MTH208: Worksheet 11

Data Collection

"Data" is one of the keys ingredients for any data analysis or data science study. It is then critical to understand

- what exactly is data
- what does it represent
- how was it collected
- what are its limitations

There are two categories of data sources:

- **Primary Data:** Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source - where the data originally originates from and are regarded as the best kind of data in research.
    - e.g.: Interviews, survey, case-studies, questionnaires, etc
- **Secondary Data:** Secondary data are data that has previously been collected by someone else but has been made available for use for others. They were probably previously primary data, but when they are reused by a third party, they become secondary.
    - books, diaries, letters, government records, etc.

For details of these data collection methods, read this reference material.

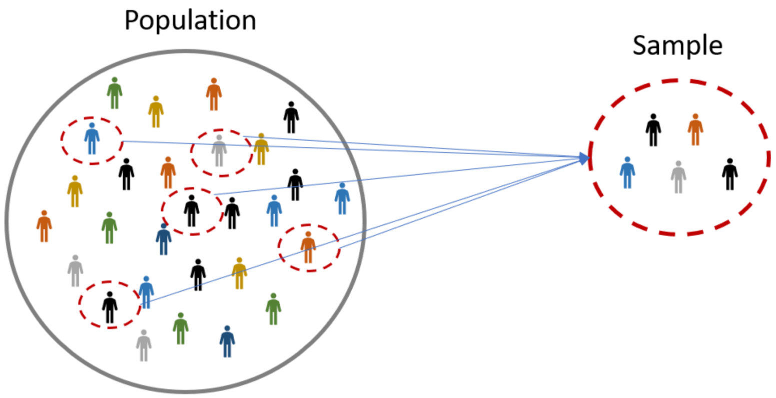
Population and samples

- In statistics, we are interested in obtaining information about a total collection of elements, which we refer to *population*.
- **Example:** suppose we are interested in the average order price and average number of orders in Hall 4 canteen in 2022. An exhaustive method of obtaining the data would be to record the data on each order placed in Hall 4 canteen in each day of 2022.

  This is an example of a finite size population.
- **Example:** suppose that your local grocery store uses a weighing machine to sell rations. Naturally, even the best weighing machines have some small errors in them. For a 1kg measurement, you are interested in knowing what is the true measurement of the weighing machine. Thus your population is the characteristic of the weighing machine in the shop.

This is an example of an infinite size population, since you can keep using the weighing machine infinite times.

- We try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called *sample*. Sampling is often more economical, accessible to the researchers, and practical and effective.
- The sample should be collected in such a way that it is representative of the underlying population. In the above example, if one collects information on sale of Hall 4 Canteen only on institute holidays, the sample data so collected will not be representative of the entire population.



Properties of representative samples:

- If a sample is representative of a population, then statistics calculated from sample data will be close to corresponding values from the population.
- Samples contain less information than full populations, so estimates from samples about population quantities always involve some uncertainty.
- *Random sampling*, in which every potential sample of a given size has the same chance of being selected, is one of the best way to obtain a representative sample.
- Thus, it is important to understand both how to conduct a random sample in practice and the properties of random samples.
- However, it often impossible or impractical to obtain a random sample.

Population parameters

After defining a population, we typically also understand the *distribution* of the population. There are two important population parameters are the population mean, μ, and the population variance, σ^2.

Finite population

Consider the Hall 4 example. Our population is

- the number of orders in Hall 4 in 2022

Let $x_1, x_2, \dots, x_N$ be the number of orders of Hall 4 in 2022, and each of these is the actual number of orders on the i th day. Then the population average number of orders an the population variance of the number of orders are:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} \qquad \text{and} \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Infinite population

Consider the shop weighing machine example. Then our population is the characteristic of the weighing machine in the shop. The population can then be described by a distribution indicating the behavior of *any randomly chosen draw* from that distribution.

For example, let X denote a randomly chosen measurement of the weighing machine. Then

- $X \sim N(1, .00001)$ is a possible assumption on the distribution or even
- $X \sim \text{Exp}(1)$ maybe

Then the average measurement of the machine is $\mu = \mathbb{E}(X)$ and the variance on the measurement if $\sigma^2 = \text{Var}(X)$. Although infinite populations will be more common for us, we will discuss the following sampling techniques for finite populations.

