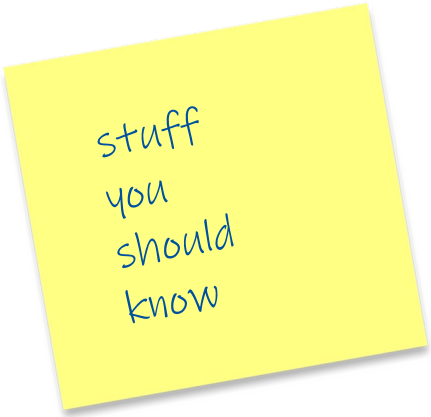
ENV 710: Lecture 1

descriptive statistics

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

learning goals

- what are different types of data and examples of each?
- key terms: population, sample, parameter, etc.
- what are measures of location and spread, and how are they calculated? pros and cons of each?
- how is the shape of a data distribution described?
- how are outliers defined, and how to deal with them?



research steps

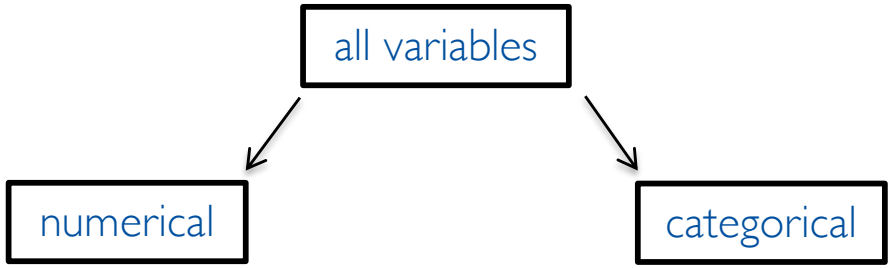
- determine your question
- design the study
- collect the data
- describe the data
- infer from the sample to the population

summary statistics & figures

Attention! we are starting in the middle of the research process

types of data

a **variable** is a characteristic or measurement that differs from individual to individual

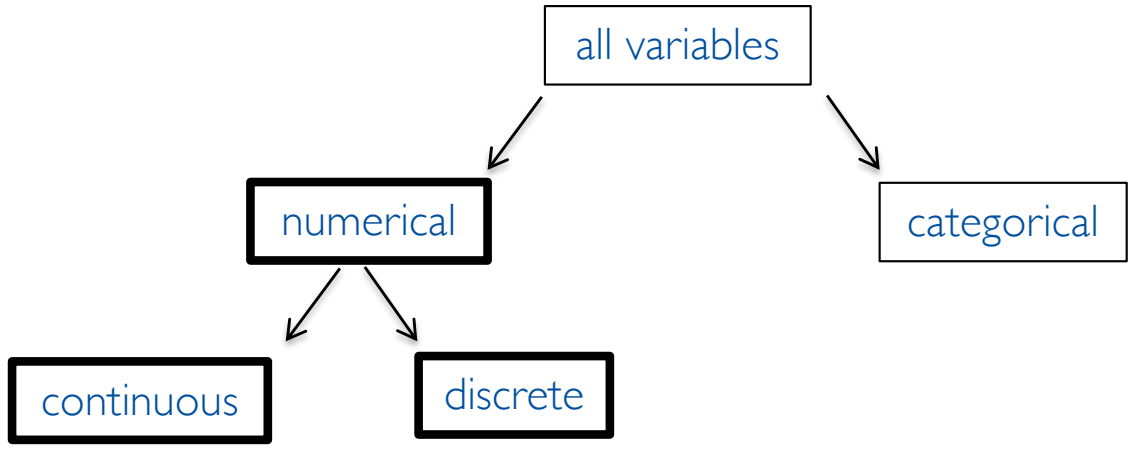


numerical values that can be added, subtracted, etc.

take on limited number of distinct categories: can be numbers, but not sensible to do arithmetic operations

