Probability, Inference, and Sample Size

(Also available in CODAP)

Students explore sampling and probability as a mechanism for detecting patterns. After exploring this in a binary system (flipping a coin), they consider the role of sampling as it applies to relationships in a dataset.

Lesson Goals	Students will be able to… • Understand the connection between probability and inference • Understand the need for random samples • Understand the role of ***sample size*** • Take random samples from a population
Student-facing Lesson Goals	• Let's explore what random sampling has to do with seeing trends
Prerequisites	• Simple Data Types • Introduction to Data Science • Contracts: Making Tables and Displays • Bar and Pie Charts
Materials	• PDF of all Handouts and Page • Fair Coins Starter File • Expanded Animals Starter File • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• *Optional Project: Food Habits [rubric]* • *Optional Project: Time Use [rubric]* • Additional Printable Pages for Scaffolding and Practice

Glossary

bias :: prejudice in favor of or against one outcome, person, or group compared with another, usually in a way considered to be unfair.

null hypothesis :: the claim that there is no difference or relationship in the larger group(s) from which we sampled. For a single variable, this claims that a summary like the mean or proportions of a population equals a proposed value. For multiple variables, this claims that no relationship exists between those variables in the larger population

random sample :: a subset of individuals chosen from a larger set, such that each individual has an equal probability of being chosen

sample size :: the number of individuals (people or things) for which data is gathered in a study

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

How to Spot a Scam *20 minutes*

Overview

Students consider a classic randomness scenario: the probability that a coin will land on heads or tails. From a data science perspective, this can be flipped from a discussion of *probability* to one of *inference*. Specifically, "based on the number of coin flips we observed, what can we conclude about whether a coin is fair or not?"

Launch

A stranger on the street invites you to play a game of chance. They'll flip a coin repeatedly. On each flip, the stranger gives you a dollar if it comes up tails. If it comes up heads, you pay them a dollar.

"It's a pure game of chance", they tell you, "we each have equal odds of winning".

If you decide to play the game, how could you then decide if the stranger's coin is fair, or if the stranger is scamming you?

- For a fair coin, what are the chances of it landing heads? Tails?
  - *A fair coin has a 50% chance of landing heads and a 50% chance of landing tails.*
- How do you know if a coin is fair or not?
  - *Flip it! The more flips you make, the more accurately you can assess if it is fair or not.*

Investigate

A fair coin should land on "heads" about as often as it lands on "tails": half the time.

In general, we assume that in the long run, an ordinary coin will land on "heads" 50% of the time. Our assumption that there is no bias towards "heads" or "tails" is our ***null hypothesis***. A weighted coin, on the other hand, might be heavier on one side, creating a ***bias*** toward one side! And since we lose money on heads, we're worried about bias in favor of heads.

*So how do we test the **null hypothesis**?*

Open Fair Coins Starter File, and complete Finding the Trick Coin

Have students share back their sample results, and their predictions after 5 samples and then 20 samples.

Do any samples seem to undermine the null hypothesis?

In Statistics and Data Science, samples like these don't **prove** anything about the coins! Instead, they either *produce enough evidence to reject the null hypothesis, or fail to do so*. If the null hypothesis is actually false, larger samples give us a better chance of producing evidence to reject it.

The chances of getting "heads" from a fair coin three times in a row aren't too small: 1-in-8! Maybe it was just the luck of the draw, and the coin is still fair.

Should we suspect a scam if the stranger's coin flipped heads 10 times in a row? The probability of a fair coin getting no tails in 10 flips is $1/2^{10}$, or roughly 0.001. So here's what we'd have to say about our hypothesis test:

"If the coin was fair, the probability of getting so few (zero) tails in 10 flips is just 0.001."

Statisticians would say it slightly differently: "If the null hypothesis were true, then the probability of getting sample results at least as extreme as the ones observed is 0.001."

Going Deeper: p-value

Describing what the number 0.001 is talking about in the example above is a mouthful, because we're have to express it as an "If…then…" outcome.

Statisticians use formal language to express the probability of obtaining sample results at least as extreme as the ones observed, under the assumption that the null hypothesis is true for the population. They call this probability a "p-value", and typically report it as a decimal.

Most of us say…	Statisticians say…
"There's a 1-in-10 chance of this"	"The p-value is 0.1"
"There's a 1-in-100 chance of this"	"The p-value is 0.01"
"There's a 2-in-100 chance of this"	"The p-value is 0.02"
"There's a one-in-a million chance"	"The p-value is 0.000001"

But of course, there **is** a way. It's just… *incredibly unlikely*.

Common Misconceptions

Students may think that *any* sample from a fair coin should have an equal number of heads and tails outcomes. That's not true at all! A fair coin *might* land on "tails" three times in a row! The fact that this is possible doesn't mean it's *likely*. Landing on "tails" five times in a row? Still possible, but much less likely.

This is where arithmetic thinking and statistical thinking diverge: it's not a question of what is *possible*, but rather what is *probable or improbable*.

