UNIVERSIDAD
CENTRAL

Ingeniería de Sistemas
Data Analytics
Prof: Hugo Franco

Session N° 12 | Statistical Inference I
Sampling

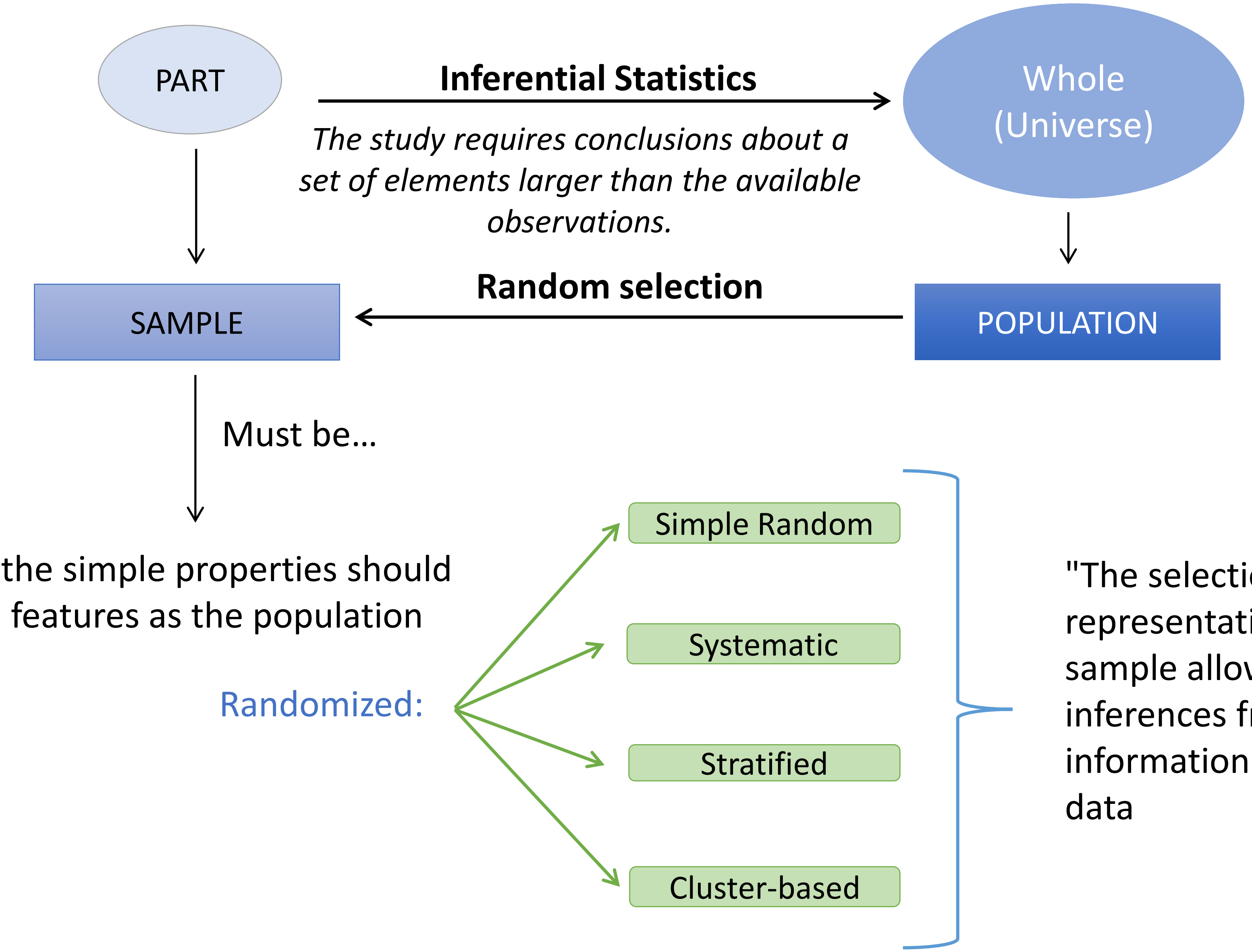
Bogotá D.C., October 22, 2022

Review: Variables

- A variable is an observable characteristic that varies between different subjects in a population. The information that we have about each subject is summarized in variables.

