STATS — FSDS 2.0

Defn: Statistics is the science of collecting, organizing and analyzing data

Data = "facts or pieces of information"

Eg: Heights of students in classroom
IQ of students
Daily Activities
Weight of people, Age.

Types of Statistics

Data scientist most of the time work on inferential stats

① Descriptive Stats

Defn: It consists of organizing and summarizing data

① Measure of Central Tendency
{mean, Median, Mode}

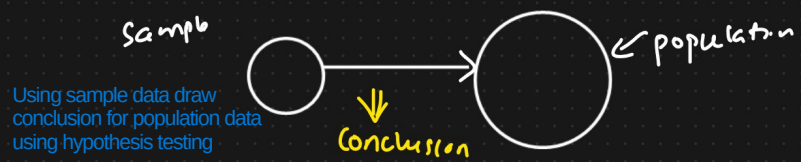
② Measure of Dispersion
{Variance, Standard deviation}

③ Different types of Distribution of data
Eg: Histogram, pdf, pmf, cdf

CLT

② Inferential Stats

Defn: It consists of data you have measured to form conclusion



Using sample data draw conclusion for population data using hypothesis testing

C.I, K, pvalue Hypothesis Testing

① Z-test
② t-test
③ Chi Square Test
④ ANOVA
⑤ F Test

⇒ Conclusion of sample on population.

Eg: Let say there are 20 classes in your college and you have collected the heights of student in the class.

Heights are recorded [175cm, 180cm, 140cm, 135cm, 160cm, 170cm]

Descriptive What can be question based on descriptive stats

"What is the average heights of the students in the classroom"

Mean = 160cms

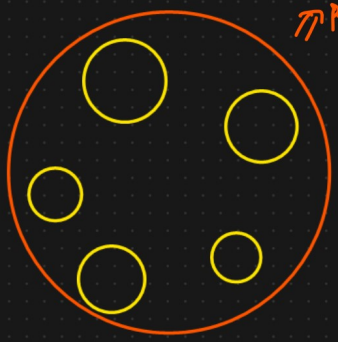
Influential What can be question based on inferential stats

"Are the height of the students (sample) in the classroom similar to what you expect in the college" → population

Population (N) And Sample (n) data

Exit Poll

Exit poll example where a sample of people is selected from different regions and based on that conclusion for whole population is made which is called exit poll.

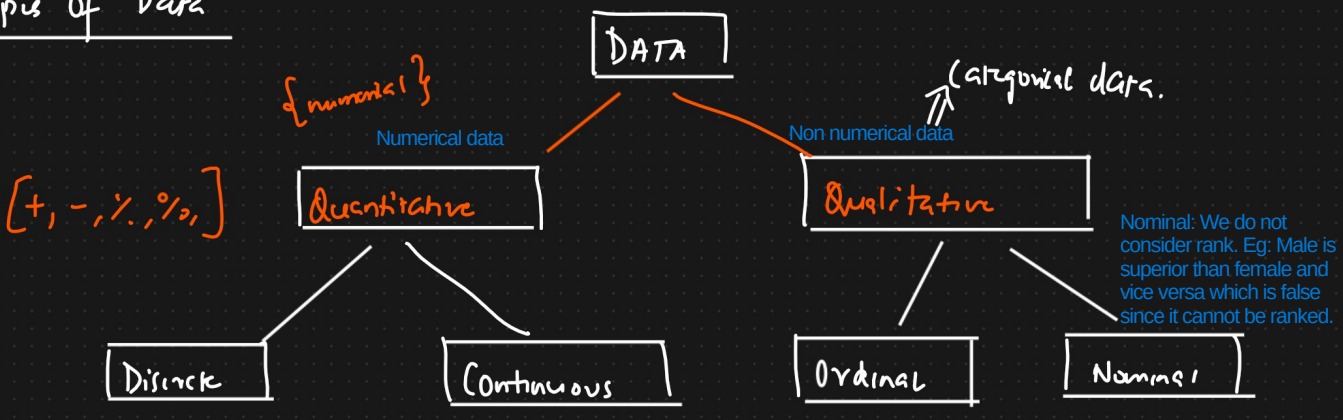


Sample size = 500

Population is super set whereas, sample is subset of population

Exit poll ex. Most of these polls come out to be false. Reason be the way sampling/selection of people done was not appropriate. Selction should cater diversity which not doing so can make it baised.

