Lecture 7

Descriptive Statistics: → Describes and summarizes data
[But do not allow us to draw conclusions about the whole population from which we took the sample]

The two keywords "population" and "sample" are quite widely used in the statistical world.

Ex: Suppose we want to estimate the stress level of students in Section-G for the DSMD final exam.

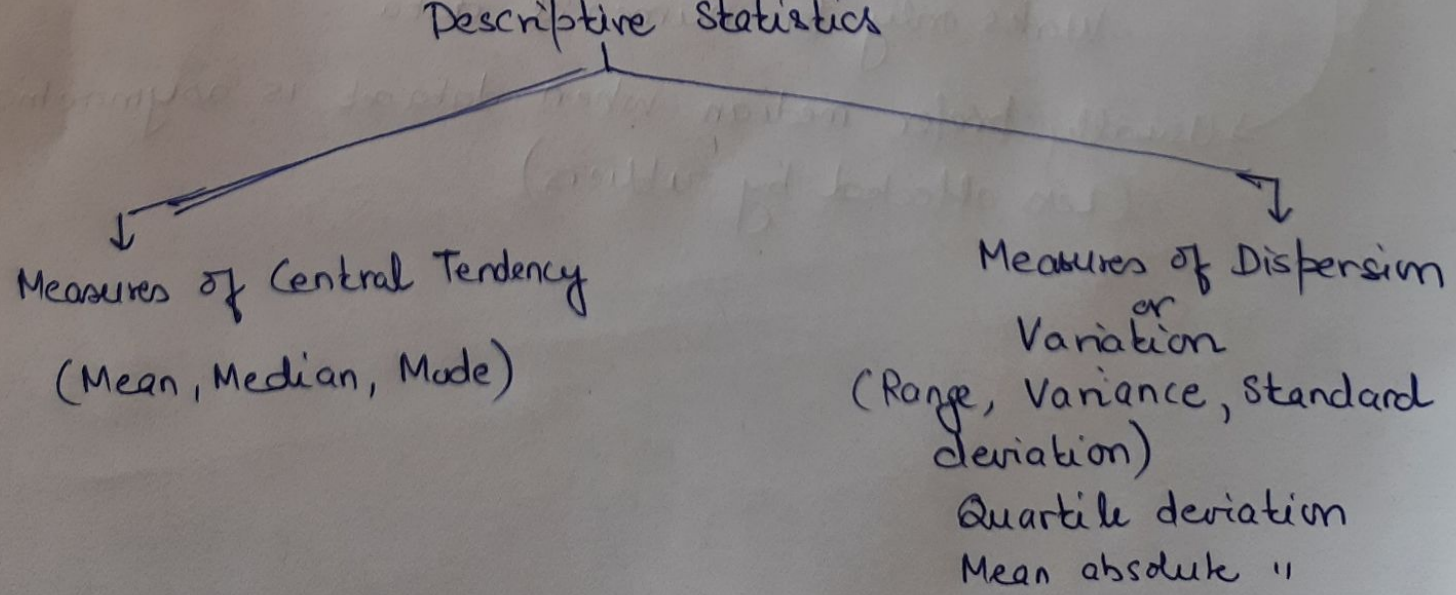
Population: All the students in Section-G

Sample: We select a few random students from the entire pool and estimate/measure their stress levels.

Getting estimate of the entire population is expensive and time consuming. So sampling is needed.

Using descriptive statistics, one can summarize the data for a sample.

There is another classification of statistics, known as the inferential statistics, where we use statistics to test a hypothesis, draw conclusions and make predictions about a whole population, based on your sample.



* Central tendency (also called measures of location or central location)
  - ↳ describes what's typical for a group of data
    - ↳ i.e. it doesn't tell you what's typical about each one piece of data, but gives an overview of the whole picture of the entire dataset.

Mean: Average of the entire dataset
- ↳ Sum all the entries and divide by the total number of points / observations / measurements.
- ↳ Can be used to find both continuous and discrete numerical data

Mode: The mode of a set of data is the item in the set that occurs most often.
- ↳ Works for both numerical and categorical data

Median: → Middle value in the dataset.
- ⟶ First listing the data in a numerical order
- ⟶ Second Locating the value in the middle of the list.

