



190F Foundations of Data Science

Spring 2020

Lecture 09

Statistics

Announcements

Statistics

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

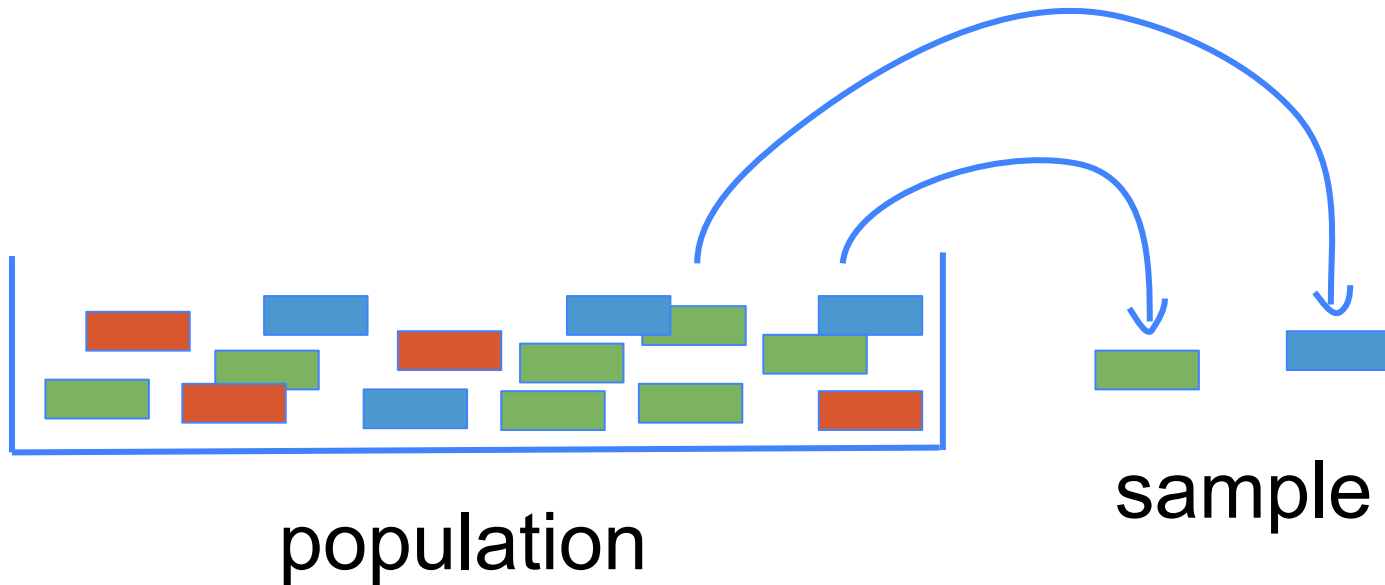
Terminology

- **Population:** A collection of individuals
 - All flights out of SFO last summer
 - **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”

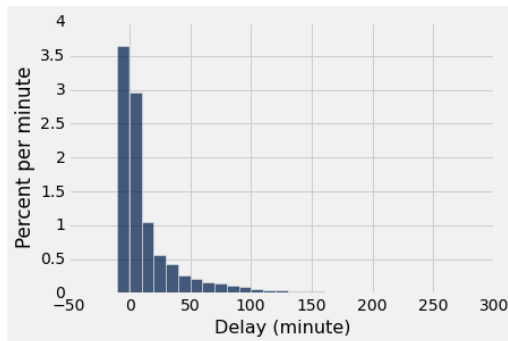


Best way to draw the sample

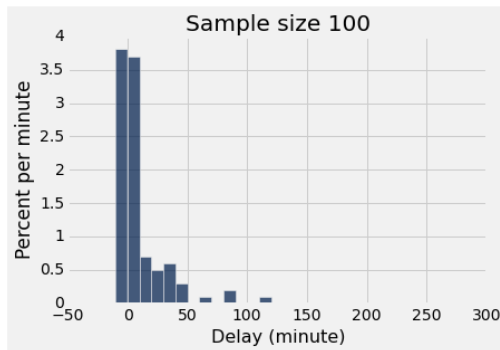
At random!

Two distributions

distribution of the
population



empirical distribution
of a sample



Why sample at random?

The empirical distribution
of a large random sample
is very likely to be close
to the distribution of the population.

That's
why.