⑥ Types of Data



Nominal: We do not consider rank. Eg: Male is superior than female and vice versa which is false since it cannot be ranked.

Whole numbers with some range

Eg: No of bank accounts of people

No. of children in a family

Discrete data:
I have 3 bank accounts

Any value

Any value without any range

Eg: Weight, Height, Age, Temperature, Speed, Salary

Continuous data:
I am 24.5 years old. I am 69.5 kg etc etc

Eg: Ranks

[3 Good, 2 Better, 1 Best]

Ordinal: Ranks (Good, Better, Best) considered. Eg: Best day to work is Friday and worst day is Monday.

We can change nominal data into ordinal. For Eg: Days in a week is nominal data but if we ask best and worst day to work, it becomes ordinal.

Eg: Gender

M, F

Blood group

Color of Hair

Pincode

Why pincode is qualitative/categorical and not quantitative?

Ans: Because its not appropriate to find mean of pincodes.

④ Scales of Measurement

How can we measure data?

① Nominal Scale Data
② Ordinal Scale Data
③ Interval Scale Data
④ Ratio Scale Data.

① Nominal Scale Data

1) Qualitative / categorical Data.

Eg: Gender, Colors, Labels

2) Order does not matter

Asking sample of 10 people about their favorite colors?

Eg: Favorite Color

Red → 5 → 50%

Blue → 3 → 30%.

Orange → 2 → 20%.

Total → 10

② Ordinal Scale Data

① Categorical Data

② Ranking and order matters

③ Difference cannot be measured

Ordinal Scale data:
Eg: If we are provided with data of 3 student who scored first, second and third. These ranked are ordinal and differences cannot be measured until marks of each student is specified based on which ranking is defined.

Eg:
- Best → 1
- Good → 2
- Bad → 3

Race

1st	3 min
2nd	4 min
3rd	5 min

1st Rank → 90

2nd Rank → 70

3rd Rank → 40

③ Interval Scale Data

① The order matters

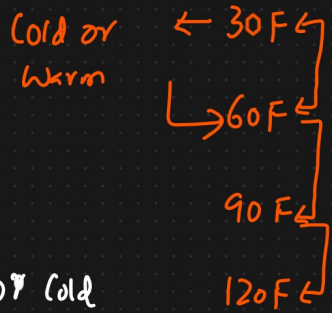
② Difference can be measured

③ Ratio cannot be measured

④ No "0" starting points

Eg: Temperature

Here temp is ordered since it's arranged in ascending order. Anything arranged in ascending/descending order is ordered data.



60:30 = 2:1

If one says that he/she will feel double hot or cold in region with 60F as compared to the region with 30F then this is completely illogical or false as there will be other external factor like humidity etc to determine the same. Hence, ratio cannot be measured.

Room 1 → 30F

Room 2 → 60F

0

-15

-30

④ Ratio Scale Data

Eg: Student marks in class

① The order matter {sort this numbers} . 0, 30, 45, 60, 90, 95, 99

② Differences are measurable including ratios

③ Contain a 0 starting point

Example

① Marital Status [Nominal Scale Data]

② Favourite food based on Gender? [Nominal]

③ IQ measurements [Ratio Scale].

⇓

Ordinal

Descriptive Stats

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency: mode median

① Measure of Central Tendency

① Mean ② Median ③ Mode.

