Confidence intervals, effect size, power

Data for today

- Marriages.sav

Inferential statistics

- Your hypothesis predicts a difference between means
- If you look at the means, you can tell whether there is a difference or not but you cannot tell whether this difference is due to chance
  - because there is chance variation across people
  - every time you test the SAME people under the SAME circumstances, you'll have slightly different results by chance
- Inferential statistics will tell you how likely it is that the difference you observe is due to chance
  - In other words: how likely it is that one mean comes from the same population than the other mean
  - if the difference is probably not due to chance, you have **statistically significant** results

What do we mean by "probably"

Statistical test tells you how likely it is that your observed means come from the same population

- Highest probability: 100% (p = 1)
  - meaning: there is a 100% chance that the difference between the means is due to chance
  - you definitely cannot reject the null hypothesis

- Low probability: 0.00000000001% (p = .0000000001)
  - meaning: there is an extremely low chance that the difference between the means is due to chance
  - you can definitely reject the null hypothesis

What do we mean by "probably"

We have to decide on a cut-off point (critical value, alpha)

- Depends on how much risk of Type I error you are willing to take
- High stakes research: alpha = .001
  - you take your results to be statistically significant if $p < .001$
  - this means that there is a 1% chance that the difference between the means is due to chance (Type I error)

- Low stakes research: alpha = .05
  - you take you results to be statistically significant if $p < .05$
  - this means that there is _____ chance that the difference between the means is due to chance. (Type I error)

One-tailed or two-tailed

- A one-tailed hypothesis: when the difference is predicted to be in a certain direction
- Two-tailed hypothesis: when the difference is predicted to be in either direction
- If your cut-off point is .05 for a one-tailed hypothesis, what is it for a two-tailed hypothesis?

Statistical Decision Process

	Reject Null Hypothesis	Retain Null Hypothesis
Null Hypothesis is True	Type I Error (False alarm)	Correct Decision (Correct rejection)
Null Hypothesis is False	Correct Decision (Hit)	Type II Error (Miss)

Hypothesis Testing

Identify Population and comparison group, state assumptions

Define the Null Hypothesis

Define the Research Hypothesis or Alternative hypothesis

- Define the Research and Control Group
- Define the Dependent and Independent Variables

Decide whether your prediction is one-tailed or two-tailed

State relevant characteristics of comparison distribution

Determine critical cutoff values

Calculate statistic

Reject or Fail to Reject the NULL Hypothesis

Assumptions for Parametric Tests

Dependent variable is a scale variable → interval or ratio

- If the dependent variable is ordinal or nominal, it is a non-parametric test

Participants are randomly selected

- If there is no randomization, it is a non-parametric test (and likely a flawed experiment, or one limited in generalization)

The shape of the population of interest is approximately normal

- If the shape is not normal, it is a non-parametric test

normal distribution?

Options:

1. Draw a histogram with a normal curve
   (SPSS: Analyze > Descriptive Statistics > Frequencies > Charts
2. Run a test of normality
   (SPSS: Analyze > Descriptive Statistics > Explore > Plots: Normality plots with tests
   1. Kolmogorov-Smirnov
   2. Shapiro-Wilk

   If $p < .05$, distribution is significantly different from normal and you cannot use parametric tests (??)

Exercise: Tests of normality

- Are the scores normally distributed:
  - Do you look at the entire sample?
  - Height, Facebook friends, alcohol consumption, sleep, etc.
  - Shopping time and distance

confidence intervals

beyond the p value

The mean vs. a confidence interval

- Point estimate: summary statistic – one number as an estimate of the population
  - E.g., 32 point difference b/w Boys and Girls maths test score

- Interval estimate: range of sample statistics we would expect if we repeatedly sampled from the same population

- Confidence interval: an interval in which the population mean is likely to fall
  - e.g. people rate phrases on a scale of 1-10 of annoyingness
  - the most annoying phrase is “whatever”, with a mean rating of 8.5 in our sample
  - based on the mean and the standard error, we can calculate a confidence interval: say the population mean will probably fall between 7.5 and 9.5
- 95% Confidence interval: an interval in which the population mean will fall with 95% probability
- 99% Confidence interval: _____

