DATA ANALYSIS

Week 2: Fitting Models to Data (Central Tendencies & Errors/Variation)

recap

- what we covered:
  - summarizing data (frequency tables / ranks & percentiles)
  - visualizing data (distributions, histograms, bar graphs)
- your to-dos:
  - *prep*: video tutorial: [Summarizing data]
  - *apply*: problem set 1 (chapter 2 problems)
  - *prep*: read Chapter 3 from the Gravetter & Wallnau (2017) textbook.

today's agenda

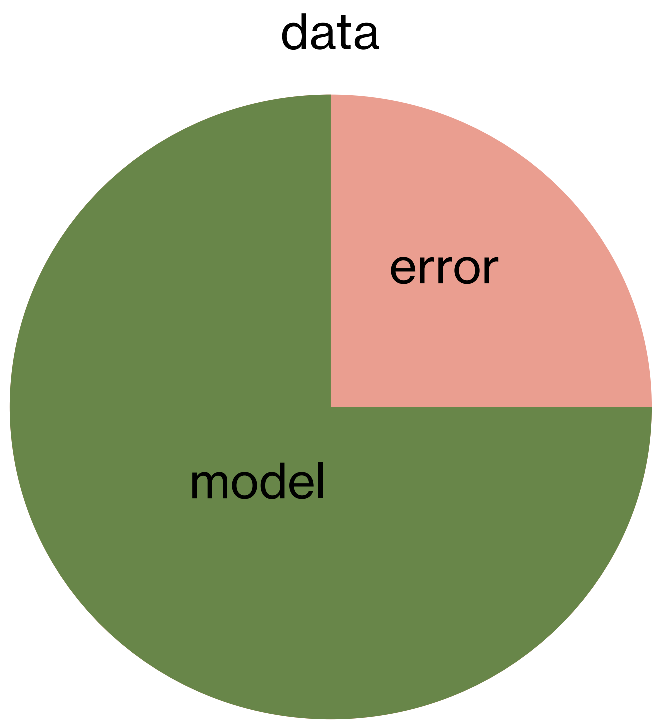
- what is a model?
- fitting models to data

what is a model?

- when you hear the word model, what do you understand?

data = model + error

- the goal of statistics is to find a simple explanation to the observed data, i.e., build a *model* of the data that approximates/explains it as well as possible
- what is a *good* model? one that represents the data really well
- best model?
- how do we start building models? we could start with a single estimate: one number that tells us something about our data



