

MODULE 7

Estimation



PERCENTILES

DEFINING PERCENTILES

- **A percentile is the value at a particular rank.**
- Let p be a number between 0 and 100. The p th percentile of a collection is the smallest value in the collection that is **at least as large as $p\%$** of all the values.
- Practically, suppose there are n elements in the collection.

To find the p th percentile:

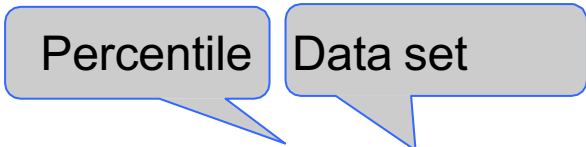
- Sort the collection in increasing order.
 - Find $p\%$ of n : $(p/100) \times n$. Call that k .
 - If k is an integer, take the k th element of the sorted collection.
 - If k is not an integer, round it up to the next integer, and take that element of the sorted collection.
-

COMPUTING PERCENTILES

The **Xth** percentile is the first value on the sorted list that is **at least as large as X%** of the elements.

Example: `s = [1, 7, 3, 9, 5]`

`s_sorted = [1, 3, 5, 7, 9]`



Percentile Data set

`percentile(80, s)` is 7 The

The diagram consists of two grey speech bubbles with blue outlines. The left bubble contains the word 'Percentile' and the right bubble contains the words 'Data set'. Arrows from both bubbles point towards the function call 'percentile(80, s)' in the text below.

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

- 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 in a set that is at least as large as $p\%$ of the elements in the set
 - Function in the ***datascience*** module:
`percentile(p, values)`
 - `p` is between 0 and 100
 - Returns the p th percentile of the array (Demo – notebook 7.1, Percentiles)
-

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 – notebook 7.1,
Percentiles in class)

ESTIMATION

INFERENCE: ESTIMATION

- How do we calculate the value of an unknown 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 – Notebook 7.1, Estimating Median -
Sample Median)

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
- Big question:
 - How different would it be if we did it again?

(Demo – Notebook 7.1, Variability of the Estimate)

QUANTIFYING UNCERTAINTY

- The estimate is usually not exactly right:

$$\text{Estimate} = \text{Parameter} + \text{Error}$$

- How accurate is the estimate, usually?
- How big is a typical error?
- When we have a census, we can do this by simulation

(Demo – Notebook 7.1, Quantifying
Uncertainty)

WHERE TO GET ANOTHER SAMPLE?

