DSFA

Spring 2019

Lecture 19

Confidence Intervals

Announcements

- Project 2: Part 1 due today
- Project 2: Final submission April 16

- Prelim 2: In-class. Tuesday, April 16
  - Sample questions will be posted
  - Study guide and personal cheat sheet

Percentiles

Computing Percentiles

The 80th percentile of a set of numbers is the smallest value in the sample that is at least as large as 80% of the sample

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

The 80th percentile is ordered element 4: `(80/100) * 5`

Percentile

Size of set

For a percentile that does not exactly correspond to an element, take the next greater element instead

The `percentile` Function

- The *p*th percentile is the smallest value at least as large as *p*% of the values in the sample
- Function in the `datascience` module:

  `percentile(p, values)`

- **`p`** is between 0 and 100
- Returns the *p*th percentile of the array

Discussion Question

Which are `True`, when `s = [1, 7, 3, 9, 5]`?

```
percentile(10, s) == 0

percentile(39, s) == percentile(40, s)

percentile(40, s) == percentile(41, s)

percentile(50, s) == 5
```

(Demo)

Estimation (Review)

Inference: Estimation

- What is the value of a population parameter?

- If you have a census (that is, the whole population):
  - Just calculate the parameter and you're done

- If you don't have a census:
  - Take a random sample from the population
  - Use a statistic as an **estimate** of the parameter

(Demo)

Variability of the Estimate

- One sample ➜ One estimate
- But the random sample could have come out differently
- And so the estimate could have been different
- Main question:
  - **How different could the estimate have been?**
- The variability of the estimate tells us something about how accurate the estimate is:

  estimate = parameter + error

(Demo)

Where to Get Another Sample?

- One sample ➜ One estimate
- To get many values of the estimate, we needed many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

The Bootstrap

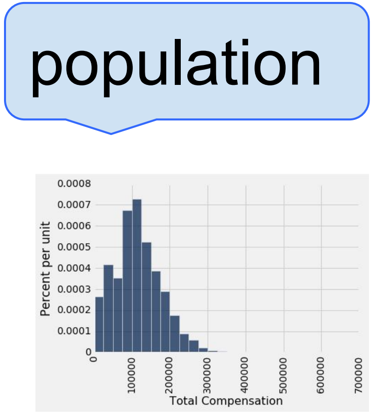
The Bootstrap

- A technique for simulating repeated random sampling

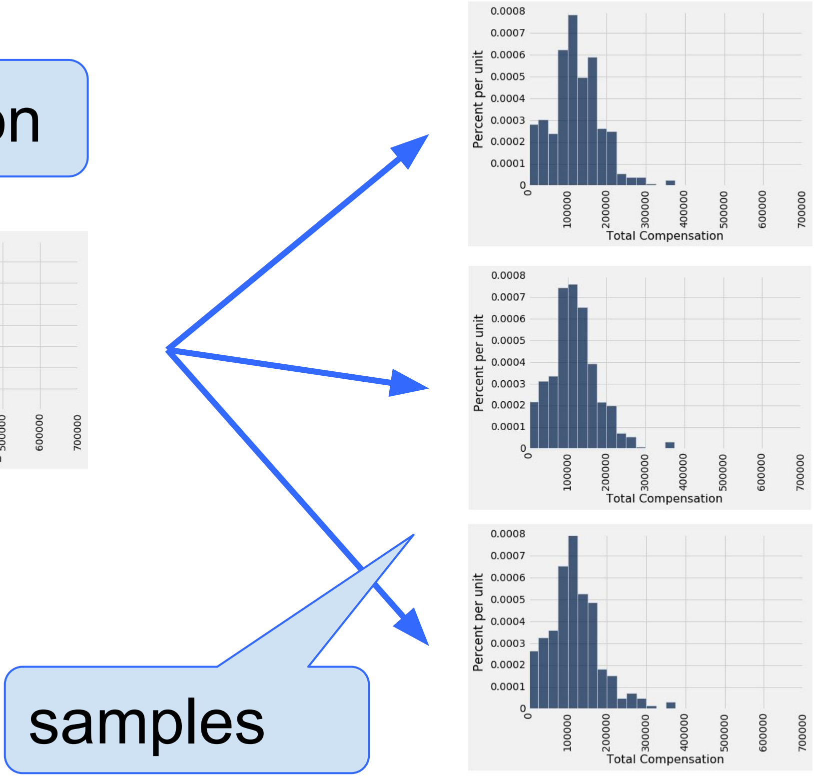
- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population

- So we sample at random from the original sample!

