

# CogLab: Odds and Ends

WEEK 14

# what's coming up

14	Wednesday, November 29, 2023	<b>Project Milestone #7 (Analyses) Due</b>
14	Thursday, November 30, 2023	W14 continued...
14	Sunday, December 3, 2023	<b>Project Milestone #8 (Poster Draft) Due</b>
15	Tuesday, December 5, 2023	<u>W15: Wrapping Up</u>
15	Thursday, December 7, 2023	<b>Project Milestone #9 (Poster Symposium) Due</b>
16	Sunday, December 17, 2023	<b>Project Milestone #10 (Final Report) Due</b>

# November class survey

- last extra credit survey, due Dec 3
- link on Canvas
- BCQs will be completed on Tuesday (Dec 5) in class
- last class agenda:
  - final check-in / Qs about final milestones
  - survey responses / discussion
  - memer of the semester survey!!

# today's agenda

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



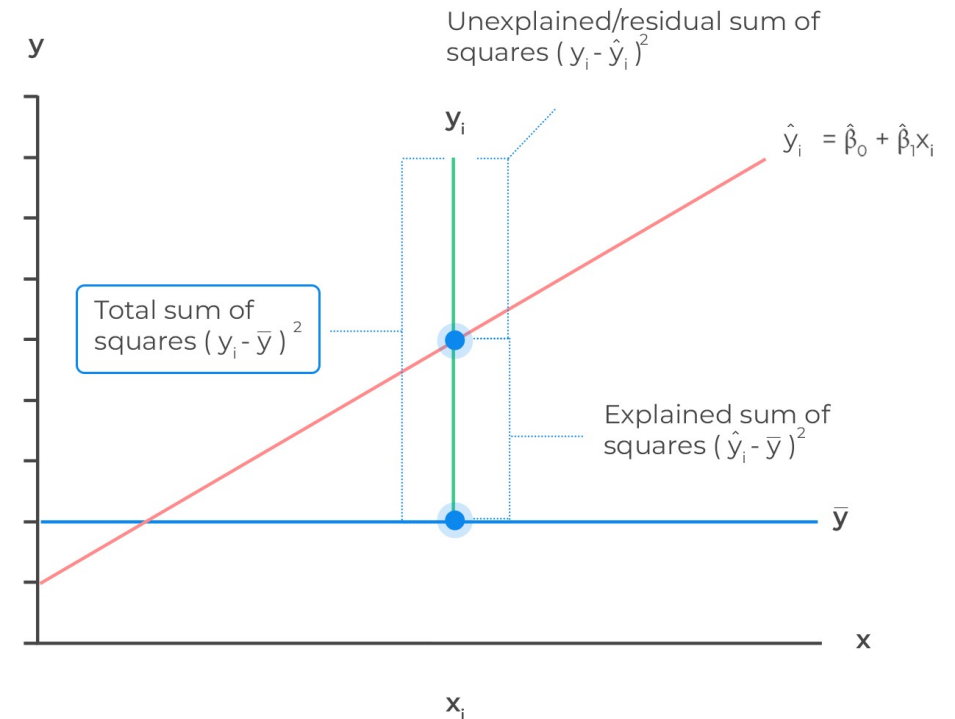
data collection + analyses check-in

# 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**
  - examples of central tendencies?
  - examples of variation?
- 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 ( $SS_{\text{between}}$ ) and noise ( $SS_{\text{within}}$ )
- regression?
  - fits the line  $y = \beta_0 + \beta_1 x$  that minimizes sums of squares
  - Total SS =  $SS_{\text{explained}} + SS_{\text{residual}}$