- Qualitative Variables
  If their values (modalities) cannot be naturally associated with a number (algebraic operations cannot be performed on them)
  - Nominal: If their values cannot be sorted
    - Sex, Blood Type, Religion, Nationality, Smoking (Yes/No)
  - Ordinals: If their values can be sorted
    - Improvement within a medical treatment, degree of satisfaction, intensity of pain

- Quantitative or Numerical Variables
  If their values are numeric (it makes sense to do algebraic operations with them)
  - Discrete: If they take integer values
    - Number of children, Number of smoked cigarettes a day, age (usual representation in years)
  - Continuous: If between two values, infinite intermediate values are possible.
    - Height, Intraocular pressure, Dosage of medication administered, age



PART

Inferential Statistics

The study requires conclusions about a set of elements larger than the available observations.

Whole (Universe)

SAMPLE

Random selection

POPULATION

Must be…

Representative: the simple properties should exhibit the same features as the population

Randomized:

Simple Random

Systematic

Stratified

Cluster-based

"The selection of a representative, randomized sample allows to make inferences from the available information contained by the data

Steps in a statistical study

1. Formulate hypotheses on a certain *population*

- **E.g.: smokers request more sick leaves at work than non-smokers**
- More frequently? More time?

2. Select data to be acquired (Experimental Design)

- **Inclusion criteria:** Which subjects will be *included* in the study (*sample*)
  - e.g., working-age smokers and non-smokers.
- **Exclusion criteria**: how will the subjects be selected?
  - e.g., those workers suffering from chronic diseases must be excluded?
- Which specific data (*variables*) will be acquired
  - e.g., sick leaves requested, duration of each sick leave, sex, age, economic sector, etc.