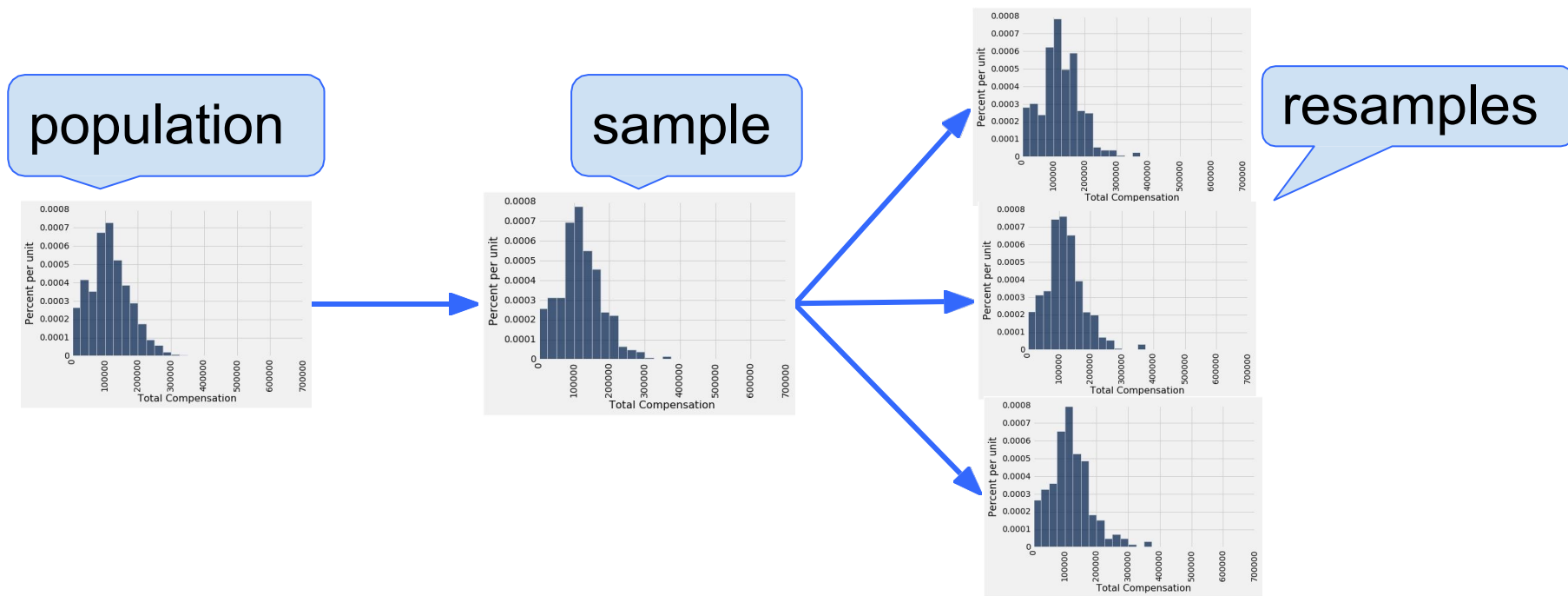
- We want to understand errors of our estimate
 - Given the **population**, we could simulate
 - ...but we only have the **sample**!
 - To get many values of the estimate, we needed many random samples
 - Can't go back and sample again from the population:
 - No time, no money
 - Stuck?
-

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!
 - AKA, *resampling*
-

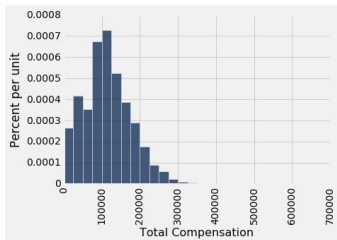
WHY THE BOOTSTRAP WORKS



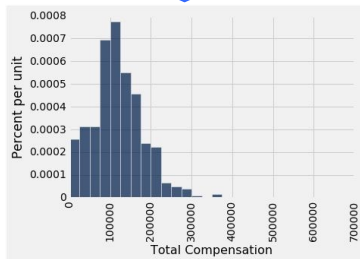
All of these look pretty similar, most likely.

WHY WE NEED THE BOOTSTRAP

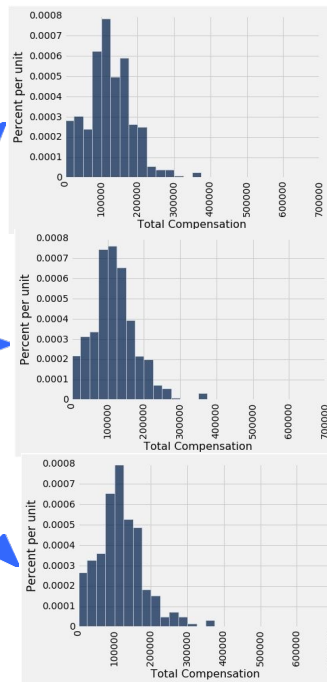
population



sample



resamples



What we
wish we
could get

What we
really
get

REAL WORLD VS. BOOTSTRAP WORLD

Real world:

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

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

THE BOOTSTRAP PRINCIPLE

- The bootstrap principle:
 - **Bootstrap-world** sampling \approx **Real-world** sampling
 - Not always true!
 - ... but reasonable if sample is large enough
 - We hope that:
 - a. Variability of bootstrap estimate
 - b. Distribution of bootstrap errors

...are similar to what they are in the real world
-

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**

(Demo – notebook 7.1, Bootstrap)

CONFIDENCE INTERVALS

DO THE ESTIMATES CAPTURE THE PARAMETER?