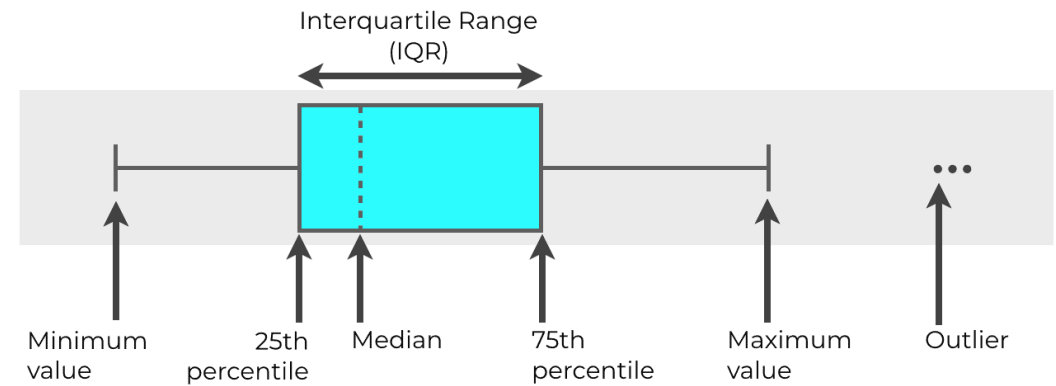
# so how do we visualize variation?

- 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 (25<sup>th</sup> percentile)
  - the median
  - the third quartile (75<sup>th</sup> percentile)
  - the maximum
- implicit measures:
  - IQR: 25<sup>th</sup> to 75<sup>th</sup> percentile
  - minimum:  $Q1 - 1.5 * IQR$
  - maximum:  $Q3 + 1.5 * IQR$
- what is an “outlier” based on a boxplot and where have we used this definition?



# bar plots

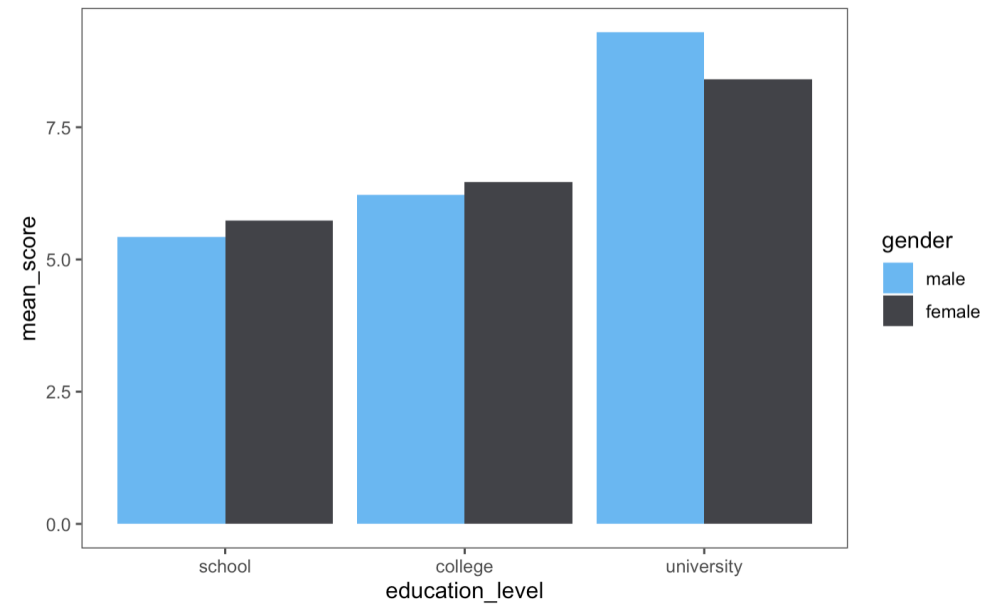
- bar plots often display the **means** for the relevant conditions in psychological studies
- but what about **variability**??
- a few different options:
  - **error bars** that denote some type of variation
    - what could this be?
  - an **overlay of original data points** in each condition

# open your RStudio project

- open the project and your .Rmd file
- **load** the packages by running the relevant chunk
- DON'T run all chunks – no need today!

