UNIVERSITY OF WATERLOO

Faculty of Mathematics
Waterloo, Ontario N2L 3G1

Centre for Education in
Mathematics and Computing

Grade 6 Math Circles

March 24, 2021

Intro to Statistics

Introduction

Statistics provides a set of tools that help us learn about the world around us. In particular, we use different statistical tools to help us collect data, organise data, analyse data, and then communicate our findings. Statistics is everywhere—games use probability, science uses statistical studies, and anything you see in the news about surveys, polls, or censuses are all statistics!

In this lesson, we'll be focusing on some of the basic ideas that make up the foundation of how we navigate statistics.

Populations & Samples

Often, we'll find ourselves in a position where we want to find out more about a group of individuals. (These individuals could be any person, place, or thing!) This entire group of individuals is called the **population**. Depending on the context, a population ranges in size—for example, you could be collecting information about students in your class, students in your school division, or every student in Canada.

Sometimes, we collect data from every individual in a population. This type of survey is called a **census**.

- **population**—the entire group that we want information about
- **census**—a survey that collects information from every individual in a population

Example 1: As this week's class captain, it's your job to decide on a game for Friday's gym class, and you want to ask everyone in the class what their top choice is. Here, your population is every student in your class, and you are conducting a census, since you're collecting information from every student in your class.

A real-life example of a large-scale census is the census run by the Canadian Government every five years to learn about Canada's population. This census is sent out to every household in the country, and it takes a huge amount of people, time, and money to carry out.

When your population is small, it's easy to conduct a census and be certain of the answer to your question. When the group gets bigger or less clearly-defined, however, it becomes more practical to collect information from only *part* of the group instead. This part of the group is a **sample** of the population. Then, we can use the information we have to make a best guess about the group as a whole.

- **sample**—a subset of individuals selected from the population that we actually collect data from

Samples are useful because collecting data from an entire population is expensive, time-consuming, and challenging. It's often impossible to maintain a perfect list of every individual in a population, and it's even harder to access and collect data from all of them!

Example 2: When testing COVID-19 vaccines prior to distribution, companies want to get an idea of how the vaccine might affect everyone in the **population**—that is, everyone who would take the vaccine. However, they obviously can't test it on every person on the planet! Instead, they carry out testing on a group of volunteers, which is a **sample** of the population.

Example 3: To make sure their products are up to par, factories carry out quality control testing throughout production. In this case, they want to learn about the entire **population**, which is everything that rolls off of the production line. However, they only check a portion of their products, a **sample**, to get a decent idea of what's going on with the rest of the batch. It's just infeasible to examine every item for imperfections—there are too many!

Statistics is using data from a sample to make a 'best guess' about the population.

Question 1: Curious about the nutritional value of the lunches in your grade, you slip a survey into every fifth locker in your class's hallway. What is the population and sample in this scenario?

Solution: The population is all of the lunches in your grade, and the sample is the lunches of students assigned to the lockers you distributed a survey to.

Sample Surveys

Whereas a census is a special kind of survey that collects information from the entire population, a **sample survey** is a survey that only collects information from a sample. Sample surveys are carefully designed to collect information that is helpful in constructing a best guess about the population as a whole.

Example 4: It's your turn to be class captain again, but this time you've forgotten to do a survey of your entire class! Instead, before gym, you ask four students sitting close to you about their top choice for the game. You use the most popular answer from this **sample survey** as a best guess for what the population—the whole class—would want to play.

In Examples 2 & 3, the processes of collecting information from only a sample of the population are both sample surveys as well!

Variables

There are different kinds of characteristics of a group that we can try to collect information about. We might be interested in the age of individuals across a population, or people's favourite colours. We call these characteristics **variables**.

We divide variables up into two main types: **quantitative** and **categorical**.

Quantatative variables are characteristics that take on numerical values that we can directly calculate with. Some examples of quantitative variables are height and length of time. It makes sense, for instance, to calculate an average height or length of time within a group.

Categorical variables are characteristics where we can place each individual into a label, or category. For example, a person's favourite colour puts them into a category, and if we survey several people about their favourite colour, then we'll have some number of categories with some number of individuals in each. Other examples of categorical variables are postal code, brand, and species. Unlike quantitative variables, it doesn't make sense to perform calculations on the variables—an "average" species doesn't mean much at all!