data are measurements of variables

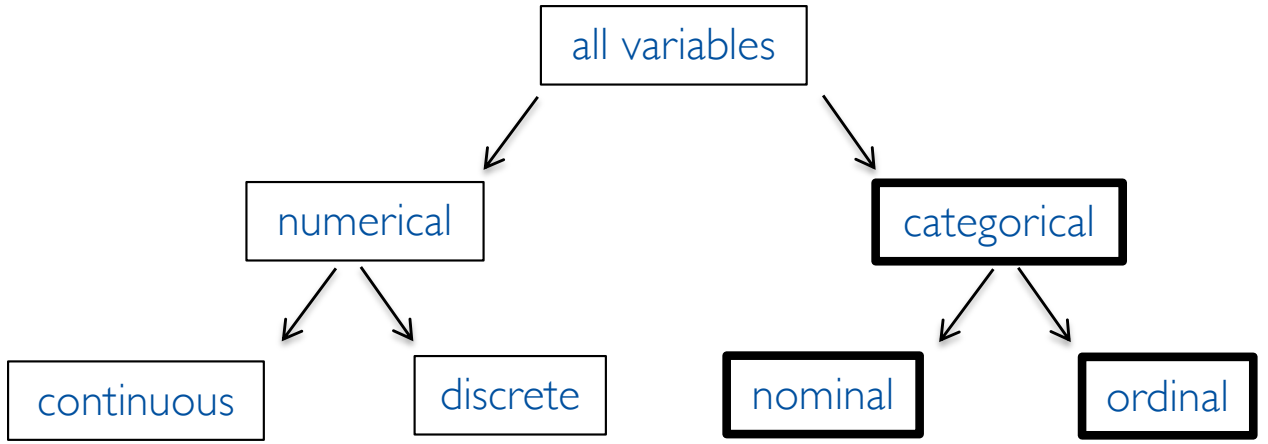
types of data



any of an infinite number of values within a given range

one of a specific set of numeric values

types of data



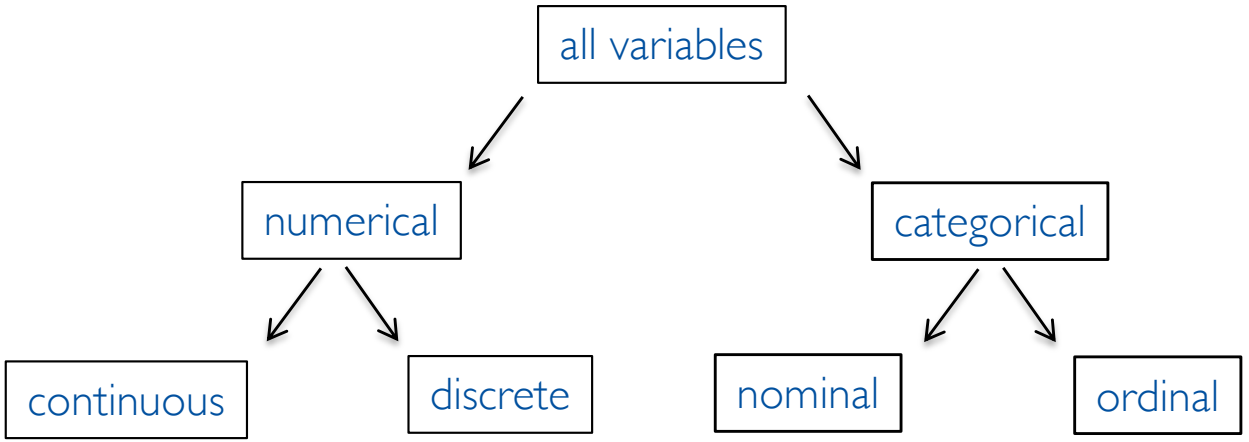
any of an infinite number of values within a given range

one of a specific set of numeric values

descriptive, discrete

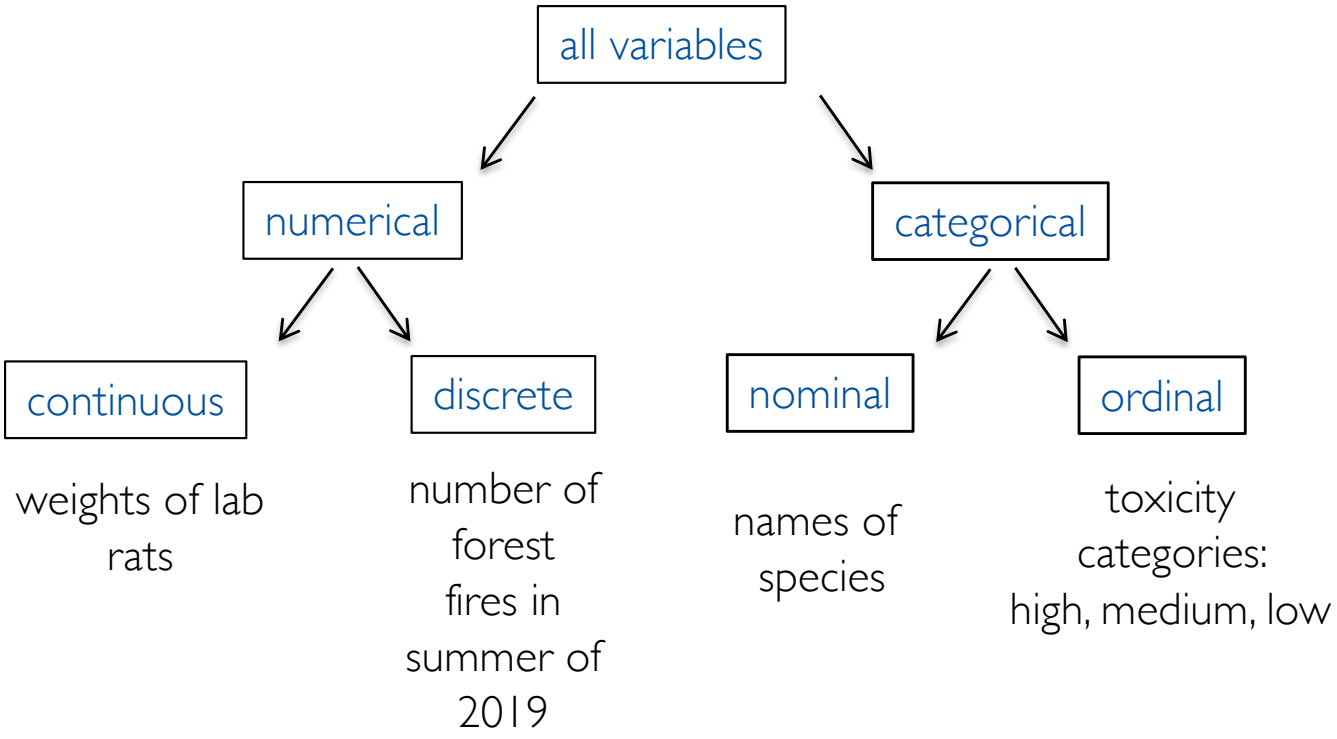
levels have an inherent ordering; distance between values cannot be determined

types of data

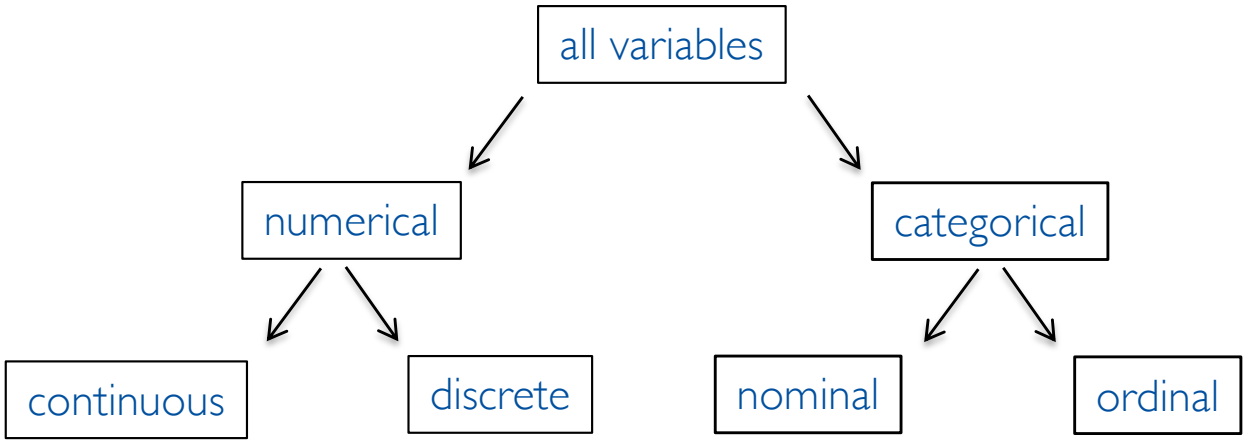


think of examples of each data type...