# exercise 1: reproduce this plot!

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

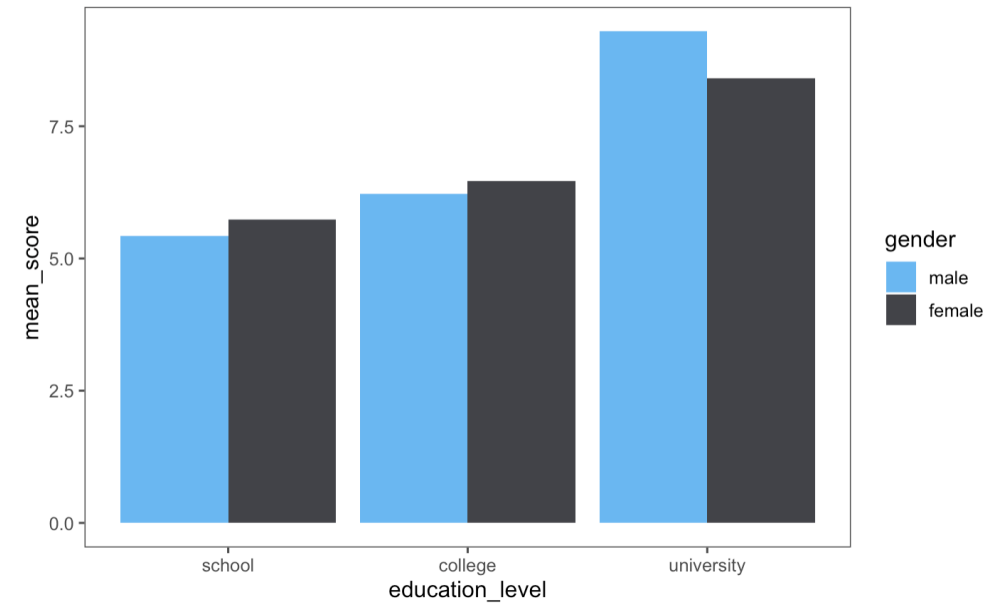


# exercise 1: reproduce this plot!

- 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()
```

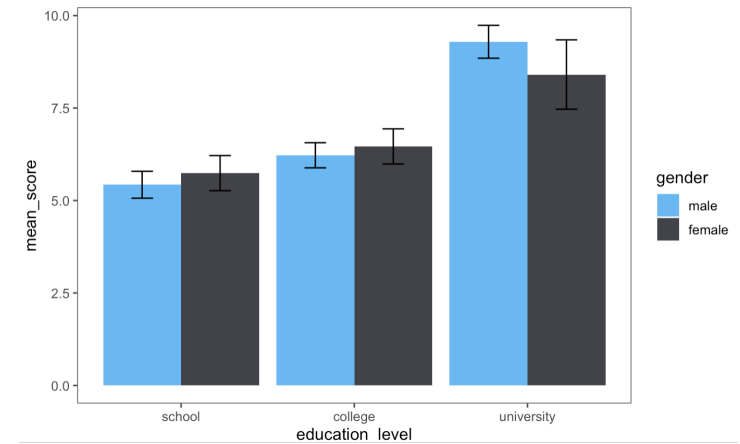


## exercise 2: adding variation

- 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

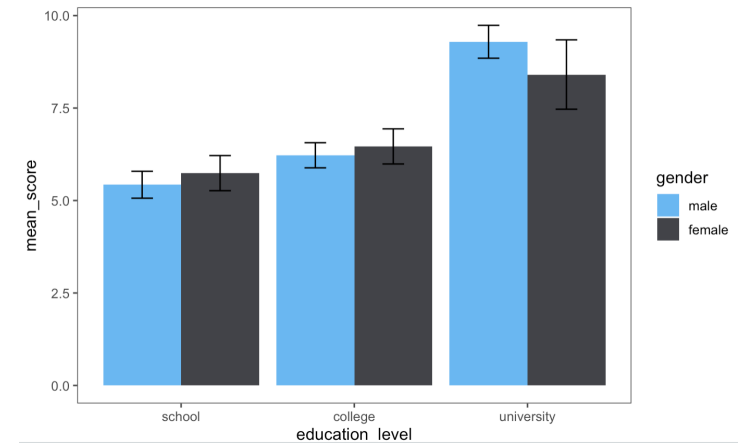


