ML | stats | EDA & FE | DL   Brushed
4   1   2   3   5

①

Task 1 → PYTHON — DAILY at Ⓗ & W at Ⓢ & Ⓢ → 3, 3

Task 2 → STATS

Task 3 → EDA & Feature Engineering.

Task 4 → ML

INTRO TO STATS:

1. Descriptive STATS

* Measure of central Tendency.
* Measure of Dispersion
↓
Summarizing the data.

Probability, Permutation, Mean
Mode, variance, median, SD
standard Deviation.

1. Gaussian Distribution
2. Lognormal "
3. Binomial "
4. Bernoullis "
} – PY
5. Pareto " (Power law)
6. Standard normal Distr
7. Transformation & Standardization.
8. Q-Q Plot

2. Inferential stats.

→ Z-test — PY
→ T-test — PY
→ ANOVA
→ CHISQUARE
→ HYPOTHESIS Testing.
↓
NULL Hypothesis
&
Alternate Hypothesis
→ confidence Intervals.
→ Z table, t-table, Chisquare table.

What is statistics?

②

* Stats is a science of collecting, oraganizing and analysing data.

* It helps out how to used the data in perfect way.

* Better decision making.

Define data?

Pieces of information that can be ~~dot~~ a measured.

Eg: The IQ of a class.

{98, 97, 60, 55, 75, 85}

Ages of students of a class.

{30, 25, 24, 23, 27, 28} - DATA.

Types of statistics:

① Descriptive stats.

It consists of organizing and summarizing data.

② Inferential stats.

It is technique where we used the data that we have measured to form conclusions.

1. Descriptive stats.

Eg: * class room of Maths Student (20)

Marks of the 1st sem

84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97, - - -

2. Inferential stats.

* What is the average ~~age~~ marks of the student in the class?

Mean, Median, Mode (or) SD.

Eg: Are the ~~age~~ marks of the students of this classroom similar to the age of the maths class room in the college?

1 $\boxed{\text{Sample}}$ ↓ taking only one sample here

5 $\boxed{\text{Population}}$ → Entire class rooms. ↓ we have 5 class rooms.

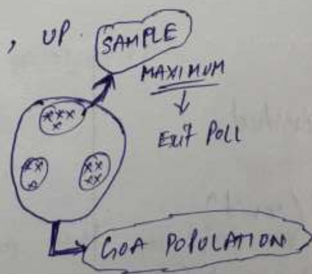
POPULATION AND SAMPLE

POPULATION (N)

SAMPLE (n)

Elections → Goa, UP.

Exit Poll

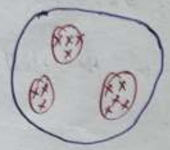


1) Here Goa Population is big.

2) So media will take samples from different locations.

SAMPLING TECHNIQUES:

1) Simple Random Sampling : Every member of the Population (N) has an equal chance of being selected for your sample (n).



2) STRATIFIED SAMPLING: Where the Population (N) is split into non-overlapping groups (strata).

BP (1)

Eg1: GENDER → MALE
→ FEMALE

Eg2: Based on Age

(0-10) (10-20) (20-40) (40-100)

Eg3:

3) Systematic Sampling:

(N) → n^{th} individual.

Eg: Mall ← Survey (covid)
↳ 8^{th} Person → survey.
↳ Every 1^{st} Person → survey.

In whole Population (N). Here we select sample (n) in a systematic way.

4) convenience sampling: Doing survey on specific topic.

Survey will be based on domain knowledge (or) interest.

$\boxed{\text{Data Science}}$
↓
we will survey only who has knowledge about data science.

$\boxed{\text{Stack overflow}}$
↓
we will survey only developers.

Eg: EXIT POLL

RBI → survey with house hold.
↓
what sampling.
↓
WOMEN (convenience sampling (or) strata).

Eg: DRUG TEST
↓
what kind of sample?
↓

VARIABLES:

A variable is a property that can take any value.

