DSC 10, Spring 2018
Lecture 15

Simulation and Bootstrapping

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Review:
Statistics and Estimation

Terminology

Parameter

A number associated with the population

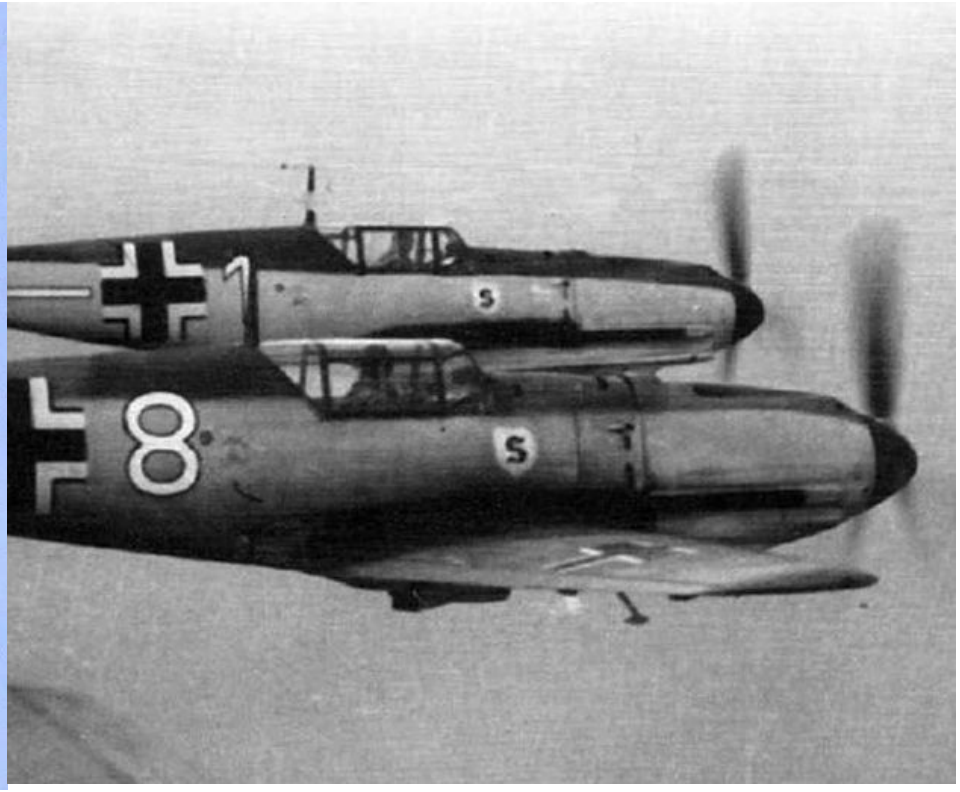
Statistic

A number calculated from the sample

A statistic can be used as an **estimate** of a parameter

(Demo)

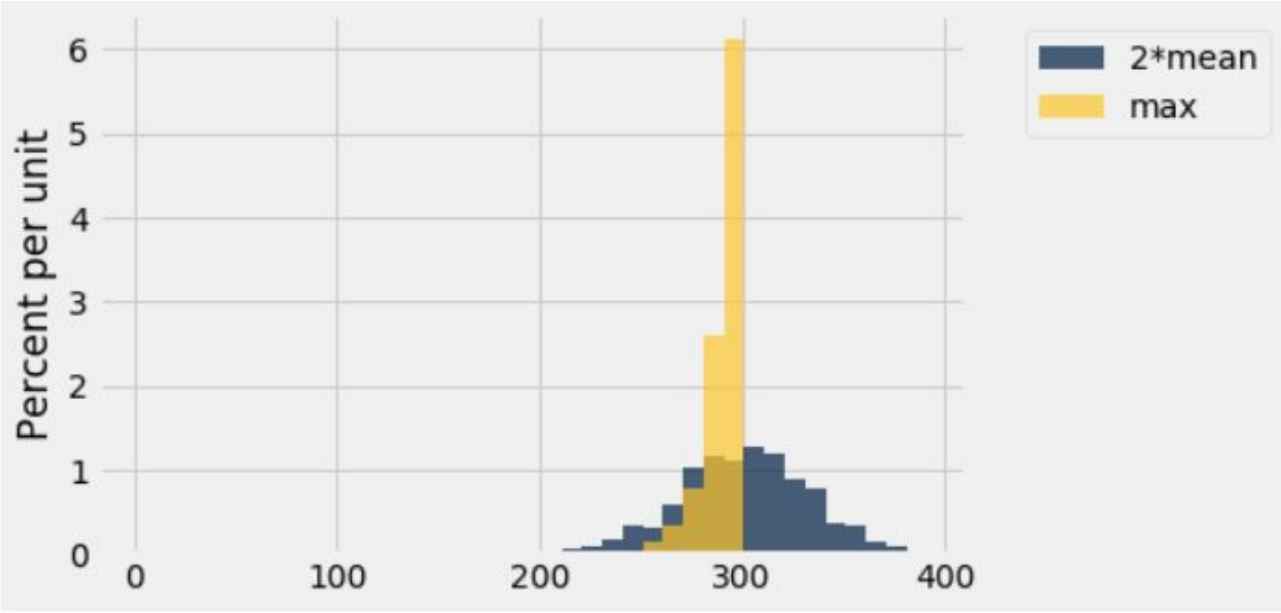
How many enemy planes?



