Basic terms of Statistics

Population

All the members of a group about which you want to draw a conclusion.

EXAMPLES: All U.S. citizens who are currently registered to vote, all patients treated at a particular hospital last year, the entire daily output of a cereal factory's production line.

Sample

The part of the population selected for analysis.

EXAMPLES: The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient-satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

Parameter

A numerical measure that describes a characteristic of a population.

EXAMPLES: The percentage of all registered voters who intend to vote in the next election, the percentage of all patients who are very satisfied with the care they received, the average weight of all the cereal boxes produced on a factory's production line on a particular day.

Statistic

A numerical measure that describes a characteristic of a sample.

EXAMPLES: The percentage in a sample of registered voters who intend to vote in the next election, the percentage in a sample of patients who are very satisfied with the care they received, the average weight of a sample of cereal boxes produced on a factory's production line on a particular day.

Tabulation

The process of placing classified data into tabular form is known as **tabulation**. A table is a symmetric arrangement of **statistical** data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements

Class Boundary

When we have different **classes** of data, there **is** always an upper and a lower **class** limit for it i.e. the dataset has a smallest and largest value. **Class boundary is** the midpoint of the upper **class** limit of one **class** and the lower **class** limit of the subsequent **class**.

Introduction to Data Types

Having a good understanding of the different data types, also called measurement scales, is a crucial prerequisite for doing Exploratory Data Analysis (EDA), since you can use certain statistical measurements only for specific data types.

1. Categorical Data

Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.

a) Nominal Data

Nominal scales are used for labeling variables, without any quantitative value. "Nominal" scales could simply be called "labels." Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. Example

Gender: a) Male b) female

Book colour: a) red b) orange c) green d) blue

b) Ordinal Data

An ordinal scale is one where the order matters but not the difference between values.

Examples of ordinal variables include:

- socio economic status ("low income","middle income","high income"), education level ("high school","BS","MS","PhD, satisfaction rating ("extremely dislike", "dislike", "neutral", "like", "extremely like").

2. Numerical Data

a) Discrete Data

Discrete data is information that we collect that can be counted and that only has a certain number of values. **Examples of discrete data** include the number of people in a class, test questions answered correctly, and home runs hit. Tables and graphs are two ways to show the **discrete data** that you collect

b) Continuous Data

Represents measurements and therefore their values **can't be counted but they can be measured**. An example would be the height of a person, which you can describe by using intervals on the real number line.

c) Interval Data

Interval scales are numeric scales in which we know both the order and the exact differences between the values. The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

d) Ratio Data

Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.

Frequencies:

A frequency is the number of times a data value occurs. For **example**, if ten students score 80 in **statistics**, then the score of 80 has a **frequency** of 10. **Frequency** is often represented by the letter f.

Proportion:

A **proportion** refers to the fraction of the total that possesses a certain attribute. For example, suppose we have a sample of four pets - a bird, a fish, a dog, and a cat. ... Therefore, the **proportion** of pets with four legs is 2/4 or 0.50

You can easily calculate the proportion by dividing the frequency by the total number of events. (e.g how often something happened divided by how often it could happen)

Group and ungroup data

Ungrouped data is the **data** you first gather from an experiment or study. The **data** is raw — that is, it's not sorted into categories, classified, or otherwise **grouped**. An **ungrouped** set of **data** is basically a list of numbers.

Grouped data is **data** that has been organized in classes after its analysis.

Grouping of Data

1. The marks obtained by 40 students of class VIII in an examination are given below:

16, 17, 18, 3, 7, 23, 18, 13, 10, 21, 7, 1, 13, 21, 13, 15, 19, 24, 16, 2, 23, 5, 12, 18, 8, 12, 6, 8, 16, 5, 3, 5, 0, 7, 9, 12, 20, 10, 2, 23

Divide the data into five groups, namely, 0-5, 5-10, 10-15, 15-20 and 20-25, where 0-5 means marks greater than or equal to 0 but less than 5 and similarly 5-10 means marks greater than or equal to 5 but less than 10, and so on. Prepare a frequency table for the grouped data.

Solution:

Arranging the given observations in ascending order, we get them as

0, 1, 2, 2, 3, 3, 5, 5, 5, 6, 7, 7, 7, 8, 8, 9, 10, 10, 12, 12, 12, 13, 13, 13, 15, 16, 16, 16, 17, 18, 18, 18, 19, 20, 21, 21, 23, 23, 23, 24

Thus, the frequency distribution may be given as under:

Marks	Tally Marks	Frequency
0 - 5	𝍸 I	6
5 - 10	𝍸 𝍸	10
10 - 15	𝍸 III	8
15 - 20	𝍸 IIII	9
20 - 25	𝍸 II	7
	Total	40