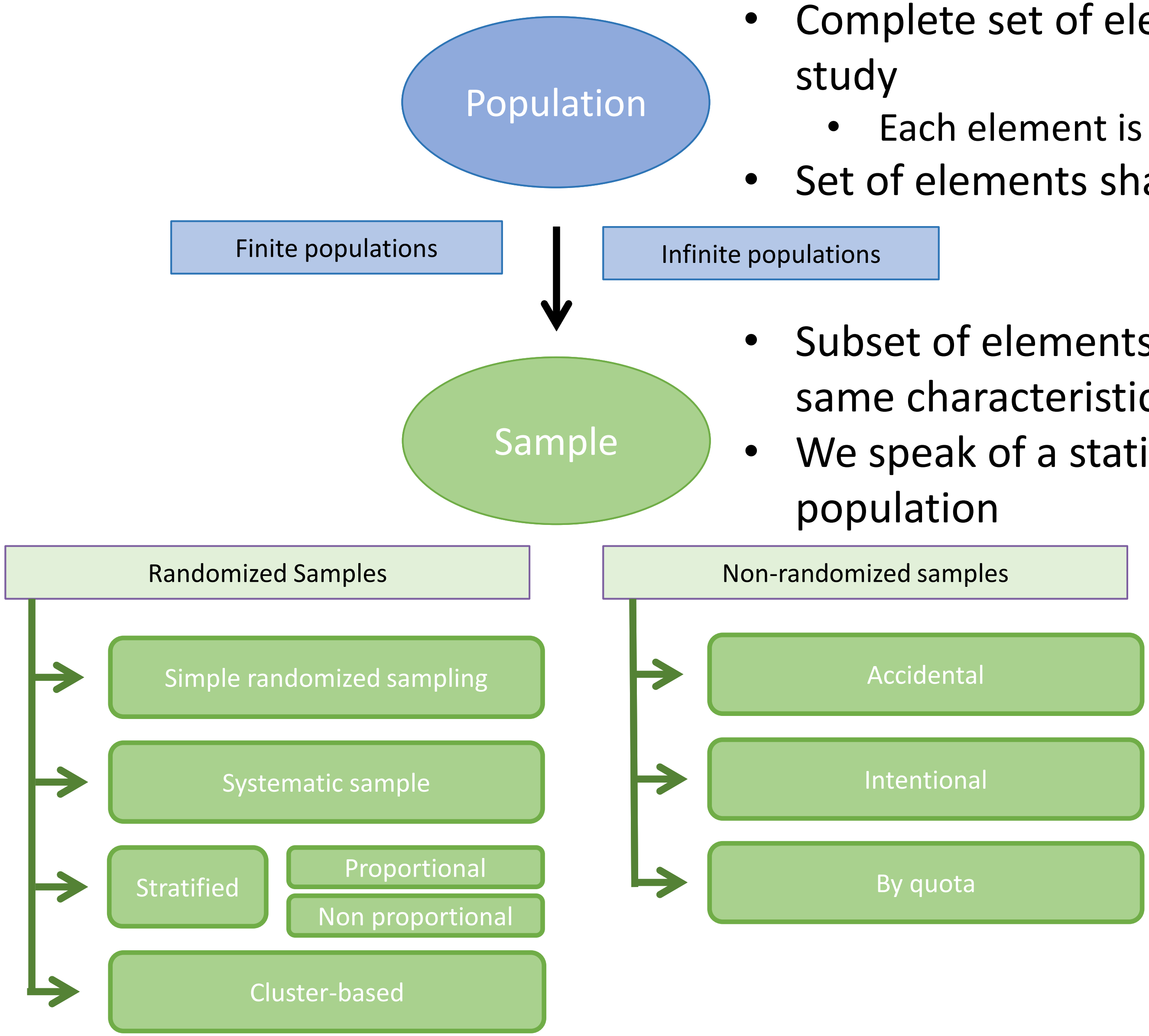
3. **Acquire data (*sampling*)**
    - Stratified? Systematic?

4. **Describe (summarize) the acquired data**
    - Leave meantime among smokers and non-smokers (*statistics*)
    - % of sick leaves per smoking habits, sex, age, etc. (*frequency analysis*), plots, tables...

5. **Obtain a significant inference over the population**
    - Smokers are on leave at least 10 days/year more on average than non-smokers.

6. **Quantify the inference reliability (confidence)**
    - *Confidence level (e.g., 95%)*
    - *Contrast significance: (e.g., p=2%)*

Populations and samples



Population

- Complete set of elements in the scope of the statistical study
  - Each element is called a **statistical unit**.
- Set of elements sharing at least one well-defined feature

Finite populations

Infinite populations

Sample

- Subset of elements of the population that maintain the same characteristics.
- We speak of a statistical sample when it is at least 5% of the population

Randomized Samples

- Simple randomized sampling
- Systematic sample
- Stratified
  - Proportional
  - Non proportional
- Cluster-based

Non-randomized samples

- Accidental
- Intentional
- By quota

Randomized Sampling

- Obtained by any system ensuring that the selection process includes randomness.
- All the elements in the universe then have a known probability of being extracted (and this probability is different from zero or one)
- Randomized samples allow the calculation of a sampling error, so it can be generalized.

Statistical inference: Extrapolation to the population.

To properly generalize assertions on the population behavior, it is necessary to work with randomized samples. They allow the contrast of explanatory, correlational and descriptive hypotheses.

Simple randomized sampling

- It is one where all the elements of the group have the same probability of being chosen and this probability is different from zero and one.
  - *"A simple random sample is the one that results from applying a method by which all possible samples of a certain size have the same probability of being chosen"* (Webster, 1998)
- It has the condition of equiprobability implicit.
- Process:

a) Define the study population.

b) List all the units of analysis that make up the population, assigning them an identity or identification number (sampling base).

c) Determine the optimal sample size for the study.

d) Select the sample using a procedure that guarantees randomness.

