Sampling from a Population

Javier E. Flores

January 25, 2019

Puzzle Analogy 2.0

Consider the following two (incomplete) puzzles:

Which gives a better sense of the whole picture? Why?

Puzzle Analogy 2.0

What about between these two?

Connection to Statistics

Each of these incomplete puzzles can be thought of as different **samples** from a **population**.

Each **sample** consists of a different collection of pieces taken from the entire set that makes up the puzzle (i.e. the **population**).

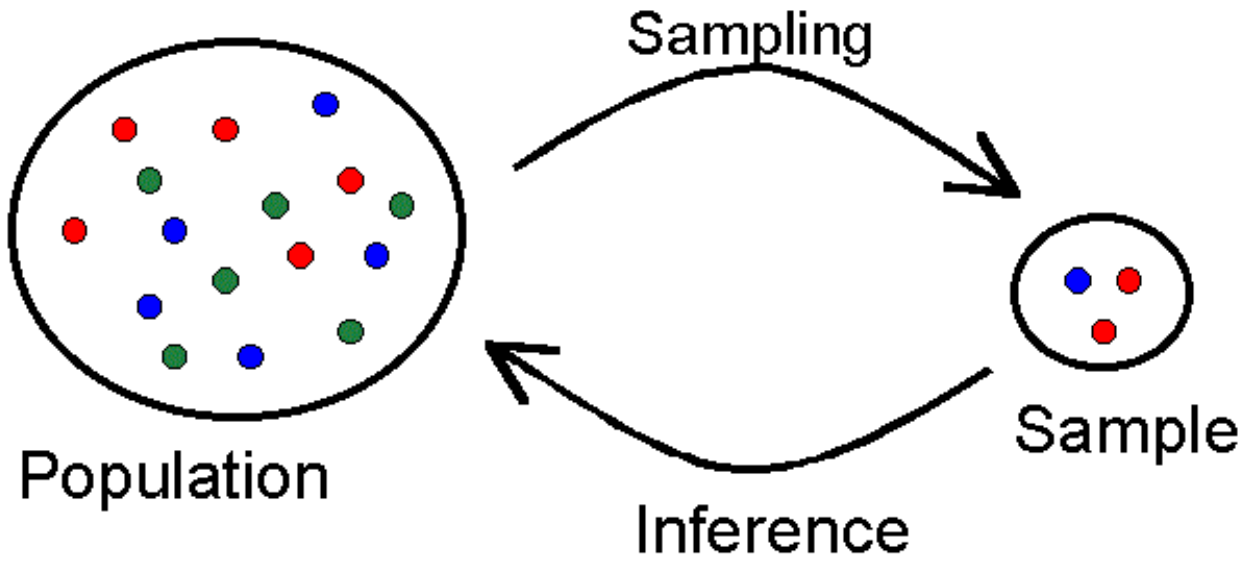
With each sample, we are able to make a guess about the puzzle's picture.

The accuracy of our guess is affected by how many pieces we have (i.e. the **sample size**) as well as how **representative** they are of the overall picture.

Connection to Statistics

More broadly speaking, statisticians use the information in a sample to make reliable statements (guesses) about a population.

This process of using data from a sample to reach conclusions about a population is called **statistical inference**.



Practice

In a study on hand washing, researchers in several cities across the United States pretended to comb their hair in public restrooms while observing whether or not people washed their hands after going to the bathroom.

They found that 85% of the 6,000 individuals they observed washed their hands.

What is the population in this study? What is the sample?

- The population is all people in the US that use public restrooms.
- The sample is the 6,000 people observed.

Why Sample?

You might be wondering...

"If we have access to the entire population, why do we use samples?"

Certainly, *If* (keyword) we had access to the entire population there would be no need to sample or draw inference.

But that's a huge "if".

- In the hand washing study, could they have observed all people in the US all times they used a public restroom?
- What if we wanted to try out a new cancer treatment? Could we (or should we) administer it to all cancer patients?

How should we sample?

Since we usually don't have access to the population, you might next be wondering...