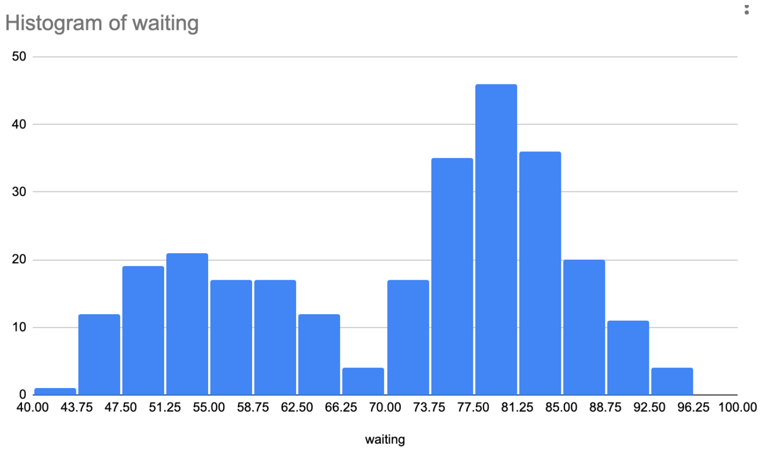
a dataset of geyser eruptions

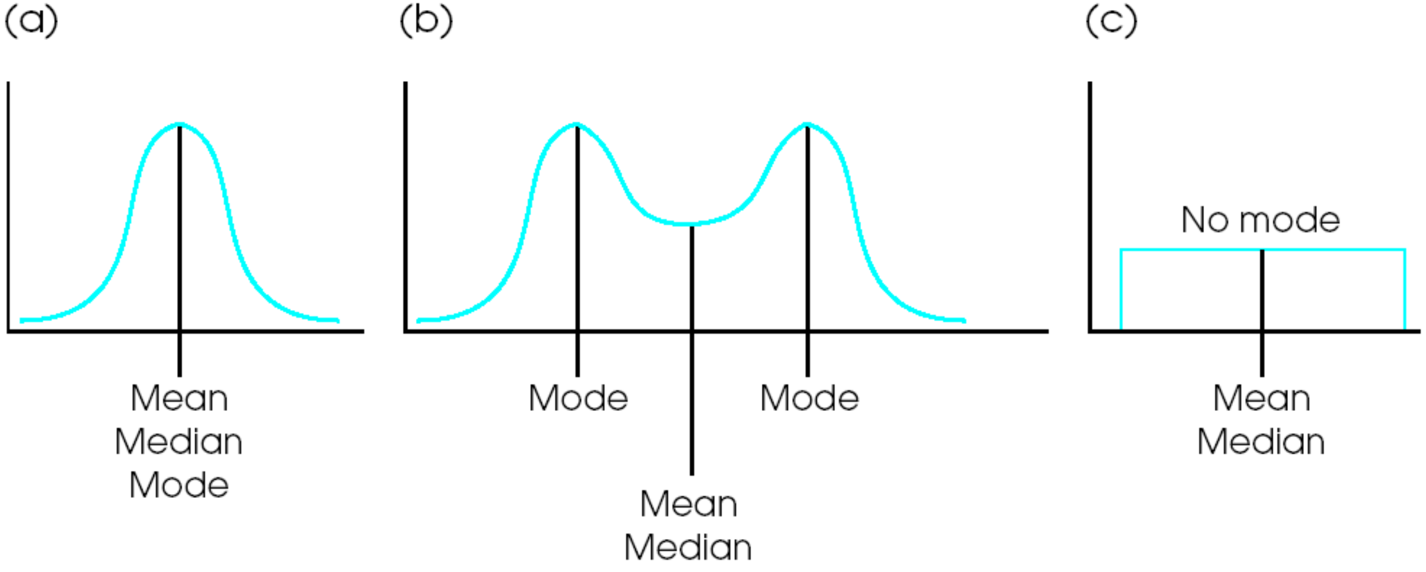
- Old Faithful geyser in Yellowstone National Park
- dataset records eruption time and waiting time in minutes [from R]
- how many rows and columns in this dataset?
- let's build a model of waiting time, i.e., how can I summarize the distribution of waiting times?

eruptions	waiting
3.6	79
1.8	54
3.333	74
2.283	62
4.533	85
2.883	55
4.7	88
3.6	85

model 1: mode

- the most “common” / frequent value in the dataset
- useful in describing the shape of a distribution





model 1: mode

- the most “common” / frequent value in the dataset
- how do we find it? by building a frequency table!
- sheets formula: =MODE(range)
- what is our statistical model?
- data = mode + error

waiting	COUNT of waiting
	0
43	1
45	3
46	5
47	4
48	3
49	5
50	5
51	6
52	5
53	7

fx =Mode(B2:B273)

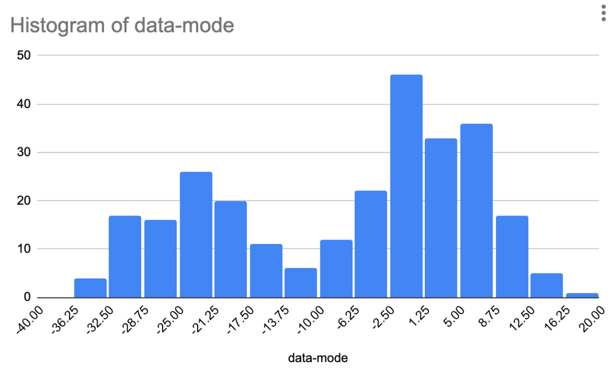
A	B	C
eruptions	waiting	Mode
3.6	79	78
1.8	54	

is the mode a good model?

