

CogLab: Communicate

WEEK 12

where are we?

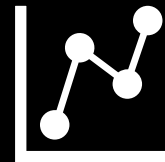
- literature review
- asking questions
- experiment creation
[HTML/jsPsych]

design



- R & Rstudio
- describe data
- infer from data

analyze



- pre-registration
- poster
- short report

communicate



where are we?

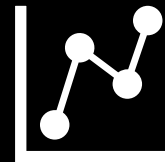
- ~~literature review~~
- ~~asking questions~~
- ~~experiment creation~~
~~[HTML/jsPsych]~~

design



- ~~R & Rstudio~~
- ~~describe data~~
- ~~infer from data~~

analyze



- ~~pre-registration~~
- poster
- short report

communicate



today's agenda

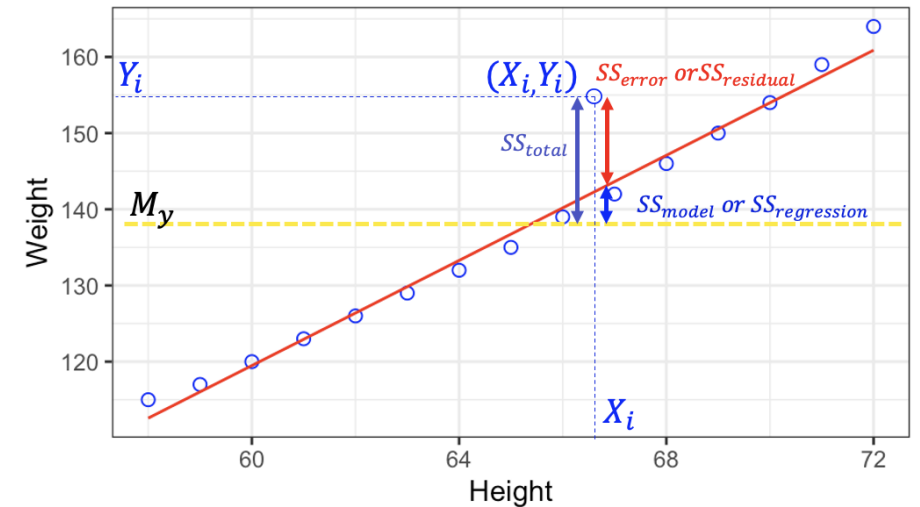
- data collection + analyses check-in
- understanding variation
- poster design principles

what is variation?

- statistical analyses such as t-tests and ANOVAs often tend to emphasize means between conditions
- but variation is fundamental to these tests and often a core part of the underlying machinery
- data = a combination of central tendency and variation
- data = model + error
- variation refers to the spread of data points around the central tendency for any set of data

variation in common statistical tests

- most statistical tests care deeply about the variation in data points (as they should)!
- t-tests
 - standard deviations used to calculate the t statistic
- ANOVAs
 - sums of squared (standard) deviations (SS) differentiate between the signal (SS_{between} or SS_{model}) and noise (SS_{within} or SS_{error})
- regression?
 - fits the line $y = \beta_0 + \beta_1 x$ that minimizes sums of squares
 - Total SS = $SS_{\text{explained}} + SS_{\text{residual}}$

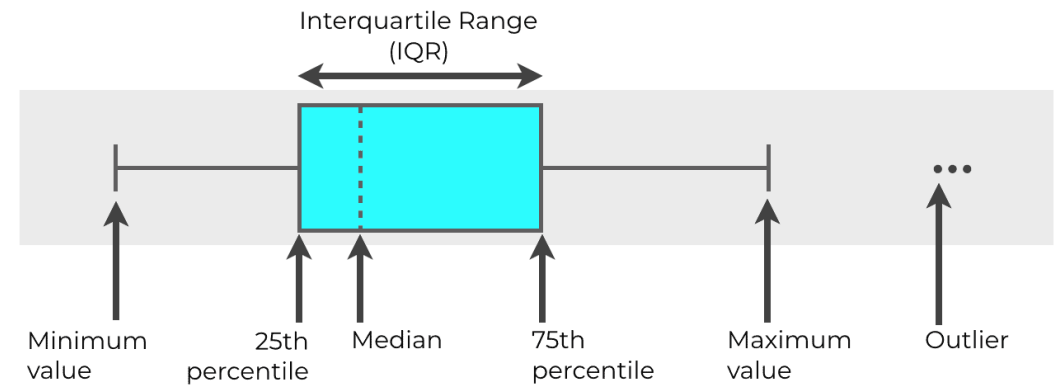


so how do we visualize variation?