"How is it that I should sample from a population so that I'm able to draw reliable conclusions?"

From our puzzle analogy, we know that **sample size** is important. Ideally, we want to obtain larger samples.

Related to our sample's size, we also know that our sample should be **representative** of the population of interest.

But how do we obtain a representative sample?

Activity: $$ Texas Sample 'Em $$

I've brought with me a bag containing exactly 100 poker chips. There are white chips (each $5) and red chips (each $10) in the bag. Your goal is to determine the total dollar amount contained in the bag.

Each group will sample 20 chips from the bag. Groups will sample (as I dictate) either one of two ways:

M1: Roll a dice. Select white chips from the bag until you have as many as three times your dice roll. Without looking, select additional chips as needed (to reach a total of 20).

M2: Without looking, pick out 20 chips from the bag one by one. After each selection, replace the chip.

Each group will then:

1) Enter data into Minitab. Create two variables: SelectedChip and ChipValue.
2) Discuss how these data may be used to estimate the total value of the bag, and report an estimate. (Hint: Ideally, samples should be representative of the population of interest).

Predictions?

How can we get an estimate of the bag's total value?

Which sampling method will provide estimates closer to the true total?

Simple Random Sampling

Without surprise to the well-trained statistician, the second sampling method (M2) generally gave better estimates.

This sampling method, commonly referred to as **simple random sampling**, ensures that each unit of the population (i.e. every individual chip) has an equal chance of being selected, regardless of what's already been chosen. As a result, the obtained sample is more representative of the whole population.

On the other hand, the first sampling method was **biased**. This method biased the sampling of white chips.

Bias and Variability

Aside from bias, **variability** may be another reason that a sample is not representative of a population.

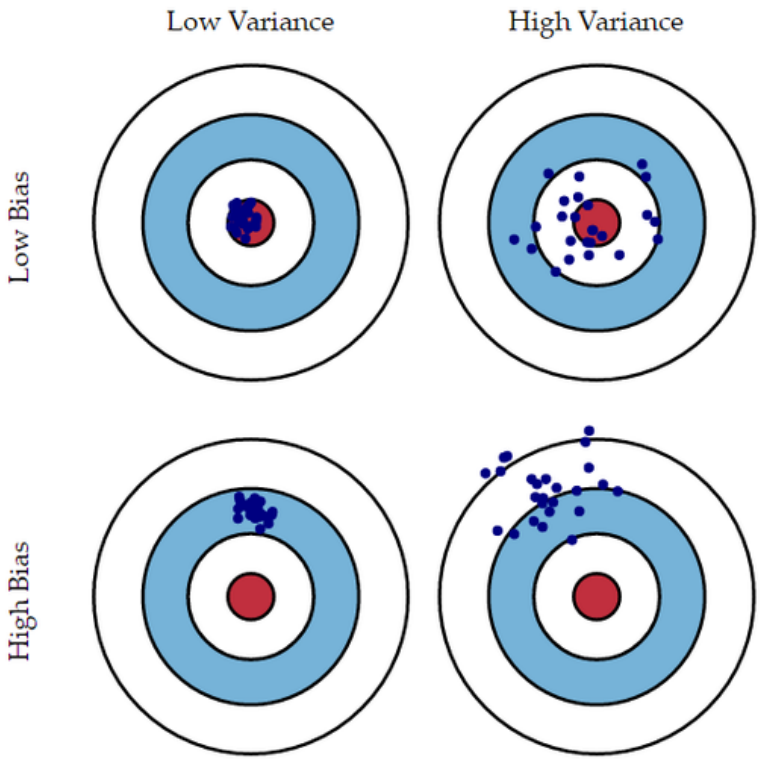
To see this, think about what would happen if we had five different-valued chips as opposed to two.

- Assuming you could draw only 10 chips, do you think it's likely you'd draw representative sample?
- What if you drew 50 chips instead?

Bias and Variability

The variability of a sample decreases as your sample size (typically denoted by the letter "n") increases.

However, the bias of a sample is not improved by increasing the sample size.