Calculating confidence intervals

1. Get the standard normal distribution
2. Choose confidence level
3. With 95% confidence, you expect the mean to fall in the middle 95% of the scores: between a z-score of -1.96 and a z-score of 1.96
4. Convert the z scores to raw scores using the standard error

Standard Error:

$$SE = \frac{SD}{\sqrt{N}}$$

Lower bound: $-z(SE) + M_{sample}$

Upper bound: $z(SE) + M_{sample}$

(Or use SPSS)

Factors affecting the confidence interval

The length of the confidence interval is influenced by sample size.

The larger the sample, the narrower the interval (because the Standard Error decreases with an increase in sample size).

- This makes sense: the more people we measure, the more scores will be close to the mean

But that does not influence the confidence level.

Confidence intervals in SPSS

Descriptive Statistics > Explore

For our sleep data:

On Tuesday Mean: 7.22 , CI_{95} = 6.81 – 7.63

On Saturday Mean: 7.66 , CI_{95} = 7.15 – 8.18

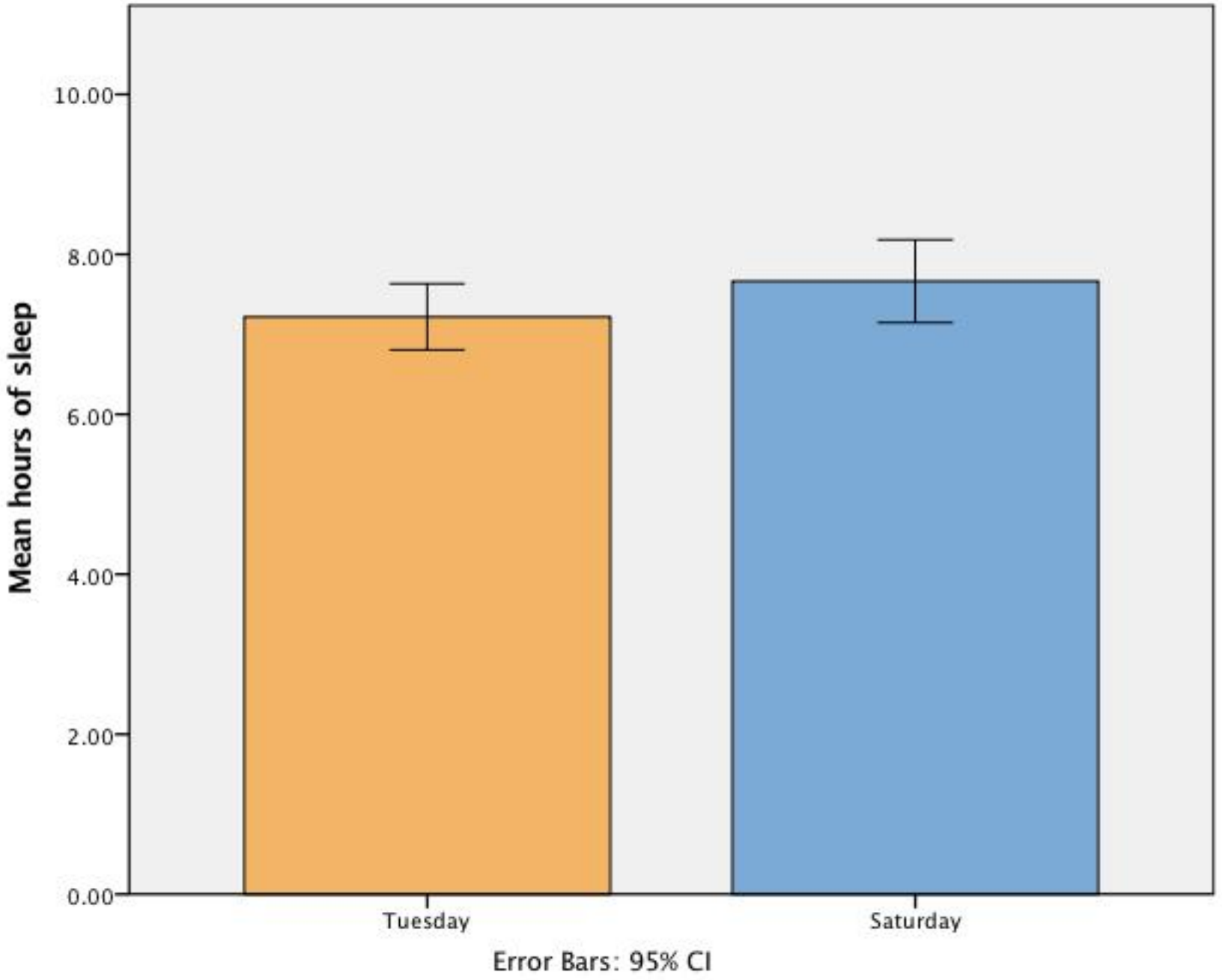
95% CI is often reported as: M = 7.22[6.81,7.63]