Systematic Sample

- Practically, it is a simple randomized sampling, yet the elements of the universe are extracted according to a system
    - Usually, nothing more than a fixed period among sample indexes.
        - e.g.,

          $\{x_r , x_{r+k}, x_{r+2k}, x_{r+2k}, x_{r+3k}, \ldots, x_{r+(\mathrm{n}\,-\,1)k}\}$
          where *r* is the random seed and *k* is the selected inter-sample interval.

Stratified Sampling

- Determines the **strata** contained in a study population to select and extract the sample from them. Usually, this method is applied while working with categorical variables or attributes distributed according to categories.
- It is useful when the population is susceptible to being divided into categories or strata where there is an analytical interest and which, for theoretical and empirical reasons, present differences between them (marital status, age, sex).

Stratum: any subgroup of analysis units that differ in the characteristics to be analyzed in an investigation. It is an exhaustive and exclusive category of the population, where the *units* that compose it are very similar within themselves but different from each other.

- Types:
  - **Proportional:** It is one whose categorical structure replicates the same share (e.g., percentage) of the Universe
  - **Non-proportional:** It is one where the percentage structure of the universe is not applied, but rather the same number of people from each collective stratum is taken with the aim of making comparisons possible.

Steps to Select a Proportional Stratified Sample

1. Define the study population
2. Determine the required sample size
3. Establish the strata or subgroups
4. Determine the total sampling fraction per stratum, dividing the size of the stratum by the size of the study population.
5. Multiply the total sampling fraction per stratum by the sample size to obtain the number of analysis units from each stratum that will be integrated into the sampling unit.
6. Selection and extraction of the sample applying the simple random sampling procedure.

Cluster sampling

- Useful when conducting research with extremely large universes such as countries, nations, etc., where it is practically impossible to get or build the sampling base. It is used when the researcher is limited by factors of time, distance, and funding, among others.
- The analysis units are encapsulated or enclosed in certain physical (e.g., geographical) places called *clusters*.

Cluster: an exhaustive and exclusive subset of the population where the elements within the same cluster are similar to each other, yet quite different from those elements in different clusters.

Example: select a sample of 20 students from a population of 600

SIMPLE RANDOMIZED

-A student is chosen at random (probability of choosing him 1/600)
-It is returned to the population, and another is chosen (probability of choosing 1/600)
-Must be returned or the probability of the second student changes (probability 1/599)
-The problem is that the same student can be chosen twice

STRATIFIED

-If we want our sample to be representative, we must know how many students there are per course: First Medium 200, Second Medium 150, Third Medium 150 and Fourth Medium 100 students.

Course	Population	Fraction	Sample
First	200	0.33333	7
Second	150	0.25	5
Third	150	0.25	5
Fourth	100	0.16666	3
Total	600	1	20

SYSTEMATIC

-Since we must choose 20 out of 600, that is, 1 in 30, we proceed as follows:
-The students are ordered and numbered, one is chosen at random, for example student 27.
-From this, the others are chosen from this interval of 30 students.

CLUSTER

Suppose we need a sample of students from all over a country, which is difficult to have the total population, but we know that they are grouped into Types of schools, Schools and levels.

So, we randomly select some types of schools, then some schools, and finally some courses.

Finally, by simple random choices, we select some students.

* Clusters are large and heterogeneous units.

