Day_1 Statistics

Statistics :-

Q- what is Statistics ?

↳ is the science of collecting, organizing and analyzing data. { Better Decision Making }

• Data: facts or pieces of information that can be measured.

ex: Ages of student of a class

{ 30, 25, 24, 22, 28, 27 }

Type :-

Descriptive Stats: It consist of organizing and summarizing data.

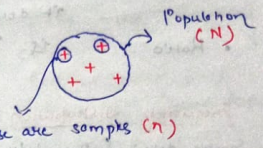
Inferential Stats: Technique we use the data that we have measured to form Conclusions!

ex: Q. what is the average age of student of a class : Descriptive

Q. Are the ages of the student of this classroom similar to the Maths classroom in the College ? Inferential

* Population And Samples:

for ex. UP, Goa Population we have to predict Exit poll



Sampling Techniques

• Simple Random Sampling

↳ Every member has equal chance of being get selected.

• Stratified Sampling

↳ where the population is split into non-overlapping groups

for ex: → Male
↳ female } survey

Age group: (10-20) (20-30) (30-40)

• Systematic Sampling

(N) → (nth) individual

eg Mall → survey (Covid)

↳ Every 8th person

- Convenience Sampling
  ↳ survey ← for ex: Data Science } only those who interested and have knowledge of it.

★ Variables: A variable is a property that can take on any value.

ex. Height = {182, 172, 168, 150, 160}
Weight = {60, 55, 75, 82}

Two kinds of Variables:-

- Quantitative Variable: → Measured Numerically. {Add, sub, multiply}
- Qualitative / categorical Variable:- Gender { Male, Female }
  ex: Blood group

Four Types of Measured Variables:

- Nominal data {categorical data} → classes
- Ordinal → Order of the data matters, value does not — for ex rank of students.
- Interval → order matters, value also matters but it does not include zero value
- Ratio — ex: Temperature {Fehrenheit}.

Frequency Distribution:

Sample data: {Rose, Lily, Sunflower, Rose, Lily, Rose, Lily}

ex

Flower	Frequency	Cummulative fre
Lily	3	3
Rose	3	6

★ Bar graph Vs Histogram
Discrete — continuous

used for

$$\mu = \frac{1+1+2+2+3 \cdots}{5} = \frac{9}{5}$$

Arithmetic Mean for Population (N) and Sample (n)
Average

$X = \{1, 1, 2, 2, 3, 3, 4, 5\}$

$$\mu = \sum_{i=1}^{N} \cdot \frac{x_i}{N}$$

$$\bar{x} = \sum_{i=1}^{n} \cdot \frac{x_i}{n}$$

focus on Notation used

Central Tendency

① Mean ② Median ③ Mode

Refers to the measure used to determine the centre of the distribution of data.

ex $\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, \underbrace{100}\}$ → These are called {outliers}

Mean = $\frac{32+100}{10+1} \Rightarrow \mu = 3.2$

$\Rightarrow \frac{132}{11} \Rightarrow \mu = 12$

- Median: first step sort the data.

{Median works good with outlier well}

As we add outlier but it will only show slight change rather than mean

- Mode: we can use this with categorical variable

we can replace our data which is missing to more reccuring data

Measure of Dispersion → means Spread {how well your data is spread}

- Variance:

Ques:

Population Variance

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

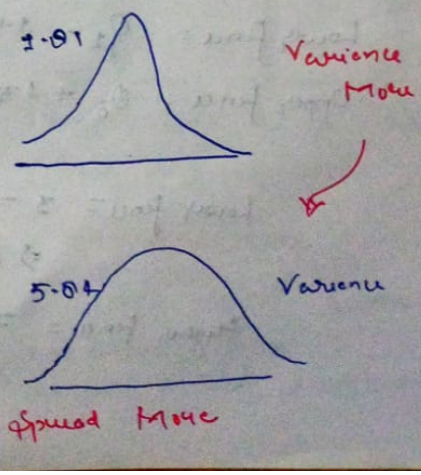
Sample Variance

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

ex:

x	μ	$x-\mu$	$(x-\mu)^2$
1	2.03	-1.03	3.34
2	2.03	-0.03	0.0009
2	2.03	+0.03	0.0009
3	2.03	0.17	0.03
4	2.03	1.17	1.37
5	2.03	2.17	4.71
			10.04