Is there a clear overlap? What does that mean?

Does the day of week likely to have an effect on hours of sleep?

Graphs:

Comparing two groups with bar chart showing confidence intervals


Mean hours of sleep
10.00
8.00
6.00
4.00
2.00
0.00
Tuesday
Saturday
Error Bars: 95% CI

Within-group vs. between-groups designs

Compare last year's class with this year

Sleep data for last year:

On Tuesday Mean: 6.52 [5.95 – 7.08]

Sleep data for this year:

On Tuesday Mean: M = 7.22[6.81,7.63]

Exercise (See Field Ch1 Task 9 for names of celebrities): create SPSS file, test for normal distribution, find outliers, get means, confidence intervals, draw graph (you can generate an ID variable through Transform -> Compute Variable -> $Casenum

Length of marriage in days			
Celebrities		**Non-celebrities**	
143	240	210	15591
72	144	13901	1339
30	26	16662	1453
2	150	16672	21963
14	150	8222	8898
1657		19543	

effect size

Effect Size

Having a statistically significant difference between means does not mean that the difference is large!

To measure the size of the difference, we calculate effect sizes: the difference between the two means divided by the standard deviation.

It is affected by

- the distance between means of two distributions (larger distance means larger effect size)
- the standard deviations of the two distributions (a smaller standard deviation means a larger effect size)

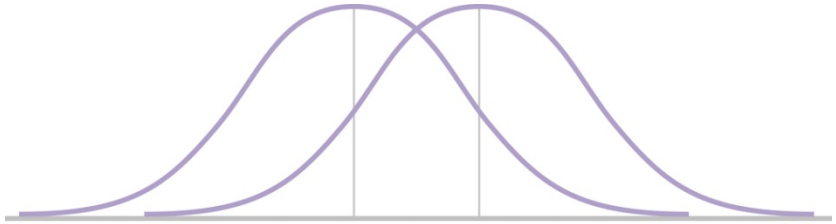
Unaffected by sample size

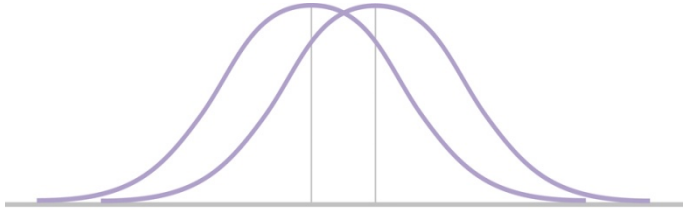
If we use standardised effect sizes, we can compare studies using different measurement scales (Meta-analyses)

