UNIVERSITY OF WATERLOO

Faculty of Mathematics
Waterloo, Ontario N2L 3G1

Centre for Education in
Mathematics and Computing

Grade 7/8 Math Circles

February 17, 2021

Random Sampling

Introduction

Random samples are an essential tool in every statistician's toolkit, and they're an important building block in designing and carrying out experiments in every scientific field. In this week's lesson, we're going to learn about what they are, why they're used, and how to generate them in different ways!

Populations & Samples

The following vocabulary helps statisticians clearly communicate about the data and numbers they are working with. They'll also come in useful later in this lesson!

- **population**—the entire group that we want information about
- **sample**—a subset of individual units selected from the population that we actually collect data from

Example: 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: 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: 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.

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

- **parameter**—a number that describes a characteristic of the entire *population*
- **statistic**—a number that describes a characteristic of a *sample*

Example: If a city wants to know what fraction of their residents has a peanut allergy, they could try to obtain a *parameter* by asking the whole *population* about their allergies. However, this would be very difficult—cities have a lot of people, and in the end, if you can't keep track of getting exactly one answer from every individual, it wouldn't be the correct number describing the peanut allergies of the city's populuation anyway! A much more practical approach is to estimate the fraction of the residents with a peanut allergy by collecting a *sample* of people to survey. You would then know the proportion of your *sample* with a peanut allergy—which is your *statistic*—which you can then apply as a reasonable estimate for the entire city. 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: You're planning a pizza party for the Grade 7's & 8'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 10 classes between the two grades. 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. How many of each pizza should you order? How is this number connected to the statistics you collected from your sample?

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

Random Samples

As highlighted above, it's extremely difficult to collect data from every single unit in a population—thus, we collect information from a sample of the population to make educated guesses about parameters.

When selecting samples, it's important to try to create a *representative* sample. A representative sample should proportionally reflect the overall characteristics of the population.

For example, if you wanted to estimate the mean height of students at your school, it would make no sense to only measure the heights of people in your class, since students from only one grade wouldn't be *representative* of the variety of different ages in an elementary school. A classroom of eighth-graders is a lot taller than a classroom of first-graders! To take a more representative sample, we can explore some common sampling methods below:

- **random sample**—a sample of a population generated using random chance

It's good practice to use *random samples* when collecting data to approximate a population. It's not foolproof, but it's the key to getting reasonably close to a representative sample without having all the information about a population.

- **simple random sample**—a random sample chosen so that every individual has an equal chance of being selected

The result of simple random sampling is the same as if you placed the name of every individual in a hat and picked out the number of names you want without looking. It's also equivalent to writing down every possible sample of individuals you could select, and then randomly choosing one. This also means that every possible sample of the desired size has an equal chance of being selected!

Example: You want to generate a simple random sample of 40 students from your school for your science project. One way to do this would be to collect an attendance sheet from every homeroom, assign every name its own number, and then randomly select 40 numbers that then correspond to the names in your sample.

You can explore a simple random sample generator here: `http://digitalfirst.bfwpub.com/stats_applet/stats_applet_13_srs.html`, and a random number generator here: `https://www.calculatorsoup.com/calculators/statistics/random-number-generator.php`. If you type "random number generator" into Google, you can even find a built-in generator!

- **stratified random sample**—a random sample chosen proportionally based on defined strata in your population
- **strata**—groups of similar individuals in a population

Strata are categories in your population that you define that are important to be proportionally represented in your random sample, such as age, family income, or right/left-handedness. To get a stratified random sample, you start by categorising your population into those *strata*—groups of similar individuals—and choose a separate simple random sample from each *stratum* (singular strata). You can then combine these simple random samples together to get your stratified random sample!

For example, in a survey about everyone's favourite movies and books, different age groups will probably have different interests. In a case like that, it would be useful to be able to have proportional representation of each age group in your sample, so that individual age groups aren't overrepresented or underrepresented (that is to say, we want the sample to be representative) when collecting information for your final dataset. For instance, if you used a simple random sample, there's a chance that your sample could completely leave out people younger than 30, which would totally throw off your data!

This method is useful if you have information about the size of the strata in your population—so, if you know how many students are in each grade, you could generate a stratified random sample based on grade level. The way you define your strata is up to you!

Example: There are 800 students at Math Circles Elementary School, 100 in each grade. Since students in each grade will be similar in height, you define your strata as each grade from 1 to 8. You want to collect a stratified random sample of 40 students to estimate the mean height of students at your school.

$\frac{100}{800} = \frac{1}{8}$

$\frac{1}{8} \times 40 = 5$ students from each grade.

