Principles of Data Collection

Lesson objectives

- I can distinguish between a population and a sample
- I understand why sampling a population can give information about that population
- I understand that when sampling data the results can vary
- I can sample data in various ways

1.1

Lesson objectives | Teachers' notes | Lesson notes

MHR Page 210 #s 1 - 8

Warm up

What if your job were to test the quality of french fries? At a french fry factory, quality control experts measure and assess batches of fries. Why do you think this is done?

Test for crispiness / cooked properly.
Check for consistent thickness of fries.

In the months leading up to an election, media outlets often predict which party will get the most votes from the **population**. They use the results published by polling firms, which contact a **sample** of the registered voters to try to predict who will win.

How do the polling firms choose who they are going to contact?

Stratified sample based upon the population. Keep ratios the same based upon gender, age range, income brackets, location etc.

Definitions

Population

- All the individuals in a group that is being studied

Sample

- A group of items or people selected from the population

Variability (in samples)

- Shows how samples are different from each other
- The more similar the samples are to each other, the lower the variability and the more accurately the samples represent the population

> For a sample to be representative of a population, each member of the population must have an equally likely chance of being selected in the sample. So, selection for the sample must be random. If the sample is not random, it is biased and not as reliable. There are different types of sampling methods. Some are better than others.

Type of Sample	Example
Simple Random • randomly choose a specific number of people • examples: stratified samples and systematic samples	Put all the names in a population into a hat and draw one or several names. Each person has an equal chance of being chosen.
Systematic • put the population in an ordered list and choose people at regular intervals	Order all the patients of a doctor in some way (e.g., alphabetically) and choose one randomly. Select the rest of the data at regular intervals from the original starting point (e.g., every tenth name after the original).
Stratified • divide the sample into groups with the same proportions as those groups in the population • time- and cost-efficient to conduct	Survey factory employees about new safety initiatives. There are 1000 employees in the factory, of which 633 are women and 367 are men. Randomly select 63 women and 37 men to take the survey.
Cluster • divide the population into groups, randomly choose a number of the groups, and sample each member of the chosen groups	Survey Little League Canada baseball players. Randomly select five districts in each province and give the survey to every player in those districts.
Multistage • divide the population into a hierarchy and choose a random sample at each level	Conduct an employee wellness survey by randomly selecting 10 stores. Randomly select three departments in each store, and randomly select 10 employees in each of those departments.
Convenience • choose individuals from the population who are easy to access • can yield unreliable results since it inadvertently omits large portions of the population • often very inexpensive to conduct	To get the public's input on a new pet by-law, a local politician goes to a local park and asks people their opinion.
Voluntary • allow participants to choose whether or not to participate • often the only people who respond are either heavily in favour or heavily against what the survey is about	Conduct an online poll asking people whether banning junk food in schools will fight obesity.

Example

Types of Samples

For each situation, identify the type of sample and discuss whether each member of the population is equally likely to be chosen.

a) A teacher wishes to get feedback from the class about a recent presentation. He plans to draw five students' names out of a hat. All 30 students' names will be in the hat.

b) A telephone company wants to determine whether a fitness centre would be used by its 3000 employees. The company plans to survey 300 employees by interviewing every tenth person on the payroll list.

c) A chain store is trying to decide whether to open a store in Huntsville, Ontario. The company decides to survey 25% of the population of Huntsville and three nearby towns. The table shows the population of each location.