```
mean_scores = jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarise(mean_score = mean(score),
            sd_score = sd(score))

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  $\text{mean} \pm \text{standard deviation}$ )
  - `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!



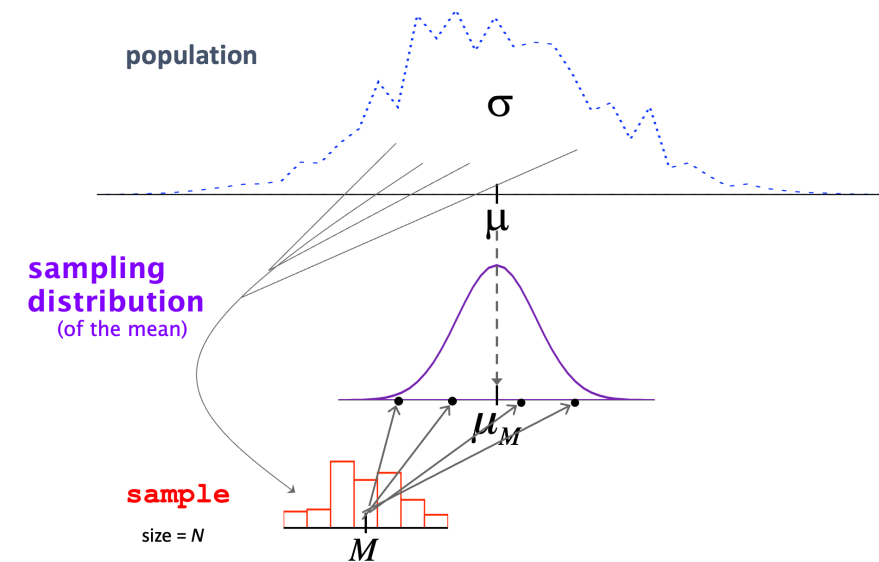
```
mean_scores = jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarise(mean_score = mean(score),
            sd_score = sd(score))

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 reliable is your mean, based on the sample size
  - $SE = sd / \sqrt{n}$
  - 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

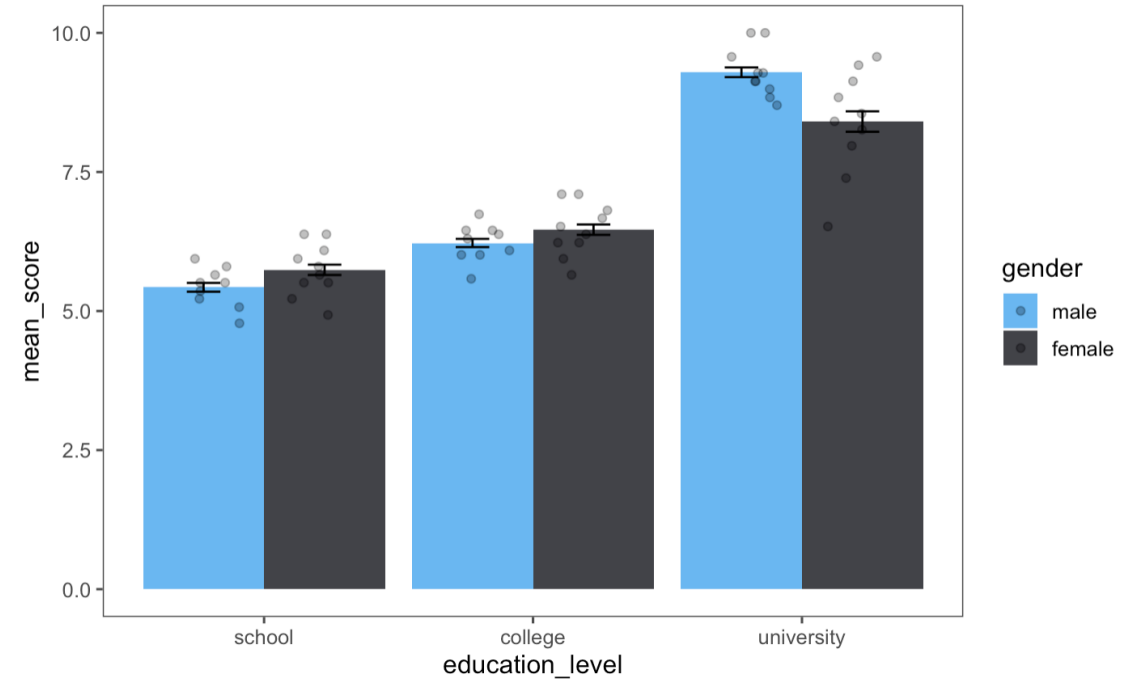
- sometimes, we can combine the power of **SE/confidence intervals** (accuracy/reliability) with **variation** using two elements (error bars and points)
- involves:
  - calculating **SE** (which requires the “n” in each condition)
  - calculating **confidence intervals** based on underlying distribution