- our analyses do incorporate variation in different ways (based on the statistical test)
- our visualizations, however, sometimes lack in displaying the full spread of the data
- what kinds of plots have we seen so far and which of them show any form of variation?

boxplots

- “five-number summary”
 - the minimum
 - the first quartile (25th percentile)
 - the median
 - the third quartile (75th percentile)
 - the maximum
- implicit measures:
 - IQR: 25th to 75th percentile
 - minimum: $Q1 - 1.5 \cdot IQR$
 - maximum: $Q3 + 1.5 \cdot IQR$



bar plots

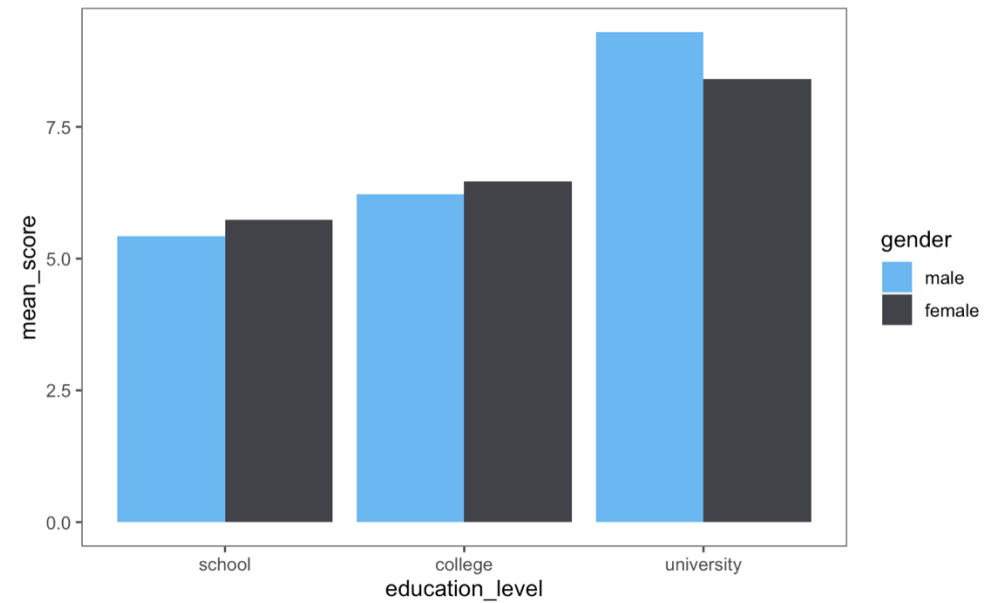
- bar plots often display the **means** for the relevant conditions in psychological studies
- but what about **variability**??
- a few different options:
 - **error bars** that denote some type of variation
 - what could this be?
 - an **overlay of original data points** in each condition

open your RStudio project

- open the first-jspsych project and your .Rmd file
- **load** tidyverse and ggthemes
- DON'T run all chunks – no need today!

exercise 1: reproduce this plot!