- **quantitative variable**—a characteristic of an individual that takes on a numerical value that makes sense to do math with
- **categorical variable**—a characteristic of an individual that places the individual into a category based on some quality

The type of variable that we are aiming to collect information about affects how we choose to analyse and communicate our data.

Activity: Follow this link: `https://www.geogebra.org/m/ddewd9we`. Sort the variables as either quantitative or categorical.

When it comes to quantitative variables, one frequent calculation that we will do is called finding the **mean**. The mean is one kind of average. To find it, you add up all the numbers in a list, and then divide by the number of numbers you added together.

Example 5: We can find the mean of a list of numbers: 3, 3, 5, 6, 9, 10. To do this, we'll add up all of the numbers, and then divide by 6, the number of numbers.
$\frac{3+3+5+6+9+10}{6} = \frac{36}{6} = 6$.
Thus, we have that the mean of the numbers 3, 3, 5, 6, 9, 10 is 6.

Question 2: Find the mean number of minutes that it takes people to solve a puzzle given to them: 1, 2, 3, 4, 4, 7, 7, 8, 9.

Solution: $\frac{1+2+3+4+4+7+7+8+9}{9} = \frac{45}{9} = 5$ minutes.

When working with categorical variables, we often want a different measure: the **proportion** of individuals in our group with a certain characteristic, or in a certain category. The proportion of individuals in a category is the number of individuals in that category, divided by the number of individuals in the whole group. It can be expressed as a fraction, decimal, or percentage.

Example 6: A survey of your class reveals that out of 20 students, 10 prefer dogs, 8 prefer cats, and 2 don't like either. The proportion of dog-lovers is $\frac{10}{20} = \frac{1}{2} = 0.5 = 50\%$. The proportion of cat-lovers is $\frac{8}{20} = \frac{2}{5} = 0.4 = 40\%$. The proportion of students that like neither is $\frac{2}{20} = \frac{1}{10} = 0.1 = 10\%$.

Inference: Putting It All Together

The goal of sample surveys is to collect information from a sample, and then use that information to draw, or **infer**, conclusions about the entire population.

Example 7: If a city wants to know what proportion of their residents has a peanut allergy, they could try to answer their question by conducting a census (surveying everybody in the population). However, this would be very difficult—cities have a lot of people!

A much more practical approach is to estimate the proportion of the residents with a peanut allergy by conducting a **sample survey**. You would then know the proportion of your sample with a peanut allergy, and you can use this as a reasonable estimate for the entire city. We can also note that "whether a person has a peanut allergy" is a **categorical variable**.

If you survey 1000 people about their allergies and 20 have a peanut allergy, there's a pretty good chance that somewhere around $\frac{20}{1000} = 2\%$ of your city has a peanut allergy too!

Question 3: You're planning a pizza party for the Grade 6's, so you want to get a good idea of what toppings you should order. The options are pepperoni, pineapple, and cheese. Each pizza serves 5 students, there are 20 students in each class, and there are 5 classes in the grade. To estimate the popularity of each topping, you decide to survey your class at lunch: 5 people choose pepperoni, 6 people choose pineapple, and 9 people choose cheese.

What are your population and sample? Are we examining a quantitative or categorical variable? Based on this sample survey, how many of each pizza should you order?

Video Solution: `https://youtu.be/gAYgVqMOk78`

Question 4: To choose a movie to watch during the pizza party, you ask some classmates to rate three different movies, each out of 5 stars. Not everyone replies to your text, but that's alright! You'll choose the movie with the highest mean (average) rating from the responses you got. The responses are summarised below:
Movie A: 1, 2, 4, 1, 2, 2
Movie B: 5, 3, 5, 5, 4, 4, 5, 5
Movie C: 3, 5, 5, 2, 2, 3, 5, 1
Are we examining a quantitative or categorical variable? What is the mean rating of each movie? Based on this sample survey, which movie should you watch?

Video Solution: `https://youtu.be/klKC5L4H6Jk`

It's important to remember that sample surveys are done with *samples*, which means that you still only get part of the picture to work with. That said, using information about variables from a sample survey is an incredibly useful technique that statisticians use every day to draw conclusions about populations, and sample surveys are more often than not the only reasonable way to collect information about a population.