Sample size estimation

Mean-based size (quantitative)

- **Known population *N***

$$n = \frac{N \times z^2 \times \sigma^2}{z^2 \times \sigma^2 + d^3(N-1)}$$

Where:

N: population size

σ^2: population variance

z^2: *significance-error level* factor

If σ is unknown, the sample size can be estimated from the sample variance s^2 by using the unbiased *sample quasivariance*

$$s_C^2 = \frac{s^2 \times n}{(n-1)}$$

- **Unknown (or infinite) population**

$$n = \frac{z^2 \times \sigma^2}{d^2}$$

Proportion-based size (qualitative)

- **Known population *N***

$$n = \frac{N \times P \times Q \times z^2}{z^2 \times P \times Q + d^2(N-1)}$$

Where:

N: population size

P: relevant case proportion.

$Q = (P-1)$: complementary (rejected) case proportion

z^2: *significance-error level* factor

- **Unknown (or infinite) population**

$$n = \frac{P \times Q \times z^2}{d^2}$$

d is the estimated difference between the estimator and the parameter:

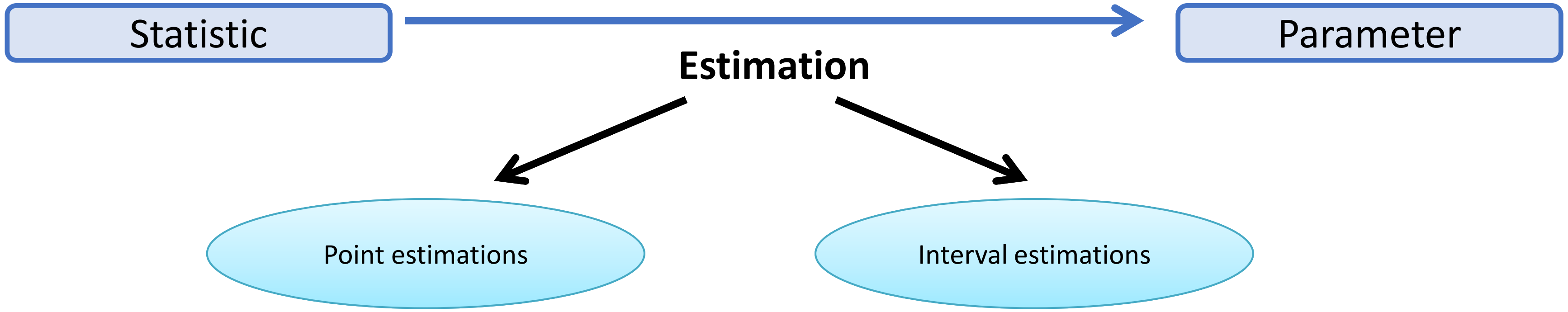
$$d = z\frac{\sigma}{\sqrt{n}} \text{ (quantitative)} \qquad d = z\sqrt{\frac{P \times Q}{n}} \text{(qualitative)}$$

Parameter estimation

Parameters «In statistics, parameters are the values or measures that characterize a population such as, for example, the mean and the standard deviation of a population (...)
They are indeterminate, constant or fixed quantities with respect to a condition or situation, which characterize a phenomenon at a given moment that occurs in a population ”(Sierra Bravo, 1991).

Statistics Value obtained from the sample values. Sample mean and variances are representative examples.

Estimation Operation intended to determine the value of a parameter, using incomplete data from a sample.



Margin note: Law of Large Numbers

- “The result obtained for a variable from many trials of the same experiment (observation, measure) converges to the *expected value* of the variable
  - The larger the number of performed trials, the closer to the expected value is the result

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i}{n} = \overline{X}$$

Parameter estimations: Point estimation

- A single numerical value is required to estimate the parameter, that is, it directly assigns the value obtained for the statistic to the parameter