- data = mode + error
    - error = data – mode
- each data point will produce its own error relative to the model
- how do we calculate the error?
    - subtract the mode from each data point
- distribution of errors?
- sum of errors = total error?
- “average” error? =AVERAGE(data range)

fx =B2-C2

A	B	C	D
eruptions	waiting	Mode	data-mode
3.6	79	78	1
1.8	54		-24



fx =AVERAGE(D2:D273)

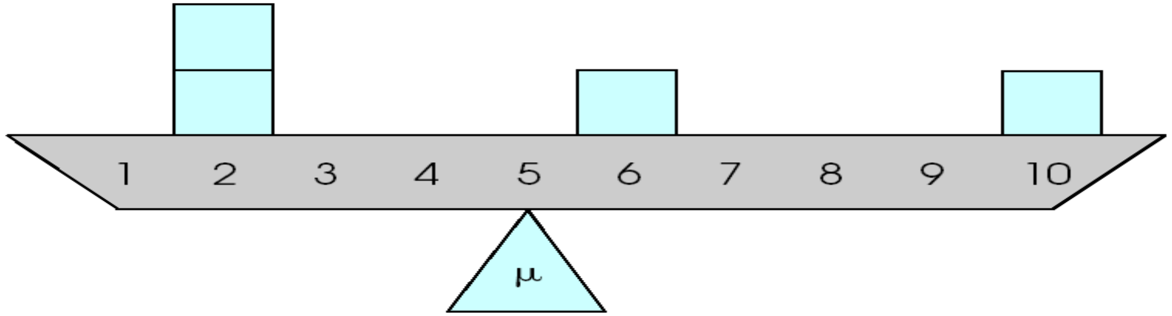
A	B	C	D	E
eruptions	waiting	Mode	data-mode	Avg_mode_error
3.6	79	78	1	-7.102941176
1.8	54		-24	

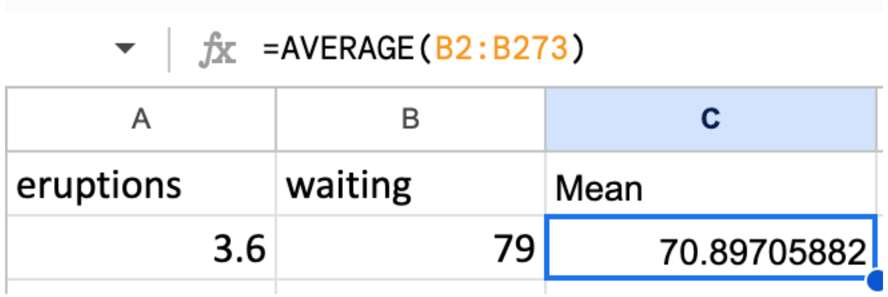
questions?

- groups of 3-4, review the “mode” sheet and see what questions are coming up!

model 2 = mean

- the arithmetic mean is the sum of scores divided by the number of scores: a balance point
- how do we find it?
  - add up all scores and divide by total number of observations
  - population mean: $\mu = \frac{\sum X}{N}$
  - sample mean: $\bar{X} = M = \frac{\sum X}{n}$
  - sheets formula: =AVERAGE(data range)




=AVERAGE(B2:B273)

A	B	C
eruptions	waiting	Mean
3.6	79	70.89705882

some properties of the mean

$$\mu = \frac{\sum X}{N}$$

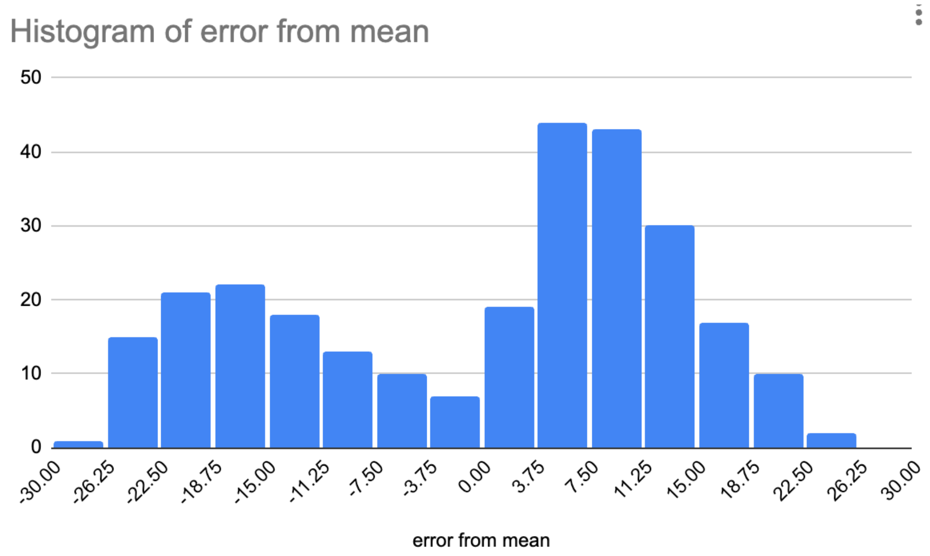
- the calculation of the mean includes **all** values, so changing a score will change the mean
- adding a new score or removing a score will ***usually*** change the mean
  - unless the new score is the mean itself
- adding/subtracting/multiplying/dividing a constant value from each score will lead to applying the same operation to the mean

is the mean a good model?

