Basic of Statistics

Basic concepts

Noara Razzak

April 2, 2020

Page 1

Basics of Statistics

Today, we will cover the basics of statistics such as Population, Sample, Parameter, Statistic, Variable and Data.

Population

Population is a collection of persons, things, or objects under study.

Population all possible observations relevant to the research question.

Example: If studying the average height of all adult women in a country, the population is every adult woman in that country.

Types of Populations:

- Finite Population (countable, e.g., all students in a school).
- Infinite Population (uncountable, e.g., all possible outcomes of rolling a die infinitely).

Sample

To study the population, we select a sample. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population.

By studying the sample, statisticians estimate population parameters while accounting for sampling error.

Sampling error

Sampling error refers to the difference between a sample statistic (e.g., sample mean) and the true population parameter (e.g., population mean) that occurs because the sample is not a perfect representation of the entire population

Sampling Error

Reducing Sampling Error

- Increase sample size (n).
- Use random sampling (avoids bias).
- Stratified sampling (ensures subgroups are represented).

Random Sampling

Simple Random Sampling

- **Description:** Every individual has an equal chance of selection.
- **Example:** Drawing names from a hat or using a random number generator.
- **Pros:** Unbiased, easy to implement.
- **Cons:** Requires a complete population list; may not represent subgroups well.

Systematic Sampling

- **Description:** Select every k-th element from a list, where $k = \frac{N}{n}$.
- **Example:** Choosing every 10th student from a school register.
- **Pros:** Simple and evenly spread.
- **Cons:** Risk of periodicity bias if the list has a hidden pattern.

Random Sampling

Stratified Sampling

- **Description:** Population is divided into subgroups (strata), and random samples are taken from each.
- **Example:** Separating employees by department and sampling proportionally.
- **Pros:** Ensures subgroup representation.
- **Cons:** Requires prior knowledge of strata.

Cluster Sampling

- **Description:** Population is divided into clusters, and entire clusters are randomly selected.
- **Example:** Randomly selecting 5 schools out of 50 and surveying all students.
- **Pros:** Cost-effective for large populations.
- **Cons:** Higher sampling error if clusters are not diverse.

Non-random Sampling

Convenience Sampling

- **Description:** Selecting the most easily accessible subjects.
- **Example:** Surveying people in a shopping mall.
- **Pros:** Quick and inexpensive.
- **Cons:** Highly biased; not generalizable.

Purposive (Judgmental) Sampling

- **Description:** Researcher handpicks participants based on specific criteria.
- **Example:** Interviewing only doctors for a medical study.
- **Pros:** Useful for specialized research.
- **Cons:** Subjective and prone to bias.

Non-random Sampling
Quota Sampling

- **Description:** Similar to stratified sampling, but selection within strata is non-random.
- **Example:** Surveying 50 men and 50 women, chosen conveniently.
- **Pros:** Ensures subgroup representation.
- **Cons:** Still biased due to non-random selection.

Snowball Sampling

- **Description:** Existing participants refer others (used for hard-to-reach groups).
- **Example:** Studying homeless populations via referrals.
- **Pros:** Useful for niche populations.
- **Cons:** Highly biased; not representative.

Parameter

A parameter is a number that is a property of the population. If we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

Statistic

A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic.

Bringing everything together

Example:

Research Question: What is the average income of software engineers in the U.S.?

- Population: All software engineers in the U.S.
- Sample: A randomly selected group of 1,000 software engineers.
- Parameter: True average income (μ).
- Statistics: Sample average income ($\bar{x}$).

Variable

A variable, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be numerical or categorical.

Data

Data are the actual values of the variable. They may be numbers or they may be words.

Frequency

Frequency refers to how often a value or category occurs in a dataset.

Next we discuss the types of frequency.

1. Absolute Frequency (f_i) The count of how many times a value x_i appears in the dataset.

$$f_i = \text{Number of occurrences of } x_i$$

2. Relative Frequency (rf_i) The proportion of the absolute frequency to the total number of observations (N).

$$rf_i = \frac{f_i}{N}$$

3. Cumulative Frequency (F_i) The sum of absolute frequencies up to a certain value x_i (used in ordered data).

$$F_i = \sum_{j=1}^{i} f_j$$

4. Percentage Frequency (pf_i) Relative frequency expressed as a percentage.

$$pf_i = rf_i \times 100\%$$

Example Consider the dataset: $[3, 5, 3, 7, 5, 3]$. Here, $N = 6$.

Value (x_i)	Absolute (f_i)	Relative (rf_i)	Cumulative (F_i)	Percent (pf_i)
3	3	$\frac{3}{6} = 0.5$	3	50%
5	2	$\frac{2}{6} \approx 0.333$	5	33.3%
7	1	$\frac{1}{6} \approx 0.167$	6	16.7%

Next class we will cover measures of the central tendency of data, skewness, mean, median and mode.