$$E(X) = \mu$$

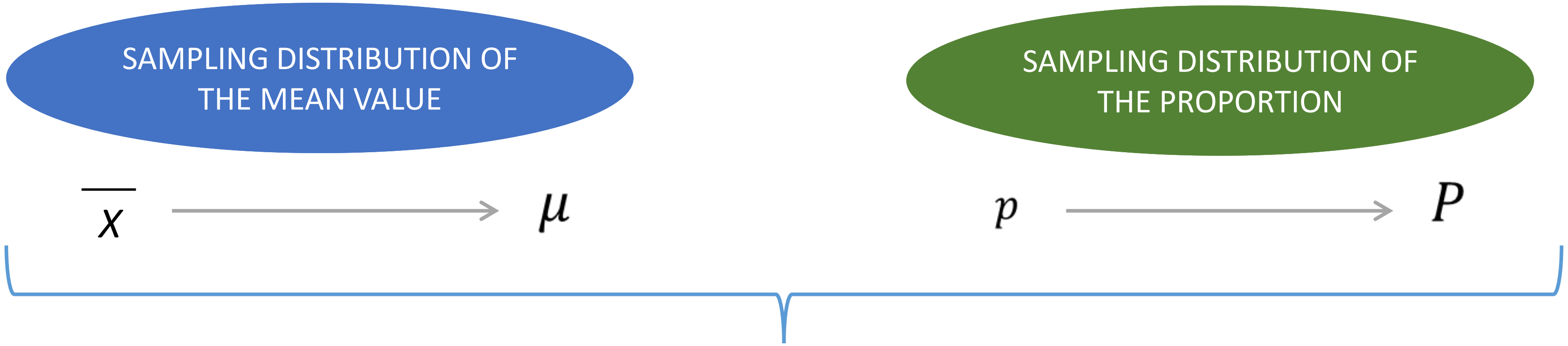
- It is the simplest inference that can be made: to assign the parameter the value of the statistic that best serves to estimate it.
- Conditions for a good estimator:
  - **Lack of Bias:** An estimator will be unbiased if its expected value coincides with that of the parameter to be estimated.
  - **Consistency**: An estimator will be consistent if, as the sample size increases, its value approaches that of the parameter
  - **Efficiency:** Given two possible estimators, we will say that the first is a more efficient estimator than the second if it is true that the first estimator has a lower *variance* than the second.
  - **Sufficiency:** An estimator will be sufficient if it uses all the available sample information

Parameter estimations: Interval estimation

- An interval estimate is a range or band of values within which the parameter is said to be with a pre-established *significance level*.
- It provides an interval, a range of values between which the parameter will be located with a certain probability (i.e., within the interval [0, 1]).
- Point estimation is rarely used, as we do not have enough data to indicate the degree of reliability of the sample data we have taken.
    - It makes more sense to ask the probability that the mean or proportion of the population *lies* within a given ***interval***.
- Understanding the rationale for interval estimates requires a mastery of the concept of a sampling distribution, specifically, the sampling distribution of the mean (MMD).
- Given a sample, the **Sampling Distribution of the mean** can be calculated where, with some certainty, the population mean that is sought will be.

Sampling distributions

- Statistic analysis provides information about the behavior of population parameters such as the mean (μ), the variance (σ), or the proportion (p).
- Thus, a random sample is drawn from the population and the value of a corresponding statistic is calculated, e.g., the sample mean (X), the sample variance (s) or the sample proportion (p).
- The value of the statistic is random because it depends on the elements chosen in the selected sample and, therefore, the statistic has a probability distribution which is called the Sampling Distribution of the statistic.



