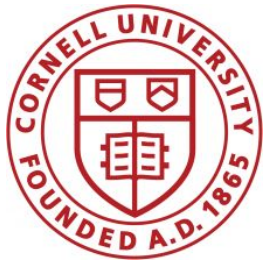


DSFA

Spring 2018

Lecture 23

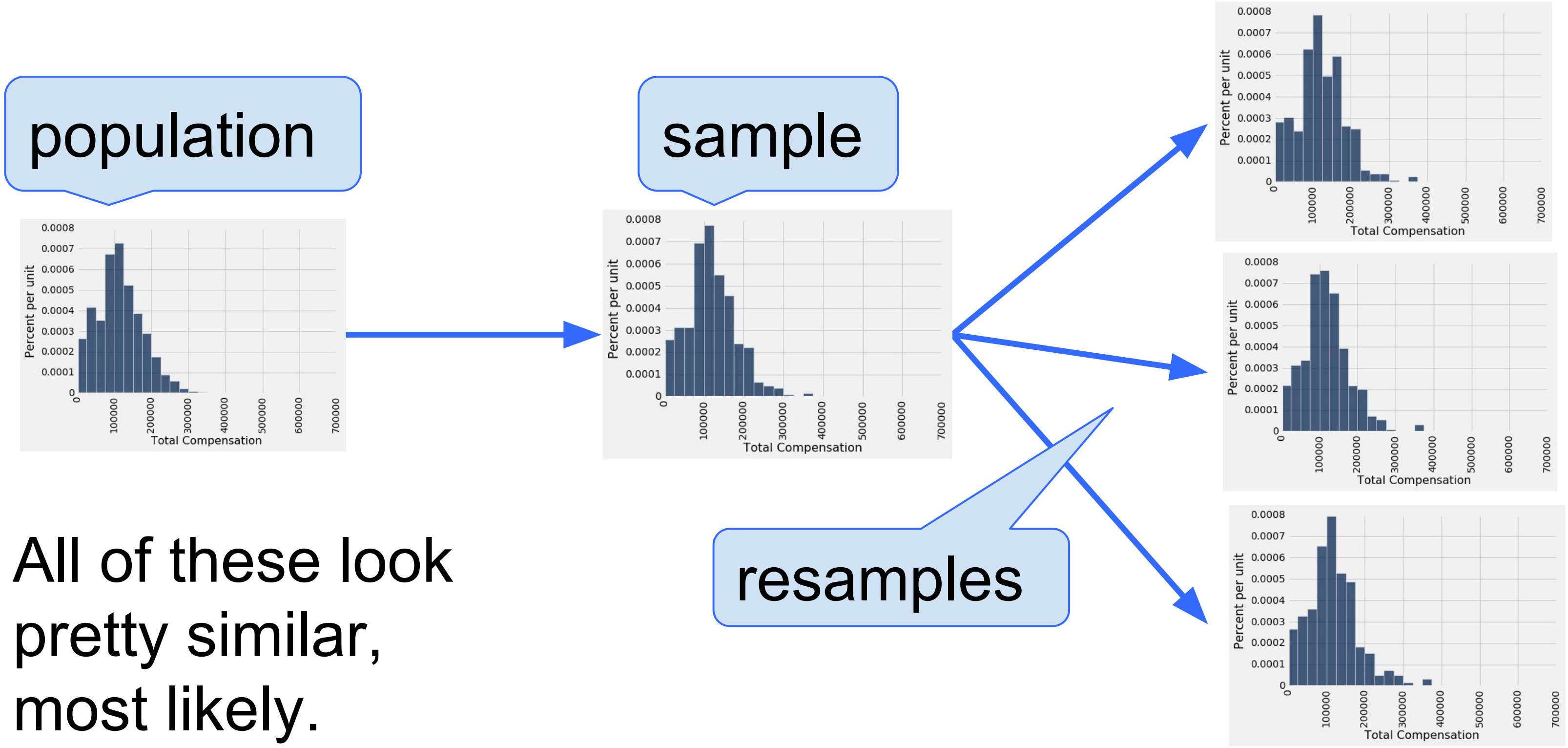
Interpreting Confidence

Announcements

- Project 2
  - checkpoint due 9pm today;
  - due 9pm M 3/26
- Lab 7 posted today
- Prof. Martin Wells (Statistics) will give next four lectures