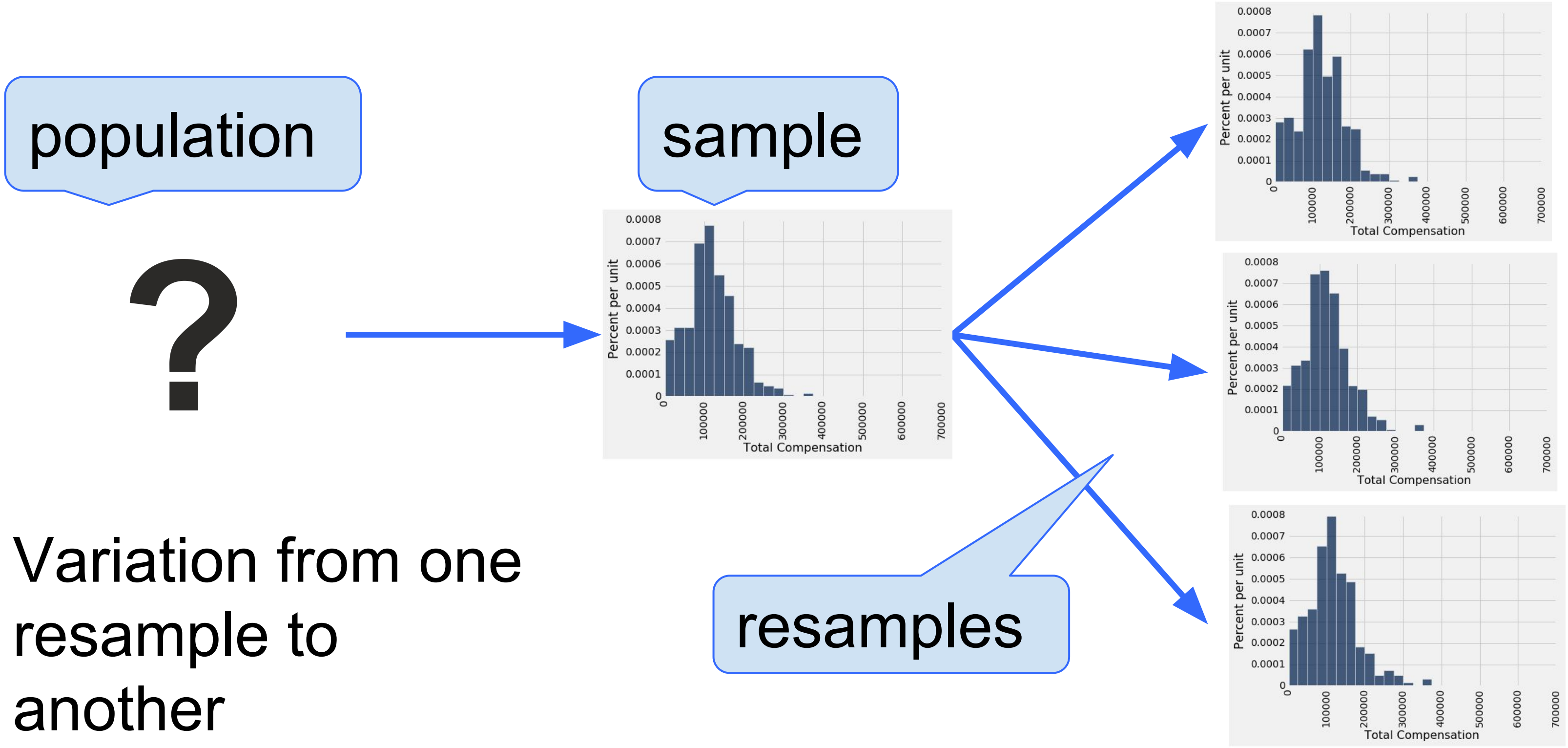
Repeated Sampling



Variation from one sample to another



The Bootstrap



Variation from one resample to another

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- Confidence level: typically 95%
    - Could be any percent between 0 and 100
    - Bigger means wider intervals
- The interval contains the parameter about 95% of the time **in repeated sampling**

(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.

