CompSci 116: Sampling & Hypotheses

Jeff Forbes

February 21, 2019

Plan and announcements

- Sampling
  - Probability Distributions
  - Empirical Distrubutions
  - Using sampling to compute a statistic

- Assignments
  - HW 3
  - Project 2
  - Start thinking about final project groups & topics

- Day of Data

CSbyUs Day of Data at Duke!

CSbyUs is a Bass Connections project designing and sharing digital era education with under-empowered learning environments.

We are hosting Durham middle school girls at Duke TODAY, February 22, for a day-long exploration of data science.

Students: please help our middle schoolers collect data for their projects by filling out these short surveys from 11:30am-1:20pm:

- tiny.cc/csbyusdayofdata1
- tiny.cc/csbyusdayofdata2
- tiny.cc/csbyusdayofdata3
- tiny.cc/csbyusdayofdata4
- tiny.cc/csbyusdayofdata5

North Carolina School of Science and Mathematics

OSPRI OPEN SOURCE PEDAGOGY, RESEARCH + INNOVATION

Duke UNIVERSITY EHD Bass Connections Education & Human Development

CSbyUs

Day of Data at Duke!

And… join us for a showcase of their completed projects this afternoon!

Come support amazing middle school girls who, alongside their NC School of Science and Math and Duke mentors, will present their fantastic data science projects to the Duke community during a drop-in showcase.

Questions? Please contact Aria Chernik at aria.chernik@duke.edu

Friday, February 22
2:30 - 3:00 PM
068 Brodhead Center

light refreshments served

North Carolina School of Science and Mathematics

OSPRI
OPEN SOURCE PEDAGOGY, RESEARCH + INNOVATION

Duke UNIVERSITY EHD Bass Connections Education & Human Development

http://bit.ly/FoDS-brookings

> According to a survey conducted in August and made public on the Brookings Institution website, a plurality of college students polled, 44 percent, believed that hate speech was not protected by the First Amendment.

- Do you believe this result?
- What questions do you have about the survey?
- What would you want to know to consider the implications?

Sampling

- Probability: Compute what will happen when you run an experiment
- Statistics: Look at the outcome of the experiment and try to reason about the world
- Sampling: Take the outcome of an experiment and use the rules of probability to reason about how it might have come out differently

Probability Distribution

- Random quantity with various possible values
- “Probability distribution”:
  - All the possible values of the quantity
  - The probability of each of those values
- In some cases, the probability distribution can be worked out mathematically without ever generating (or simulating) the random quantity

Empirical Distribution

- Based on observations
- Observations can be from repetitions of an experiment
- “Empirical Distribution”
  - All observed values
  - The proportion of counts of each value

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

fixed

Use the data to guess the value of an unknown number

depends on the random sample

Create an **estimate** of the unknown quantity

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 do we test a hypothesis?

- Chocolate has no effect on cardiac disease.
- Yes, chocolate has some effect on cardiac disease.

- This jury panel was selected at random from eligible jurors.
- No, it has too many people with college degrees.

- Create a model for our set of assumptions about the data

How do we assess a model?

- Simulate data according to the assumptions of the model
  - Learn what the model predicts.

- Compare the predictions to the data that were observed.

- If the data and the model's predictions are not consistent, that is evidence against the model.

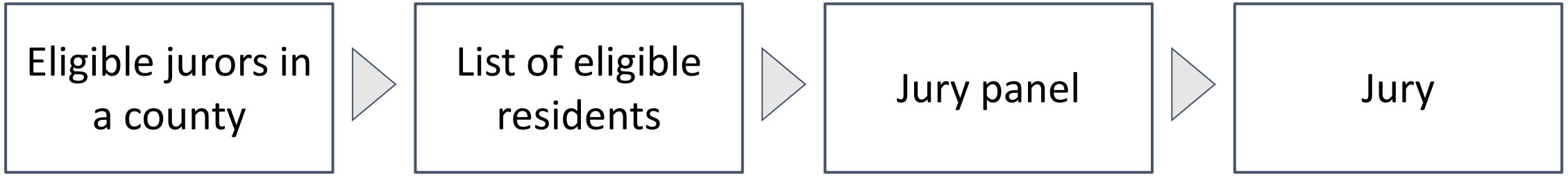
Robert Swain v. Alabama

1965 Supreme Court case about jury selection

- In Talladega, Alabama, 26% of residents were black
- In Swain's jury panel, 8 of 100 panelists were black
- All 8 were struck from the jury by the prosecution (using peremptory challenges)

Ruling: "The overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of [black men]."

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.

Sampling from a Distribution

- Sample at random from a categorical distribution

```
sample_proportions(sample_size, pop_distribution)
```

- Samples at random from the population
  - Returns an array containing the distribution of the categories in the sample

Perfect information

- You want to know how many US voters support a particular policy.
- You could ask everyone. That works.
- But, sometimes we can't afford to do that. So, instead, we could ask some of them, and draw inferences about the general population.