The Bootstrap

Why the Bootstrap Works



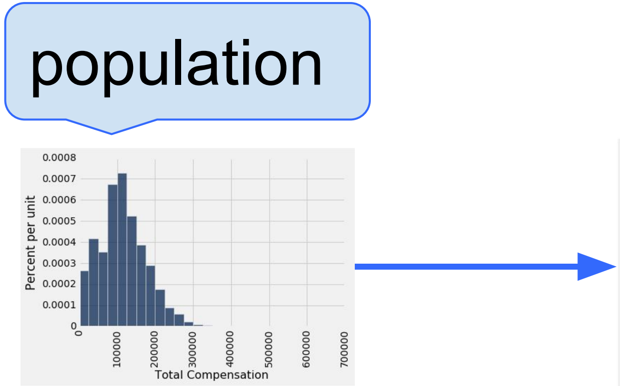
All of these look pretty similar, most likely.

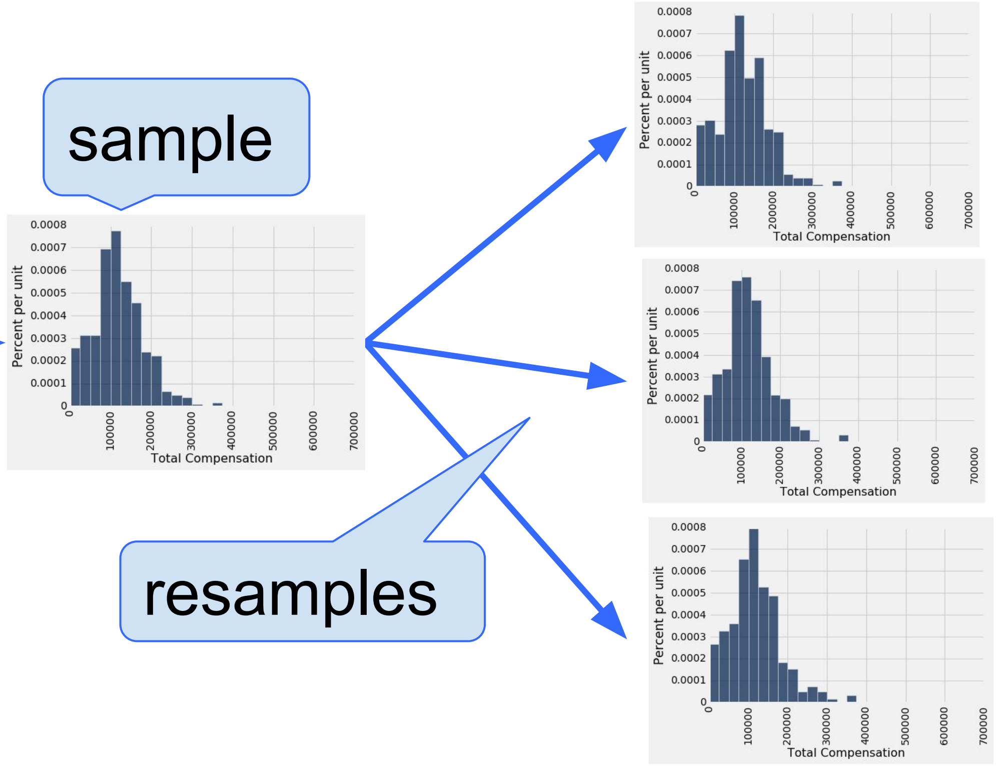
Key to Resampling

- 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

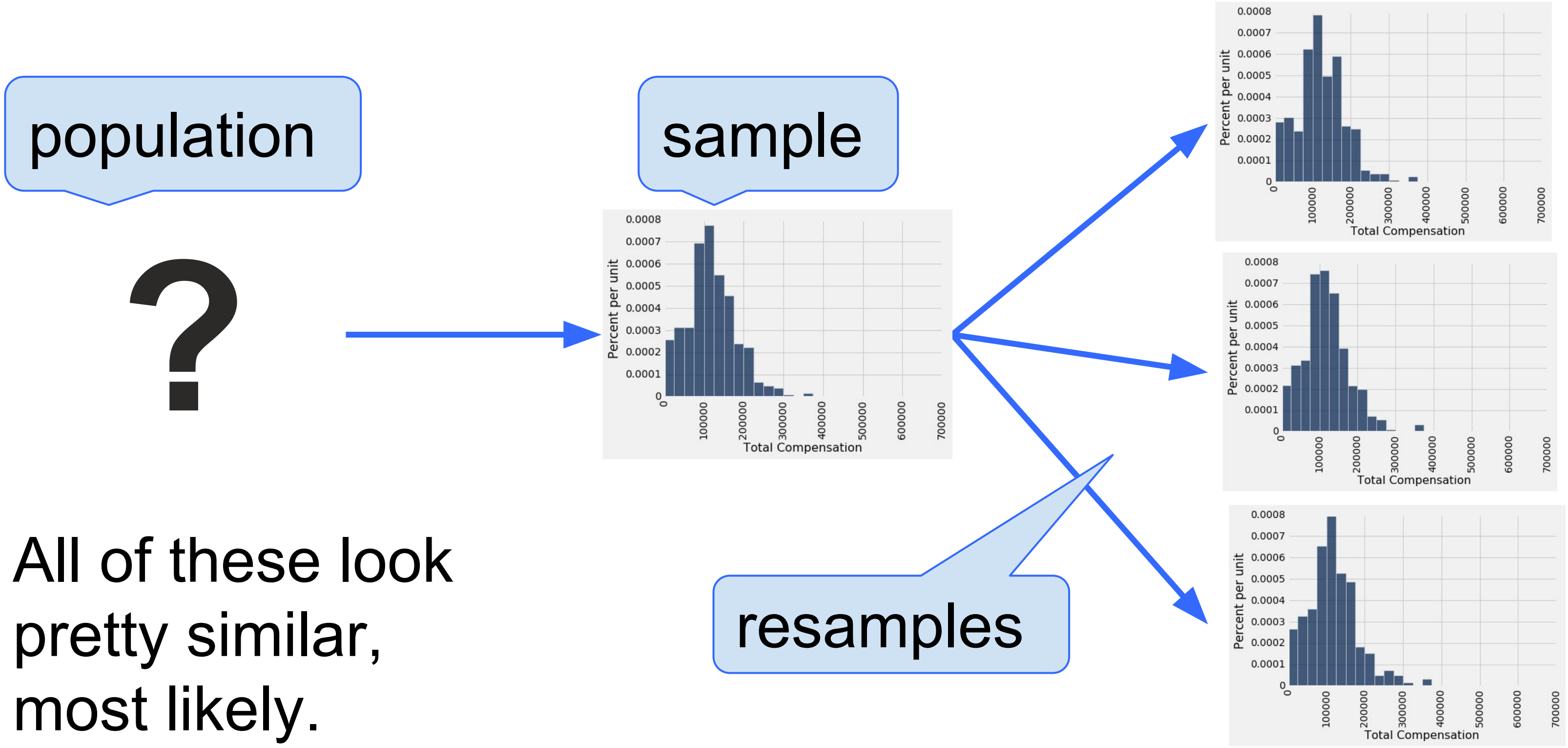
Why the Bootstrap Works





All of these look pretty similar, most likely.

Inference Using the Bootstrap



All of these look pretty similar, most likely.

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
  - Bigger means wider intervals
- The **confidence is in the process** that generated the interval:
  - It generates a “good” interval about 95% of the time.

(Demo)

Use Methods Appropriately

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 (~15)

(Demo)

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.

(Demo)

Is This What a CI Means?

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:

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

Answer: False. It's not a probability; that's either true or false.

