What is Statistics?

Science of collecting, organising and analyzing data → facts or peices of information.

Types of Statistics

Descriptive Stats → organising and summarizing data

Inferential Stats → consist of using data you have "measured" to form conclusion

Descriptive	Inferential
① Measure of central tendency (Mean, median, mode)	① Z-test
② Measure of dispersion (Variance, Std Deviation)	② t-test
③ Different type of distribution of data. (PMF, CDF, Histogram)	③ Chi Square Test
	④ Hypothesis testing Pvalue, significance

Population and Sampling

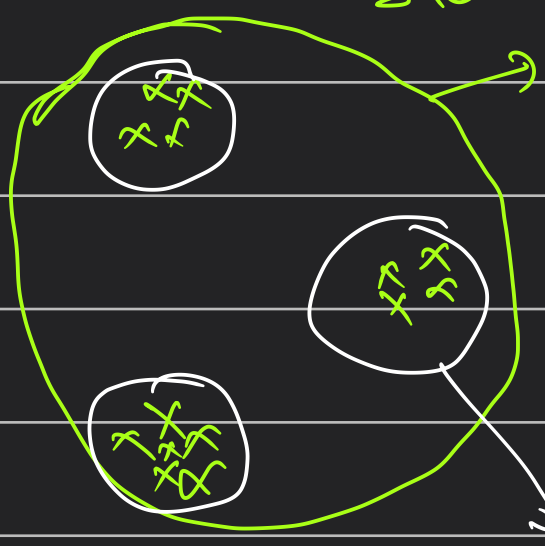
Population :- the group we are intrested in studying

Sample :- a subset of population.

for Eg - Exit poll {Media} ⇒ will try to predict who will win?

Suppose a state 'A' is having election.

State A-1 precinct



The media went to these region and asked who did you vote based upon response from these regions media will predict who will win? These regions can be called Samples.

Remember one thing Exit polls are not always true....

- Types of Sampling Techniques

Techniques used to create a sample that is representative of entire population these techniques are called sampling techniques.

[Population is represented by "N" and sample is represented by "n"]

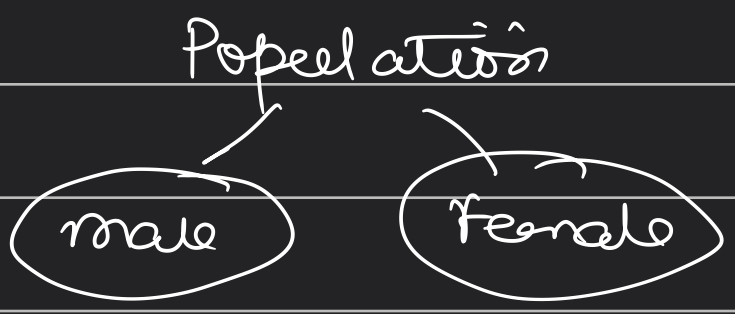
① Simple Random Sampling

We randomly select a sample "n" from population "N".

When performing this, every member of population (N) has an equal chance of being selected for your sample (n).

② Stratified sampling

↳ layering ② Non overlapping groups



Let's say I wanna do a survey where I need info from male group. (Separate grouping, where groups dont have common thing to overlap) then we can use Stratified sampling.

③ Systematic Sampling

Involves selecting k^{th} element from popul^n to be included in sample. Where k^{th} is the sampling interval or skip interval.

Ex Select^n every 4th guy in a mall to fill a survey.

④ Convenience Sampling

↳ Voluntary response sampling

sample is selected based on convenience & accessibility. Involves choosing things which are readily available & easy to reach.

[We use different sampling technique based upon our different use cases]

What are variables and its types?

is a property that can take on many values.

Variable ⇒ Singular Mode.

Ages = [2, 10, 12, 14] ⇒ not a variable.

- Types of Variable

Quantitative Variable

- **Discrete**
  - ↓
  - measurable number
  - or
  - No of bank accounts,
  - No of children in family
- **Continous**
  - is continous in nature. can be upto infinity
  - Eg Height 153.7, 165.5. or weight

Qualitative / Categorical

- ↓
- Based upon some property we classify the variable
- eg Gender — male, female
- Types of flower — Rose, Lily, Cactus
- We divide these things based upon some characteristic

Measure of Central Tendency - Mean, Median & Mode

Central tendency refers to measure used to determine the "center" of distribution of data.

$\{1, 2, 3, 4, 5, 6\}$

Mean = Average

Population Data (N)

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

↓ population mean

Sample (n)

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

↓ Sample mean

Median = Central value (more called measure of central tendency)

1. Sort all numbers
2. Find central element
   - odd length = middle element
   - even length = Average of two central elements.

Mode :- most frequent elements

[1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 17, 9] = 5 is repeated most so 5.

Use Cases

① Outliers ⇒ median

② Categorical features ⇒ mode, based upon if some category is repeating more.

Measure of Dispersion - Variance & Standard Deviation.

Measure of Dispersion talks about Spread.

