Statistical Sampling

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

Topics

Introduction

- Until now we have spoken about characterising samples & some important probability concepts.
- We want to be able to *draw conclusions* about populations from data.
- Now we are going to focus on how to use data to make informed decisions.

Populations and Samples

- Investigations usually focuses on a well-defined collection of objects defining a **population** of interest.
- When desired information is available for all objects in the population, we have what is called a **census**.
- Often very difficult and inefficient to conduct a census.
- A **sample** is used to represent the population instead.

Descriptive Statistics

- The task of **descriptive statistics** is to characterise (describe) the sample
  - It is not meant to gain any insights about the population.
  - Important to become acquainted with the data.
  - Examine data quality (very important for inferential statistics).

Goal of Samples

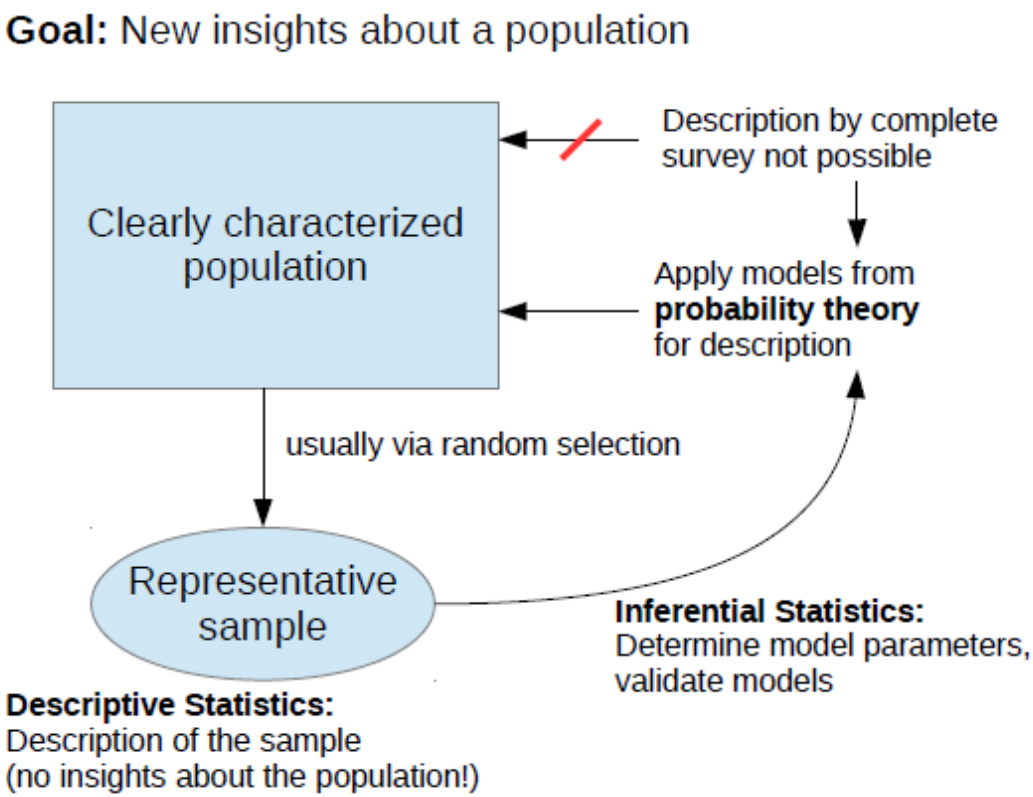

Figure: *source:* (1)

- The goal is to obtain some (new & important) insights about this population.

Inferential Statistics

- The objective of **inferential statistics** is to draw conclusions about the population from the sample
  - Estimate unknown parameters of assumed population distributions.
  - Test the validity of previously made assumptions about the population.

Statistical Sampling

- Sampling approaches are important to drawing accurate conclusions about populations.
- Feasible to survey every single internet user in Canada?
- Obtain sufficient information to draw valid inferences about a target population.

Sample Design

- First step of sampling is to design a sampling plan that will yield representative samples of the target population.
- Formally: a **sampling plan** is a description of the approach that will be used to obtain samples from a population prior to any data collection activity.
- The population **frame** is the list that the sample is selected from.

Sampling Plan

- A Sampling plan contains all of the following:
  - States the objective(s) of the sampling activity.
  - The target population (which population is being studied).
  - The population frame (where are we getting our sample(s)).
  - The method of sampling.
  - The operational procedures for collecting the data.
  - The statistical tools that will be used to conduct the analysis.

Sampling Objective

- Some potential sampling objectives:
  - Key population parameters (mean, proportion, or variance).
  - Determine if significant differences between two or more groups exists.
  - Understand potential impacts of a *treatment* on a population.