$\mu = 2.03$



Percentiles And Quartlies {find Outliers}

↳ Is a value below which a certain percentage of observation lie.

ex: Dataset: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, $\boxed{10}$, 11, 11, 12 $\quad n = 20$

what is the percentile of 10?

$$\text{Percentile Rank of } x = \frac{\text{No. of values below } x}{n} \times 100$$

$$\Rightarrow \frac{16}{20} \times 100 = 80\%$$

Q. what value exists at percentile ranking of 25%?

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$\Rightarrow \frac{25}{100} \times (21) = \underbrace{5.25}_{\text{This is index value not value}}$$

Five Number Summary {for Removing Outlier}

- Minimum
- First Quartile (Q_1)
- Median
- Third Quartile (Q_3)
- Maximum

- Removing the Outliers

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, 27}

[Lower fence ⟷ Higher fence]

Lower fence = $Q_1 - 1.5\,(IQR)$

Upper fence = $Q_3 + 1.5\,(IQR)$

IQR ⇒ Interquartile Range

$Q_3 - Q_1$
(75%) (25%)

Lower fence = $3 - 1.5(4)$
$\Rightarrow 3 - 6 = -3$

Higher fence = $7 + 1.5(4)$
$7 + 6$
$= 13$

} Remove other data

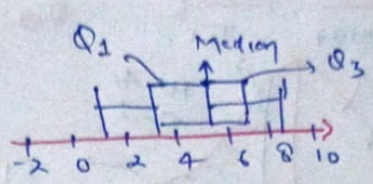
Remaining data:

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, ~~24~~

Minimum = 1
$Q_1 = 3$
Median = 5
$Q_3 \Rightarrow 7$
Max = 9

→ 5 Number Summary

Box Plot



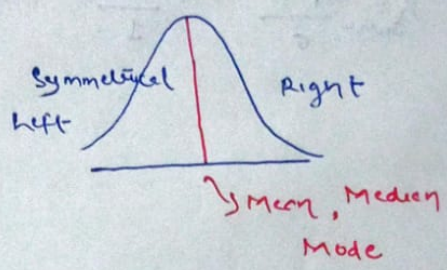
Application of Box Plot

It gives a visualisation to see where an outlier is present

Guassian / Normal Distribution

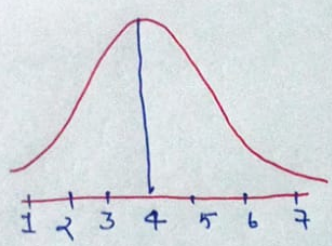
Empirical formula

68 - 95 - 99.7 % Rule



e.g: Height
- weight
- IRIS dataset

$\mu = 4 \quad \sigma = 1$



To find how far from μ

$$\boxed{Z\text{score} = \frac{x_1 - \mu}{\sigma}}$$

$\Rightarrow \frac{4.75 - 4}{1} = 0.75$ sd

how many sd left or right

Practical Application

Dataset

Age (years)	Salary (Rs)	Weight (kg)
24	40k	70
25	50k	80
26	60k	55
27	70k	65

To bring all data in single form as we have two different units

Standardization

⇒ we can apply zscore and convert into Standard Normal distribution

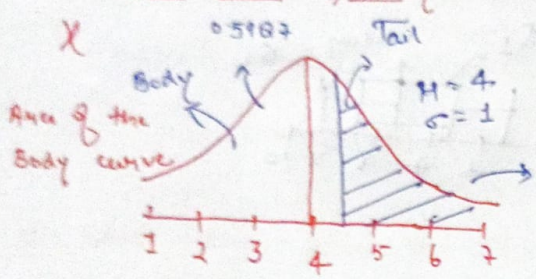
$\{\mu = 0, \sigma = 1\}$

Normalization:

↳ convert data in range (0 to 1)

MinMax Scaler → (0 to 1)

★★ Stats Interview Question



Ques: what percentage of scores falls above 4.25?

(1 − left area)

⇒ 1 − 0.5987

⇒ 0.4013 ⇒ 40%

$$Z_{score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$