Bias-variance trade-off

- **Max** has low variability, but it is biased.
- **2*average** has little bias, but it is highly variable.
- Life is tough.

(Demo)



Comparing Samples to the Population

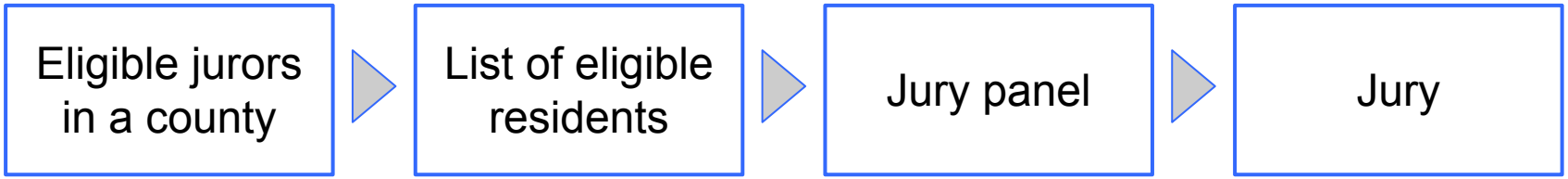
Jury Selection in Alameda County

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California

October 2010

Jury Panels



Section 197 of California's Code of Civil Procedure: All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court.

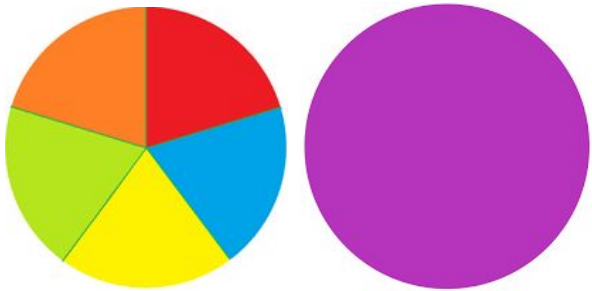
Sixth Amendment to the US Constitution: … the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the state and district wherein the crime shall have been committed. (Demo)

Total Variation Distance

- For each category compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum and divide by two.

Question

	purple	red	orange	green	yellow	blue
Distribution 1	0	1/5	1/5	1/5	1/5	1/5
Distribution 2	1	0	0	0	0	0



What is the total variation distance between the two distributions above?

A: 1/5
B: 3/5
C: 1
D: 7/5
E: 2

Inference: Estimation

Inference: Estimation

- How big is 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)

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

Where to Get Another Sample?

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

The Bootstrap

- Need another random sample that looks like the population
- 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!
- A technique for simulating repeated random sampling

The Bootstrap

Questions

What should be the size of your new sample?

A. 25% of the original sample
B. 50% of the original sample
C. 75% of the original sample
D. 100% of the original sample
E. Depends on the problem

How should we obtain this new sample?

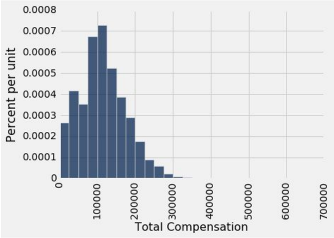
A. **with** replacement
B. **without** replacement
C. Depends on the problem

Key to Resampling

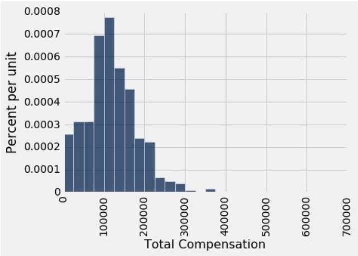
- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained
- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

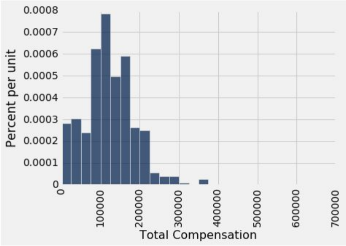
Why the Bootstrap Works

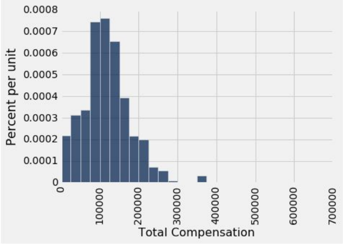
population

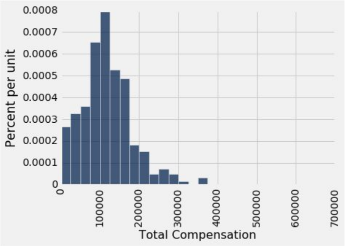


sample









resamples

All of these look pretty similar, most likely.

(Demo)