types of data

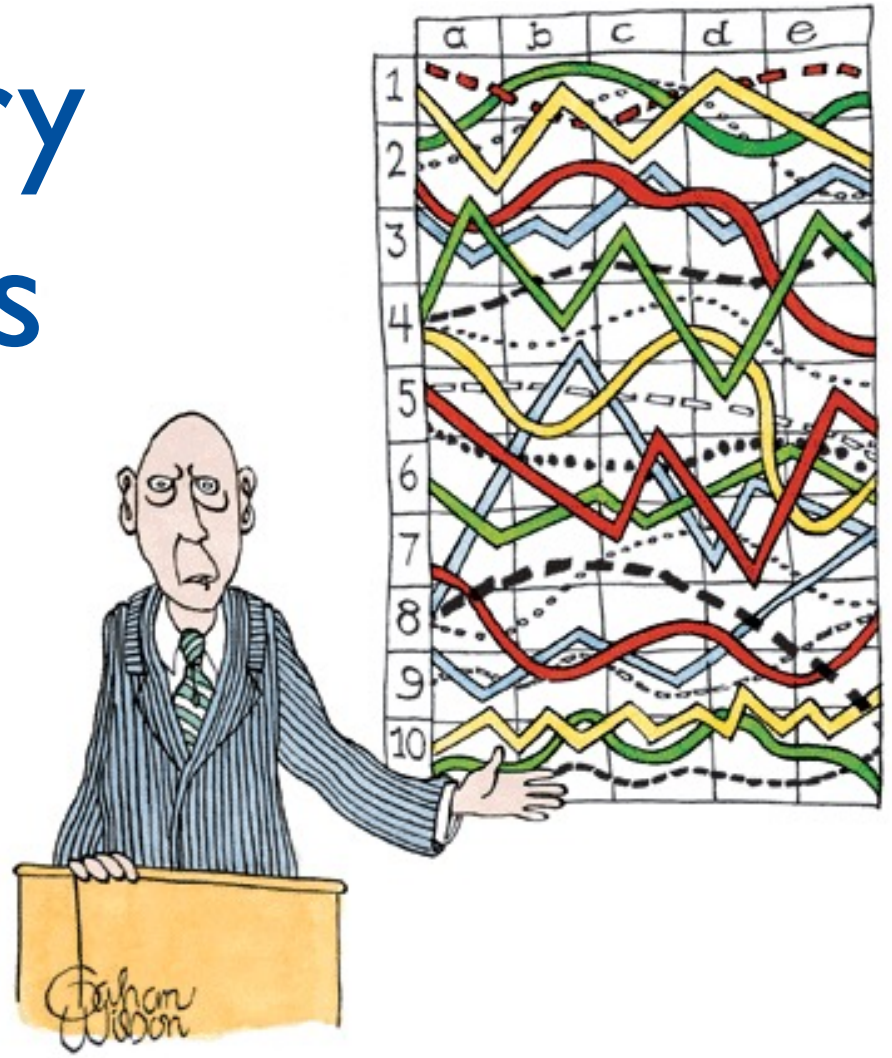


types of data



probability distributions of different types of data are different,
therefore we model them in different ways

summary statistics



"I'll pause for a moment so you can let this information sink in."

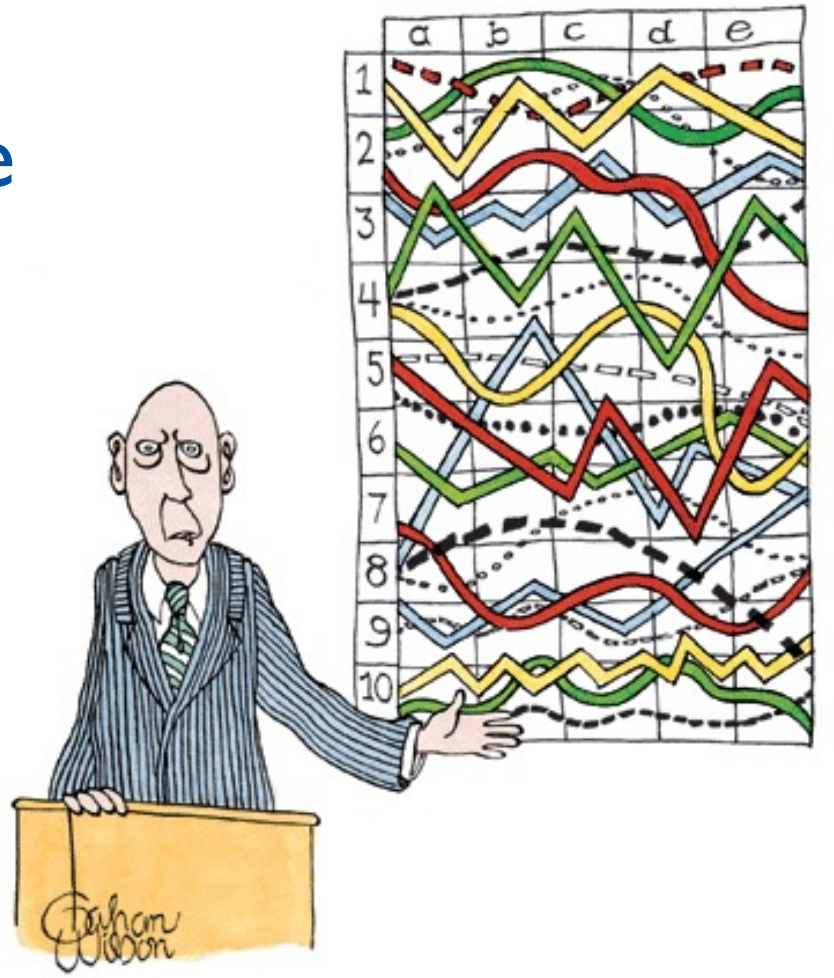
explore and summarize

summarize your data

- summary statistics (e.g., mean, standard deviation, etc.)
- 5-point summary

graph your data

- boxplots, histograms, etc.
- graph, graph, graph

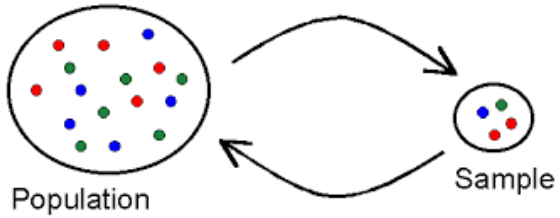


"I'll pause for a moment so you can let this information sink in."