- data = mean + error
  - error = data – mean
- each data point (datum) will produce its own error relative to the model
- calculate the error?
  - subtract the mode from each data point
- histogram of errors?
- “average” error from the mean? =AVERAGE(range)
  - 0?!

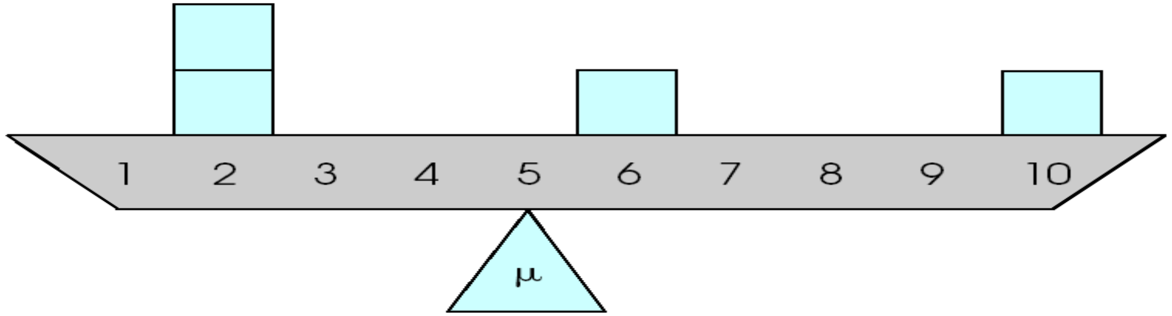
fx =B2-C2

A	B	C	D
eruptions	waiting	Mean	error from mean
3.6	79	70.89705882	8.102941176
1.8	54		-16.89705882



why is the error zero?!

- the mean is a balance point in the sense that "errors"/"distances" above the mean must have the same total as the errors below the mean
- the mean, by definition, is the middle point where errors above and below the mean cancel each other out!



why is the error zero?! a proof

- average error = average (data – mean)

$$\frac{\sum_{i=1}^{n}(X_i \; - M)}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} M}{n}$$

$$= M - \frac{\sum_{i=1}^{n} M}{n}$$

$$= \; M - \frac{n\,M}{n}$$

$$= \; M - \; M$$

$$= 0$$

re-calculating errors

- positive and negative errors canceling out is problematic: it de-emphasizes/washes out the differences between data points and the model and suggests that the mean produces no error!
- we could take the absolute value of errors? square the errors?
- turns out, squaring has several mathematical advantages over taking the absolute values

65.65765571

=D2*D2

B	C	D	E	F
waiting	Mean	error from mean	Average_mean_	squared_errors
79	70.89705882	8.102941176	0	=D2*D2
54		-16.89705882		285.5105969
74		3.102941176		9.628243945
62		-8.897058824		79.15765571

re-calculating errors

- after squaring, how do we get a single estimate of the error?
  - sum of squared errors (SSE or SS)
    - depends on the number of observations
  - mean of squared errors (MSE)
    - not in original units of the data
  - root mean squared error (RMSE)
    - error is in same units as the original data!

65.65765571

=D2*D2

B	C	D	E	F
waiting	Mean	error from mean	Average_mean_	squared_errors
79	70.89705882	8.102941176	0	=D2*D2
54		-16.89705882		285.5105969
74		3.102941176		9.628243945
62		-8.897058824		79.15765571

re-calculating errors for the mean

- sum of squared errors (SSE or SS): $\sum_{i=1}^{N}(X_i - \mu)^2$
- mean of squared errors (MSE): $\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{SS}{N}$
- root mean squared error (RMSE): $\sqrt[2]{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = \sqrt{MSE}$

questions?

- groups of 3-4, review the “squared_errors_for_mean” sheet and see what questions are coming up!