Eg: Height = {182, 172, 180, 190}
weight = {78, 75, 85, 90}

Two kinds of variables.

1) Quantitative variables.
2) Qualitative variable (or) categorical variable.

1) Q V → Measured Numerically

Eg: Age, weight, Height

2) C V → It should be converted to numeric data.

Eg: Gender → M, → F { Based on some characteristics we can derive categorical variable }.

Eg : T-shirt sizes.

L, XL, M, S

Quantitative

Discrete variable (QDC)

Continuous variable

Q CV

Eg: Whole Number

No. of Bank Accounts.

Eg: 2, 3, 4, 5, 6, 7, 8

Eg2: Total number of children

~~1~~ 2, 3, 4, 5, 6, 7, 8.

→ Float, whole,

Eg: Height = 172.5, 162, 163.5

Weight = 100, 99.5, 98.

1) What kind of variable Gender is? Q CV (Categorical)

Marital status? CV "

River length? Q C V

Song length ! Q CV

Variable Measurement scales :

4 types of measured variable.

1) Nominal Data.

categorical data → colors, Gender, Type of flower.

2) ordinal data

order of the data matters. But value does not

Eg: students (marks)

students (marks)	RANK
100	1
96	2
85	3
88	4

→ ordinal data.

3) Interval data :

Both order of data & value matters.

Eg: Temperature.

*) Fahrenheit

70-80   80-90   90-100   0

You will some range of values & order also matters.

*) 4) RATIO DATA :

FREQUENCY DISTRIBUTION:

sample data set: Rose, lilly, sunflower, Rose, lilly, lilly, Rose sunflower.

Frequency distribution table

flower type	frequency
Rose	3
lilly	3
Sunflower	2

Total number of flowers → cummulative frequency

3

6

8

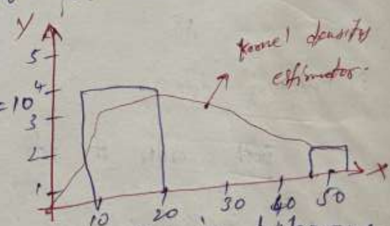
1) BAR GRAPH

Discrete variables.

2) HISTOGRAMS

→ continuous variables.

Ages = {10, 12, 14, 18, 50, 80, 88}



we do bins in histograms.

BAR VS HISTOGRAMS.

PDF: Probability Density function.
↓
smoothing of histograms.

1) ARITHMETIC MEAN for POPULATION & sample

Mean (Average)

Population (N) Sample (n)

1, 1, 2, 2, 3, 3, 4, 5, 5, 6.

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

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

$$\mu = \frac{1+1+2+2+3+3+4+5+5+6 \rightarrow (n)}{10 \longrightarrow (N)}$$

$$\mu = \frac{32}{10} = 3.2 \rightarrow \text{Average}.$$

Sample.

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

$= 3.2$

Mean = is a part of central tendency.

central Tendency:

→ It refers to the measure used to determine the centre of the distribution of data.

→ centre of the data.

1) Mean 2) Median 3) Mode.

2) Median.

Data = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6}, 100 → New element

↳ we consider those as outliers

$$\text{Mean} = \frac{32+100}{11} \Rightarrow \frac{132}{11} \Rightarrow 12 \quad \boxed{\mu = 12}$$

① $\mu = 3.2$

② $\mu = 12$ } huge difference

* we must be careful with outliers.

MEDIAN: {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112

Median → 3, 4

outliers → 100, 112

↳ odd = 12

~~Mode~~ Median $= \frac{3+4}{2} = 3.5$

to do

1st thing → 1) Sort the numbers in median.

odd number = 11

Median = 3

when you got 2 outliers Median works better than mean.

MEDIAN WORKS WELL WITH OUTLIERS

MODE: $\{1, \underbrace{2, 2}_{2}, 3, 4, 5, \underbrace{6, 6, 6}_{3}, 7, 8, 100, 200\}$

Measure of central tendency.