A common scenario

- You have to make a decision based on incomplete information.
- The quality of your decision is affected by
  - the information that you have
  - the information that you don’t have

- So, before making the decision, it is worth examining why and how your information came to be incomplete.

Terminology

- **Population:** A collection of individuals
  - All United flights out of SFO in Summer 2015

- **Variable:** Something that varies in the population
  - airline *(categorical variable)*
  - amount of delay in departure *(quantitative variable)*

- **Sample:** A subset of the population

Why take a sample?

- You want to understand the variable in the population,

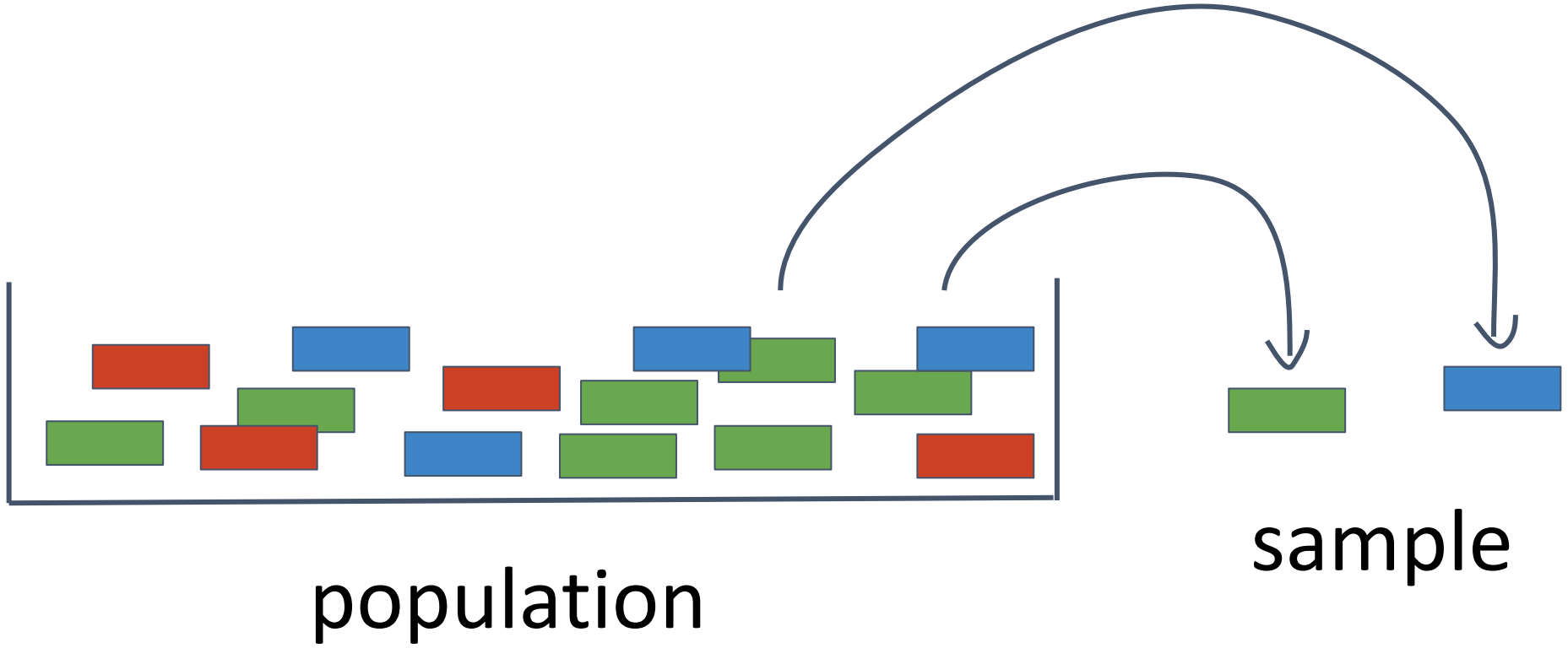
but

- you don't have the resources to measure the variable on all the individuals in the population,

so

- you just measure it on a subset of them.

“Tickets in a box”