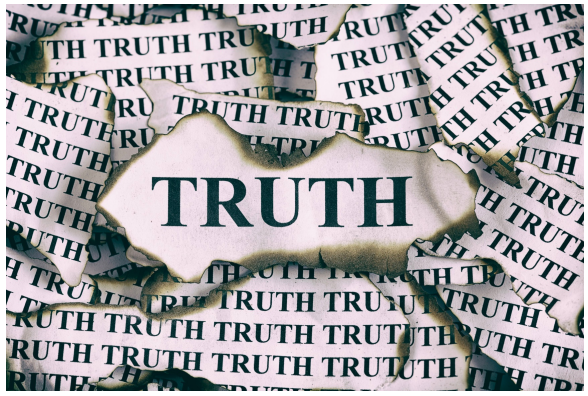
summary stats

- population is the total set of observations
- sample is a portion of a population

- parameter is any numerical quantity that characterizes a given population or some aspect of it (truth)
- statistics are estimates of population-level parameters (approximation)



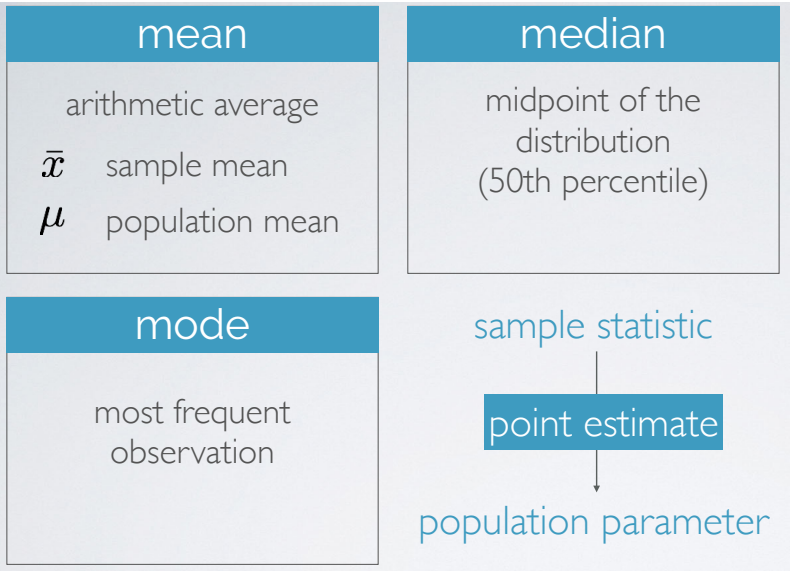
fundamental assumption: there is a true value for each parameter



https://www.jcu.edu.au/__data/assets/pdf_file/0008/115478/Basic-Statistics-6_Sample-vs-Population-Distributions.pdf
https://nednote.com/art-tells-the-truth/

summary stats

- **measures of location**: where most of the data are located



when to use them?

mean: means of large samples of random variables conform to a normal distribution

median/mode: better when distributions of observations cannot be fit by a standard probability distribution, and when there are extreme observations

- arithmetic, geometric, and harmonic means are sensitive to extreme observations

other measures of location

trimmed mean: reduces effects of outliers

- trim a % of the observations and calculate mean

geometric mean: describes multiplicative processes (growth rates)

- normalizes the range being averaged so a given percentage has the same effect
- use when numbers are multiples of each other

harmonic mean: average of rates

```
Yi <- c(10,10,10,10,1000)
```

$$GM_Y = e^{\left[\frac{1}{n}\sum_{i=1}^{n}\ln(Yi)\right]}$$

```
exp(mean(log(Yi)))
```

$$H_Y = \frac{1}{\frac{1}{n}\sum\frac{1}{Y_i}}$$

```
1/mean(1/Yi))
```

```
mean(c(Yi))
```

10		10
10		10
10		10
10		10
1000		0.1
25.1	geometric mean	4.0
12.5	harmonic mean	0.5
208	mean	8.0

income gap in the US

average compensation in the US climbed from $35,977 (adjusted for inflation) in 1984 to $50,000 in 2018

what's the problem?

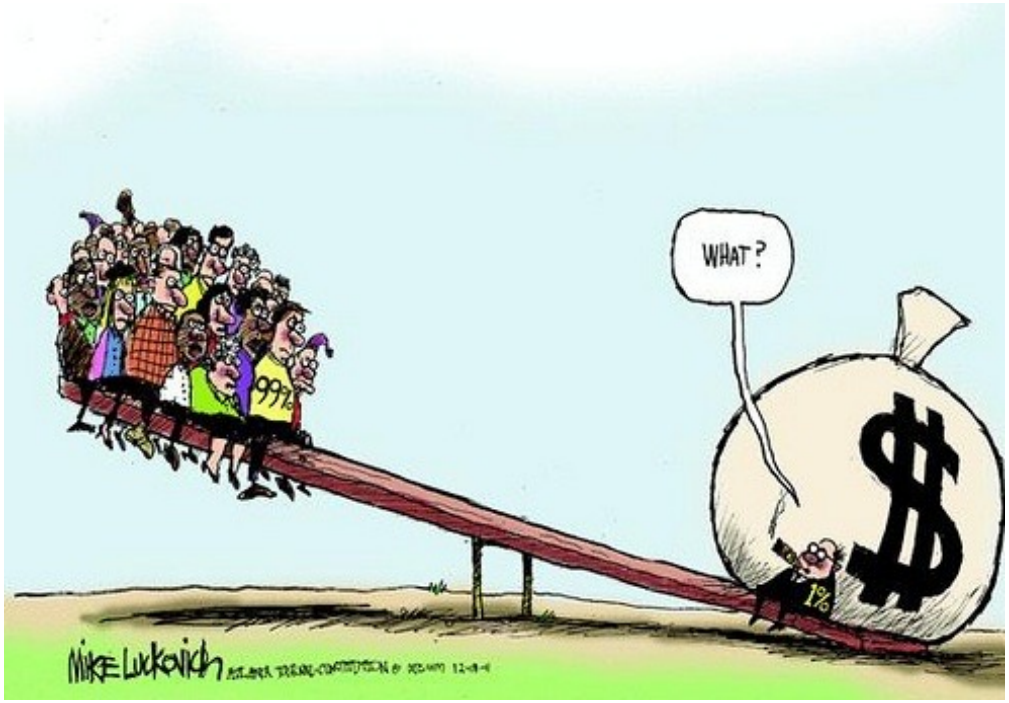
Bernie Sanders, Presidential Candidate 2020