Synthesize

- Suppose we are rolling a 6-sided die. How could we tell if it's weighted or not?
    - *We could record how many times the die landed on each number after rolling many times. If the die is fair, we should see that it lands on each number approximately equally.*
- Could a coin come up "heads" twice in a row, and still be a fair coin? Why or why not? What about 10 times in a row? 20?
    - *The coin could be fair in all of these intsances! Heads 20 times in a row, however, is extremely unlikely.*
- What is the relationship between how weighted a coin is, and how many samples you need to figure it out?
    - *A fair coin should land on heads about 50% of the time.* If a coin has been designed to land on heads *100%* of the time, it wouldn't take long to figure out that something was up! A trick coin designed to come up heads *60%* of the time, however, would need a much larger sample to be detected. The smaller the bias, the larger the sample we need to see it. A small bias might be enough to guarantee that a casino turn a profit, and be virtually undetectable without a massive sample!

Probability v. Inference *35 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.

Launch

Probability reasons forwards.

Because we know that the chance of coming up heads each time for a "population" of flips of a fair coin is 0.5, we can do probability calculations like "the probability of getting all three heads in three coin flips is $0.5 \times 0.5 \times 0.5 = 0.125$." Likewise, we can say the probability of getting three of a kind in a randomly dealt set of five cards is 0.02.

"Based on what we know is true in the population, what's the chance of this or that happening in a sample?" *This is the kind of reasoning involved in probability.*

Inference reasons backwards.

In the coin-flip activity, we took samples of coin flips and used our knowledge about *chance* and *probability* to make ***inferences*** about whether the coin was fair or weighted.

In other words, we looked at sample results and used them to decide what to believe about the population of all flips of that coin: *was the overall chance of heads really 0.5?*

"Based on what we saw in our sample, what do we believe is true about the underlying population?" *This is the kind of reasoning involved in inference.*

Statistical inference is used to gain information in practically every field of study you can imagine: medicine, business, politics, history; even art!

Suppose we want to estimate what percentage of all Americans plan to vote for a certain candidate. We don't have time to ask every single person who they're voting for, so pollsters instead take a *sample* of Americans, and *infer* how all Americans feel based on the sample.

Just like our coin-flip, we can start out with the null hypothesis: assuming that the vote is split equally. Flipping a coin 10 times isn't enough to infer whether it's weighted, and polling 10 people isn't enough to convince us that one candidate is in the lead. *But if we survey enough people* we can be fairly confident in inferring something about the whole population.

Sample size matters!

- We're taking a survey of religions in our neighborhood. There's a Baptist church right down the street, so we could get a nice big sample by asking everyone there…right?
  - *Sampling this population would reveal to us that everyone in the neighborhood is Baptist, which might not be the case!*
- Taking a sample of whoever is nearby is called a *convenience sample*. Why is a convenience sample a problem in this example?
  - *Everyone at the church is Baptist, but the entire neighborhood might not be.*
- 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?
  - *Calling only certain segments of the population will not reveal the way an entire population will vote.*

Bad samples can be an accident - or malice!

When designing a survey or collecting data, Data Scientists need to make sure they are working hard to get a good, random sample that reflects the population. Lazy surveys can result in some really bad data! *But poor sampling can also happen when someone is trying to hide something, or to oppress or erase a group of people.*

- A teacher who wants the class to vote for a trip to the dinosaur museum might only call on the students who they know love dinosaurs, and then say "well, everyone I asked wanted that one!"
- A mayor who wants to claim that they ended homelessness could order census-takers to only talk to people in verified home addresses. Since homeless people don't typically have an address, the census would show no homeless people in the city!
- A city that is worried about childhood depression could survey children to ask about their mood… but only conduct the survey at an amusement park!

Can you think of other examples where biased sampling has been used - or could be used - to harm people?

Investigate

The main reason for doing inference is to guess about something that's *unknown* for the whole population.

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.

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?

Let's see what happens if we switch from smaller to larger sample sizes.

Divide the class into groups of 3-5 students.

- Open the Expanded Animals Starter File, save a copy and click "Run".
- Complete Sampling and Inference, sharing their results and discussing with the group.
- *Optional:* complete Predictions from Samples

Random samples help avoid bias, and larger samples get closer estimates of what's true for the whole population.

Common Misconceptions

Many people mistakenly believe that larger populations need to be represented by larger samples. 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.

Extension

In a statistics-focused class, or if appropriate for your learning goals, this is a great place to include more rigorous statistics content on sample size, sampling bias, etc.

Synthesize

- Were larger samples always better for guessing the truth about the whole population? If so, how much better?
- Why is taking a ***random sample*** important for avoiding bias in our analyses?

Project Options: Food Habits / Time Use

Optional Project: Food Habits *[rubric]* and *Optional Project:* Time Use *[rubric]* are both projects in which students gather data about their own lives and use what they've learned in the class so far to analyze it. These projects can be used as a mid-term or formative assessment, or as a capstone for a limited implementation of Bootstrap:Data Science. Both projects also require that students break down tasks and follow a timeline - either individually or in groups. Rubrics for assessing the projects are linked in the materials section at the top of the lesson.

(Based on the projects of the same name from IDS at UCLA)