Working with samples in R

Kaloyan Ganev

2022/2023

Lecture contents

Samples

Sampling: A Brief Intro

Population

The population consists of all elements of a set that are subject to analysis.

Sample

A sample is a portion of a population subject to analysis.

- Whole populations are very rarely analysed
- Typically we deal with samples

Parameters vs. Statistics

- Measures such as mean, standard deviation, mode, median, etc., when they describe a **population**, are called parameters
- When they describe a **sample**, are called statistics

Types of Sampling

- Random/probability sampling:
  - All items in the population have the chance to be selected
  - If carried out properly, the sample is representative of the population (size matters!)
- Non-random/judgement sampling:
  - Personal opinion and knowledge are used to select items for the sample
  - Often the judgement sample is used as a pilot (trial) sample; helps in designing the random sampling carried out later
  - Easier to make as the complicated statistical analysis underlying random sampling is skipped
  - If the sample lacks a significant degree of representativeness, it might mislead

Methods of Random Sampling

Four major methods:

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

Simple random sampling

- Each element has an **equal chance** of being selected
- Each possible sample has an **equal probability** as any other sample of the same size
- Works on finite and infinite populations
- How is it implemented?
  - Generate random numbers and pick the elements using them
  - Use paper slips, balls in bowls, etc.

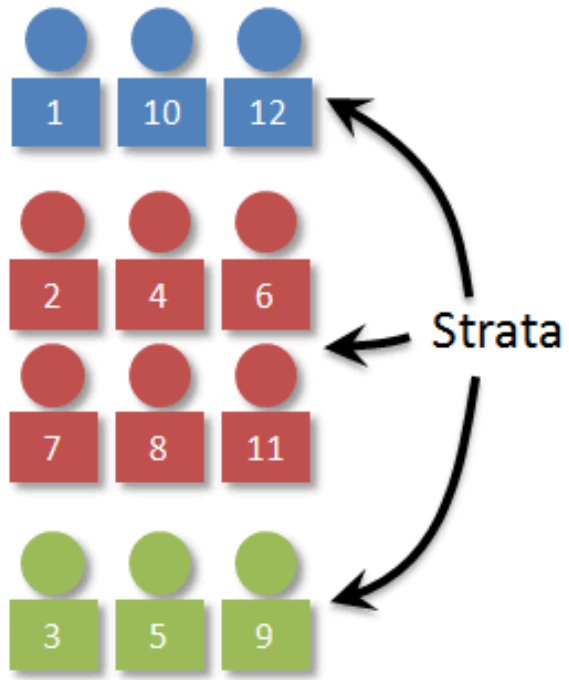
Systematic Sampling

- Items are selected at an uniform interval
- The interval might be measured in time, space, or order

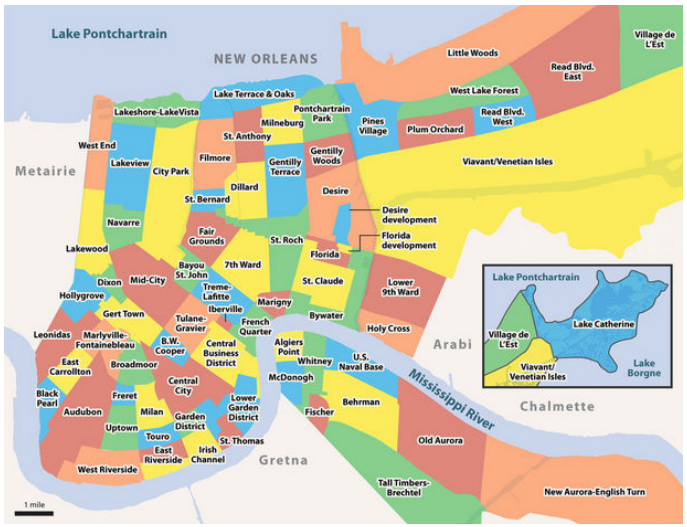
- Unlike simple random sampling, each sample does not have an equal probability, although each element has an equal chance of being picked
- This might introduce error and bias in the sampling process
- Advantages: requires less time, implies lower costs than simple random sampling

Stratified Sampling

- The population is split into homogeneous groups called strata
- Then, two approaches are possible:
  1. Elements are selected randomly from each group, their number is proportional to group size
  2. An equal number of elements is selected from each group (stratum), then the results are weighted for group size
- Advantage: if properly implemented, achieves a better reflection of population's characteristics



Cluster Sampling



- The population is divided into subgroups (usually non-homogeneous) called clusters
- Each cluster is representative of the population as a whole
- Example: A TV company wants to study watching behaviour of citizens
- Several blocks (clusters) are selected for study, then every element in the block is studied

To sum up:

Stratified sampling is appropriate when there is not much variation within groups but significant variation among groups. Cluster sampling is appropriate when there is significant variation within groups but not much variation among groups.

What Is a Sampling Distribution?

- The sample statistics' values usually differ from sample to sample
- This means that the statistics themselves are random variables, with their own distributions
- If we list all possible values that a statistic can take, with the corresponding probabilities, then we have the sampling distribution of the statistic
- Of course, listing is not always an option, but writing the corresponding mathematical law is

Standard Errors

- Even if sampling is well designed and implemented, sampling error occurs merely by chance
- In other words, the sample will almost never be completely representative of the population
- Therefore, the calculated statistics will almost never coincide with the values of the corresponding parameters
- As a result, we observe variability of statistics across samples, around the true population parameters
- We can define

Standard Error

The standard error of a statistic is the standard deviation of the distribution of that same statistic.

Sampling from (Non-)Normal Populations

- If we sample from a normal distribution with parameters μ and σ^2, then the sampling distribution will have a mean equal to the population mean:

$$\mu_{\overline{x}} = \mu$$

- The standard error of the sample mean equals:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}},$$

  where n is sample size
- If we sample from a non-normal distribution, we still have $\mu_{\overline{x}} = \mu$
- We cannot readily tell what the standard error looks like

The Central Limit Theorem

- One of the most important theorems in statistical science
- Broadly stated:

Central Limit Theorem

As sample size increases, the sampling distribution of the mean approaches normal.

- Allows to use sample statistics to make inferences about population parameters without knowing anything about the shape of the population distribution

Sampling in R

The notion of random sampling in R

- Realized through the `sample` function
- If, for example, you want to sample 10 numbers at random from 1 to 100, you write:

```
sample(1:100, 10)
```

- Default behaviour of the function is sampling without replacement
- To have replacement, the option `replace = TRUE` is added:

```
sample(1:100, 50, replace = TRUE)
```

Some quick examples

Examples

- Simulate tossing a coin 100 times:

```
faircoin <- c("Heads", "Tails")
sample(faircoin, size = 100, replace = TRUE)
```

- Count the number of heads and tails:

```
table(sample(faircoin, size = 100, replace = TRUE))
```

- If you want to take a sample from a dataset:

```
mydataset <- read.xlsx("zzz.xlsx", sheet = "Sheet1")
mysample <- mydataset[sample(1:nrow(mydataset), 10, replace =
    TRUE),]
```

- Pay attention to the comma before the closing square bracket!!! We have a data frame, the comma shows that we are taking all columns