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.
-

Probability distribution of a statistic

- Values of a statistic vary because random samples vary
 - “Sampling distribution” or “probability distribution” of a statistic consists of:
 - All possible values of the statistic
 - Corresponding probabilities
 - Can be hard to calculate
 - Need math
 - Or generate many random samples
-

Empirical distribution of a statistic

- Empirical distribution of the statistic:
 - Based on simulated values of the statistic
 - Consists of all the observed values of the statistic,
 - and the proportion of times each value appeared
- Good approximation to the probability distribution of the statistic *if the number of repetitions in the simulation is large.*

Simulating a Statistic

- Figure out the code to generate one value of the statistic
- Create an empty array in which you will collect simulated values
- For each repetition of the process:
 - Simulate one value of the statistic
 - Append this value to the collection array
- The array will contain all of the simulated values

(Demo)

Assessing Models

Steps in Assessing a Model

- Come up with a statistic that will help you decide whether the data support the model or an alternative view of the world.
 - Simulate the statistic under the assumptions of the model.
 - Draw a histogram of the simulated values. This is the model's prediction for how the statistic should come out.
 - Compute the observed statistic from the sample in the study.
 - Compare this value with the histogram.
 - If the two are not consistent, that's evidence against the model.
-

Jury Selection

Swain vs. Alabama, 1965

- Talladega County, Alabama
 - Robert Swain, black man convicted of crime
 - Appeal: one factor was all-white jury
 - Only men 21 years or older were allowed to serve
 - 26% of this population were black
 - Swain's jury panel consisted of 100 men
 - 8 people on the panel were black (8%)
-

Supreme Court Ruling

- About disparities between the percentages in the eligible population and the jury panel, the Supreme Court wrote:

“... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of [blacks]”

- The Supreme Court denied Robert Swain's appeal
-

Sampling from a Distribution

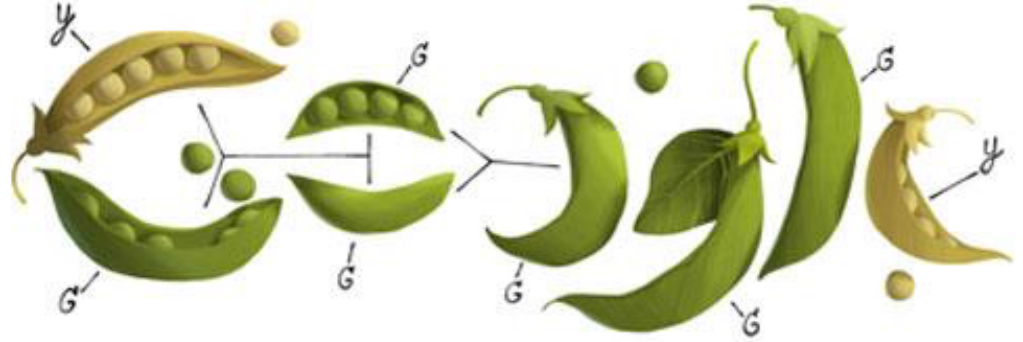
- Sample at random from a categorical distribution:

`sample_proportions(sample_size, pop_distribution)`

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

A Genetic Model

Gregor Mendel, 1822-1884



A Model

- Pea plants of a particular kind
 - Each one has either purple flowers or white flowers
 - Mendel's model:
 - Each plant is purple-flowering with chance 75%, regardless of the colors of the other plants
 - Question:
 - Is the model good, or not?
-

Choosing a Statistic

- Start with percent of purple-flowering plants in sample
 - If that percent is much larger or much smaller than 75, that is evidence against the model
 - ***Distance*** from 75 is the key
 - Statistic:
 - | sample percent of purple-flowering plants - 75 |
 - If the statistic is large, that is evidence against the model
-

Testing Hypotheses

Choosing One of Two Viewpoints

- Based on data
 - “Chocolate has no effect on cardiac disease.”
 - “Yes, it does.”
 - “This jury panel was selected at random from eligible jurors.”
 - “No, it has too many people with college degrees.”
-

Estimation

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.
-

If you have the entire population...

- Formulate a question you want to answer (a parameter of the population).
 - Visualize the data (the population).
 - Compute the answer.
 - Interpret the results, and explain them in language without statistical jargon.
-

If you don't...

- Formulate a question you want to answer (a parameter of the population).
 - Select a method of inference.
 - Visualize the data (the sample).
 - Calculate the statistic on your sample, then apply the method to estimate the population parameter.
 - Interpret the results, and explain them in language without statistical jargon.
-