MODE = { MOST FREQUENT ELEMENT }

$\boxed{\text{MODE} = 6}$

Type of flower	Petal length	Petal width	DATASET
ROSE			
LILLY			
SUNFLOWER			
—			
—			
—			

→ Missing value → can be Replaced with Most frequent occuring element.

↓

work well with categorical Variable.

Ages of students.

Age

25

26

=

=

=

32

34

38

⇒ we use Mean

MEASURE of Dispersion :

Dispersion → $\boxed{\text{SPREAD}}$ → How Good your data is spread.

1) VARIANCE .

2) STANDARD DEVIATION .

1) VARIANCE

POPULATION VARIANCE

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

SAMPLE VARIANCE

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

x	μ	$x - \mu$	$(x_i - \mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.68
2	2.83	+0.83	0.68
3	2.83	0.17	0.03
4	2.83	1.17	~~0.03~~ 1.37
5	2.83	2.17	4.71

$\mu = \frac{17}{6}$

$\mu = 2.83$

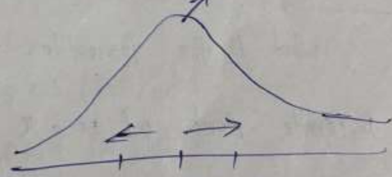
10.84

$\rightarrow \frac{10.84}{6}$

$\rightarrow 1.81$

which will have more variance (Spread) 2

BELL CURVE



$SD = \sqrt{\text{variance}}$

$= \sqrt{1.81}$

$= 1.345$

1

2.83

1.34

4.17

Percentiles and Quartiles {1st step to find outliers}.

Percentage: 1, 2, 3, 4, 5.

% of the numbers that are odd?

$$\Rightarrow \% = \frac{\text{\# of numbers that are odd}}{\text{Total numbers}}$$

$$\% = \frac{3}{5} = 0.6 = 60\%$$

Percentiles:

A percentile is a value below which a certain percentage of observation lie.

Data set: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12.

what is the percentile ranking of 10?

Percentile Rank of x

$$x = 10 \Rightarrow \frac{\text{\# of values of below } x}{n} \times 100$$

$$= \frac{16}{20} \times 100$$

$$= 80\%$$

$$\Rightarrow x = 11 \Rightarrow \frac{17}{20} \times 100$$

$$\Rightarrow 85\%$$

Five number Summary :

1) MINIMUM 2) First Quartile (Q1)

2) Median 4) Third Quartile (Q3)

5) MAXIMUM

* This is used to $\boxed{\text{remove outliers}}$.

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

-50

outlier

When ever you want to remove outlier ?

[Lower fence ⟷ Higher fence]

Lower fence = Q1 − 1.5 (IQR)

upper fence = Q3 + 1.5 (IQR) Interquartile Range .

(IQR) = Q3 − Q1 ⇒ Q3 = (75%)

Q1 = (25%)

1) Distributions: To have idea about dataset.

Age: {24, 26, 27, 28, 30, 32 - - - - -}

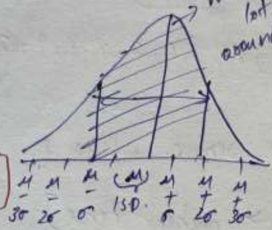
i) Gaussion | Normal distribution:

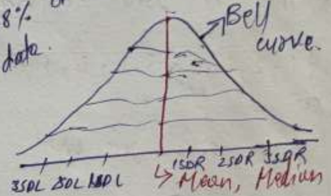
Emperical formula

68-95-99.7% Rule.

It follows

$\boxed{68-95-99.7\% = \text{Rule}}$





SD towards left | Standard deviation Right.

1 SD = 68% data is Present from entire data

2SD = 95% data is present from entire data.

3SD = 99.7% data is Present from entire data.

Eg: Height → Normally distributed.

Domain expert → {Doctor}

Weight, IRIS DATASET.