NATI HARNIK/ASSOCIATED PRESS/ASSOCIATED PRESS

income gap in the US

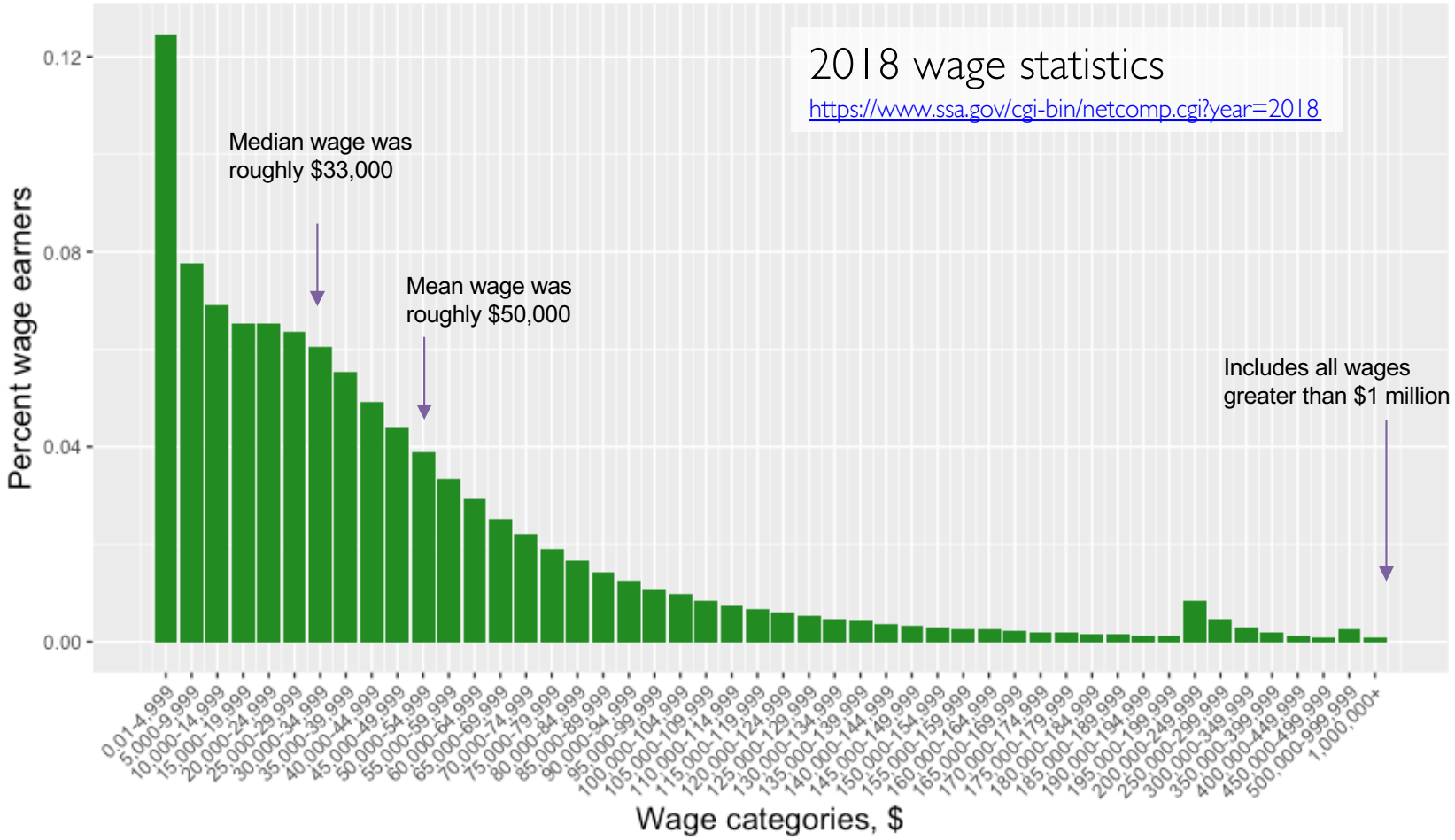
average compensation in the US climbed from $35,977 (adjusted for inflation) in 1984 to $50,000 in 2018

what's the problem?