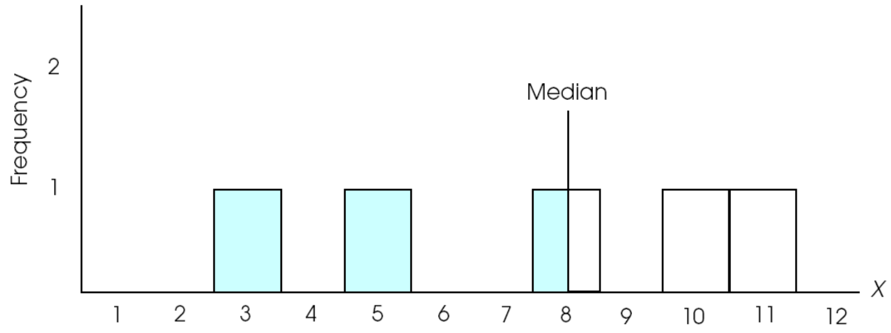
re-calculate errors for the mode

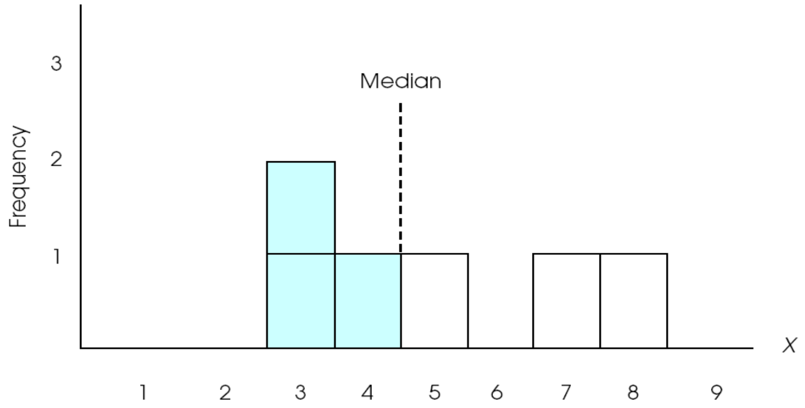
- previously, we calculated the average “error” for the mode without squaring
- we could now calculate SS, MSE, and RMSE for the mode
- between the mode and the mean, which is the better model?
- which model has the lowest RMSE?
- can we do better??

model	RMSE
mode	15.32
mean	13.57

model 3: median

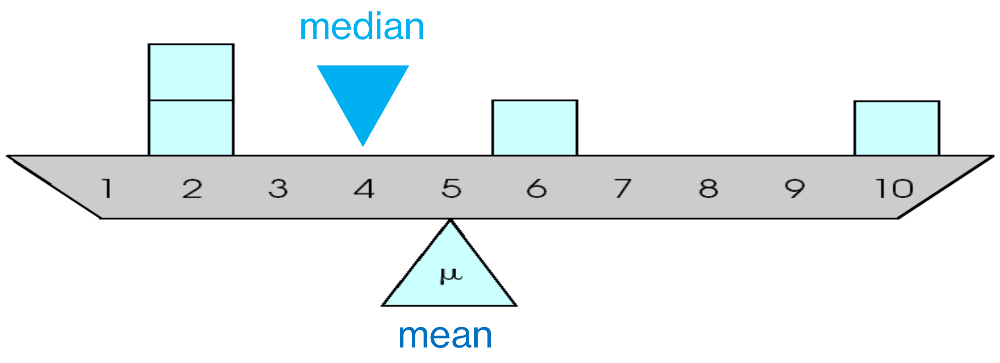
- divides the distribution exactly in half: value of the median is equivalent to the 50th percentile
- how do we find it?
    - put the scores in order (ascending or descending) and find the middle value
- if N is odd, the median is the middle score (when the scores are in order)
- if N is even, the median is the mean of the middle two scores (when the scores are in order)





mean vs. median

- both are balance points, but in different ways
- the mean is trying to find the point that balances the errors/distances above and below it, but it may not always be at the "center" of the scores; easily swayed by extremes
- the median is not worried about the errors and is literally trying to find the center in terms of the scores



is the median a good model?

- we could now calculate SS, MSE, and RMSE for the median
- between the median, mode and the mean, which is the better model?
- which model has the lowest RMSE?