Z-score:

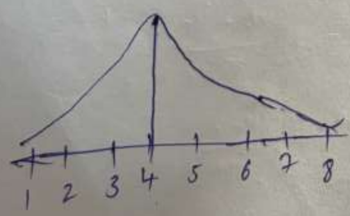
$$Z\,score = \frac{x_i - \mu}{\sigma}$$

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

+ve value.

Z score is measured in terms of SD from the mean.

Eg: $\mu = 4$, $\sigma = 1$



After applying Z-score.

Data = {1, 2, 3, 4, 5, 6, 7, 8}

$z(1) = \frac{1-4}{1} = -3$, $z(2) = \frac{2-4}{1} = -2$, $z(3) = \frac{3-4}{1} = -1$

It is easy to calculate S.D. using Z-score.

SD = {-3, -2, -1, 0, 1, 2, 3, 4}

{1, 2, 3, 4, 5, 6, 7, 8} After Applying Z-score {-1, -2, -3, -4, 0, 1, 2, 3, 4}
↳ This is normal distribution. ↳ standard normal distribution

PRACTICAL APPLICATION:

our mean should be $\mu = 0$, & SD = 1

$\boxed{\mu = 0, \sigma = 1}$
↳ Convert this to standard normal distribution using Z-score.
↓
This process is called as
STANDARDIZATION
↓
Inside Z-score will be applied

DATASET:

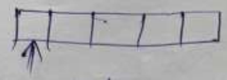
AGE	SALARY	WEIGHT
Years	RS	Kg
24	40k	70
25	80k	80
26	60k	85
27	80K	99

NORMALIZATION:

↳ MINMAX SCALER → (0 to 1)

Normalization gives you the process where you can define lower and upper bond.

CNN → Image classification



Normalization.

Practical eg:

In 2021

$\mu = 250,\ x_i = 240,\ \sigma = 10$

In 2020

$\mu = 260,\ x_i = 245,\ \sigma = 12$.

DAY-3

1) Distributions
   - ↳ Normal dist
   - ↳ standard

{India vs SA}

1) ODI Series 2021

Score Average 2021 = 250 ⇒ μ

standard deviation = 10 ⇒ σ

Rishbpant ~~final~~ Avg score = 240

2020

Score Average 2020 = 260

SD = 12

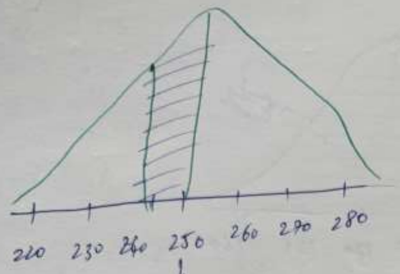
Rishbpant ~~final~~ Avg score = ~~28~~ 245

compare both the series in which year Rishab Pant final score was better?

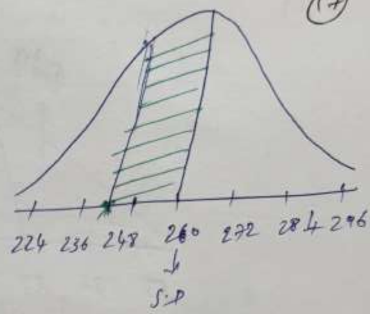
→ Here we apply z-score for 2021

$$z\ score = \frac{x_i - \mu}{\sigma} = \frac{240 - 250}{10} \Rightarrow \cancel{\frac{180}{10} = 18} \quad \frac{-10}{10} = -1$$

$$z\ score = \frac{x_i - \mu}{\sigma} = \frac{245 - 260}{12} \Rightarrow \cancel{\frac{192}{12}} \Rightarrow = \frac{-15}{12} = -1.25$$



z score = −1



Z-Score = −1.25.

Example for Z-score :

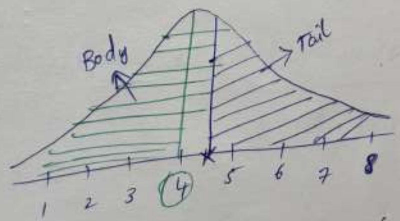
stats interview question.

