Statistics Day - 1

Job
data analyst, Business analyst, Data scientist

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

Data: facts or pieces of information

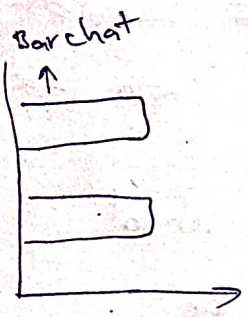
Eg: Ages of students in a class room

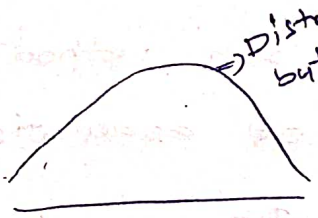
{24, 25, 26, 28, 30, 42} ⇒

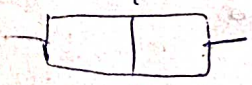
Statistics

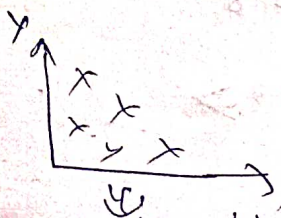
Descriptive stats

① It consists of organizing and summarizing the data.

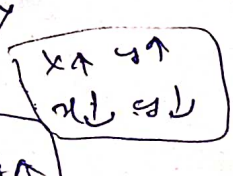


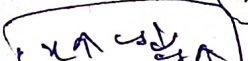






(height) weigh





Inferential stats

* It consists of collecting sample data and making conclusion about population data using some experiments ↓

Hypothesis testing

University → 500 people

classA → 60 people

⇓

Sample data ⇒ Age ⇒ Avg age of entire university

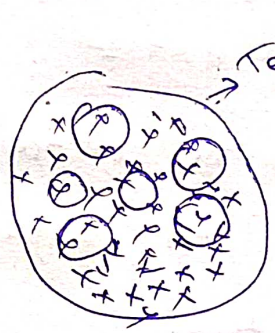
⇓

Hypothesis testing

C.I ⇒ Confidence Interval

8.

Sample data Vs Population data



Hypothesis testing
P_value
① z-test
② t-test
③ chi square test
④ F test

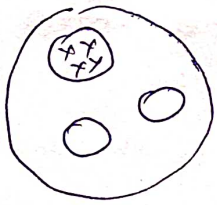
Sample data → conclusion → Population data
Hypothesis Testing

Choose a sample data:

Sampling Techniques Population(N) sample(n)

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

n=1000



② Stratified Sampling: is a method of sampling that involves the division of population into smaller subgroups known as strata

Examples:

Gender { male, Female

eduction { High school, master, phd

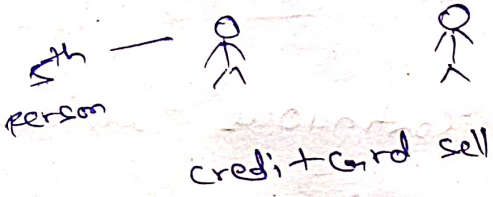
Blood group

Vote { above >18, below 18 }

③ Systematic Sampling :- is a probability sampling method where researches select member of the population at a regular interval

Ex :- by selecting every 15th person on a list of population

* Select every nth individual out of population (N)



④ Convenience Sampling :- only those who are interested in the survey will only participate

Data science survey → General AI survey

* Survey regarding New technology

---

① Variable :- A variable is a property that can take any value.

Eg :- age = 14
age = 25
age = 100

variable
ages = [24, 25, 26, 27, 29]

Two different Types of variables

① Quantitative variable :-

→ measured Numerically & mathematical operation

Ex :- age, weight, height, rainfall (cm), temperature

② Qualitative variables :- Categorical variables

∴ Based on some characteristics they are grouped together

Eg :- Gender, Types of flowers, Types of movies

Quantitative variables

Discrete variables

Eg :- Whole number → fixed

* No. of Bank accounts
  {1, 2, 3, 4, 5}
* No. of childern

Continuous variables

Eg :- Continuous → decimal value

* Height, weight, ages
  Rainfall, speed
* marital — { married, not married }

∴ categorical variables

Statistics Day-2

Agenda:

1. Histograms
2. measure of Centeral Tendency
3. measure of Dispursion
4. Percentiles and Quartiles
5. 5 number Summary (Box plot)

Histogram: is a bar graph-like representation of data that buckets a range of classes into colums along the horizantal x-axis, the verical y-axis represents the number count or perceentage of occurance in the data for each column.