Sampling Frame

- An ideal sampling frame is a complete list of all members of the target population.
- We know this is not always possible:
  - All soccer players in Canada.
  - A frame could be all soccer players currently registered in an official league sanctioned by the C.S.A..
- Understanding how well the frame represents the target population is very important.
  - Maybe we are interested in barriers to playing organized soccer in Canada.
  - Does our frame do a good job?

Sampling Methods

- Sampling methods can be *subjective* or *probabilistic*.
- Subjective methods include judgement sampling and convenience sampling.
  - **Judgement sampling**: *Expert judgement* is used to select the sample.
  - **Convenience sampling**: Samples are selected based on the ease of collection.
- **Probabilistic** approaches are necessary to draw **valid** statistical conclusions.

Simple Random Sampling

- The most commonly used approach to sampling is simple random sampling.
- **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected.
- If the population (or even their identifiers) are stored in a database we can generate a random sample.

Simple Random Sampling in R

- Always set your seed (`set.seed(1000)`)!
- In base R:

```
random_order <- sample(nrow(data.frame))
random_sample <- data.frame[random_order[1:n],]
```

- Using the *dplyr* package:

```
library(dplyr)
random_sample <- data.frame %>%
    sample_n(n, replace = FALSE)
```

Example 1

- Import the *starwars* dataset from the *dplyr* package.
- Generate a simple random sample of ten ($n = 10$) characters.
  - Use base R and *dplyr*.
- **Remember to use the `set.seed()` function!**

Systematic/Periodic Sampling

- Selects items periodically from the population.
- Every possible sample of a given size in the population does not have an equal probability of being selected.
- Example: To sample 250 names from a list of 400000, selected the first name at random from the first 1600, and then every 1600^{th} name could be selected.
- Depending on the situation, this method may create significant bias in estimations.
  - Only sampling people who made purchases on Saturday.

Stratified Sampling

- Stratified sampling applies to populations that are divided into natural subsets called strata.
- Allocates the appropriate proportion of samples to each stratum.
- Example: B.C. is divided into eight different geographical regions with different populations. A **stratified sample** would choose a sample of individuals in each stratum proportionate to its size.
- Ensures that each stratum is weighted by its size relative to the overall population.
  - *May be issues of cost or with differences in mixes within identified strata.*

Stratified Sampling in R

- Always set your seed (`set.seed(1000)`)!

- Using the *dplyr* package:

```
library(dplyr)
strat_sample <- data.frame %>%
    group_by(strata.name) %>%
    sample_frac(size = percentage)
```

Example 2

- Import the *Aids2* dataset from the *MASS* package.
- Take some time to examine the dataset.
- Generate a stratified sample of 20% using *state* as the stratifying variable.
  - We need to use *dplyr*.
- **Remember to use the `set.seed()` function!**

Cluster Sampling

- Divides the population into subgroups called clusters.
- Then sampling a set of the clusters.
- Generally, a census is conducted on the selected (sampled) clusters.
- Example: Cluster TRU students by department, take a random sample of departments, and then survey every student within the selected departments.

Cluster Sampling in R

- Always set your seed (`set.seed(1000)`)!
- Using base R:

```
clusters <- sample(unique(df$cluster.variable), size =
num.clusters, replace = FALSE)
clustered_sample <- df[df$cluster.variable %in% clusters, ]
```

Sampling from a Continuous Process

- Commonly used in manufacturing and quality control.
- Commonly conducted one of two ways:
  1. Select a time at random; then select the next n items produced after that time.
  2. Select n times at random; then select the next item produced after these times.
- The different approaches apply to different objectives.

Example 3

- Import the *Aids2* dataset from the *MASS* package.
- Generate a clustered sample using *state* as the clustering variable.
- Select two random clusters to be part of your sample.

Errors in Sampling

- Sample design can lead to two sources of errors:
  1. **Nonsampling error**: when the sample does not adequately represent the target population.
     - *Generally a result of a poor sample design.*
  2. **Sampling/Statistical error**: when the sample is only a subset of the total population.
     - *Sampling error is a natural part of any sampling process.*
- Sampling error depends on the sample size (n) relative to the population size (N).

Exercise 1

- Import the *Cars93* dataset from the *MASS* package in R.
- Take some time to get to know the dataset.
- Generate a random sample, a stratified sample, and a clustered sample from the *Cars93* dataset.
  - Use your discretion to choose the stratifying & clustering variable(s).
  - Why did you choose these variables?
- Try to code you answers in both base R and using the *dplyr* package.

Exercise 2

- Think of some situations where one of the two ways of sampling from a continuous process would be more useful than the other.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.
2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- sample_n()