Location	Population
Huntsville	19 056
Kearney	841
Emsdale	2 317
McMurrich	779

d) A market research company mails surveys to all of the adult residents in a town. The survey asks about brands of consumer products. The residents are asked to mail back their responses in a prepaid envelope.

e) A reporter stops people on a downtown street to ask what they think of the city's waterfront.

f) Researchers want to investigate the use of pesticides by apple farmers in Ontario. They divide the province into 10 sections and choose five sections at random. They sample all farms within the five sections.

g) The province wants to randomly choose 250 students. It randomly selects five school boards from the 72 in Ontario. Then it randomly selects five schools in each of those boards. Finally, it randomly chooses 50 students in each of those schools.

a) Simple random sample: includes all possible students, so each student has an equally likely chance of being selected.

b) Systematic sample: employees are ordered and every 10th person is chosen. If the first person chosen is selected randomly, then everyone has an equally likely chance of being chosen.

c) Stratified sample: as long as 25% of each town is selected. Everyone has an equally likely chance of being selected if they are chosen randomly.

d) Voluntary response: residents have the choice to respond or not. In these cases, those who respond tend to have strong feelings on the issue(s). The entire population is not represented in the sample.

e) Convenience: as the data is collected on the street it is easily accessible. The entire population is not represented, since only people within that area are surveyed.

f) Cluster: population is divided into groups, sections are selected randomly, all farms within the section are selected. As the sections are selected randomly, all farms have an equally likely chance or being chosen. This is used when covering large areas as it is more efficient.

g) Multistage: since this is a hierachical structure and selections are made randomly at each tier. Students are chosen randomly, so they have an equally likely chance of being selected.

Your Turn

In each case, identify the type of sample.

a) You want to find out if your town is in favour of starting a composting pickup service. You ask everyone on your street.

b) A university is polling its students. It selects 200 students at random in the same proportions as the enrollment in each department.

c) There are 139 swim clubs in Ontario. Swim Ontario conducts a survey to vote on its new logo. The organization randomly selects 10 swim clubs and surveys every member in each of those clubs.

d) A coach puts the names of all the basketball players into a hat and draws one name for a free basketball.

e) A questionnaire is sent to every ninth person on an alphabetical list of a store's credit card customers. The first person chosen from the list is picked randomly.

f) The student council invites all students to provide ideas for activities.

g) A marketing firm wants to collect information on certain products in a city of 800 000 people. The researchers randomly select 10 neighbourhoods. In each neighbourhood they randomly select five streets, and on each street they randomly select 10 households.

a) Convenience: as the data is collected on the street it is easily accessible. The entire population is not represented, since only people within that area are surveyed.

b) Stratified sample: the 200 students are selected in the same proportions to the numbers in each department. Everyone has an equally likely chance of being selected if they are chosen randomly.

c) Cluster: population is divided into groups, sections are selected randomly, all farms within the section are selected. As the sections are selected randomly, all farms have an equally likely chance or being chosen. This is used when covering large areas as it is more efficient.

d) Simple random sample: includes all possible players, so each player has an equally likely chance of being selected.

e) Systematic sample: employees are ordered (alphabetical) and every 9th person is chosen. If the first person chosen is selected randomly, then everyone has an equally likely chance of being chosen.

f) Voluntary response: students have the choice to respond or not. In these cases, those who respond tend to have strong feelings on the issue(s). The entire population is not represented in the sample.

g) Multistage: since this is a hierachical structure and selections are made randomly at each tier. Students are chosen randomly, so they have an equally likely chance of being selected.

Key Concepts

- A population is the entire group of a set of people or things. A sample is a smaller portion of that population.
- You can learn a lot about a population by examining samples of that population, as long as all members of the population are equally likely to be part of the sample.
- When multiple samples are taken from the same population, they are different from each other. This is called variability of samples. The smaller the differences in the samples, the more likely the sample closely represents the population.
- There are many types of sampling techniques. Some types of samples work better in certain situations. A good sample is random, and each person in the population has an equally likely chance to be chosen.

R1. What are the differences and similarities between a cluster sample and a multistage sample?

They both divide the population into groups, but they do it differently. For a cluster sample, a random number of groups are chosen and each member of the group is sampled. For multistage sampling the group is further subdivided by creating a hierachy and randomly sampling at each level.

R2. If a population is either heavily in favour or heavily against a certain topic, the sample size for a survey does not need to be as large as it would if the opinions were mixed. Why might this be?

With a smaller sample size it will be evident if the population is biased in either way.