DATA 8

Spring 2020

Lecture 25

Confidence Intervals, Center, and Spread

Weekly Goals

- Monday + Wednesday
  - Estimation
  - Variability of the Estimate
  - Bootstrap
- **Today**
  - Confidence intervals
  - Describing a distribution
  - Center and spread

Confidence Interval Review

Review: Quantifying Uncertainty

- The estimate is usually not exactly right:

Estimate = Parameter + Error

Estimate	Parameter	Error
get from sample: random	unknown, fixed	unknown, ***random***

- How big is a typical error?
- What if the sample had been different?
- How can we express our uncertainty with numbers?

Review: Real vs. Bootstrap World

Real world (what we want):

- True probability distribution (**population**)
  - → Random sample 1
    - → Estimate 1
  - → Random sample 2
    - → Estimate 2
  - …
  - → Random sample 10000
    - → Estimate 10000

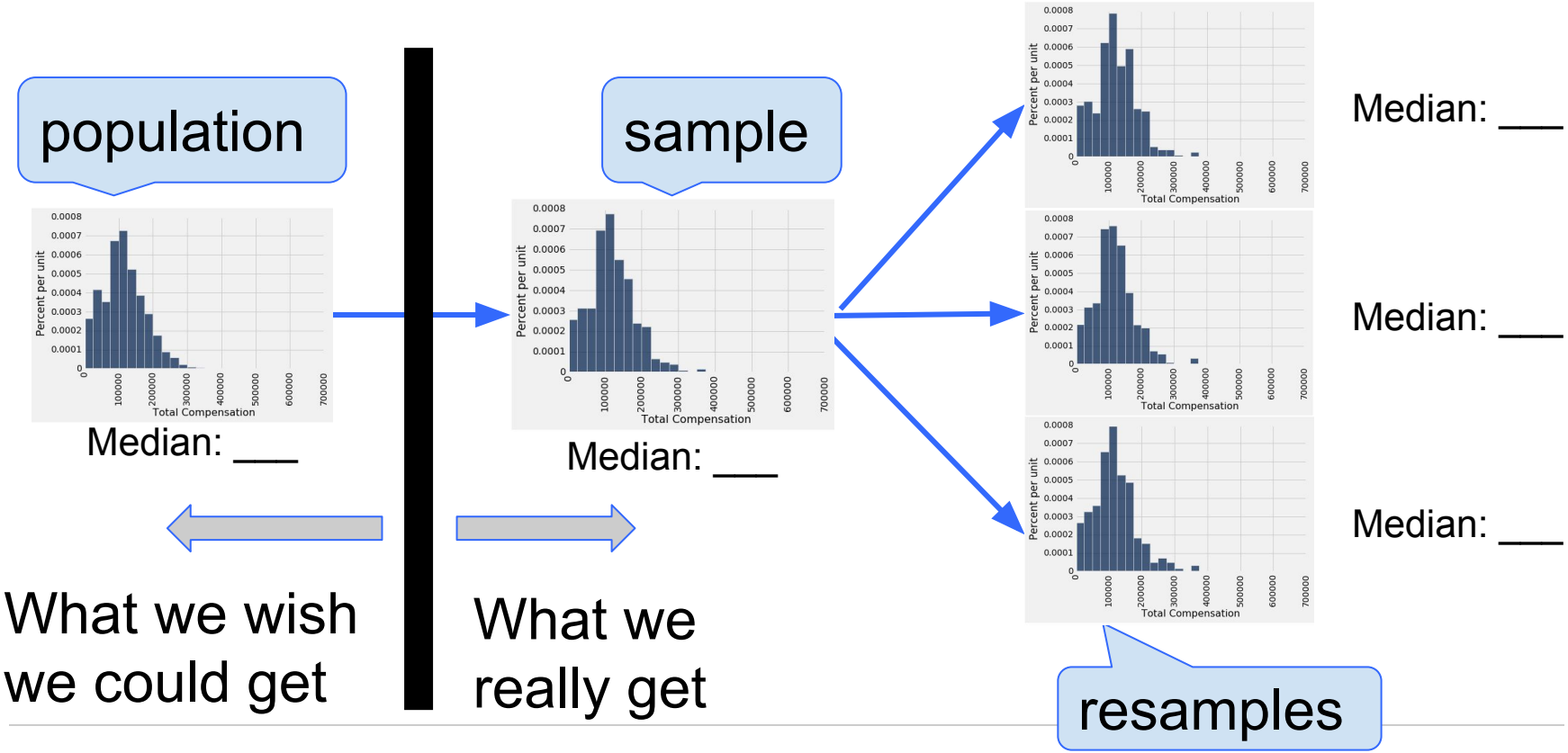
Can't get these :(

Bootstrap world:

- Empirical distribution of original sample (**"population"**)
  - → Bootstrap sample 1
    - → Estimate 1
  - → Bootstrap sample 2
    - → Estimate 2
  - ...
  - → Bootstrap sample 1000
    - → Estimate 1000

Hope: these two scenarios are analogous

Review: Why We Need the Bootstrap



Resampling Review

- From the original **sample**,
  - draw at random
  - **with** replacement
  - as many values as the original sample contained
- The size of the new sample has to be the same as the original one, so that the two estimates are comparable.
- Ideally we'd use fresh samples from the population, but this is the best we can do instead.

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- 95% is called the confidence level
  - Could be any percent between 0 and 100
  - Higher level means wider intervals
- The **confidence is in the process** that gives the interval:
  - It generates a “good” interval about 95% of the time.

(Demo)

Discussion Question

Give two different ways to make our confidence interval smaller, and mark what you'd change in the code for each:

```
our_sample = sf_pop.sample(300, with_replacement = False)
our_sample_median = percentile(50, our_sample.column('Total Compensation')

bootstrap_medians = make_array()
for i in np.arange(201):
    new_median = one_bootstrap_median()
    bootstrap_medians = np.append(bootstrap_medians, new_median)

left = percentile(2.5, bootstrap_medians)
right = percentile(97.5, bootstrap_medians)
```