↳ Determined two thing Variance & Standard deviation

We measure these things around Popelatn & sample.

Variance of Population

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

μ → population mean

Sample Variance

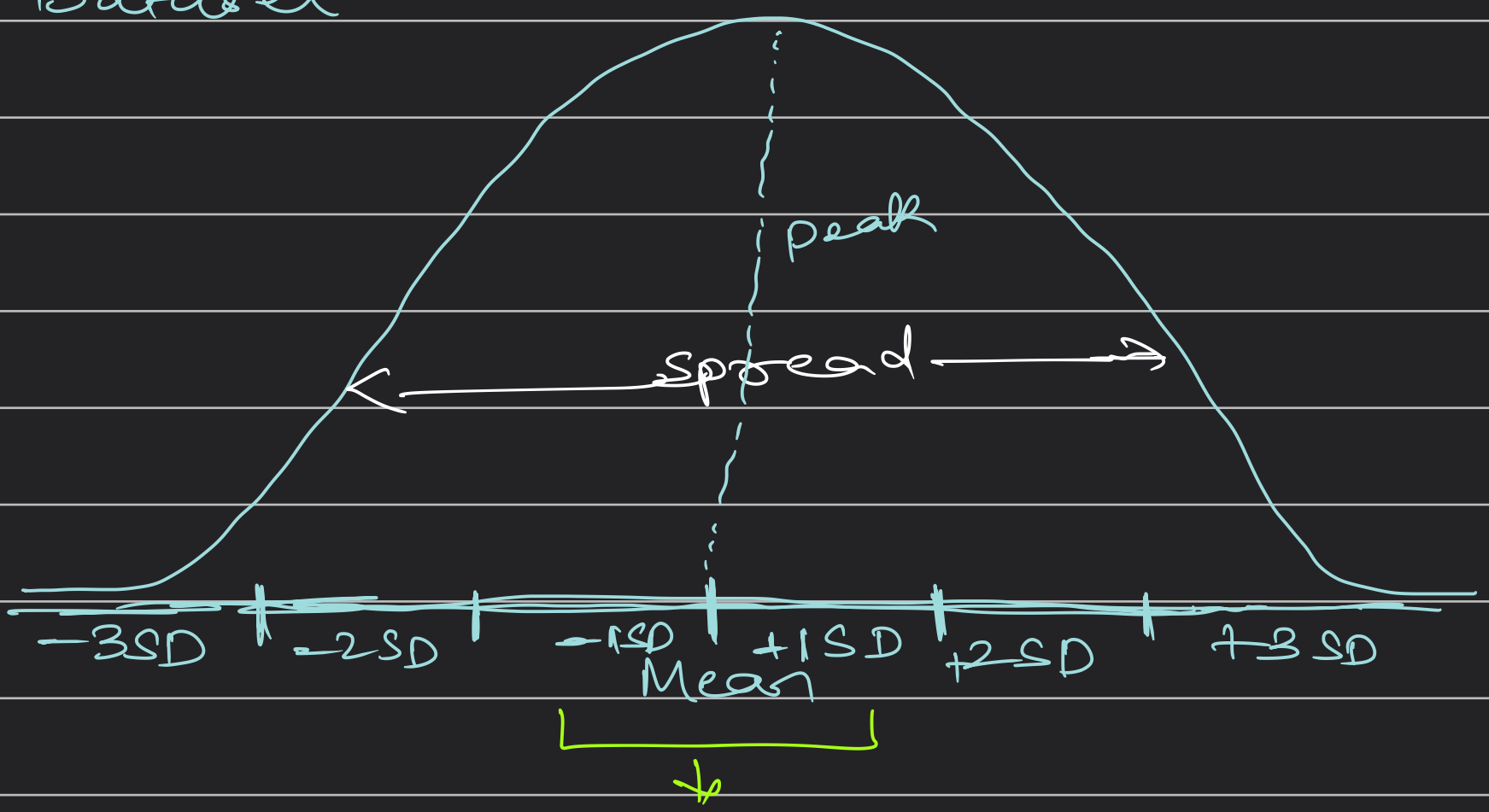
$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Bessel Correction or Degree of Freedom

Standard Deviation = $\sqrt{\text{Variance}}$

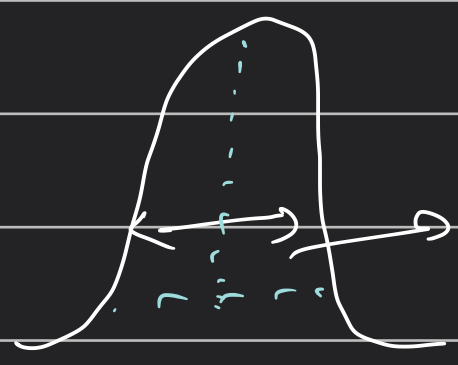
#Let understand what these mean

Dataset



Any value around +1SD -1SD of the mean will lie here.

The bigger the "Variance" the bigger the spread of data



small Variance, small spread

The less the S.D more the "height"/peak of bell curve and less the "spread".