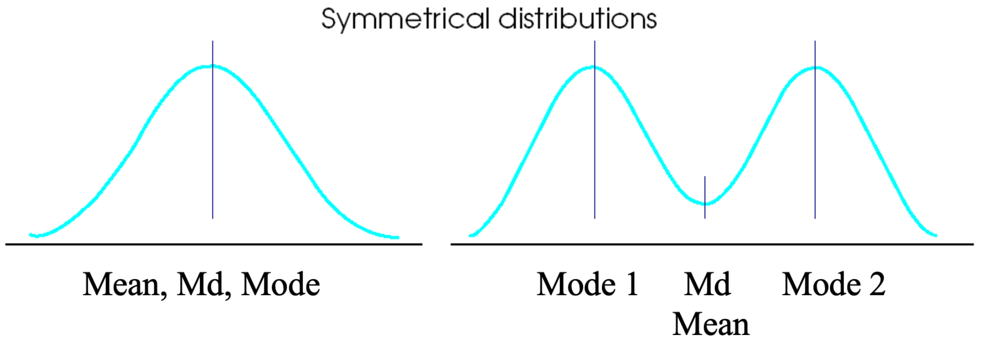
model	RMSE
mode	15.32
mean	13.57
median	14.50

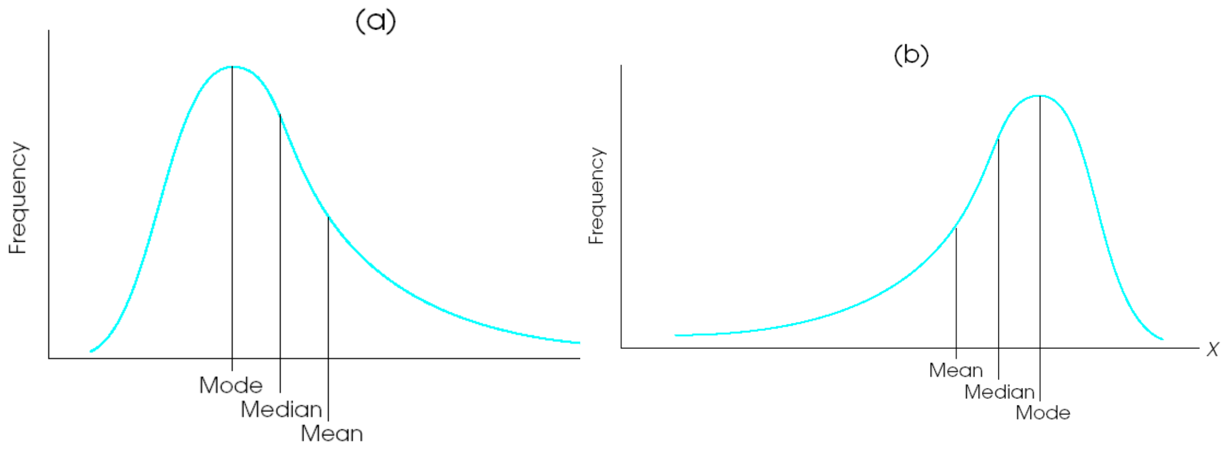
when to use which measure?

- mean, median, mode are together called measures of **central tendency**
- mean
  - most common, includes all scores, generally our “best” bet if we have nothing else available
- median
  - small number of extreme scores
  - undetermined values / open-ended distribution
- mode
  - nominal scale, only the mode can be used
  - if the “most typical case” is to be identified. mean and median often produce fractional values

mean vs. median vs. mode

- symmetric distributions
  - single mode: mean = median = mode
  - multiple modes: mean = median
- skewed distributions
  - positive skew: mode < median < mean
  - negative skew: mean < median < mode





variability

- describing data via a measure of central tendency tells only <u>half the story</u>
- we also want to know the spread of the data and how well our “model” fits this spread
- we already did this by estimating the errors!
  - variance = mean of squared errors (MSE)!
  - standard deviation = (square) root mean squared error (RMSE)!
- more next time!

next time

- **before** class
  - *try*: week 2 quiz
  - *watch:* Central Tendencies video
  - *submit*: problem set #1 (follow video tutorial for submission guidelines)
  - *apply*: optional meme / discussion post
  - *prep*: read Chapter 4
- **during** class
  - understanding variability better