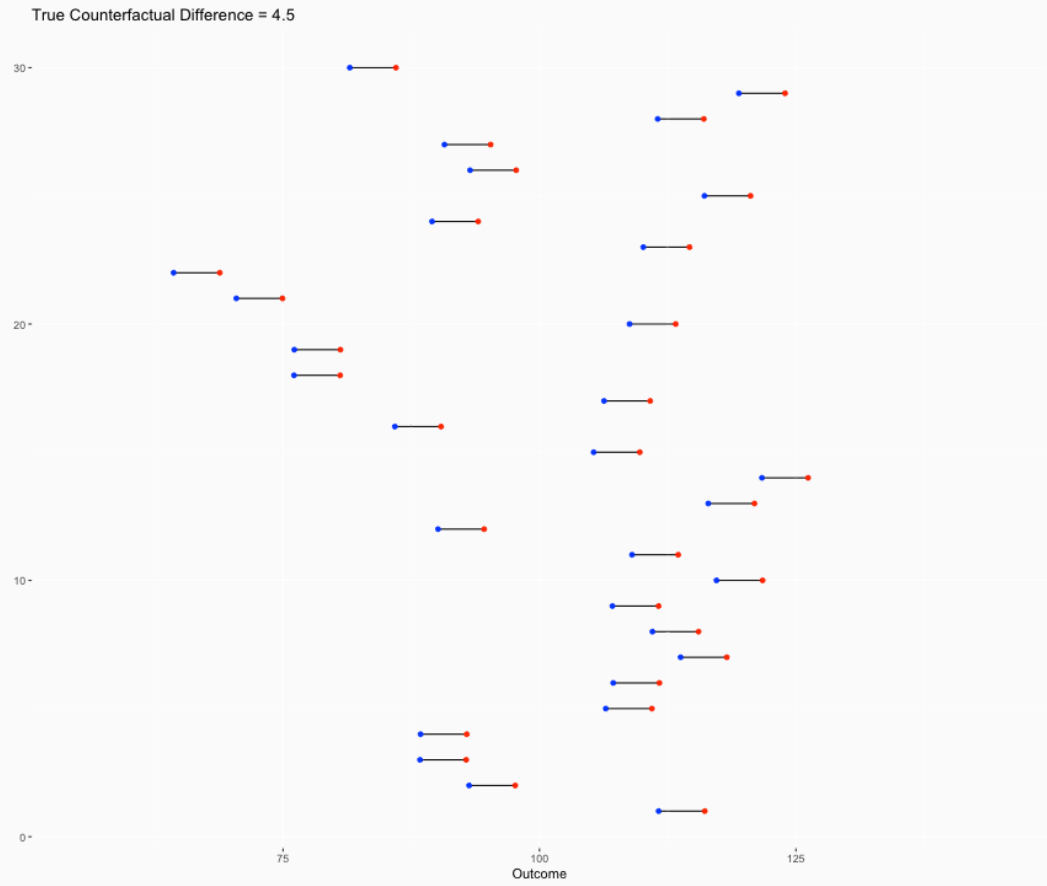
Causality

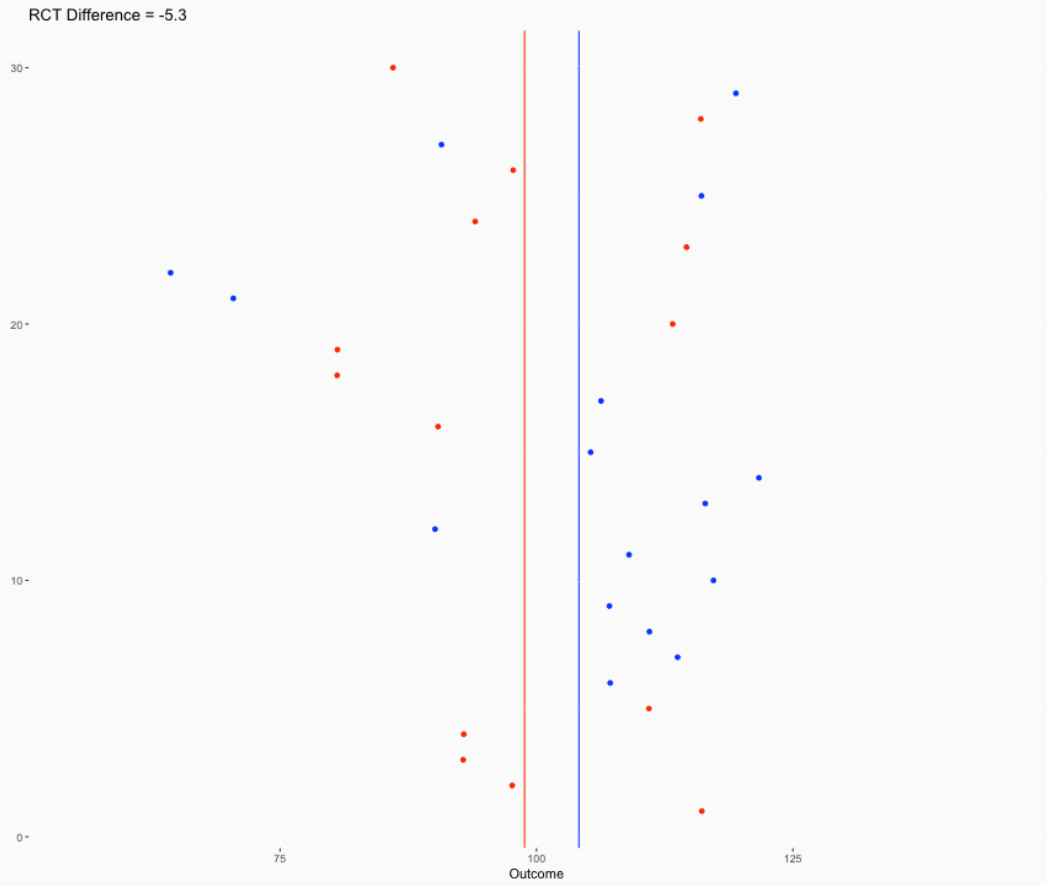

Condition A
Outcome A
Universe A
Causal Effect
(Difference in
outcomes A and B)
Research Subject
(e.g. student, patient, mice)
Condition B
Outcome B
Universe B
Time
The only difference between universes A and B is conditions A and B and, potentially, outcomes A and B.