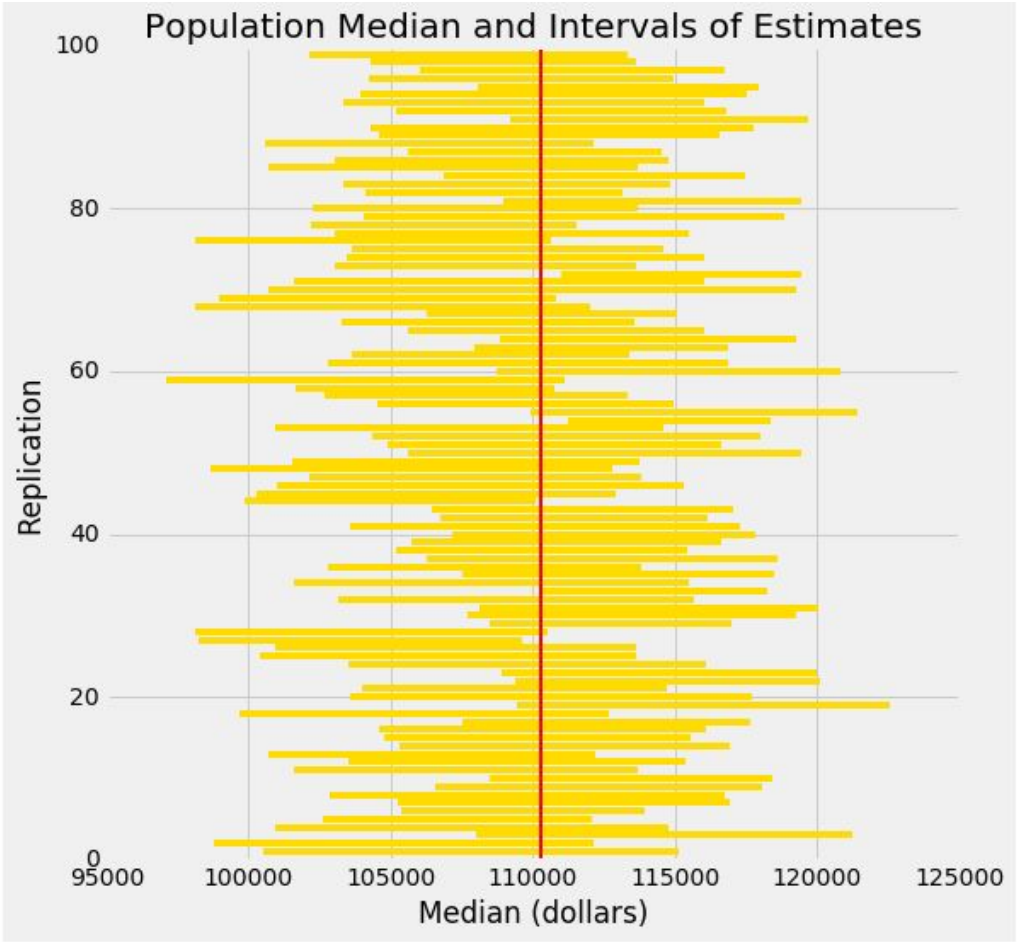
How to Interpret Confidence Intervals

(and how NOT to interpret them, too)

95% CI: Usage vs Interpretation

- **How to create it**
  - Middle 95% of the bootstrapped estimates

- **How to interpret it**
  - 95% of samples will give a 95% CI that contains the true parameter



Each line here is a 95% confidence interval from a fresh sample from the population

Can You Use a CI Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

Answer: False. We're estimating that their **average age** is in this interval.

Is This What a CI Means?

An approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

Answer: False. The average age of the mothers in the population is unknown but it's a constant. It's not random. No chances involved.

When *Not* to Use The Bootstrap

- If you're trying to estimate very high or very low percentiles, or min and max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)
- If the original sample is very small

Confidence Intervals For Testing

Using a CI for Testing

What if we want to do a hypothesis test, but we can't simulate under the null?

- Null hypothesis: **Population average = *x***
- Alternative hypothesis: **Population average ≠ *x***
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the population average
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, can't reject the null

Center and Spread

Questions

- How can we quantify natural concepts like “center” and “variability”?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

Average

The Average (or Mean)

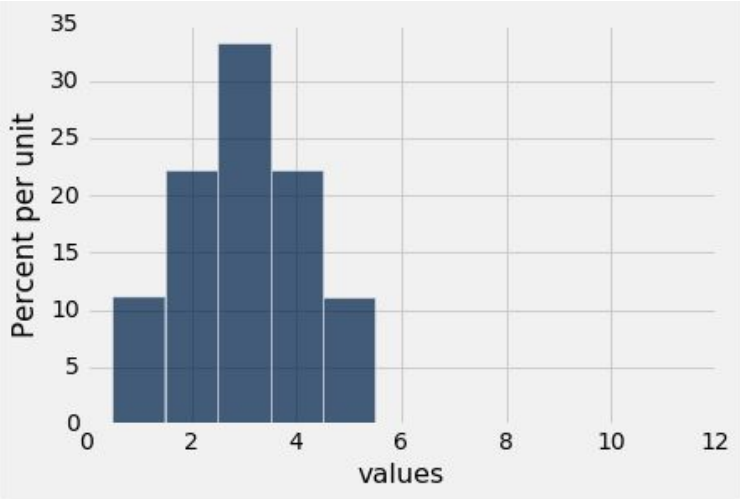
Data: 2, 3, 3, 9 **Average = (2+3+3+9)/4 = 4.25**

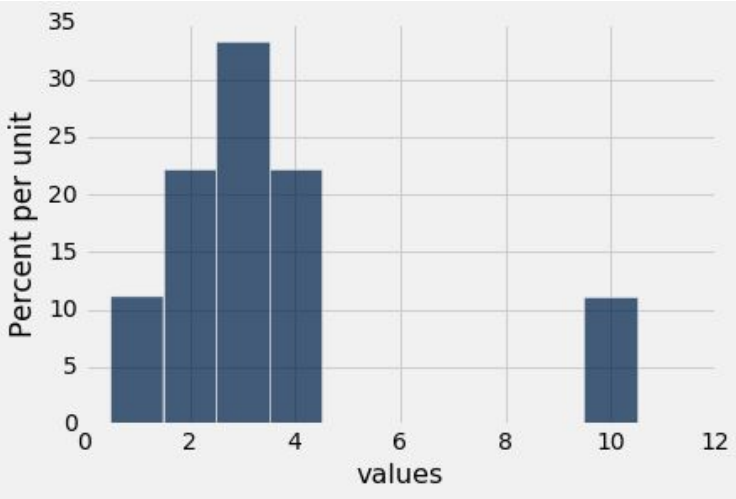
- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

(Demo)

Discussion Question

Are the medians of these two distributions the same or different? Are the means the same or different? If you say “different,” then say which one is bigger.