(I) Simple Random Sampling

Simple Random Sampling is a type of probability sampling that ensure that each sample of size n has an *equal* chance of being selected.

- In a simple random sample, all individuals are equally likely (*equal probability*) to be included in the sample.
- Estimates from simple random samples are unbiased; there is no systematic discrepancy between sample estimates and corresponding population values.
- For random samples, larger samples are typically more accurate (on average)

(II) Simple Random Sampling With and Without Replacement:

Simple random sampling with replacement (SRSWR):

SRSWR is a method of selection of n units out of the N units one by one such that at each stage of selection, each of the N unit has an equal probability of being selected, i.e., $1/N$.

Note: In SRSWR one particular unit can be sampled more than once.

Simple random sampling without replacement (SRSWOR):

SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, each of the *remaining* units have the same chance of being selected. Therefore, at the first draw each of the N components of the population has probability $1/N$ of being selected. At the second draw, each of the remaining $N - 1$ components of the population has probability $1/(N - 1)$ of being selected, and so on.

Note: In SRSWOR one particular unit, once sampled, is not considered for further sampling. Thus, none of the population units can be sampled more than once.

(III) Stratified Random Sampling:

In many situations, it is not possible to conduct simple random sampling. For example, if we are interested in the data of month household income in rural area of Uttar Pradesh (UP), the population consist of all the households residing rural areas of UP. Conducting an SRS for this population is very inconvenient administratively. It is much more convenient to split the entire region of UP into districts and conduct SRS within each district separately. This method is called *Stratified Sampling*.

- In Stratified Random Sampling (StrRS), one divides the population into k sub-populations called Strata, which are relatively homogeneous within themselves.
- The sample y consists of k different sub-samples, for $i = 1, \dots, k$, y_i is a simple random sample of the kth strata
- The samples from different strata are usually assumed to be independent.

Questions

1. For each of the following, identify the population and sample:

   a. Hall 4 canteen example

   b. In elections, television networks often declare the winner well before all the votes have been counted. They do this using exit polling, interviewing voters after they leave the voting booth. Using an exit poll, a network can often predict the winner after learning how several thousand people voted, out of possibly millions of voters.

   c. The salary of Data Scientist jobs in India. The website glassdoor.com contains average salaries of people in the role of "data scientist" in India.

   d. in your course project (think about it)

2. Set $x_i = i$, and $N = 100$. Calculate μ and σ^2 for this population.

   a. Obtain a random sample of $n = 10$ using R code.

   b. Calculate the *sample mean* and *sample variance*

   c. Repeat the above process 1000 times and calculate the average of the sample means and the average of the sample variance. Print these answers. Here's the structure of the code

```
reps <- 1000

# population
N <- 100
xi <- 1:N
mu <- ....
```

```
sig2 <- ....

means <- numeric(length = reps)
vars <- numeric(length = reps)
for(r in 1:reps)
{
  samp <- ....
  means[r] <- ....
  var[r] <- ....
}

mean(means)  # should be close to your answer for a.
mean(vars)  # should be close to your answer for b.
```

3. Suppose in a finite population, $Y_1, Y_2, \dots, Y_n$ is sample randomly chosen without replacement from the population $\{x_1, x_2, \dots, x_N\}$ that has mean μ and variance σ^2. Then notice that the probability of obtaining sample $\{y_1, y_2, \dots, y_n\}$ is

$$\begin{aligned}
\Pr(Y_1 = y_1, \dots, Y_n = y_n) &= \Pr(Y_1 = y_1) \Pr(Y_2 = y_2 | Y_1 = y_1) \cdots \Pr(Y_n = y_n | Y_{n-1} = y_{n-1} \cdots Y_1 = y_1) \\
&= \frac{1}{N} \frac{1}{N-1} \cdots \frac{1}{N-n+1} \\
&= \frac{(N-n)!}{n!} = \frac{1}{P(N,n)}
\end{aligned}$$

   We will compare the SRSWR with SRSWOR. Implement the steps of Q3 now for SRSWR and compare the sample mean and the sample variance. Which sampling technique is better?

4. Consider the Hall 4 canteen example. If we do a SRS sample, maybe that is not as reasonable, since every day of the week may have its own behavior in terms of the number of orders.

   Load the `hall4.csv` dataset provided and implement Stratified sampling to obtain an overall sample of size n.

5. What would be a good method of estimation μ when sampling is done through Stratified sampling?