① Mean :

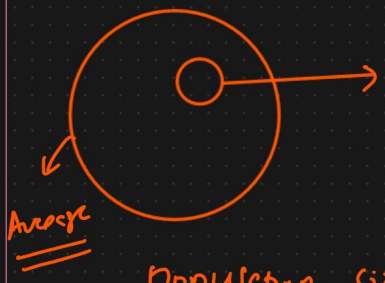
Population (N) | **Sample (n)**

In all this and below examples X is considered as an random variable

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

$$\text{Population Mean } (\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$$

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$



Population size (N)

Sample size (n)

$$\frac{1+1+2+2+3+3+4+5+5+6}{10} = 3.2$$

② Median

$X = \{4, 5, 9, 3, 2, 1\}$

Steps

① Sort the Random variable $\{1, 2, 2, 3, 4, 5\}$

② No. of elements

③ if Count == even

{1,2,2,3,4,5}

2+3/2 = 2.5 median.

if count == odd

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

3 median

Why Median ?

Mean are affected by outliers

Without outlier

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

$\bar{x} = \frac{1+2+3+4+5}{5} = 3$

Median = 3

With outlier

X = {1,2,3,4,5,100} Here 100 is outlier

$\bar{x} = \frac{1+2+3+4+5+100}{1} = \frac{115}{6} \approx 19$

Notice that as we introduce outlier mean got drastically affected whereas, median was not affected much due to the presence of outlier.

X = {1,2,3,4,5,100}

median = 3+4/2 = 3.5

-->Means are affected by outliers.

-->Median is not affected by outliers.

Conclusion :

Median is used to find the central Tendency

When outlier is present.

③ Mode : Maximum Frequency occurring element

{2,1,1,1,4,5,7,8,9,9,10}

Mode : 1

Mean, median and mode are generally used for handling missing values during EDA and feature engineering steps

Mean and median is basically used for handling numerical missing values.

Whereas, mode is basically used for handling categorical missing values.

Instead of handling missing values using mean, median and mode if one have dropped the entire row as a resolution then there would be loss of data. Hence, mean, median and mode can be used as a better alternative.

EDA and Feature Engineering

- Missing Value

Age	Weight	Salary	Gender ↓ Mode	Degree ↓ Mode
24	70	40K	M	B.E
25	80	70K	F	- B.E
27 ← Outliers	95	45K	F	- B.E
24	-	50K	M	PHD
32	-	60K	- M	B.E
- Median or Mean	60	-	- M	Master
-	65	55K	- M	BSc
40	22	-	M	B.E

② Measure Of Dispersion [Spread of the data] Imp

Spread

For visualizing variance and standard deviation follow this link:
https://evangelinereynolds.netlify.app/post/variance-and-sd-visualization/

① Variance (σ^2)

② Standard deviation (σ)

① Variance

Variance is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.
Variance always has squared units. For example, the variance of a set of weights estimated in kilograms will be given in kg squared.
Since the population variance is squared, we cannot compare it directly with the mean or the data themselves.

Population Variance

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

$x_i \Rightarrow$ Data points

$\mu \Rightarrow$ population mean

$N \Rightarrow$ populate size

Sample Variance

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

$x_i \Rightarrow$ data points

$\bar{x} \Rightarrow$ Sample mean

$n \Rightarrow$ sample size

Assignment : Why we divide sample variance by n-1?

Ans: To build an unbiased estimator of population variance.

Eg: {1, 2, 3, 4, 5}.

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

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
$\bar{x} = 3$		$\sum (x_i - \bar{x})^2 = 10$

$$s^2 = \frac{10}{4} = 2.5$$

DIFFERENCE B/W Variance and Standard Dev:

Standard deviation is the spread of a group of numbers from the mean. The variance measures the average degree to which each point differs from the mean. While standard deviation is the square root of the variance, variance is the average of all data points within a group.