Ex:

Ages = {10, 12, 14, 18, 22, 24, 28, 30, 32, 36, 40, 44, 46}

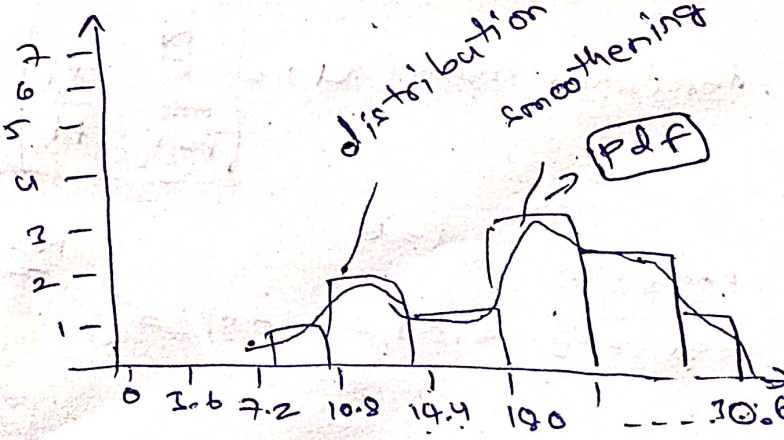
1. Sort the Numbers
2. Bins → No. of groups
3. Bins size → size of Bins

bins = 10

$$\text{bin size} = \frac{\text{max} - \text{min}}{\text{bins}}$$

$$= \frac{46 - 10}{10} \Rightarrow \frac{36}{10} \quad 3.6$$

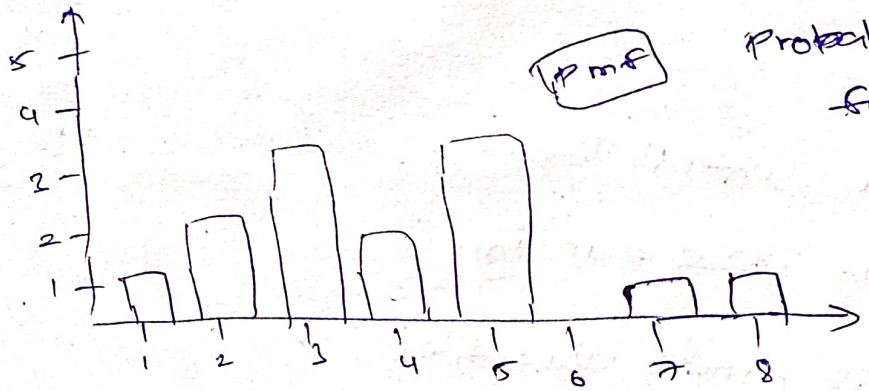
binsize => 3.6



(continuous values)

Discrete value :

No. of Bank accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



Probability mass function

Pdf ⇒ Probability density function ⇒ continuous

Pmf ⇒ Probability mass function ⇒ discrete

* Measure of central tendancy :

A measure of CT is a single value that attempts to describe a set of data identifying the central position.

(1) Mean (2) median (3) mode

(1) mean :

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

Average/mean $= \frac{1+2+3+4+5}{5} = \frac{15}{5}$

$\Rightarrow 3$

Population (N)

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

Sample (n)

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

Population $N = 6$

$N > n$

$n = 4$

Age $= \{24, 23, 21, 28, 27\}$

$\{24, 23, 21, 27\}$ Avg

Sample avg $= \{24, 21, 27\}$

Population mean $(N) = \frac{24+23+21+28+27}{6}$

$\mu = 17.5$

Sample mean $(\bar{x}) = \frac{24+21+27}{4}$

$\bar{x} = 13.5$

$M \geq \bar{x}$ or $\bar{x} \geq M$

$M = 17.5$

$\bar{x} = \frac{24+23+28+27}{4} \Rightarrow \frac{102}{4}$

$\bar{x} = 23$

np.nan ⇒ Null value

NAN ⇒ Not A Number

Mean = Sum Of Observation / Total No. Of Obser

Age	Salary
24	45
28	50
29	NAN
NAN	60
21	75
26	80
NAN	NAN
80	200

less Information

Salary ÷ 62

mean

outlyers

* median ÷

{1, 2, 3, 4, 5} ⇒ $\bar{x} = 3$

{1, 2, 3, 4, 5, 100} ⇒ $\bar{x} = 19.16$

outlier

$\frac{1+2+3+4+5+100}{6} \Rightarrow 19.16$

Steps to find out median ÷

① sort the numbers

② find the central Number

{① if the no of elements are even we find the average of central elements }

② if the no. of elements are Odd we find the central elements }

Sorted

{0, 1, 2, 3, 4, 5, 6, 7, 8, 100, 120} mean = 25.6

median = $\frac{5+6}{2}$ = 5.5 median = 5

if n is odd = media = $\left(\frac{n+1}{2}\right)^{th}$ term

if n is even: median = $\frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$