- using the `jobsatisfaction` dataset from the `datarium` package, reproduce this plot

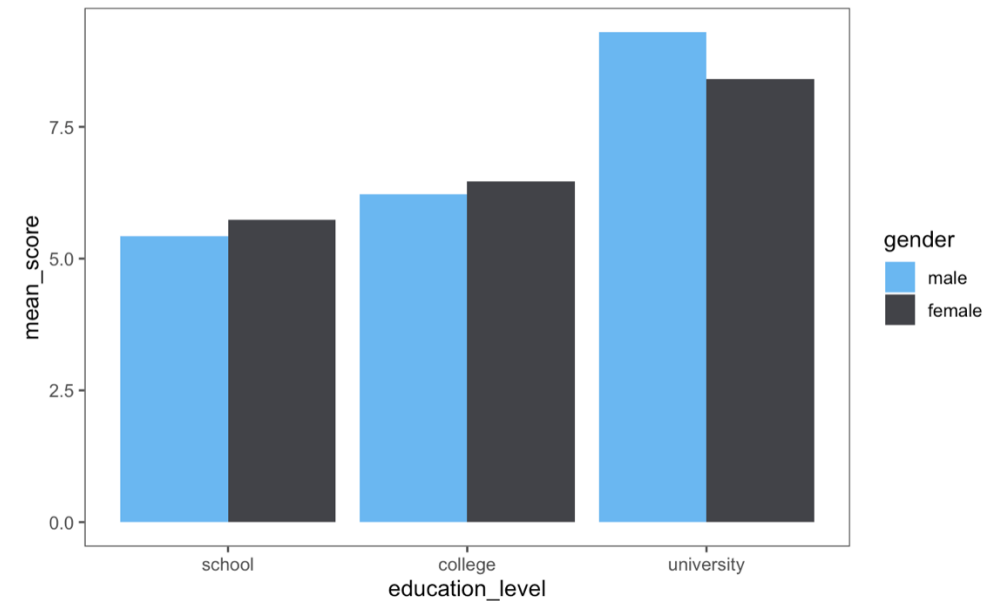


exercise 1: reproduce this plot!

- using the `jobsatisfaction` dataset from the `datarium` package, reproduce this plot

```
data("jobsatisfaction", package = "datarium")

jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarise(mean_score = mean(score)) %>%
  ggplot(aes(x = education_level, y = mean_score,
             group = gender, fill = gender)) +
  geom_col(position = "dodge") +
  scale_fill_hc() +
  theme_few()
```

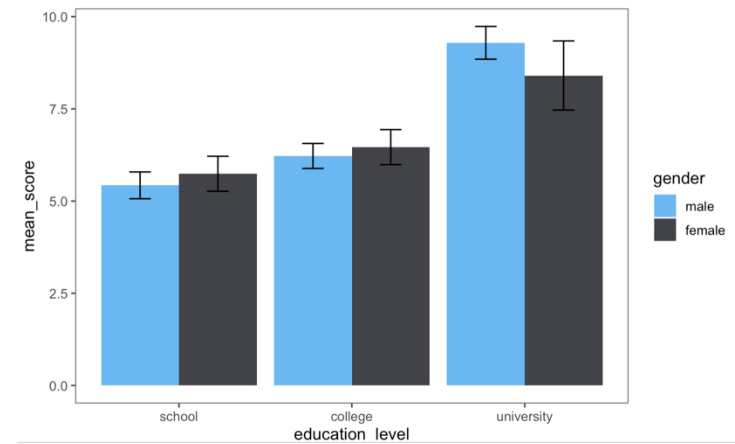


exercise 2: adding variation

- now, let's try to add some variation to our bars:
 - store the mean scores and standard deviation of scores in a dataframe

exercise 2: adding variation

- now, let's try to add some variation to our bars:
 - store the mean scores and standard deviation of scores in a dataframe
 - use `geom_errorbar` to add an error bar to each bar of your plot