Is This What a CI Means?

Based on our sample, 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. It's not a probability. Either the population average is in the interval or it isn't!

Confidence Interval Tests

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**

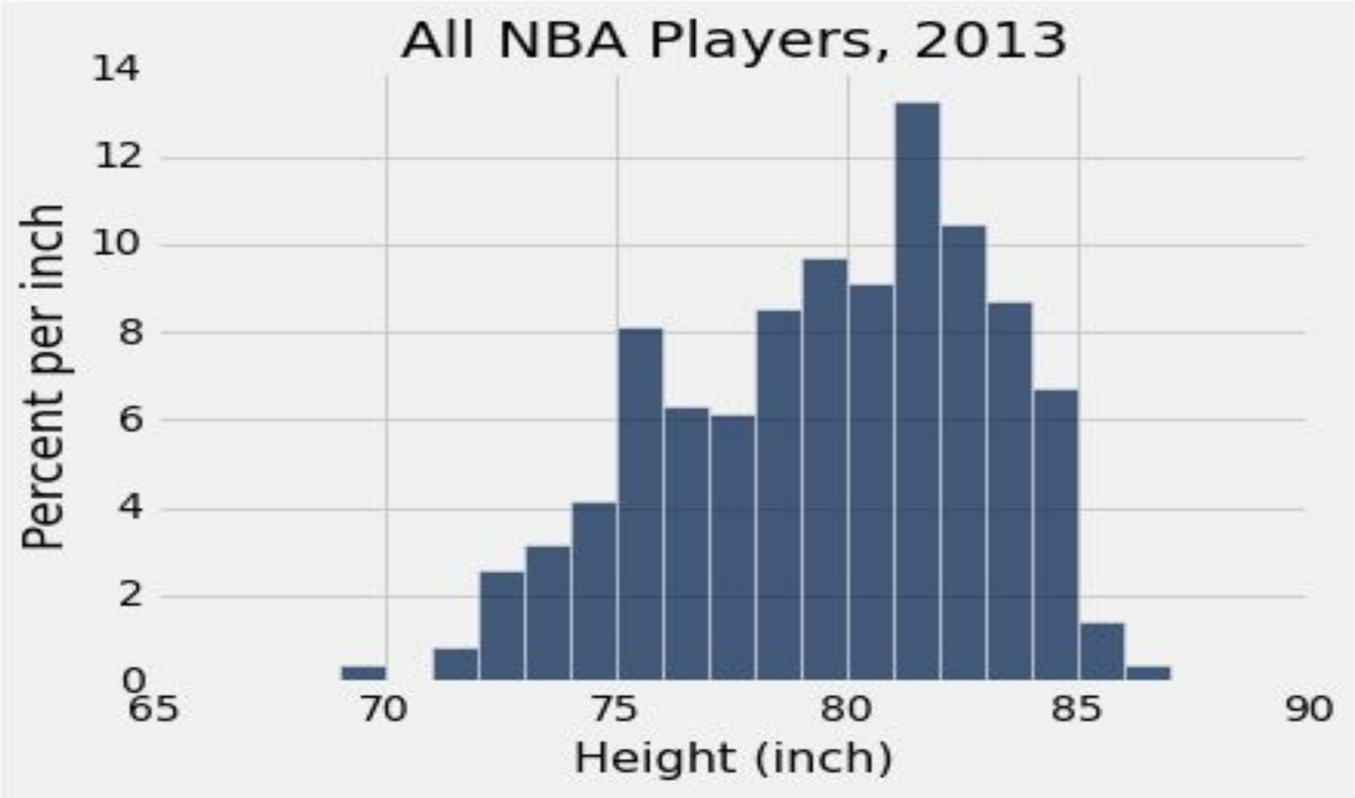
- Not a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data

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

Key to Bootstrap/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

(Demo)