Subject Statistic:-

MON TUE WED THR FRI SAT SUN

Statistic Class (Day-2)

By Krish Naik Sir

Agenda

1. Histogram
2. Measure of Central Tendency
3. Measure of Dispursion
4. Percentiles and Quartiles
5. five Number of Summary (Box plot)

1. Histogram

Eg:- Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

Steps

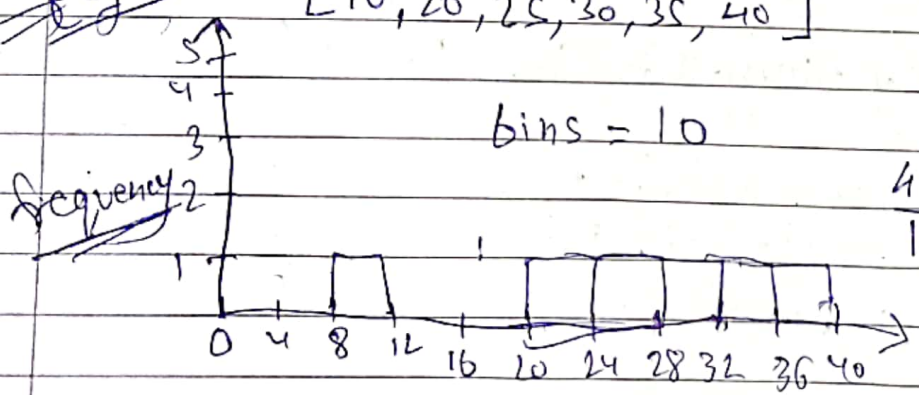
a. Sorts the Numbers
b. Bins → No. of groups
c. Bins Size → Size of Bins

Eg:- [10, 20, 25, 30, 35, 40]