1936 Presidential Election

In 1936, Franklin Roosevelt was up for reelection and running against the Republican nominee, Alfred Landon.

Being held only two years after the worst point in the Great Depression (the unemployment rate was as high as 25%), political discourse centered around the depression and how the country might be pulled out of it.

Landon and Roosevelt had differing views on what a solution might entail, and with Roosevelt as the incumbent, it would be reasonable to expect a change in leadership.

1936 Presidential Election

Enter the *Literary Digest*.

The *Literary Digest* magazine had, since 1916, successfully predicted the winners of each presidential election.

With this record of success, the *Literary Digest* was fairly confident in their prediction for the outcome of the election at hand.

Far from a random guess, their prediction - a landslide victory for Landon - was based on a (huge!) sample of 2.4 million people.

1936 Presidential Election

Turns out, the *Literary Digest* prediction was wrong. Big time.

Rather than the predicted landslide win for Landon, the election concluded with an overwhelming 62% - 38% Roosevelt victory.

So what happened? How could a 2.4 million sample taken by a magazine with a track record of successful prediction be so very wrong?

1936 Presidential Election

Selection Bias

- The *Literary Digest* mailed 10 million questionnaires to addresses they gathered from telephone directories, club membership lists, and their own subscribers.
- This disproportionately screened out the poor - only one in four households owned a telephone at the time, and club members were typically upper class.

Non-response Bias

- Earlier I mentioned that the prediction was based on a 2.4 million sample. What I didn't mention was that 10 million questionnaires were sent out - which translates to a measly 24% response rate!
- With such a low response rate, it's hard to make the claim that the respondents were representative of all those individuals polled.
- Compounded with the selection bias we discussed above, it's clear to see how the *Literary Digest* made such a bad prediction.

WWII Fighter Planes

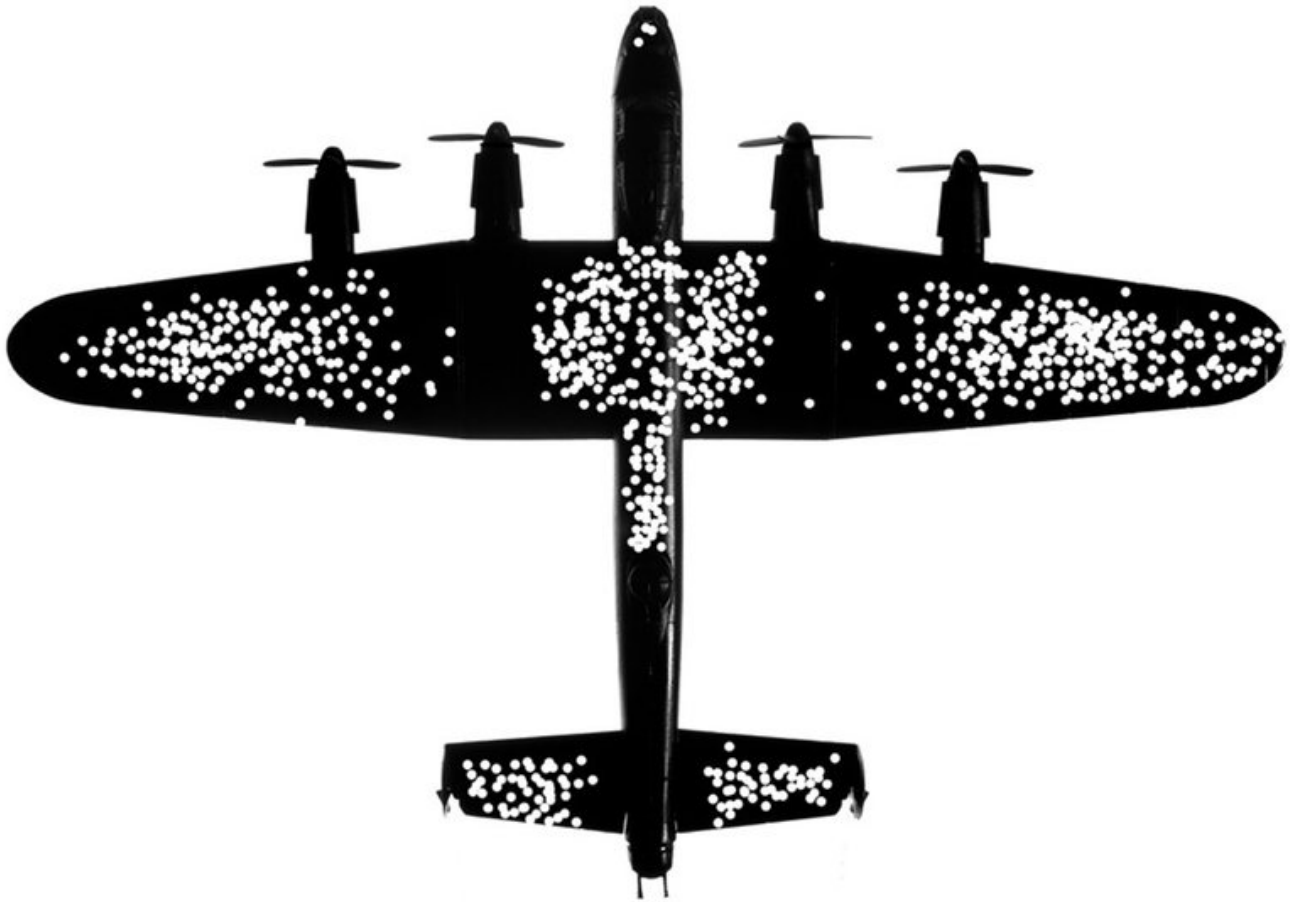
Not too long after the *Literary Digest* fiasco, statisticians had a chance for redemption.