Confidence Interval Tests

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
  - Bigger means wider intervals
- The **confidence is in the process** that generated the interval:
  - It generates a “good” interval about 95% of the time.

(Demo)

Using a CI for Testing

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

Average

The Average

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

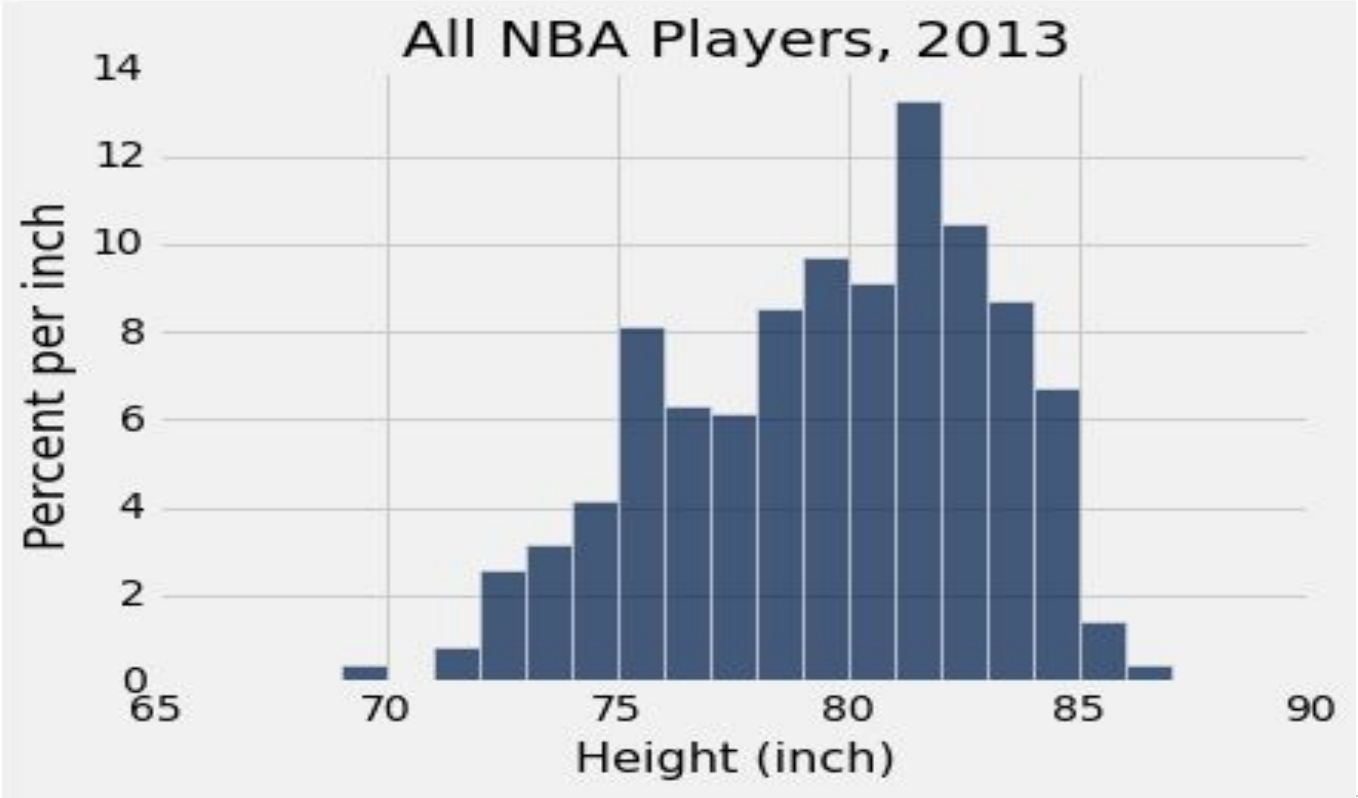
- Not a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

Discussion Question

Which is bigger?

(a) mean

(b) median



Properties of the Mean

- Balance point of the histogram
- Not the “halfway point” of the data; the mean is not the median...
- Unless the distribution is symmetric about a point, then that point 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