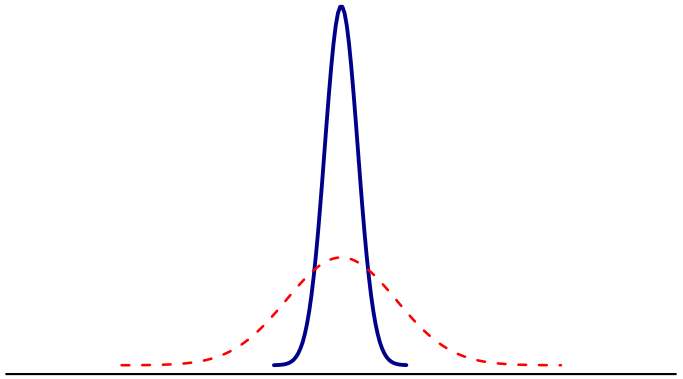
https://sevenpillarsinstitute.org/consequences-economic-inequality/

Distribution of wages in 2018 in the US

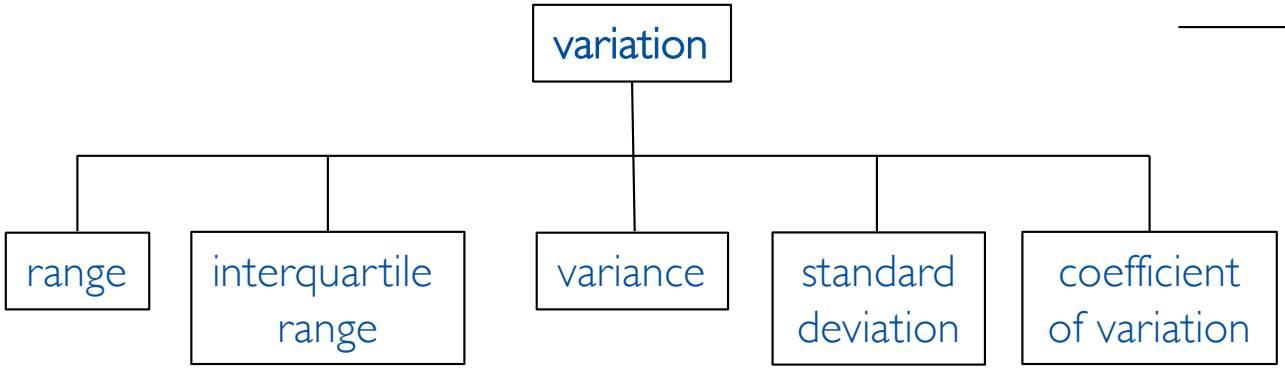


summary stats

- **measures of spread**: spread or variability of the data



same center,
different spread



variance

`var()`
gives the sample variance

average squared deviation from the mean

population variance σ^2

sample variance s^2

n	country	life exp.
1	Mozambique	31.3
2	Botswana	32.3
3	Zambia	35.3
…	…	…
180	Andorra	83.5

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$s^2 = \frac{(31.3 - 66.0)^2 + (32.3 - 66.0)^2 + ... + (83.5 - 66.0)^2}{180 - 1} = 162.8 \text{ yrs}$$

variance

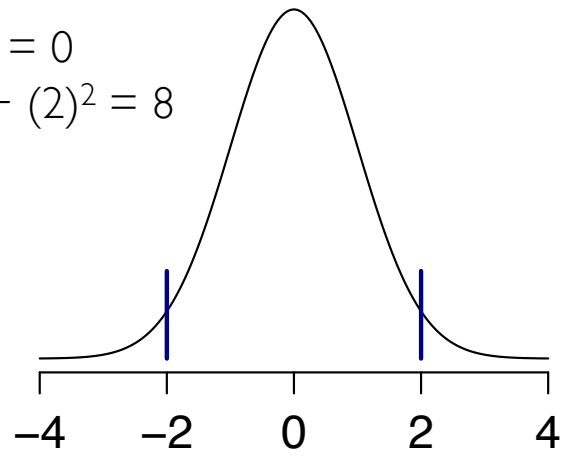
why do we square the differences?

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

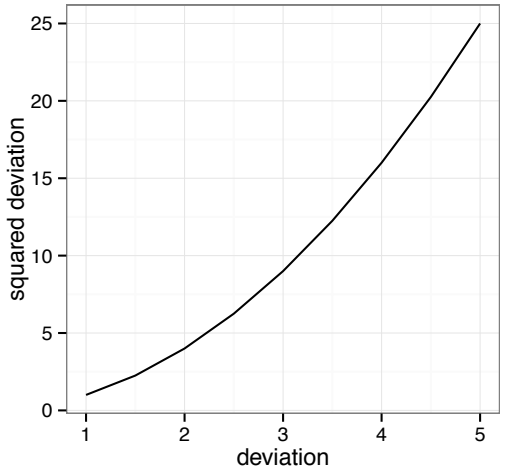
- get rid of negatives so that negatives and positive don't cancel each other out

-2 + 2 = 0

$(-2)^2 + (2)^2 = 8$



- increase larger deviations more than smaller ones so they are weighed more heavily



standard deviation

average deviation around the mean,
expressed in the same units as the data

population standard deviation σ

sample standard deviation s

$$s = \sqrt{s^2}$$

$$s = \sqrt{162.8} = 12.8 \text{ yrs}$$

`sd()`
gives the sample variance

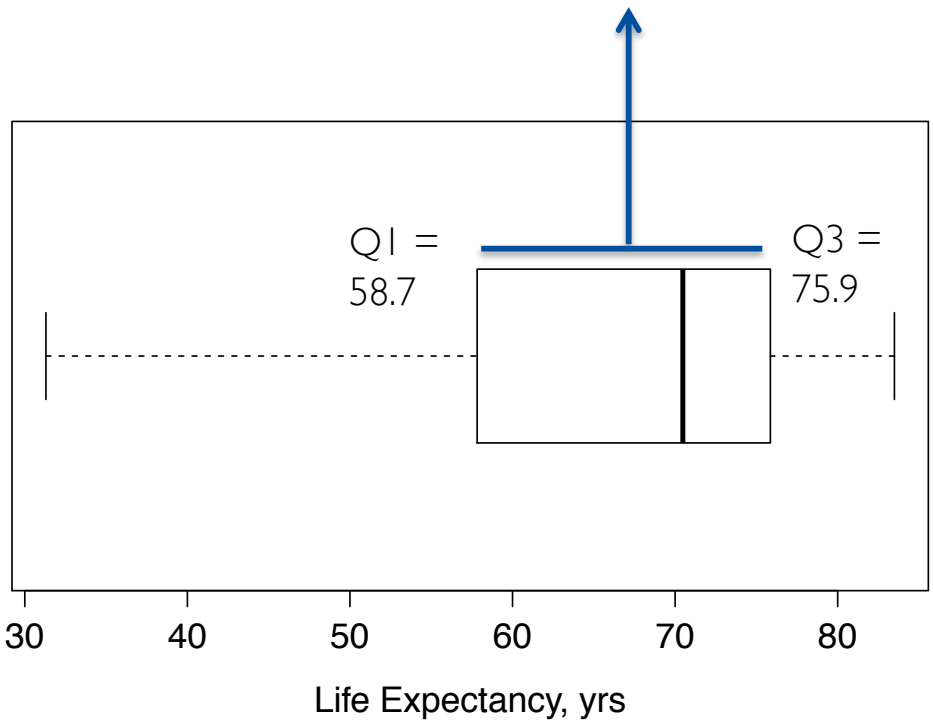
	country	life exp
1	Mozambique	31.3
2	Botswana	32.3
3	Zambia	35.3
…	…	…
180	Andorra	83.5