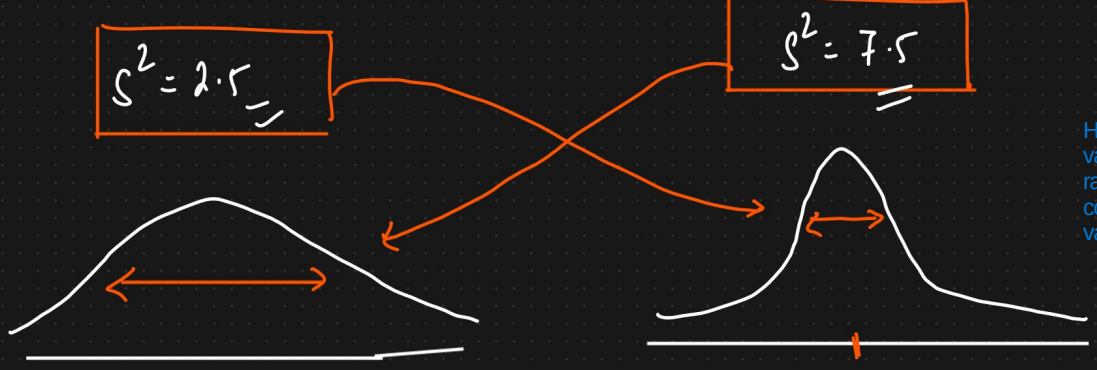
Refer image: https://media.gettyimages.com/vectors/standard-normal-distribution-standard-deviation-and-coverage-in-vector-id1213863214

X = { }

$s^2 = 2.5$

Y = { }

$s^2 = 7.5$



Horizontal spread of data ie; variance is more for Y random variables as compared to X random variable.

(4) Standard deviation

Population Std σ = $\sqrt{\text{Variance}}$

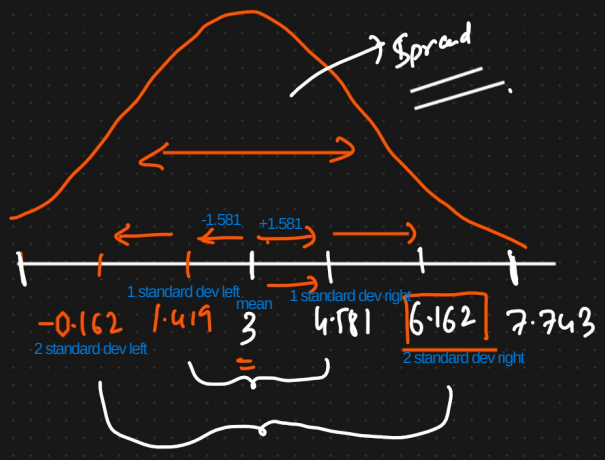
Sample Std $s = \sqrt{s^2}$

Standard Deviantion: How much far a data is from mean. Sqaure of standard deviation is variance.

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

$\bar{x} = 3$

$s = 1.581$



3.00
1.581
4.581
1.581
6.162
1.581
7.743

ⓕ Random Variable

Variables used in algebra is fixed. Random variables are different form such variables.

Linear Algebra

$$\left\{\begin{array}{ll} x+5=7 & \Rightarrow x=2 \\ 8=y+x & \boxed{y=6} \end{array}\right\} \text{Variables}$$

Random Variable is a process of mapping the output of a random process or experiment to a number.

Eg: Tossing a Coin $\{\text{Head, Tail}\} \Rightarrow$ Process

$$X=\begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$$

$X=\{$ Ages of student in a class $\}$

Eg: Rolling a dice $\{1,2,3,4,5,6\}$

$Y=\{$ Sum of rolling of dice 7 times $\}$

$\Downarrow$

$Pr(Y \geq 15) =$ ____ $Pr(Y<10)$

$Pr(40 \leq Y \leq 15) =$

⑥ Histograms And Skewness → [Frequency]

Histogram shows data distribution.