(3) mode :- {most frequent occuring elements}

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

2, 3

Date:- Types of flower :- (Categorical variables)

lily
sunflower
Rose
NAN ← Rose
Rose
sunflower
Rose
NAN ← Rose

* Measure of Dispersion

① Variance (σ^2)

② Standard deviation (σ)

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

Variance

Population Variance (σ^2)

$$\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample Variance (s^2)

$$s^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

First one

$\{1,2,3,4,5,6,7,8,9,10\}$

⇓

Variance

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

$\mu = 3$

$$\sigma^2 = \frac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$$

$$\Rightarrow \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2 = \sigma^2$$

second one

$\{1,2,3,4,50,60,70,100\}$

⇓

Variance

$\{1,2,3,4,5,6,80\}$

$\mu = 14.4$

$$\sigma^2 = \frac{(1-14.4)^2 + (2-14.4)^2 + \cdots}{7}$$

$\sigma^2 = 719.10$

Variance ↑↑ = Spread ↑↑

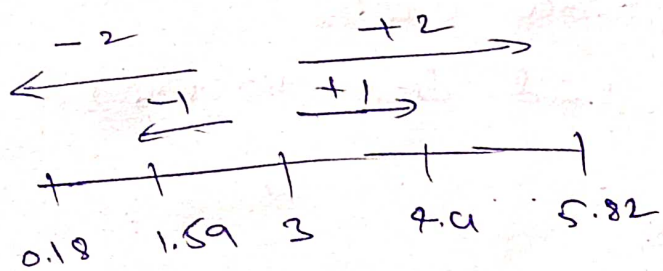
* Standard deviation $(\sqrt{\sigma^2}) \Rightarrow \sigma$

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

$\mu = 3$

$\sigma^2 = 2$

$\sigma = \sqrt{2} = 1.41$



* percentile and Quantiles:-

Percentage = $\{1, 2, 3, 4, 5, 6, 7, 8\}$

Percentage of even numbers $= \frac{\text{No of even numbers}}{\text{Total no. of numbers}} = \frac{4}{8} = 0.5 \Rightarrow 50\%$

Percentiles:- Gate, Cat, Jels.

Defn:- A percentile is a value below which a certain percentage of observation lie

99 percentage:- It means the person has got better marks than 99% of the entir students

Dakent:- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

what is the percentage rank of 10

Percentile rank of $x = \frac{\#\text{No. of value below } x}{n} = \frac{16}{20} = 80$ percent

$= 0.8$

* what is the value ~~of~~ that exists at 25 percentile

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

$$= \frac{25}{100} \times 20 = 5^{th} \text{ Index}$$

o/p = 5

* 5 Number Summary :-

1. minimum
2. first Quartile (25 percentile) (Q_1)
3. median
4. Third Quartile (75 percentile) (Q_3)
5. maximum

Boxplot ⇒ Remove the outlie -rs

outlier

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

(lower fence ⟷ Higher fence)

lower Fence = $Q_1 - 1.5\,(IQR)$

Higher Fence = $Q_3 + 1.5\,(IQR)$

$IQR = Q_3 - Q_1$

⇓

Inter Quartile Range (IQR)

$Q_1 = \frac{25}{100} \times 21 = 5.25$ Index = 3

$Q_3 = \frac{75}{100} \times 21 = 15.75$ Index = $\frac{8+7}{2} = 7.5$

lower Fence = $3 - (1.5)(4.5) = -3.65$

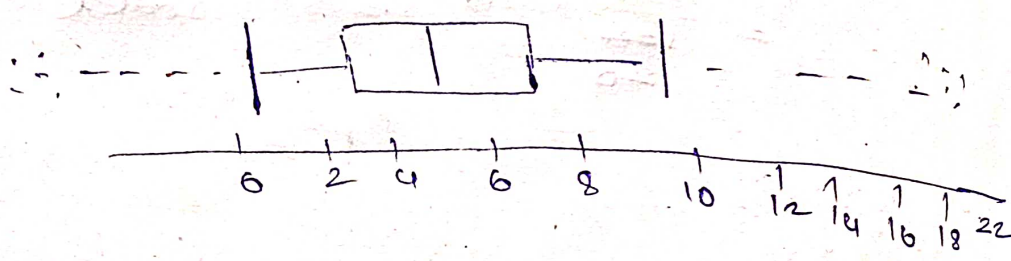
Higher Fence = $7.5 + (1.5)(4.5) = 14.25$

{1, 2, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29}

5 Number summary

1. minimum = 1
2. $Q_1 = 3$
3. median = 5
4. $Q_3 = 7.5$
5. maximum = 9

Boxplot



To Treat outliers