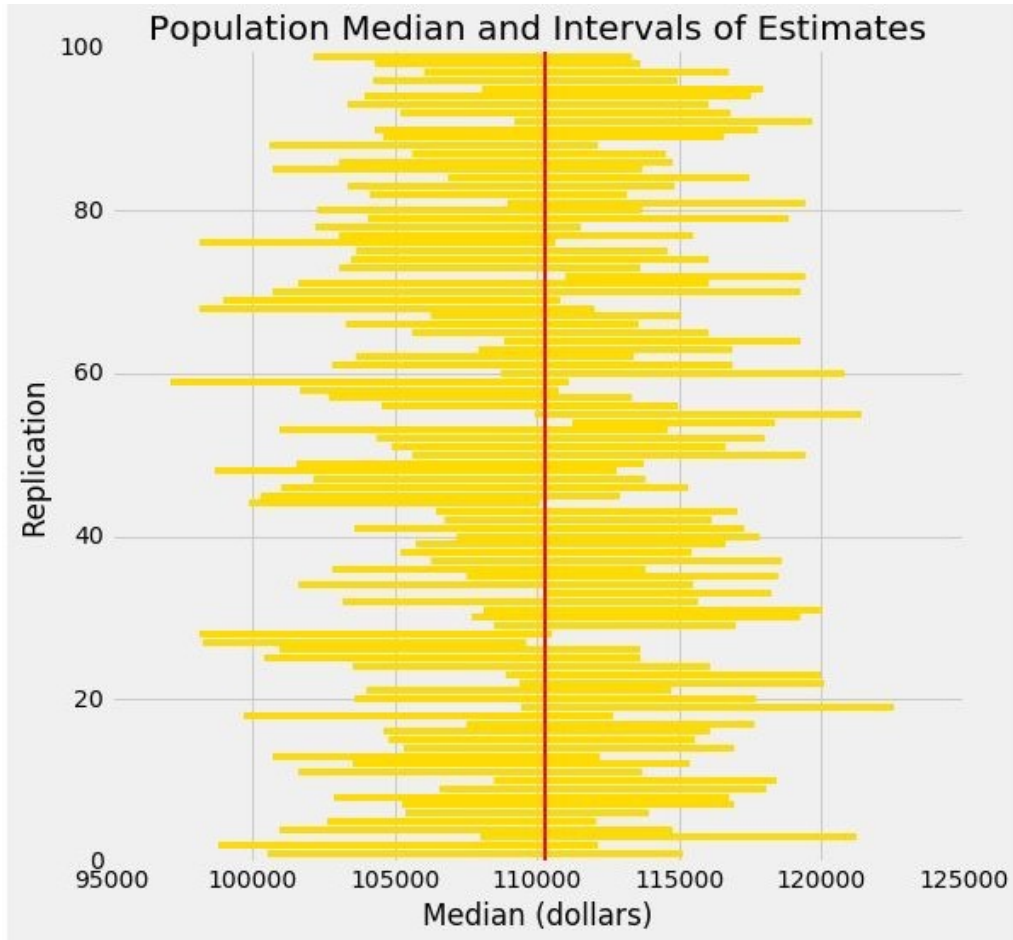
Q: How often does the empirical histogram of the resampled medians **contain our parameter?**

For instance, **how often** does the **middle 95% of the resampled medians contain our parameter?**

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 – notebook 7.1, Confidence Intervals)



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

USE METHODS APPROPRIATELY

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

- Null hypothesis: Population average = x
 - Alternative hypothesis: Population average $\neq x$
 - 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
-

QUESTIONS?