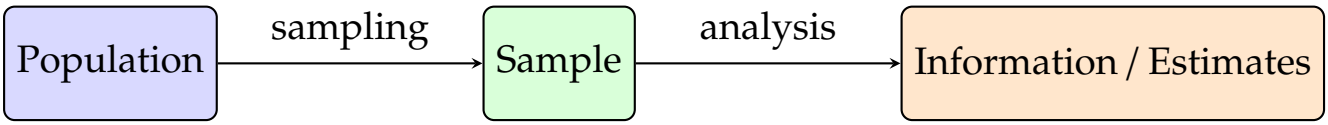
Chapter 1

Preliminaries

Often, analyses in experimental sciences (pharmacology, biology, agronomy, etc.) focus on studying the characteristics of a given population, for which complete information is not accessible.

Example 1.0.1. *To study the milk crisis in Algeria, the population of interest may be defined as dairy cows. Several characteristics could be investigated : the age of cows, the volume of milk produced, or the duration of lactation.*

In such cases, it is not feasible to collect data from the entire population. A sampling process is therefore required in order to approximate the desired information.



1.1 Sampling

Sampling is a fundamental concept in statistics, and it's essential to understand how it works, why it is used, and how to implement it.

If the population size is N, one can draw from this population several samples of size n ($n < N$), and the results will vary from one sample to another. From a single sample, we can determine an estimated value of the parameter of the studied population. The study is therefore based on estimations rather than certainties.

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

1. **Cost-Efficiency :** Studying the entire population is often too expensive and time-consuming. Sampling reduces costs while still providing useful information.
2. **Practicality :** In many cases, it is impossible to study every individual in a large or widely dispersed population. Sampling makes data collection feasible.
3. **Destructive Testing :** Sometimes data collection destroys the unit being studied (e.g., crash-testing cars, chemical analysis of food, or medical blood tests). In such cases, only a sample can be examined.
4. **Accuracy :** If the sample is well designed, it can give results that closely represent the entire population, without the need to test everyone.

1.1.1 Types of Samples and Sampling Methods

In research, it is often not possible to include every individual from a population. Therefore, a **sample** is taken. A sample is a smaller group selected from the population that should represent it. There are different **types of samples** and different **methods** of choosing them.

• **Types of Samples (General Categories)**

1. **Random (Probabilistic) Sample** Each member of the population has a known, non-zero chance of being selected.
   Example : Assigning numbers to all students in a school and using a random number generator to select 50 of them.
2. **Non-random (Non-probabilistic) Sample** The selection is based on convenience or judgment; not all members have a chance of being included.
   Example : Interviewing only the first 20 people who enter a supermarket.
3. **Sampling without replacement** Each selected element is removed from the population and cannot be chosen again.

Example : Selecting 10 balls from an urn without putting them back.

4. **Sampling with replacement** Each selected element is returned to the population and may be chosen again.
   Example : Drawing a card from a deck, recording it, and then shuffling it back before the next draw.
5. **Representative Sample** A sample that reflects the main characteristics of the population, allowing results to be generalized.
   Example : A national health survey including participants from different age groups, genders, and regions in proportion to the population.

Note : Sampling "with" or "without" replacement refers to the procedure of selection, and can apply to both random and non-random sampling.

• **Methods of Random Sampling**

Within probabilistic sampling, different methods can be applied :

1. **Simple Random Sampling :** Every member has an equal chance of being selected.
   Example : Using a lottery system or a random number generator to pick students from a class.
2. **Stratified Sampling :** The population is divided into subgroups (strata) based on specific characteristics (e.g., age, gender, income). Random samples are then taken from each subgroup.
   Example : Dividing students by grade level, then randomly selecting individuals from each grade.
3. **Systematic Sampling :** Every nth member is selected from a list after a random starting point.
   Example : Selecting every 10th person from a company's employee list.
4. **Cluster Sampling :** The population is divided into clusters (groups), and a random selection of clusters is made. All members of the chosen clusters are included in the sample.
   Example : Selecting certain schools in a city, then surveying all students within those schools.

1.2 Random sample and sample statistics

Definition 1.2.1. *An* random sample *of size n is defined by a set of n independent random variables* $(X_1, X_2, \ldots, X_n)$ *following the same law called the* parent distribution*. A realization* $(x_1, x_2, \ldots, x_n)$ *of a random sample corresponds to n independent repetitions of the same experiment.*

Remark 1. *An* random sample *of size n is called an* n-sample*.*

Definition 1.2.2. *A function S of the random sample* $(X_1, X_2, \ldots, X_n)$ *of size n is called a statistic of the sample :* $S(X_1, X_2, \ldots, X_n)$*.*

Definition 1.2.3. *"Estimator and estimation", the most important of the objectives of a statistical study, is the obtaining of reliable estimates of the characteristics of a population from a representative sample taken from that population. This is called the decision problem concerning parameters such as :*

— *The mathematical expectation :* $\mathbb{E}(X) = m = \overline{X}$ *.*

— *The variance* $\mathrm{Var}(X)$ *or the standard deviation* σ_X*.*

— *The proportion p in the case of a discrete (countable) characteristic.*

Estimating one of the parameters listed above means determining an approximate value of this parameter from the results obtained on a random and representative sample drawn from the population.

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 :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

1.2.2 Empirical Variance (Sample Variance)

The empirical variance, often denoted as S_n^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 :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Empirical variance corrected We call the corrected empirical variance, the statistic denoted by

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2,$$

Equivalently, the realized statistic is

$$\tilde{s}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2,$$

with $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$.
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.

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

$$\mu : \text{ Mean}$$
$$\sigma : \text{ Standard Deviation}$$

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 :

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