Chapitre 1

Sampling

1.1 Notions of samples

Sampling is a fundamental concept in statistics, and it's essential to understand how it works, why it's used, and how to implement it. Here are more details about sampling, along with practical examples :

Why Sampling is Used : Sampling is used in statistics for several reasons :

1. **Cost-Efficiency :** Collecting data from an entire population can be time-consuming and expensive. Sampling allows researchers to gather a smaller, more manageable set of data points.
2. **Practicality :** In some cases, it's impossible to collect data from an entire population, especially if the population is large or widely dispersed.
3. **Destruction of Items :** When the data involves destructive testing or examination, such as in medical research, it's not feasible to collect data from the entire population.
4. **Accuracy :** When done correctly, sampling can provide accurate and representative information about the entire population.

1.1.1 Types of Sampling :

There are different methods of sampling, depending on the research objectives and available resources. Some common sampling methods include :

1. **Simple Random Sampling :** In this method, each member of the population has an equal chance of being selected. This is usually done using random number generators or random selection techniques.

   Example : Suppose you want to conduct a survey of 1000 students in a school. You assign each student a unique number and use a random number generator to select 100 students from the list.
2. **Stratified Sampling :** The population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, income), and

random samples are taken from each stratum. This ensures that each subgroup is adequately represented.

Example : In a survey about a product's popularity, you divide the population into age groups (e.g., 18-24, 25-34, 35-44) and then randomly sample from each age group.

3. **Systematic Sampling :** Researchers select every nth item from the population list after randomly selecting a starting point. This method is straightforward and efficient.

   Example : In a factory with 1000 employees, you select a random starting point (e.g., the 7th employee), and then select every 10th employee thereafter.

4. **Cluster Sampling :** The population is divided into clusters, and a random sample of clusters is selected. Then, all members of the selected clusters are surveyed.

   Example : When studying healthcare in different cities, you randomly select several cities (clusters) and then survey all individuals within those cities.

1.1.2 Common Pitfalls to Avoid :

When conducting sampling, it's crucial to avoid common pitfalls, such as :

1. **Sampling Bias :** This occurs when the sampling method systematically excludes or over-represents certain groups in the population.
2. **Non-Response Bias :** If a significant portion of those selected for the sample does not participate, it can lead to a non-response bias.
3. **Sampling Error :** This is the natural variation that occurs when working with samples instead of the entire population. It can be minimized by increasing the sample size.
4. **Confounding Variables :** Failure to control for confounding variables (variables that are related to both the independent and dependent variables) can affect the results of the study.

In summary, sampling is a critical technique in statistics used to gather data efficiently and effectively from a subset of a larger population. The choice of sampling method depends on research goals and available resources. Proper sampling techniques are essential to ensure the validity and reliability of statistical analyses and research findings.

1.2 Statistics of samples : empirical mean, empirical variance

Data statistics often involve calculating various descriptive statistics to summarize and understand the characteristics of a dataset. Two essential statistics

are the empirical mean (also known as the sample mean) and the empirical variance (sample variance).

1.2.1 Empirical Mean (Sample Mean) :

The empirical mean, often denoted as $\bar{x}$, represents the average or central tendency of a dataset. It is calculated as the sum of all data points divided by the number of data points in the dataset.

The formula for the empirical mean is :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where :

$\bar{x}$: Empirical mean
n : Number of data points in the dataset
x_i : Individual data points

1.2.2 Empirical Variance (Sample Variance) :

The empirical variance, often denoted as s^2, measures the spread or variability of data points in the dataset. It quantifies how much individual data points deviate from the mean. The larger the variance, the more dispersed the data points are.

The formula for the empirical variance is :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where :

s^2 : Empirical variance
n : Number of data points in the dataset
x_i : Individual data points
$\bar{x}$: Empirical mean

The use of $n-1$ instead of n in the denominator is due to Bessel's correction, which accounts for the fact that we are estimating the population variance from a sample.