Comparing Mean and Median

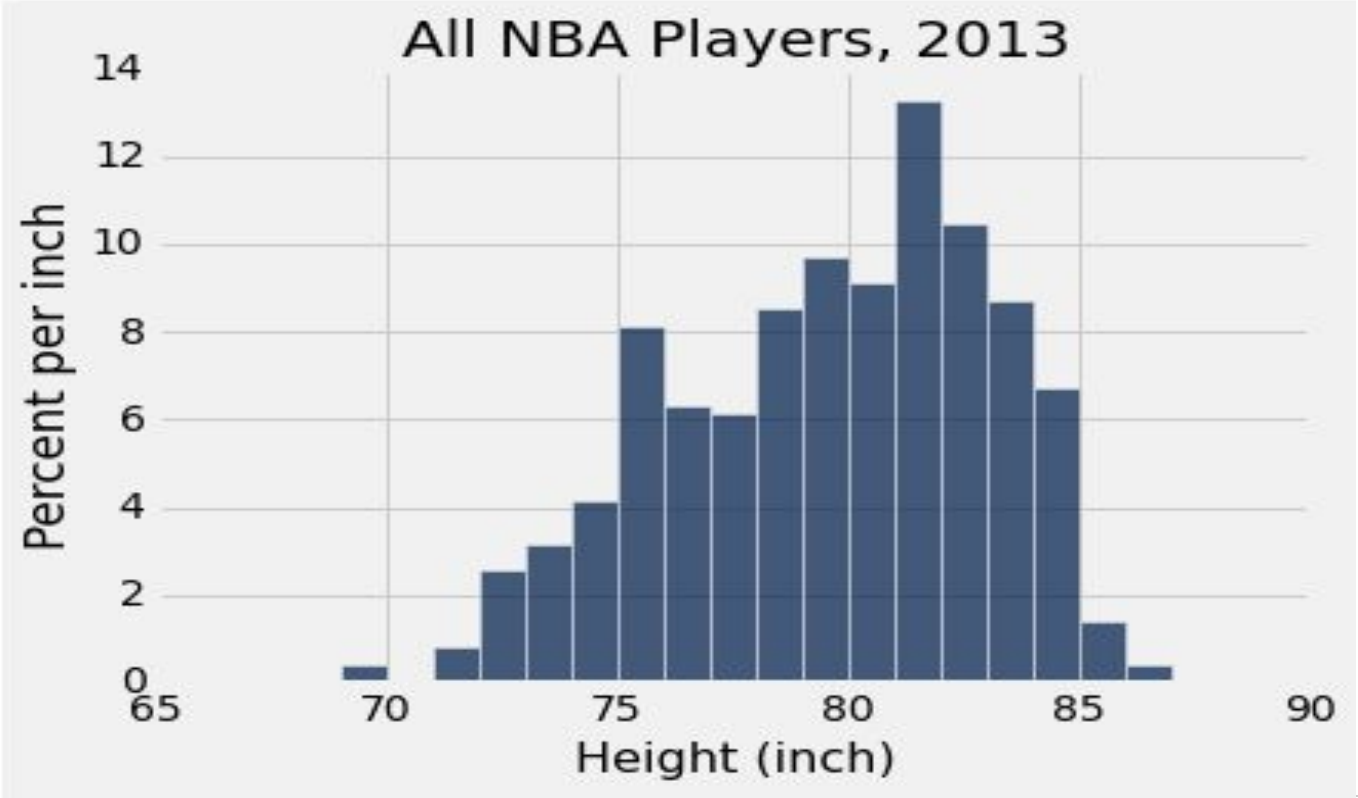
- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

Discussion Question

Which is bigger?

(a) mean

(b) median



Standard Deviation

Defining Variability

Plan A: “biggest value - smallest value”

- Doesn’t tell us much about the shape of the distribution

Plan B:

- Measure variability around the mean
- Need to figure out a way to quantify this

(Demo)

How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average
- SD = root mean square of deviations from average

  5 4 3 2 1
- SD has the same units as the data

Why Use the SD?

There are two main reasons.

- **The first reason:**

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

- **The second reason:**

Coming up in the next lecture.