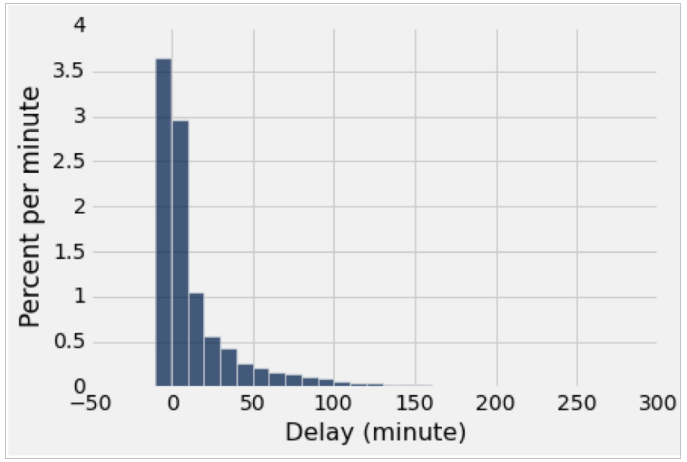


Best way to draw the sample

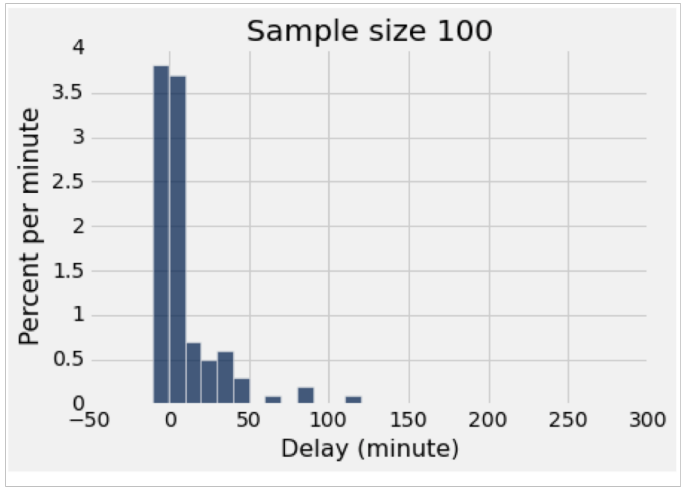
At random!

Two distributions

distribution of the population



empirical distribution of a sample



Large Random Samples

If the sample size is large,

then the empirical distribution of a *uniform random sample*

resembles the distribution of the population,

with high probability

The effect of sample size

- Larger random samples are more likely to resemble the population than smaller ones.

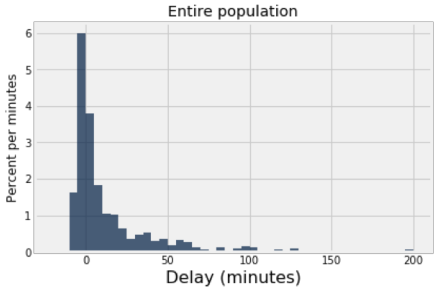
- However, if the method of sampling is not random, taking a larger sample isn't necessarily better.
  - You could just end up with a big bad sample.

More terminology

- **Parameter:** A number calculated using the values in the population
  - Median delay among all flights
  - Proportion of voters who are Republican

- **Statistic:** A number calculated using the values in a sample

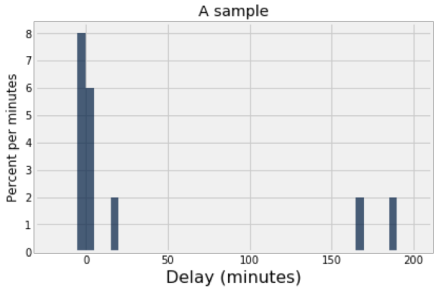
- A statistic can be used as an estimate of a parameter.

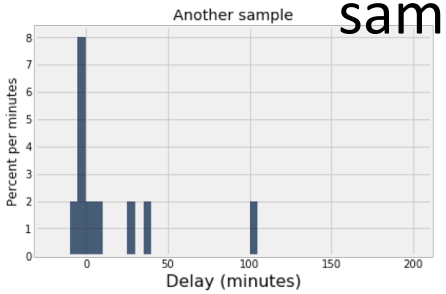


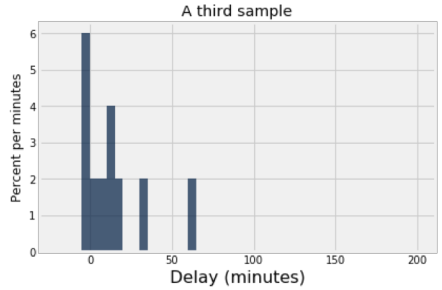
randomly sample

randomly sample

randomly sample



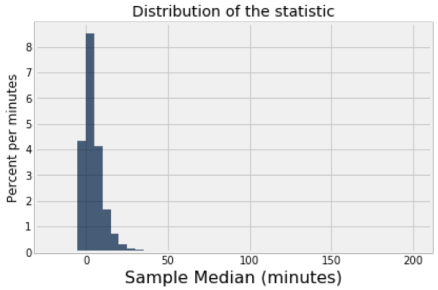




median = 2.0

median = -0.5

median = 9.5



Gary's Game

- Flip a fair (?) coin 10 times
  - **`if`** number of heads >= 5, we win
  - **`else`** Gary wins

- Play the game once
  - There's one head
  - Was the game rigged?

Final Project

- Demonstrate proficiency in techniques from class applied to dataset of your choosing
- Data!
    - Domain: ?
    - Scale: Large enough
        - Categorical and numerical variables
        - More than 100 observations
    - Accessible & Manageable