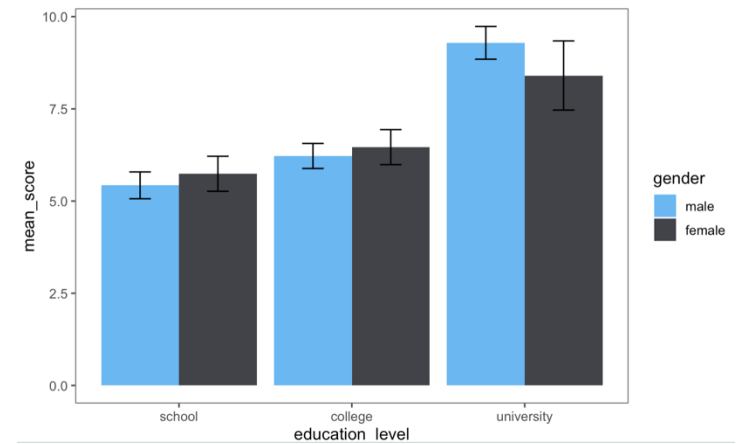


```
mean_scores = jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarise(mean_score = mean(score),
            sd_score = sd(score))

mean_scores %>%
  ggplot(aes(x = education_level, y = mean_score,
            group = gender, fill = gender))+
  geom_col(position = "dodge")+
  geom_errorbar(aes(ymin = mean_score-sd_score,
                  ymax = mean_score+sd_score),
              width = .25,
              position = position_dodge(width=0.9))+
  scale_fill_hc()+
  theme_few()
```

ggplot2::geom_errorbar()

- geom_errorbar allows you to add error bars to your lines or bar plots
- it requires:
 - **ymin/ymax**: where to start and end the bar (we can use mean \pm standard deviation)
 - **width**: how wide the error bar should be
 - **position**: where should the error bar be, need to play around with this usually
- try removing width or position and see what happens!

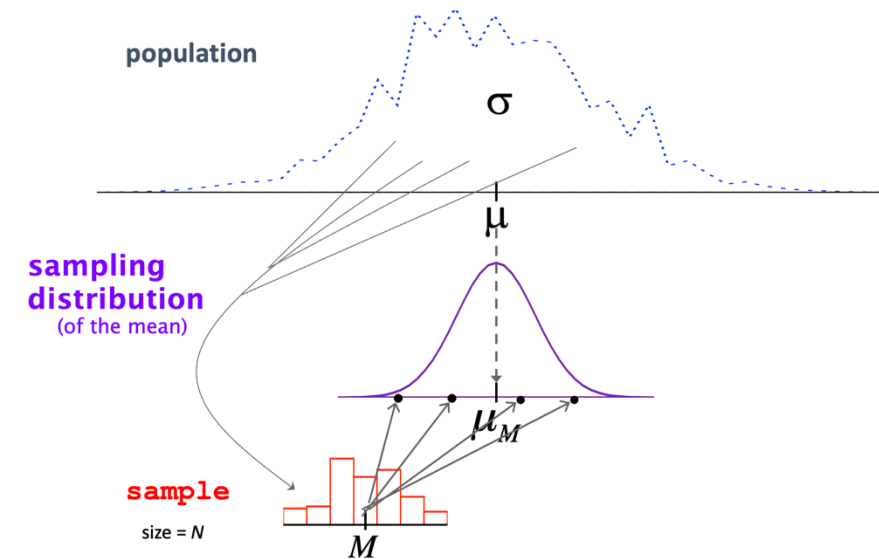


```
mean_scores = jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarise(mean_score = mean(score),
            sd_score = sd(score))

mean_scores %>%
  ggplot(aes(x = education_level, y = mean_score,
            group = gender, fill = gender))+
  geom_col(position = "dodge")+
  geom_errorbar(aes(ymin = mean_score-sd_score,
                  ymax = mean_score+sd_score),
               width = .25,
               position = position_dodge(width=0.9))+
  scale_fill_hc()+
  theme_few()
```

other forms of variation

- **standard deviation** is often used to describe the variation around the mean of a sample of data points
 - why not use variance?
- **standard errors**: an estimate of “accuracy” of the mean, i.e., how far the mean is on “average” from all other means
 - $SE = sd / \sqrt{n}$
 - higher n means lower SE, i.e., more confidence in your estimate
- **confidence intervals**
 - another way to assess the reliability of your sample: indicates how often the **true mean** is likely to be within a given interval, if repeated samples were drawn of the same size
 - $CI = \text{sample mean} \pm z * SE$
 - can also be “bootstrapped”, i.e., does not need to assume normality