SAMPLING DISTRIBUTION OF THE MEAN VALUE

$\overline{X} \longrightarrow \mu$

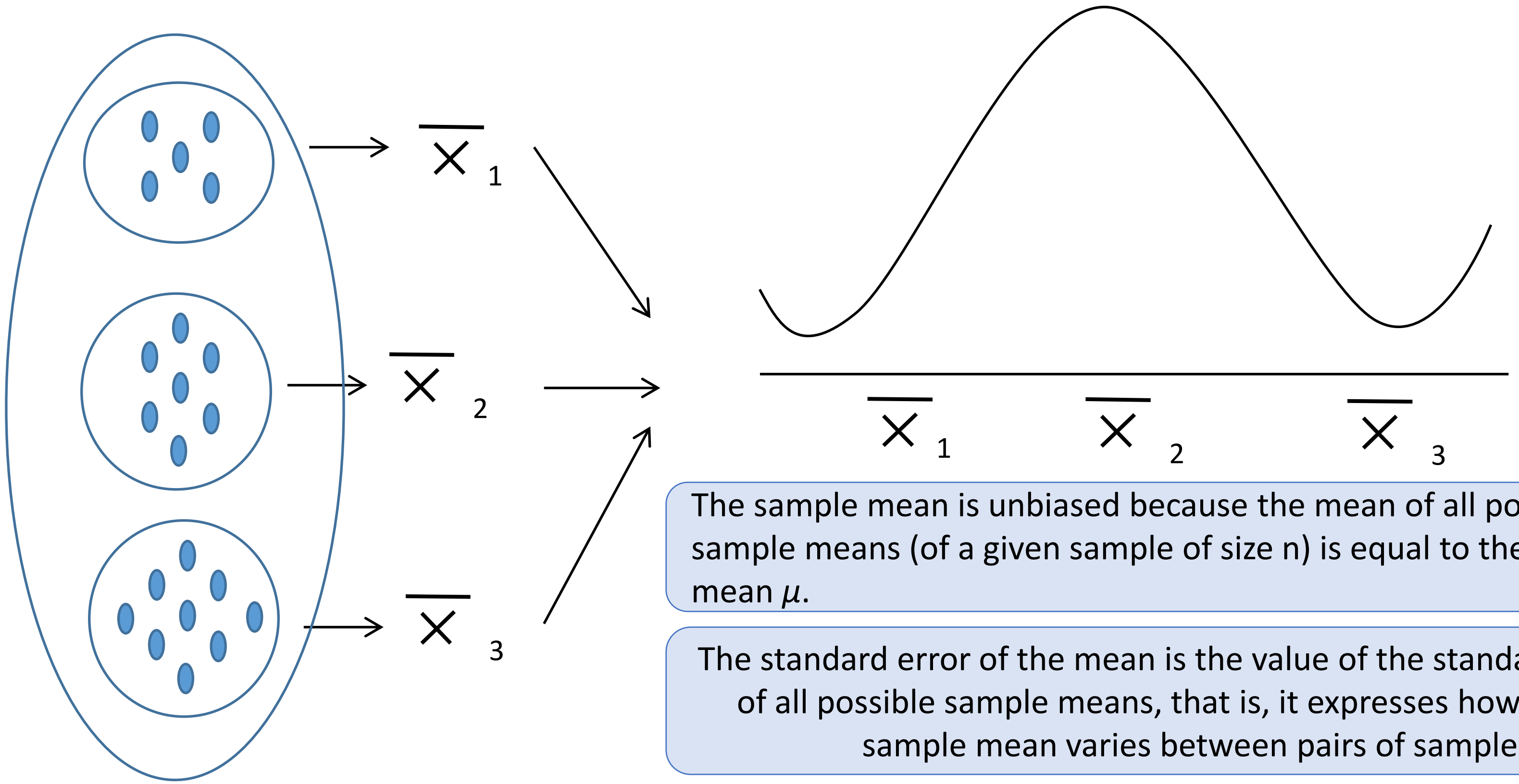
SAMPLING DISTRIBUTION OF THE PROPORTION

$p \longrightarrow P$

The sampling distribution is the distribution of the obtained results if all possible samples were effectively selected

Sampling Distributions: Mean sample distribution

- Distribution of all possible means if all possible samples of a certain size were selected, i.e., it is a frequency distribution, not of raw values, but of sample means, where each mean of the sample is based on a random sample of n raw values



The sample mean is unbiased because the mean of all possible sample means (of a given sample of size n) is equal to the population mean μ.