```
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/n,  
         ymin = mean_score - 1.96*SE,  
         ymax = mean_score + 1.96*SE)
```

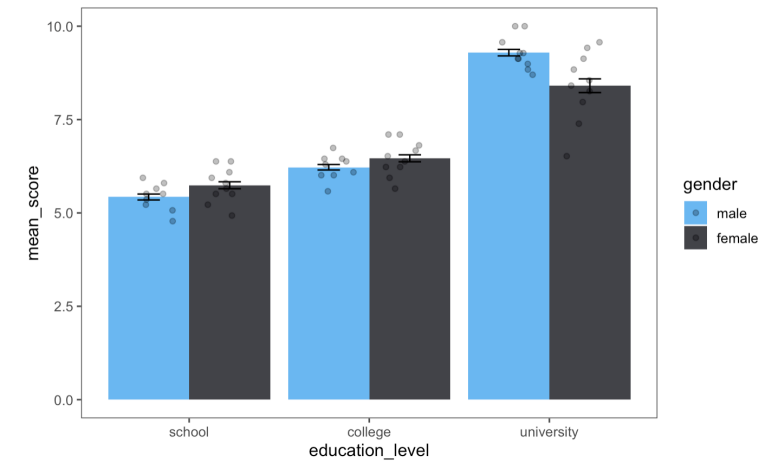
# putting it all together...