Percentiles & Quarantiles

is a value below which a certain percentage of observation lies.

Ex If say I got 95 percentile. this means that I have better marks than 95% of other students or 95% of people got marks less than me.

Dataset : 1, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 10, 11, 11, 12

what is percentile rank^ of 10?

percentile rank^ = $\frac{\#\text{No of elements below 10}}{\text{total number of elements}} \times 100$

$= \frac{16}{20} \times \frac{5}{100} = 80^{\text{th}}$ percentile

So rank of "10" is 80 percentile that means 80% of values in dataset are below 10.

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

what value of 25 percentile

$= \frac{25}{100} \times 21 = (5.25)$ → Index

5^{th} & 6^{th} value — average

Quartile → 25^{th} percentile — I^{st} Quartile

50^{th} percentile — II^{nd} Quartile

75^{th} percentile — 3^{rd} Quartile

Inter Quartile Range = 3^{rd} quartile − 1^{st} quartile

used for finding or detecing outliers.

How to construct Box plots for outliers.

5 Number Summary and Box Plot

1. Minimum
2. first quarile (25%) Q1
3. Median
4. Third quarile (Q3)
5. Maximum

These 5 component help us plot a BOX plot which in turn help us find an "Outlier".

⟶ any value that is different from dataset.

Techniques to find outlier

[Lower Fence ⟷ Higher Fence]

Any values smaller than Lower Fence & Bigger than higher fence would be considered as an outlier.

How to determine lower fence & Higher fence

⇒ Lower fence = Q1 − 1.5 (IQR)

Higher fence = Q3 + 1.5 (IQR)

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

19 values

$Q1 = \frac{25}{100} \times (19+1)$

[Data should be sorted order]

$Q1 = 5^{th}$ Index

$Q1 = 3$

$Q3 = \frac{75}{100} \times (20) = 15^{th}$ Index $= 7$

$IQR = Q3 - Q1 = 7 - 3 = 4$

lower fence $= Q1 - 1.5 \times 4$

$= 3 - 6 = -3$

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

$[-3 \longleftrightarrow 13]$ ← Outlier

One value is greater than 13 i.e 27

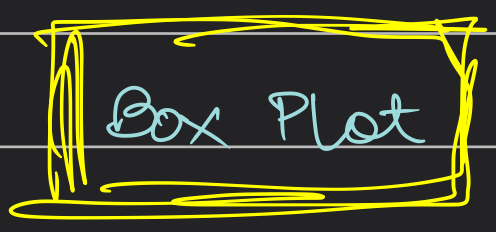
Similarly our 5 numbers will be

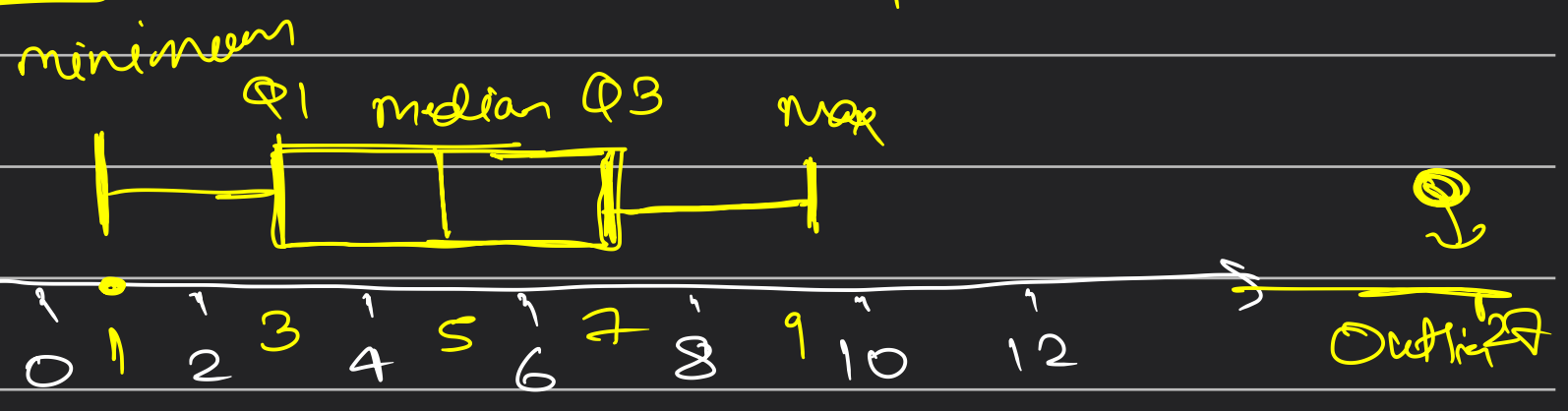
1. Minimum = 1
2. First Quartile $= Q1 = 3$
3. Median $= 5$
4. Third Quartile $= Q3 = 7$
5. Maximum $= 27 = 9$

(outlier removed)

↓

Using this we plot our Box Plot





- finding outlier using these techniques using python. [Check GitHub] Sakpal Akshay Data Science.