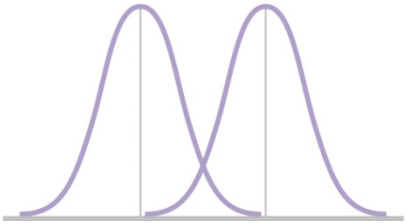
Effect Size and Mean Differences

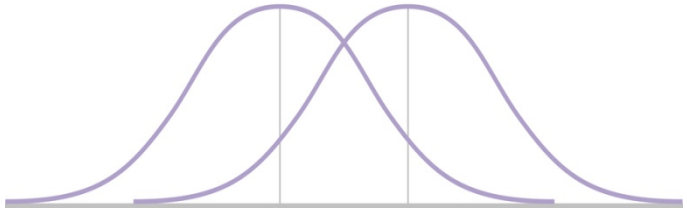
Imagine both represent significant effects

Note the Spreads and Distances: Which effect is bigger?









calculating Effect Size: a standardised method

Cohen's *d*: effect size estimate

$$d = \frac{(M_1 - M_2)}{SD}$$

TABLE 12-1. COHEN'S CONVENTIONS FOR EFFECT SIZES: *d*

Jacob Cohen has published guidelines (or conventions), based on the overlap between two distributions, to help researchers determine whether an effect is small, medium, or large. These numbers are not cutoffs, merely rough guidelines to aid researchers in their interpretation of results.

EFFECT SIZE	CONVENTION	OVERLAP
Small	0.2	85%
Medium	0.5	67%
Large	0.8	53%

SD of control group or pooled:

$$SD = \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2}{N_1 + N_2 - 2}}$$

CAUTION: The formula for the z-stat and *d* differ importantly at the denominator -- sample size matters for *z* but not *d*

$$z = \frac{(M - \mu)}{SE}$$

Other measures of effect size

- Cohen's d for t-tests
- η^2 (Eta squared) or ω^2 (Omega squared) for ANOVA
- φ (phi), Cramer's V or odds ratio for chi square
- r for non-parametric tests

But the different measures can be converted into each other.

Online calculator: http://www.campbellcollaboration.org/resources/effect_size_input.php

- We can calculate Confidence intervals for effect sizes (the online effect size calculator will do it)
  - What's a 95% Confidence Interval?

- Forest plot: effect sizes with their confidence intervals
  - Confidence intervals are affected by Standard Errors
    - A larger Standard Error means more variability, ie., less confidence in our estimate of the population parameter (effect size in this case)

A meta-analysis

- Depression in children and adolescents (Horowtiz & Garber, 2006) Table 21.3 in Howell
- Studies of intervention programmes at schools collected
- Effect sizes for treatment vs. control groups calculated