interquartile range

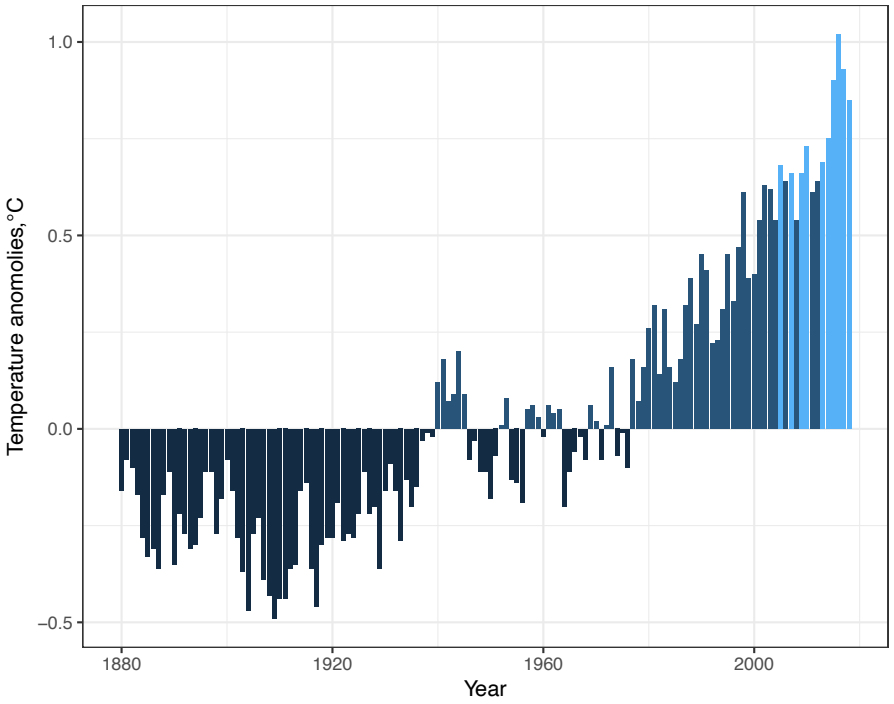
range of middle 50% of the data, distance between the 1st quartile (25th percentile) and 3rd quartile (75% percentile)

IQR = Q3 – Q1

IQR = 75.9 - 58.7.3 = 17.2



summary



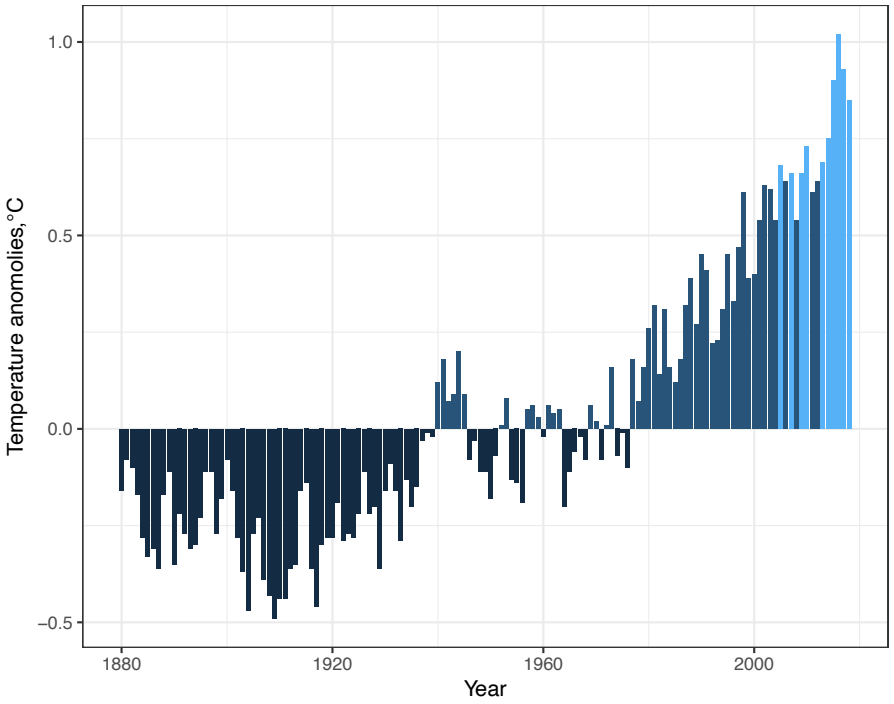
deviations from the 1951-1980 mean surface temperatures

5-point summary
minimum = -0.49
1st quartile = -0.21
median = -0.07
3rd quartile = 0.21
maximum = 1.02