The Statistical Research Group (SRG) was a classified program that was operational during WWII. (Think the Manhattan Project, but equations were developed and not nuclear bombs)

Among several other contributions (e.g. optimizing ammunition management, rocket propellant composition, etc.) SRG was famously tasked with determining an optimum in terms of armoring planes.

Based off of data from Allied aircraft returning from skirmishes across Europe, SRG believed the most efficient solution was to concentrate armor on those areas getting hit the most.

Do you agree?



WWII Fighter Planes

Survival Bias

- Abraham Wald, an SRG statistician who had immigrated to the U.S. to escape Nazi persecution, had a brilliant realization: The only aircraft with available data were those that survived combat and returned.
- Therefore it would not make sense to concentrate armor where the returning aircraft were hit, but rather where they weren't!
- Needless to say, Wald's recommendations were quickly put into effect, saving countless lives and further highlighting the power of statistical thinking.

Other Examples

For a more recent example of sampling bias, I'd encourage you all to read about the NFL CTE study.

Some other sources and examples of bias include (but are not limited to):

- Social Desirability Bias, or the tendency for respondents to answer in ways that make themselves look good.
- Confirmation Bias, or the tendency to search for, interpret, favor and/or recall information in a way that confirms preexisting beliefs.
- Leading Questions, in which questions are structured to influence the response.

Practice

With your group, read through the following scenarios and determine the target population of each sample. Then decide whether there is any bias. If so, explain.

1. A podcast host is curious about how much their listeners enjoy the show. As a result, the host ask their listeners to visit the podcast website and participate in a poll. Of the 200 respondents, 89% said that they "love" the show.

2. A senator polled 100 people randomly sampled from the phone book in order to determine how people in their state felt about internet privacy.

3. High school counselors interested in determining student smoking habits personally asked every student whether they smoked.

Wrap-Up

Right now, you should...

- Understand the difference between populations and samples.
- Know that *statistical inference* is the process of using sample information to learn about a population.
- Recognize the importance of random sampling
- Identify the target population of a sample and any potential source of bias.

These notes cover Section 1.2 of the textbook. Please read through the section and its examples along with any links provided in this lecture.

Before you go...

I know you all REALLY wanted to see the full picture of my dog from the start of this lecture, so I've included it here ☺