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

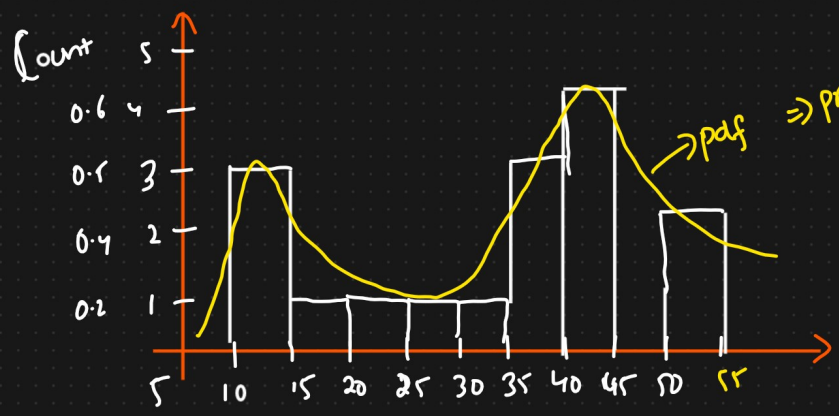
Max approx value = 50

No. of bins = 10

= 5 → bin size

No. of Bins = 10 → buckets

Below along y axis in the interval of 5 (bin size) 10 buckets are made



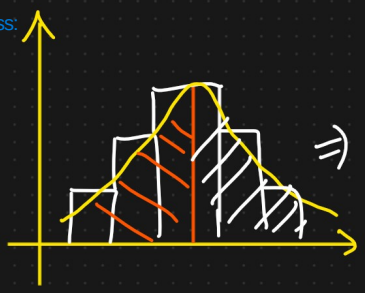
⇒ probability density function

[Kernel density estimator]

Smoothing of the curve is done through kernel density estimator.

Skewness

1. No skewness:



⇒ Normal Distribution

Symmetrical Distribution

⇓

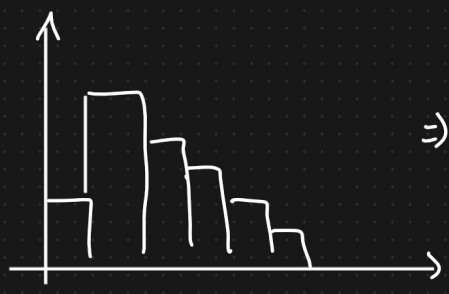
No skewness

Median = Mean = Mode

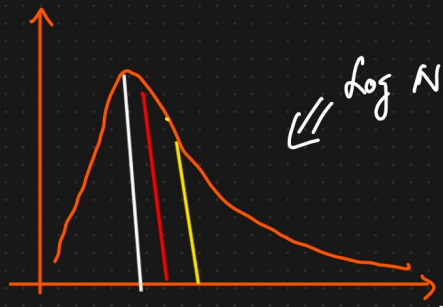
In case a no skewness distribution is normal/symmetrical/gaussian.

Also, in such a case for central region: Mean=Median=Mode.

② Right skewed



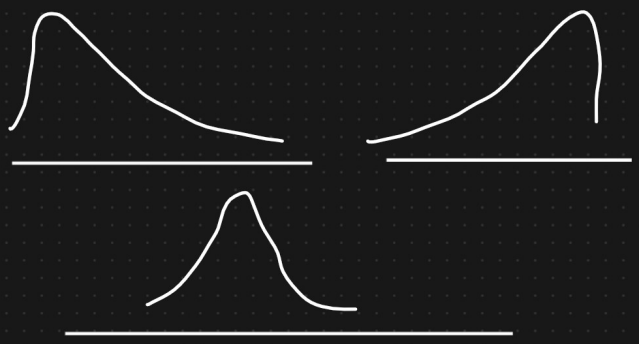
⇒ Positive skewed



mean ≥ median ≥ mode

③ Left skewed

mode > median > mean



--> Symetrical Skewness : LHS=RHS.
mean=meadian=mode

-->Right sided Skewness in smoothen graph means Right Skewed.
mean>=median>=mode.

-->Left Skewed:left side is elongated.

mode>median>mean.

Knowledge sharing ⟶ Profile Building

Make linkedin profile and start sharing knowledge that you would be learning of daily basis