$\mu = 4$

$\sigma = 1$

What percentage of scores falls above 4.25?

x.



$$z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1}$$

$$\boxed{z = 0.25}$$

z score helps you to find area of the body curve.

Look At $\boxed{\text{Z-table}}$

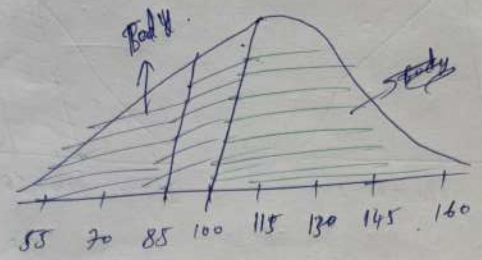
1 − 0.5987 ⇒ 0.4013

* In india the avg IQ is 100, with a S.D of 15. What % of population would you expect to have an IQ lower then 85 ?

$\mu = 100$

$\sigma = 15$

$$z \text{ score} = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$



$= -1 - 0.15886$

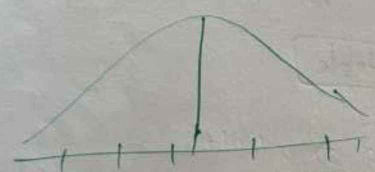
$= \frac{-1.15886}{4}$ =

Iq b/w 90 to 120.

$z = \frac{90-100}{15} \Rightarrow$

$z =$ 120

DAY - 4



SD 1 ⇒ 68%.

SD 2 ⇒ 95%.

SD 3 ⇒ 99.7%.

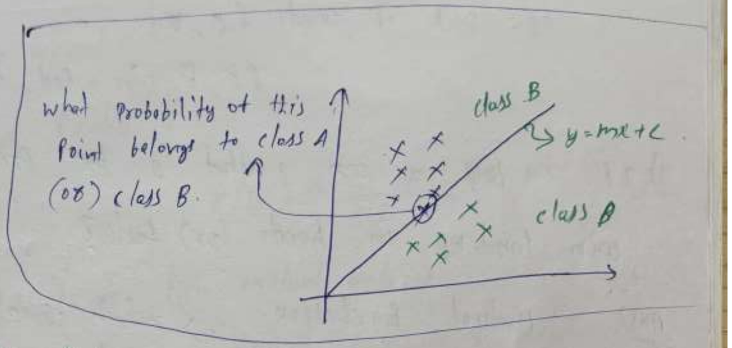
After SD3 all are outlier.

It can be treated as outlier after SD3.

$$Z\text{-Score} = \frac{x_i - \mu}{\sigma} \quad (\text{or}) \quad \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

* Probability:

Probability is a measure of the likelihood of an event.



Eg: Roll a dice {1, 2, 3, 4, 5, 6}.

what is $\text{Prob}(6) = \frac{\#\text{ of way an event can occur}}{\#\text{ of possible outcomes}}$

$$= \frac{1}{6}$$

Eg: Toss a coin. (H, T)

$$P(H) = 1/2$$

2) Addition Rule (Probability, "or")

Mutual Exclusive Events:

Two events are mutually exclusive if they cannot occur at same time.

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

Non - Mutual Exclusive

Multiple events can occur at the same time.

Eg: Deck of cards {Q, ♡}

{K, ♡, color, Red, Black}.

1) If i toss a coin, What is the Probability of the coin landing on heads (or) tails?

Ans) Mutual Exclusive. → which is also called as mutual exclusive.

Pr (H (or) T)

$$Pr(A \text{ or } B) = P(A) + P(B)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$\boxed{Pr(A \text{ or } B) = 1}$$

Roll a dice

$$Pr(1 \text{ or } 3 \text{ or } 5) = P(A) + P(B) + P(C)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$= \frac{3}{6} = \frac{1}{2}$$

$$= 0.5$$