1.3 Gaussian Samples

A Gaussian sample, also known as a normal sample, represents a set of data points that follow a Gaussian (normal) distribution. The Gaussian distribution

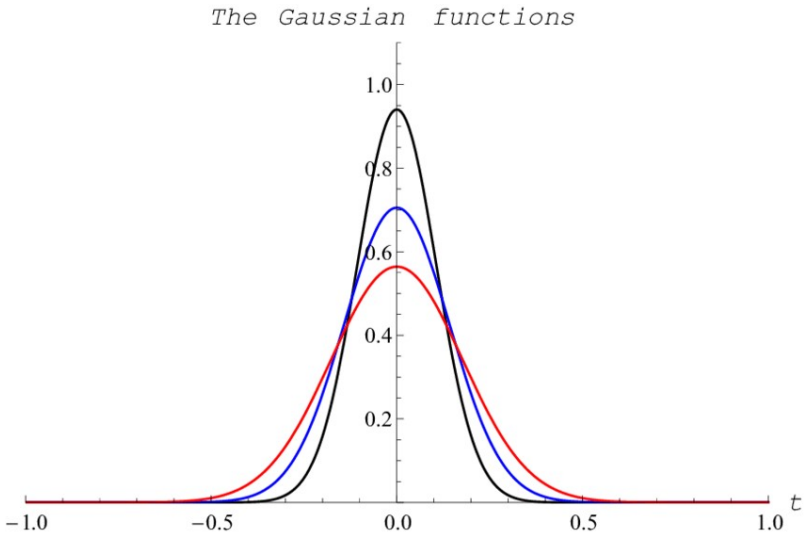

FIGURE 1.1 – Gaussian sample

is characterized by its mean (μ) and standard deviation (σ). For example, let's consider a Gaussian sample with the following properties :

μ : Mean
σ : Standard Deviation

Suppose we have a Gaussian sample of 100 data points :

$$X = \{x_1, x_2, \ldots, x_{100}\}$$

where x_i represents an individual data point.
The Gaussian distribution for this sample can be expressed as :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To provide concrete values, let's say :

$$\mu = 10$$
$$\sigma = 2$$

Then, a few data points from this Gaussian sample might be :

$$X = \{11.23, 9.75, 10.50, 9.98, \ldots\}$$

These data points are drawn from the Gaussian distribution with the specified mean and standard deviation.

1.4 Application

Exercices

Exercice 1.1. ***Simple Random Sampling***

You have a population of 1,000 students in a school. You want to select a simple random sample of 100 students to conduct a survey about their study habits. Explain how you would carry out this sampling process.

Exercice 1.2. ***Systematic Sampling***

You have a list of 500 employees in a company, and you want to select a systematic sample of 50 employees for a performance evaluation. Describe the steps you would take to choose the sample and determine the sampling interval.

Exercice 1.3. ***Stratified Sampling***

You are conducting a political opinion poll in a city with a diverse population. Outline how you would use stratified sampling to ensure a representative sample. Identify the strata and explain how you would select individuals from each stratum.

Exercice 1.4. ***Cluster Sampling***

You are conducting a survey of households in a large urban area. Describe how you would use cluster sampling to select your sample. Identify the clusters and explain how you would choose which clusters to include in your sample.

Exercice 1.5. ***Convenience Sampling***

Discuss the advantages and disadvantages of convenience sampling. Provide an example of a situation where convenience sampling might be appropriate and another situation where it might lead to biased results.

Exercice 1.6. ***Non-Probability Sampling***

Explain what non-probability sampling is and why it might be used in research. Discuss the potential limitations of non-probability sampling methods.

Exercice 1.7. ***Sampling Error***

Define sampling error and explain why it is important to consider when interpreting the results of a sample. Provide an example of how sampling error can impact the accuracy of survey findings.

Exercice 1.8. ***Sample Size Determination***

You are designing a survey to estimate the proportion of customers satisfied with a new product. How would you determine the appropriate sample size to achieve a desired level of confidence and margin of error ?

Exercice 1.9. ***Oversampling and Undersampling***

Discuss the concepts of oversampling and undersampling in the context of survey sampling. Explain when and why these techniques might be employed.