```
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()
```



# key takeaways

- variation is good, both in tests and visuals!
- use some form of variation in your plots when you make your posters
  - give the reader a window into your sample!
  - can be standard deviation, standard errors, confidence intervals, points, or ALL of this!
  - could also use boxplots + bar plots



# project milestone #8: poster draft

- general tips:
  - **de-clutter**, keep the text to a minimum
  - use **tables/figures** wherever possible (procedure, results, etc.)
  - use variation!
  - use **symmetry** and **colors** to guide the reader
  - think of what you will say and **organize** in a logical manner
- sample posters/resources up on course website

# poster contents

- introduction

- why is this topic important, what can we learn?
- background & current research question
- ideas: venn diagrams, smart art, etc.

- methods

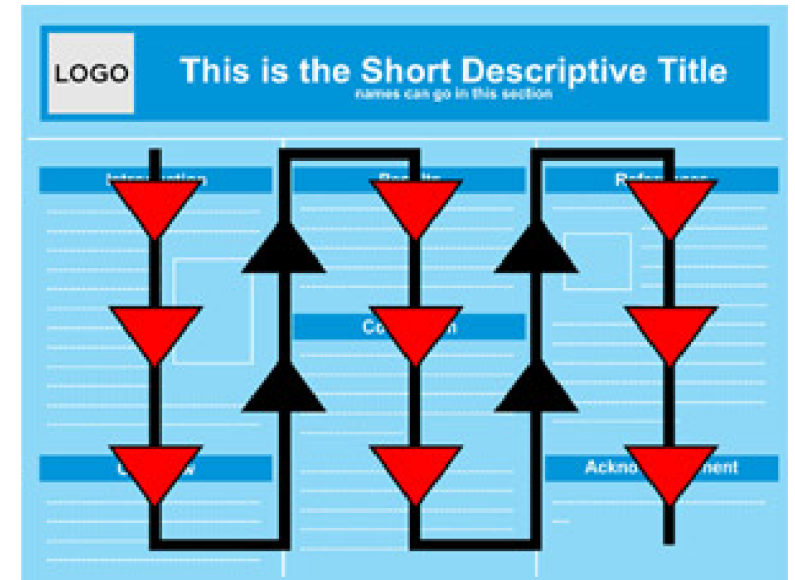
- IV/DV, items, counterbalancing etc.
- ideas: design figure, sample trial, etc.

- analysis

- statistical tests & results, inclusion/exclusion criteria, etc.
- ideas: tables, figures with p-values, regular figure

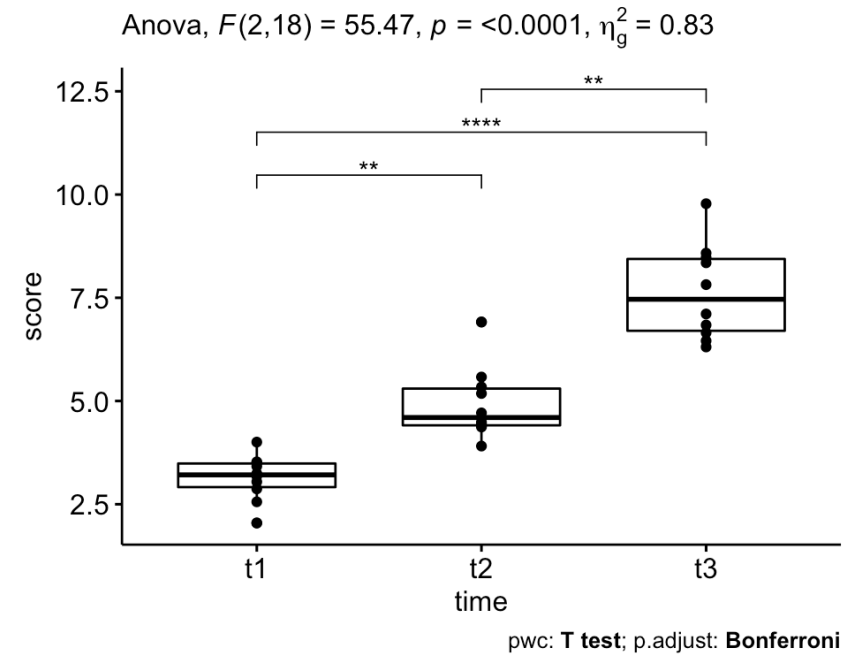
- conclusion / future steps

- small and picture: what did you learn? where do you see it going?



# statistics in posters

- **less text**, more images & numbers
- use the same format to report statistics but edit out all the text and **point to figures**



# next time

- **before** class
  - *monitor*: data collection on Sona + Prolific
  - *complete*: Week 13 quiz (inferences, due Dec 3)
  - *fill out*: class survey (due Dec 3)
  - *resubmit*: formative milestone #3 (due Dec 3)
- **during** class (Dec 5)
  - wrapping up!!