Works only for numerical data

↳ Usually prefer median when dataset is asymmetrical (Less affected by outliers)

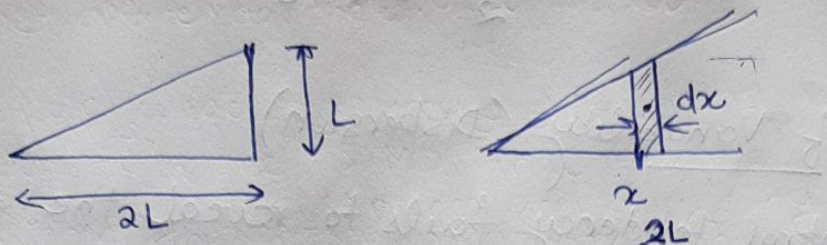
Ex: ① Discrete data

Retirement age in a company

62, 70, 60, 63, 66

↳ Mean = $\frac{320}{5} = 64$

② Continuous data: Find centre of Mass of the object below



Let ρ be the mass per unit area

$$\text{Center of Mass (x-coordinate)} = \frac{\int_0^{2L} (\rho\, dx)\, x \cdot x}{\int_0^{2L} \rho x\, dx}$$

$$= \frac{\int_0^{2L} x^2 dx}{\int_0^{2L} x\, dx} = \frac{(2L)^3}{3} \cdot \frac{2}{(2L)^2}$$

$$= \frac{4L}{3}$$

$$\text{Center of mass (y-coordinate)} = \frac{\int_0^{2L} (\rho\, dx)\, x \cdot \frac{x}{4}}{\int_0^{2L} \rho x\, dx} = \frac{L}{3}$$

∴ Center of Mass is located at $\left(\frac{4L}{3}, \frac{L}{3}\right)$.

③ Find mode of the following dataset

55, 55, 55, 56, 56, 57, 58, 58, 59, 60

↳ Frequency = 3 (Highest) → Mode is 55.

Limitations: Mode may not reflect the center of the set. In this example, center of the dataset is 57, but mode is lower.

④ Find the median of the dataset

24, 31, 28, 21, 22, 32, 24, 29, 27, 26

↳ Sort the list:

21, 22, 24, 24, 26, 27, 28, 29, 31, 32

(26, 27: Middle)

$$\text{Median} = \frac{26+27}{2} = 26.5$$

* Measures of Variability (Dispersion)

↳ Central tendency fails to reveal the extent to which the values of the individual items differ in a dataset.

Ex: Consider two datasets.

2, 2, 2, 4, 6, 6, 6 → Mean = 4, Median = 4

1, 1, 1, 4, 7, 7, 7 → Mean = 4, Median = 4

↳ Even though the second dataset is more variable.

Dispersion helps to interpret variability in the data, like in a group of very rich and very poor people, it helps us capture disparity.

Measures:

* Range: Simply the difference b/w the maximum value and the minimum value. $\boxed{\text{Range} = X_{max} - X_{min}}$

* Variance: Captures the average degree to which each point differs from the mean. (eg: used to capture market volatality).

$$\text{Sample Var}(\sigma^2) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \rightarrow \text{Bias correction term}$$

$$\text{Population Var} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Ex: Suppose the mean of the whole population is 2050, but the statistician doesn't know that, and must estimate it based on the small sample chosen randomly from the population:

2051, 2053, 2055, 2050, 2051

Sample Mean = 2052 (Mean is biased)

Suppose, one does not use the bias correction term, the variance would then be given by:

$$\frac{1}{5}\left[(2051-2052)^2 + \dots + (2051-2052)^2\right] = \frac{16}{5} = 3.2$$

And, the population variance is:

$$\frac{1}{5}\left[(2051-2050)^2 + \dots + (2051-2050)^2\right] = \frac{36}{5} = 7.2$$ > Sample Variance

(This is true always, except when sample mean = Population mean)

* Standard deviation: Square root of the variance (σ). Same unit as the original data.

* Mean absolute deviation: Mean of absolute deviations from the mean, i.e., Average of all $|x_i - \bar{x}|$.

* Quartile deviation: The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance b/w the third and the first quartiles.

Example: The wheat production (in kg) of 20 acres is given as:
1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, 1885.
Find the quartile deviation.

Sol^n: Arrange the observations in ascending order

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960

Q_1 = Value of $\left(\frac{n+1}{4}\right)^{th}$ item

= Value of $\left(\frac{20+1}{4}\right)^{th}$ item

= Value of $(5.25)^{th}$ item

= 5^{th} item + 0.25 (6^{th} item − 5^{th} item)

= 1240 + 0.25(1320 − 1240) $\Rightarrow$ $Q_1 = 1260$

Q_3 = Value of $\frac{3(n+1)}{4}^{th}$ item

= Value of $(15.75)^{th}$ item

= 15^{th} item + 0.75(16^{th} item − 15^{th} item) = 1750 + 0.75(1755 − 1750)

= 1753.75

∴ Quartile deviation, $\frac{Q_3 - Q_1}{2} = \frac{1753.75 - 1260}{2} = 246.875$

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.164$