The standard error of the mean is the value of the standard deviation of all possible sample means, that is, it expresses how much the sample mean varies between pairs of samples.

Sampling of populations without normal distribution (Central Limit Theorem)

- When the sample size (i.e., the number of values in each sample) is large enough, the sampling distribution of the mean has an approximately normal distribution.
- This holds no matter the distribution of the individual values within the population, i.e., the *sampling distribution of the mean* approaches to "normal" behavior as *n* (sample size) increases (this is usually accepted for $n \geq 30$)
  1) For most population distributions, regardless of their shape, the *sampling distribution of the mean* has *an approximately* normal distribution when samples of at least 30 items are selected.
  2) If the population distribution is "fairly symmetric", the *sampling distribution of the mean* is approximately ***normal*** for samples as small as 5-item samples.
  3) If the population has a normal distribution, the sampling distribution of the mean also has a normal distribution, regardless of the sample size.

The Hypothesis test for the population mean will be carried out using the z-scores of the normal curve.

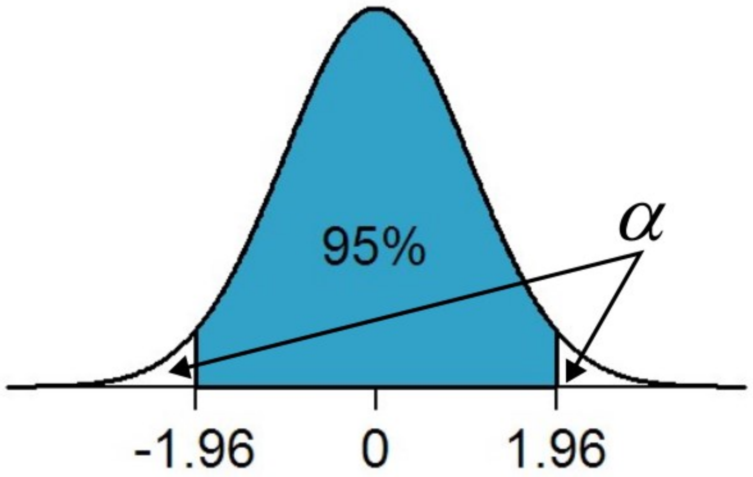
$$z = \frac{x - \mu}{\sigma}$$

Z value and confidence interval for the normal distribution

- The critical value for *z*, $z_{\alpha/2}$ is estimated according to the confidence level (1-α)
- $z_{\alpha/2}$ is the required *Z* score so that the area under the curve is evaluated to α (two-tailed version)
- The confidence interval is associated with the probability that a specific population parameter (e.g., the mean value) lies within the specified interval for a future observation

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- It represents the number of standard deviations required to reach the confidence level



If the acceptable error is α = 0.05 (5%), the confidence level will be at 95% (area under the curve of the distribution function)

Confidence level	Critical value $z_{\alpha/2}$
99%	2.576
98%	2.326
95%	1.96
90%	1.645

Distribution of the sample proportion

- In categorical variables (dichotomous or polytomous), the proportion *P* ("proportion of interest") is the ratio of elements in the entire population that have the interesting/relevant feature.
- The sample proportion, represented by *p*, is the ratio of elements in the sample that present the characteristic of interest.
- The sample proportion is used to estimate the population proportion, a parameter.

$$p = \frac{cases\ of\ interest}{total\ cases}$$

- The sampling distribution of the proportion generally follows the model of a probabilistic distribution for discrete quantitative variables: *the Binomial Distribution*.
- However when it happens that *n * P* and *n * Q* are ≥ 5, the binomial distribution can approximate the model of the normal curve and, consequently, the hypothesis test for the population proportion is carried out through the z-scores of the normal curve.

Standard proportion error:

$$\sigma p = \sqrt{\frac{P * Q}{n}}$$

Hypothesis contrast:

$$z = \frac{p - P}{\sqrt{\frac{P * Q}{n}}}$$