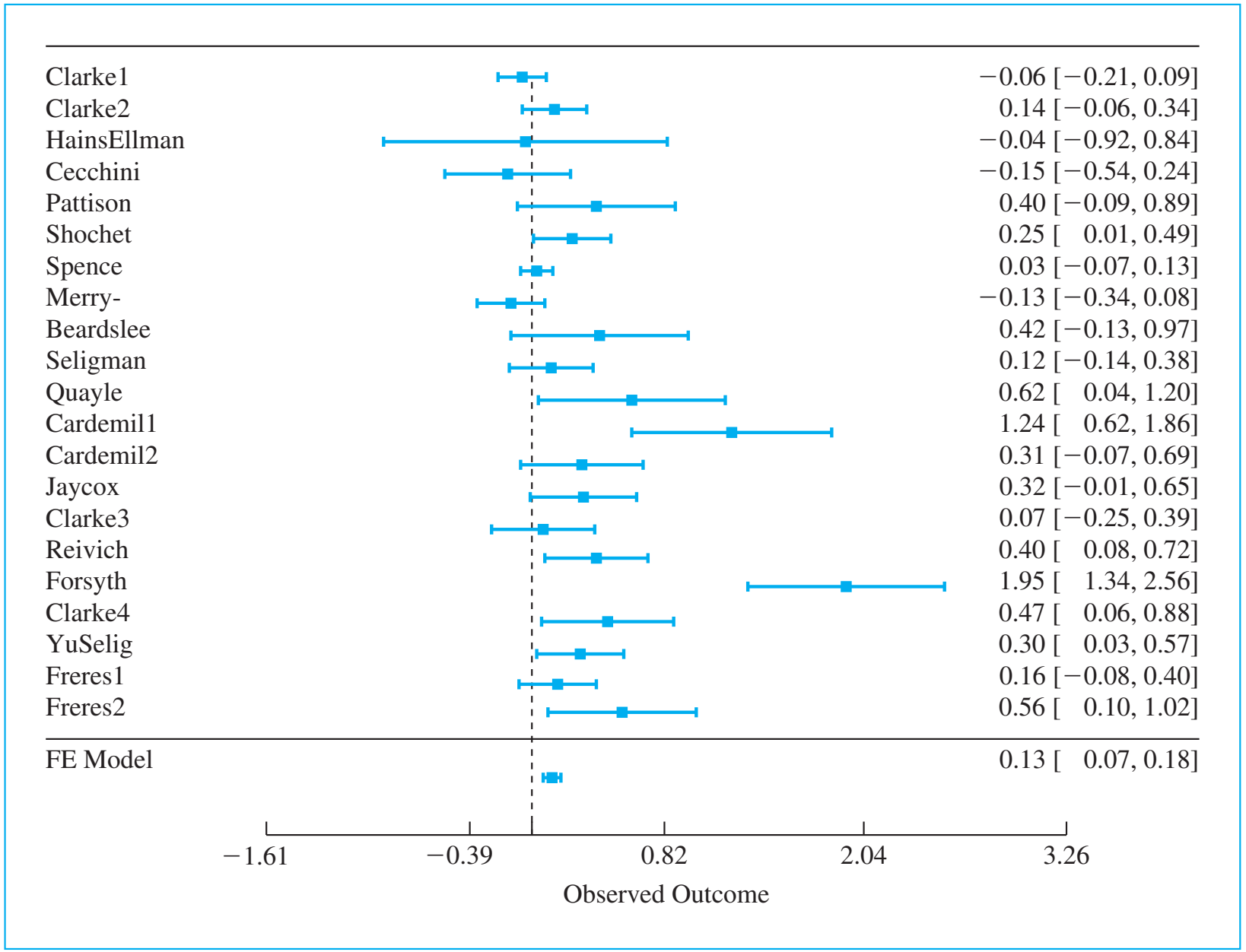


Figure 21.1
Forest plot for Horowitz and Garber's data.

Reminder: Calculating standard errors and confidence intervals

1. Get the standard normal distribution
2. Choose confidence level
3. With 95% confidence, you expect the mean to fall in the middle 95% of scores: between a z-score of -1.96 and a z-score of 1.96
4. Calculate the Standard Error:

$$SE = \frac{SD}{\sqrt{N}}$$

Lower bound: -1.96(SE) + *d*

Upper bound: 1.96(SE) + *d*

- Next step: weight each *d* by the inverse of its standard error:
  - A *d* with a large SE will have less weight (it's less reliable)
  - A *d* with a small SE will have more weight (it's more reliable)

- Then calculate the mean weighted effect size and it's confidence limits
  - This is shown under the horizontal line in Figure 21.1 (FE = Fixed Effect (without the sampling error of individual studies))

statistical power

statistical power

- The probability of a hit if the research hypothesis is true (i.e. that we don't miss a significant result)
- Power = 100 – probability of Type II error

- Power is affected by:
  - sample size
  - effect size (i.e. means and standard deviations)
  - (alpha)
  - (one or two-tailed hypothesis)

Why calculate Statistical Power?

- It can be used to estimate the required sample size:

http://www.statisticalsolutions.net/calculators.php

- How many people do we need to have at least 80% statistical power in
  - the sleep study?
  - the marriage study?

Homework

Rerun the analyses on the marriage data and write it up.

- Distribution:
  - histograms
  - test of normal distribution
  - identifying outliers, what to do with them
- Descriptive statistics
  - Means, standard deviations
- Confidence intervals
  - Calculate 95% confidence intervals
  - Draw graphs
  - Calculate Cohen's d
  - Draw conclusions