Randomized Control Trials



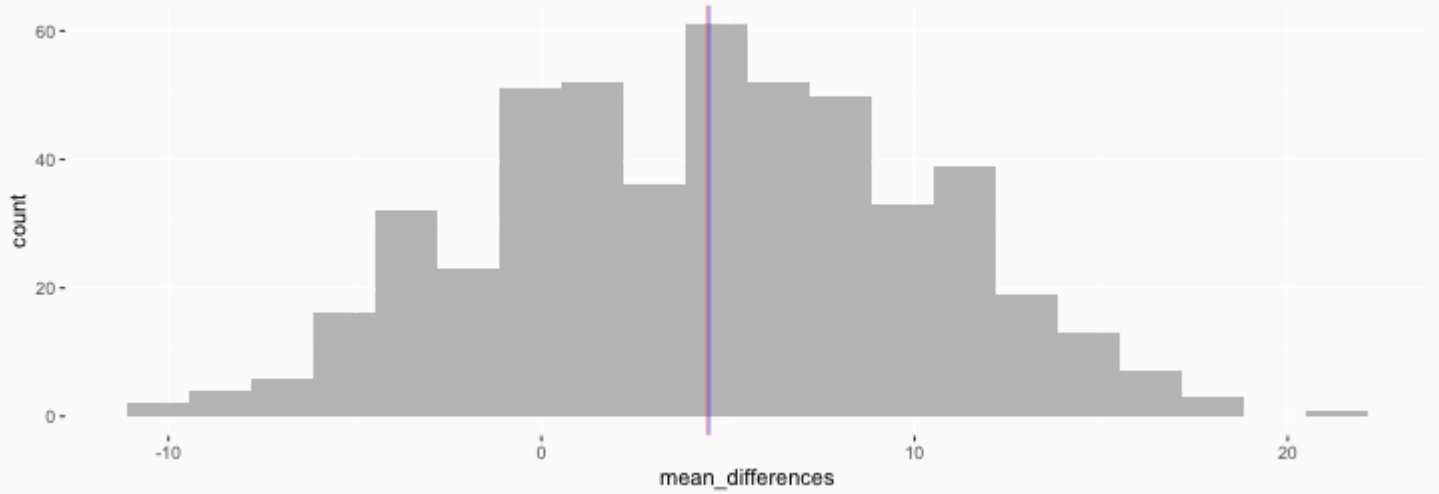
Randomized Control Trials





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)
```



One Minute Paper

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?

Scan the QR code or click here to complete the One Minute Paper