Question: Since each of the grades, which are our *strata*, are the same size in this example, how else could we have arrived at the answer of 5 students from each grade?

Solution: In this case, we could have simply used division! The calculation $40 \div 8 = 5$ gives us the same answer.

Question: To learn more about how well French lessons delivered by readings vs. videos work for your grade, you decide to survey your class so you can present the results to your teacher. You know that some of your peers transferred from a French Immersion school last year, so you suspect that over-representing or under-representing that group would probably make the results innacurate. In your grade, 20 students transferred from French Immersion, and 80 students are from your current school. There are 20 students in your class, and 3 of them transferred from French Immersion. How many students from each group should be in a stratified sample size of 20? To keep your sample representative by these strata, how big can your sample size actually be?

Video solution: `https://youtu.be/eYakm0wO59o`

- **cluster random sample**—a random sample generated by first splitting the entire population into representative groups (clusters), and then randomly selecting clusters to be part of the final sample

Depending on usage, sometimes, all the individuals in the cluster are included in the sample; other times, a simple random sample is taken within each chosen cluster. Additionally, depending on context, some experiments might call for one cluster, whereas others might include multiple clusters in their sample.

Clusters random sampling is most effective when the clusters look like the population. For instance, if a population is all students in a grade, individual classrooms would work great as clusters! Similarly, individual neighbourhoods would likely be effective clusters for a city. Individual regions in a province might not work so well, however—think of how different cities are from rural areas!

This method is often used for practical reasons, such as saving money or time. Individual classrooms are helpful clusters, because students are much easier to track down in a big school if they're from one classroom. Houses are close together in neighbourhoods, so it's much more straightforward to treat a neighbourhood as a representative sample than to run around the whole city.

Example: A school district is looking to partner with a new classroom furniture company, but they want to test how students like the company's chairs before signing a deal. (What is the population?) The district has 35 schools all around the region, so they decide to treat each school as a cluster to make the test run simpler. They then take a simple random

sample of the 35 schools to choose 3 of the lucky schools that get to try out the new chairs.

(The population is all of the students in the school district, and the sample is all of the students in the 3 randomly-selected schools!)

Example: At Math Circles Elementary, there are exactly 20 students in every homeroom. To create clusters, you could gather attendance sheets from each homeroom and assign numbers to each name based on which row they're in on the attendance sheet, from 1 to 20. Then, randomly select a number from from 1 to 20, and all the corresponding names are in your sample! This should be representative, because you're selecting one student from each class, and alphabetical order of names should have no effect on height.

Question: The principal of your school is enlisting your help to distribute a survey about phone usage at Math Circles Elementary to 60 students during an assembly tomorrow with students in Grades 6, 7, and 8. A diagram of the gym is below.
How would you select 60 students to complete the survey using:

(a) simple random sampling?

(b) stratified random sampling?

(c) cluster random sampling?

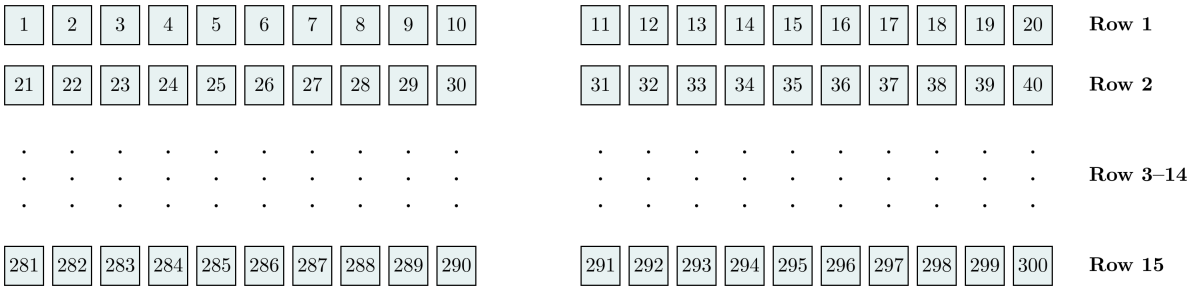

Figure 1: Grade 6: Rows 1–5; Grade 7: Rows 6–10; Grade 8: Rows 11–15

Video solution: `https://youtu.be/GACXRx9U4pI`
Note: There are many varying solutions other than those outlined in the video!

It's important to understand the difference between *strata* and *clusters*! We want each stratum to contain similar individuals and for there to be differences between the strata—"similar within, but different between." With clusters, we want each cluster to look like a smaller copy of the population and for clusters to look like each other as well—"different within, but similar between."