Randomness and Sample Size

Students learn about random samples and statistical inference, as applied to the Animals Dataset. In the process, students get a light introduction to the role of sample size and the importance of statistical inference.

Prerequisites	Defining Table Functions
Relevant Standards OK CSTA NGSS CC-Math	*Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).*
Lesson Goals	Students will be able to... • Take random samples from a population • Understand the need for random samples • Understand the role of sample size
Student-facing Lesson Goals	• Let's explore how random sampling can be used with datasets.
Materials	• Computer for each student (or pair), with access to the internet • Student workbook, and something to write with
Preparation	• Lesson slides (Google Slides) • Make sure all materials have been gathered • Decide how students will be grouped in pairs
Supplemental Resources	
Language Table	See table below

Language Table

Types	Functions	Values
Number	`num-sqrt, num-sqr`	`4, -1.2, 2/3`
String	`string-repeat, string-contains`	`"hello", "91"`
Boolean	`==, <, <=, >=, string-equal`	`true, false`
Image	`triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized`	◍△◈
Table	`count, .row-n, .order-by, .filter, .build-column`	

Glossary

statistical inference :: using information from a sample to draw conclusions about the larger population from which the sample was taken

Do Now

Students should log into CPO open the Random Samples Starter File, and save a copy.

Flip the Script: Inference v. Probability

30 minutes

Overview

Statistical inference involves looking at a sample and trying to *infer something you don't know* about a larger population. This requires a sort of backwards reasoning, kind of like making a guess about a *cause*, based on the *effect* that we see. To better understand the process of going from the sample back to the population, it helps to understand the more straightforward process of going from the population to a sample. If the sample is random, we call this process Probability! In real life we typically don't know what's true for an entire population. But this probability thought-experiment will start with a larger population with *known* properties (such as the fact that half of the entire population are males). Then we'll see what kind of behavior we tend to see in random samples taken from that population.

Launch

Inference Reasons Backwards; Probability Reasons Forwards

One of the most useful tasks in Data Science is using sample data to *infer* (guess) what's true about the larger population from which the sample was taken. This process, called ***statistical inference***, is used to gain information in practically every field of study you can imagine: medicine, business, politics, history; even art! Early on, statisticians discovered that *random* samples almost always work best.

Suppose we want to make an educated guess about who the next US president will be. We can't ask everyone who they're voting for, so pollsters instead take a *sample* of Americans, and *generalize* the opinion of the sample to estimate how Americans as a whole feel. But choosing a sample can be tricky...

- Would it be problematic to only call voters who are registered Democrats? To only call voters under 25? To only call regular churchgoers? Why or why not?
- How could we choose a representative subset, or *sample* of American voters?
- Would it be problematic to only sample a handful of voters? What do we gain by taking a larger sample?

Before we infer something *unknown* about a population from a sample, we need to know what makes a "good" sample!

Sampling is a complicated issue. The main reason for doing inference is to guess about something that's *unknown* for the whole population. But a useful step along the way is to practice with situations where we happen to *know* what's true for the whole population. As an exercise, we can keep taking random samples from that population and see how close they tend to get us to the truth. Another discovery (besides the value of randomness) that statisticians made early on was something that's perfectly consistent with common sense: Larger samples are better than smaller ones, because they tend to get us closer to the truth about the whole population.

Let's see what happens if we switch from smaller to larger sample sizes, if we're taking a random sample of shelter animals to infer what's true about the larger population...

Investigate

The Animals Dataset we've been using is just one *sample* taken from a very large animal shelter. How much can we infer about the whole population of hundreds of animals, by looking at just this one sample?

- Divide the class into groups of 3-5 students.
- Have students open the Random Samples Starter File, and click "Run".
- Have students complete Sampling and Inference (Page 40), sharing their results and discussing with the group.

Synthesize

Have students share how much better their larger samples are at guessing the truth about the whole population.

Common Misconceptions

Larger populations need to be represented by larger sample sizes. In fact, the formulas that Data Scientists use to assess how good a job the sample does is only based on the *sample size*, not the population size.

how good a job the sample does is only based on the sample size, not the population size.

Going Deeper

If appropriate for your learning goes, this is a great place to include more rigorous statistics content about sample size.

Additional Exercises

- Project: Project: Food Habits
- Project: Project: Time-Use