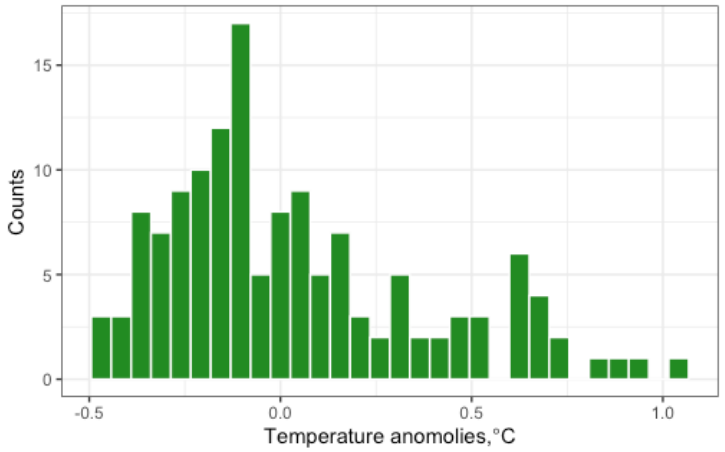
IQR = 0.21—0.21 = 0.42

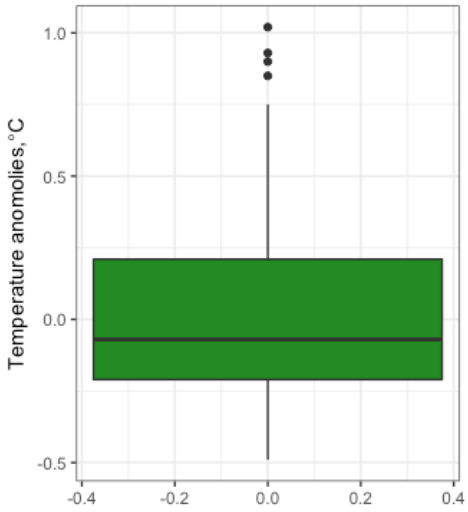
`summary()`
provides the 5-point summary (plus the mean)

summary



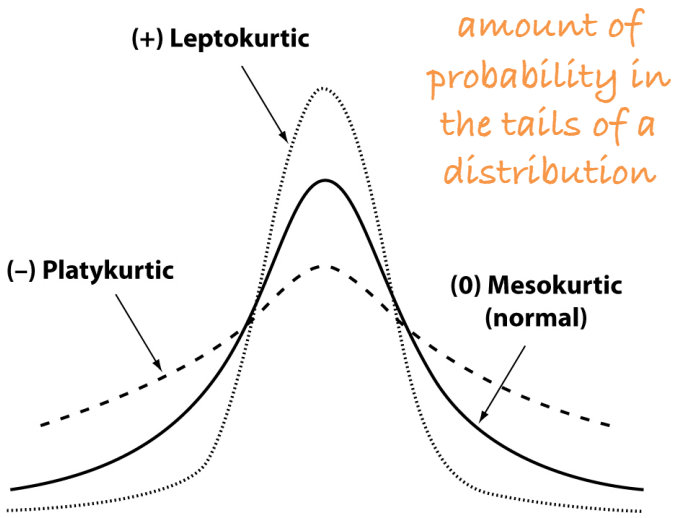
deviations from the 1951-1980 mean surface temperatures





shape of distributions

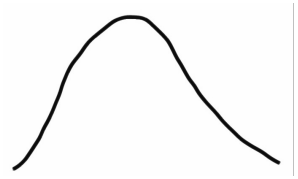
kurtosis – extent to which a distribution is distributed in the tails versus the center



forms of kurtosis

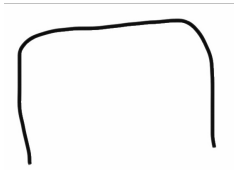
modality – describes the peak of the distribution

bimodal



unimodal

multimodal

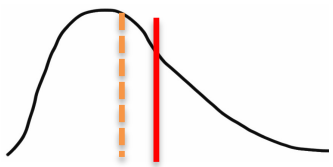


uniform

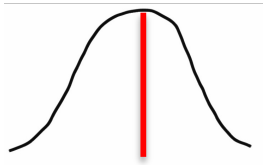
shape of distributions

skewness – how the sample differs in shape from a symmetrical distribution; measure of symmetry

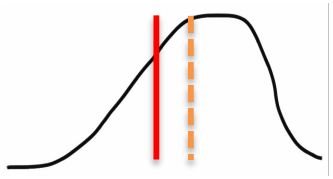
measure of symmetry



right skewed
positive skew
mean > median



symmetric
mean ≈ median



left skewed
negative skew
mean < median

outliers

"an observation in a data set which appears to be inconsistent with the remainder of the set of data." (Johnson, 1992)

"…an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." (Hawkins, 1980)

outliers

- measurement error
- data entry error
- may occur by chance
- observation generated by a different distribution, mechanism, or process

rule of thumb: data point falls outside the lower and upper fences:

- 3rd quartile + 1.5 x IQR
- 1st quartile – 1.5 x IQR

this may not be the best way!

http://clipart-library.com/thumbs-up-images.html

what do we do?

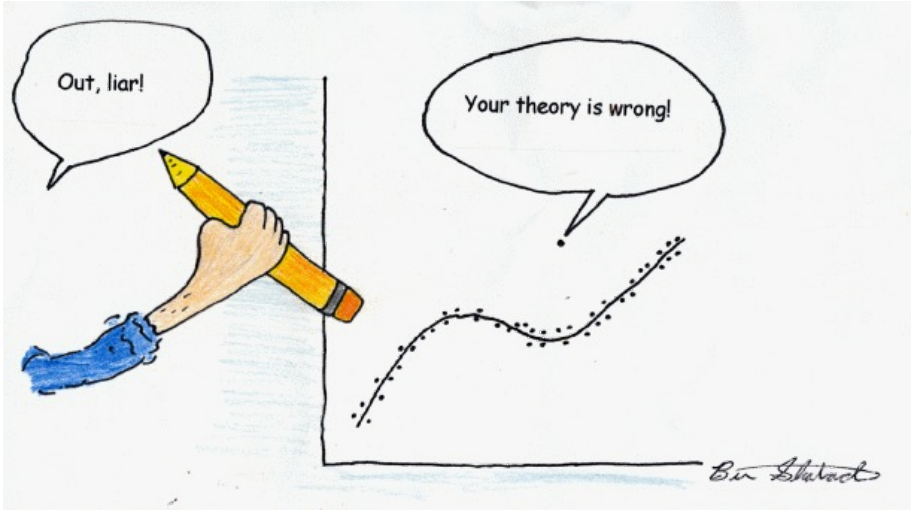
think about larger context of data and data collection

was there a mistake in measurement?

use alternative outlier criteria, e.g., Chauvenet's criterion

use estimators that are robust to outliers (median, not the mean)

consider removal, but only if defensible (in case of measurement error)



http://davidmlane.com/ben/cartoons.html

Post your questions to be answered during lecture