best of both worlds: points + bar plot

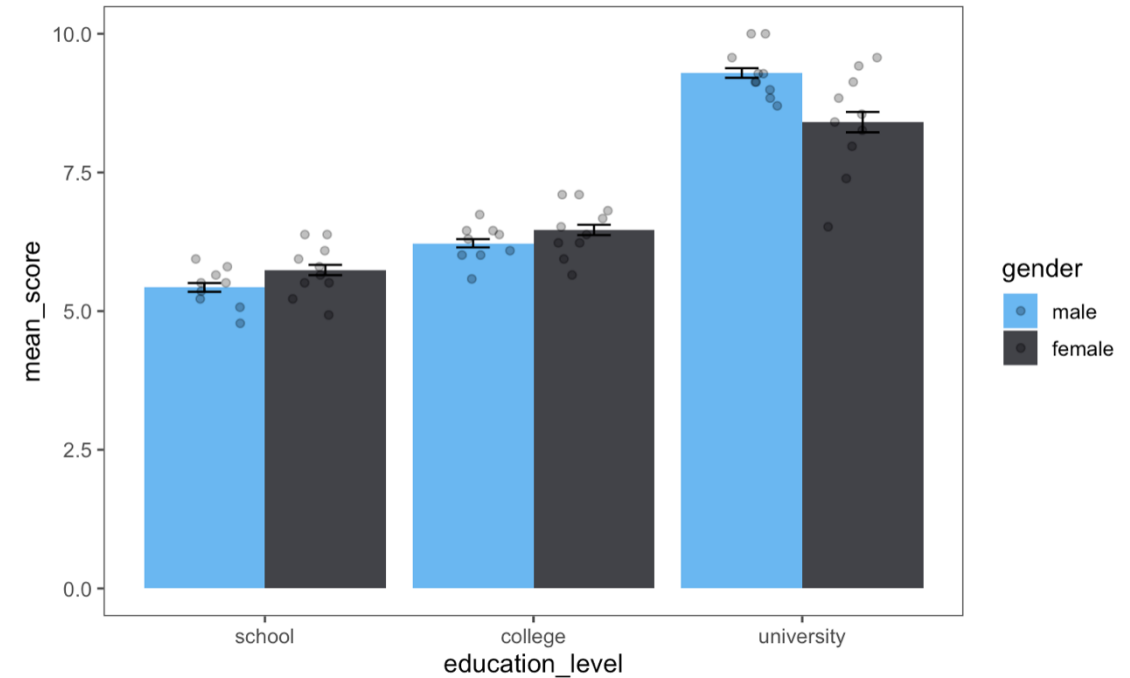
- sometimes, we can combine the power of **SE/confidence intervals** (accuracy/reliability) with **variation** using two elements (error bars and points)
- involves:
 - calculating **SE** (which requires the “n” in each condition)
 - calculating **confidence intervals** based on underlying distribution

```
counts = jobsatisfaction %>%  
  group_by(gender, education_level) %>%  
  count()
```

```
mean_scores = jobsatisfaction %>%  
  group_by(gender, education_level) %>%  
  summarise(mean_score = mean(score),  
            sd_score = sd(score)) %>%  
  left_join(counts) %>%  
  mutate(SE = sd_score/sqrt(n),  
         ymin = mean_score - 1.96*SE,  
         ymax = mean_score + 1.96*SE)
```

putting it all together...

```
mean_scores %>%  
  ggplot(aes(x = education_level, y = mean_score,  
             group = gender, fill = gender))+  
  geom_col(position = "dodge")+  
  geom_errorbar(aes(ymin = ymin, ymax = ymax),  
               width = .25,  
               position = position_dodge(width=0.9))+  
  geom_point(data = jobsatisfaction, aes(x = education_level, y = score,  
                                          group = gender),  
            position = position_jitterdodge(),  
            alpha = 0.3)+  
  
  scale_fill_hc()+  
  theme_few()
```



next time

- **before** class
 - *monitor*: data collection on Sona + Prolific
 - *work on*: project milestone #6b (analyses) and 7 (poster draft)
- **during** class
 - poster design