Statistics – from the top make it drop

**Chapter 1 – Basics and definitions**

Statistics rotates around the concept of **collecting data** and **analysis of the collected data:**

* **Design** how to obtain data
* **Describe** in a synthetic way the data
* **Infer** decisions and predictions on what the data means

Initial definitions:

* **Population** The set of all possible observations
* **Sample** The subset of the population that gives us the data
* **Parameter** A numerical summary of the population (e.g., proportions)
* **Statistic** A numerical summary of the sample
* **Statistical unit** Unit of observation for which data are collected or derived

The first part of the course is **Descriptive statistics,** which is made of techniques to capture the essence of data

**Randomness and variability**

Randomness should be present to ensure a **fair** representation of the population, it’s the **lack of pattern** in the data set

Variability is instead related to the variables per se, that provide a known pattern of the data set (e.g., parabola), also referring to how the data is spread out

**Chapter 2 – Types of data**

A variable is any characteristic observed on the subjects in a study, they mainly sit in one of two groups:

* **Categorical,** with each value belonging to one of a set of categories:
  + **Binary,** 0/1 or T/F
  + **Ordinal,** Hierarchy or order of preference implied
  + **Nominal,** No hierarchy implied
    - **KEYWORD: Percentage of observations in each category, groups**
* **Quantitative,** with observations taking numerical values that represent different extents of the variable itself
  + **Discrete,** if the set of possible values is not dense (usually integer numbers)
  + **Continuous,** if the set of possible values is dense (the values come from an interval and they usually are decimal)
    - **KEYWORD: Center and spread (location and variability), amount**

Example : Identify the variable types

1. Number of siblings in a family Quantitative, Discrete
2. County of residence Categorical, Nominal
3. Distance (in miles) of commute to school Quantitative, Continuous
4. Marital status Categorical, Nominal
5. Length of time to take a test Quantitative, Continuous
6. Number of people waiting in line Quantitative, Discrete
7. Number of speeding tickets from last year Quantitative, Discrete
8. Your dog’s weight Quantitative, Continuous

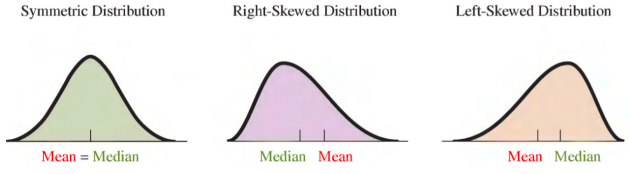
**Proportion, percentage → relative frequencies**

* Proportion = (category is a variable, see example below)
* Text

  Description automatically generatedPercentage = proportion \* 100

All the relative frequencies and observations are usually saved in a **frequency table**

**Shape of data distribution**

* **Unimodal** → the highest point is the mode (it has one peak)
* **N-modal** → multiple peaks

We must now introduce some concepts:

* **Mean** 
  + the mean is affected by extreme values
* **Median,** middle value of observations resistant to outliers (extreme values):
  + If observations are even, the median is the middle value
  + Else, the median is the sum of the two middle values divided by two
* **Mode,** most appearing value in observations
* **Range** 
  + **Text

    Description automatically generated**Max – Min (difference between outliers)

Note that when talking about samples n becomes n-1

**Measures of position**

* Percentile
  + A value s.t. *p* percent of the observations fall below or at that exact value
* Chart, histogram

  Description automatically generatedQuartile
  + Division in 4 parts, with

each part being the median, starting from the middle

* IQR – InterQuartile Range
  + Distance between third and first quartile

**The Box plot**

It is represented (horizontally or vertically oriented) by a rectangle divided into two parts, from which two segments come out

The rectangle (the "box") is bounded by the first and third quartiles, q1/4 and q3/4, and divided inside by the **median**, q1/2

Chart, box and whisker chart

Description automatically generatedThe segments (the "whiskers") are delimited by the **minimum and maximum values**

Potential outliers are represented with stars (\*)

Graphical user interface, text, application

Description automatically generated**Z-score (stantardization)**

The Z-score is the number of std deviations by which an observed value is above or below the mean, with the following properties:

* Positive z-score indicates the observation is above the mean
* Negative z-score indicates the observation is below the mean
* Mean is 0 and std deviation is 1