min = 10
max = 40

bins = 10

$\frac{40}{10} \Rightarrow 4$



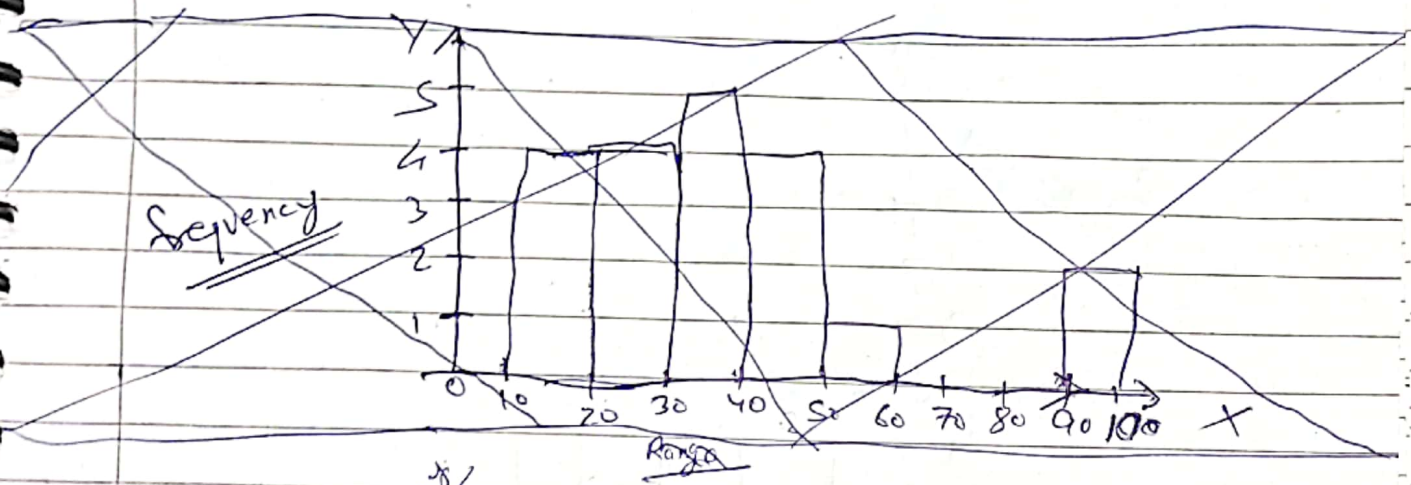
Subject _____

MON TUE WED THR FRI SAT SUN

Eg:- Ages ⇒ {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

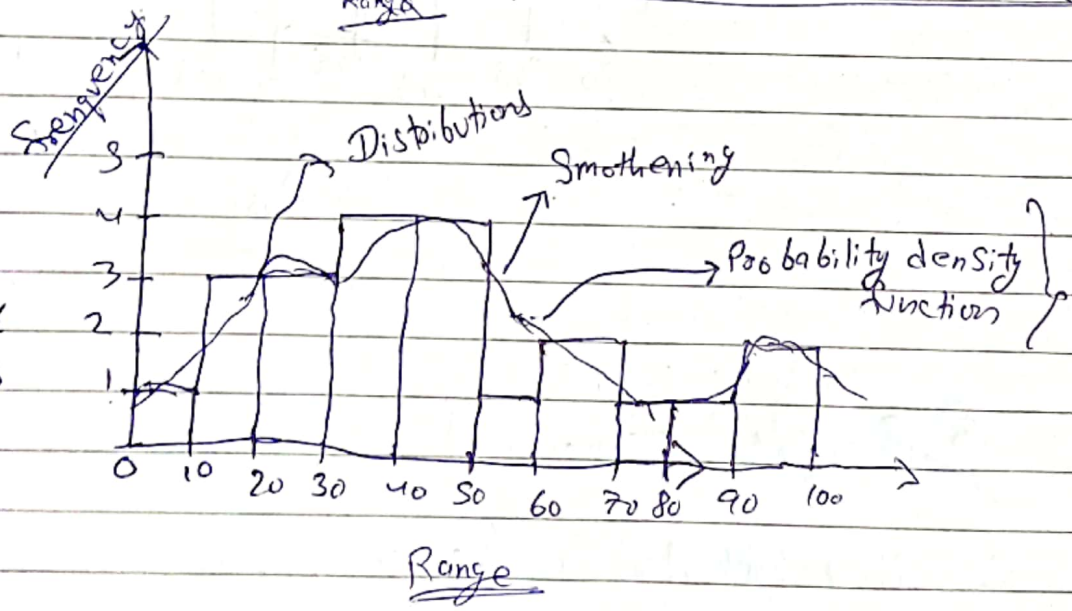
bins = 10

bins Size = $\frac{100}{10}$ = 10



Bins =

Bin Size ⇒ $\frac{Max}{bins}$



your writing partner

Subject _____

MON TUE WED THR FRI SAT SUN

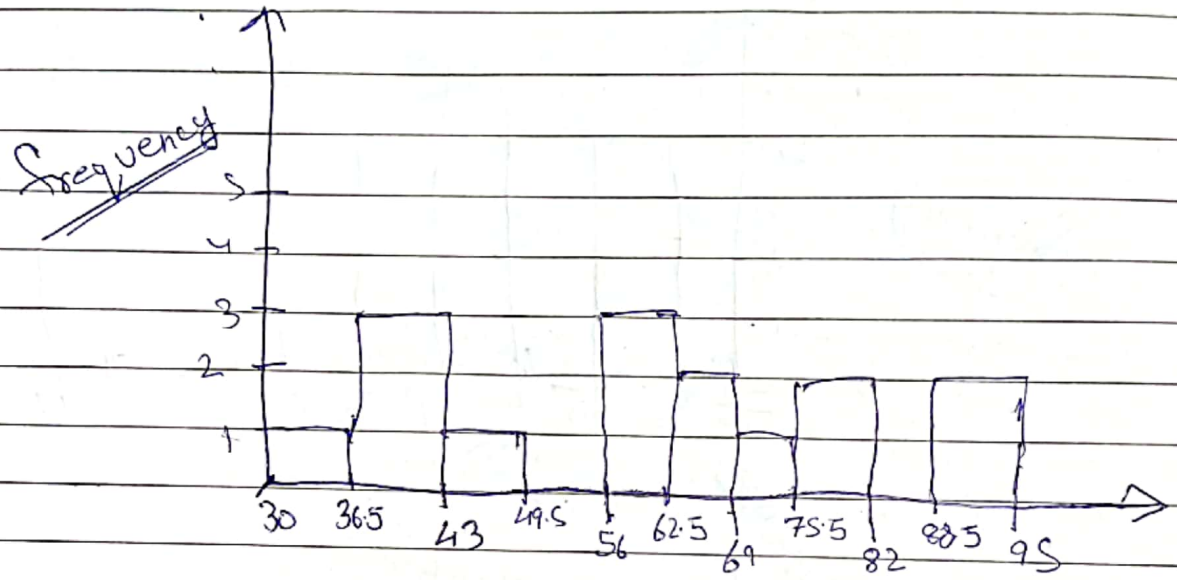
Eg:-

Weights = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

(Continuous Variable)

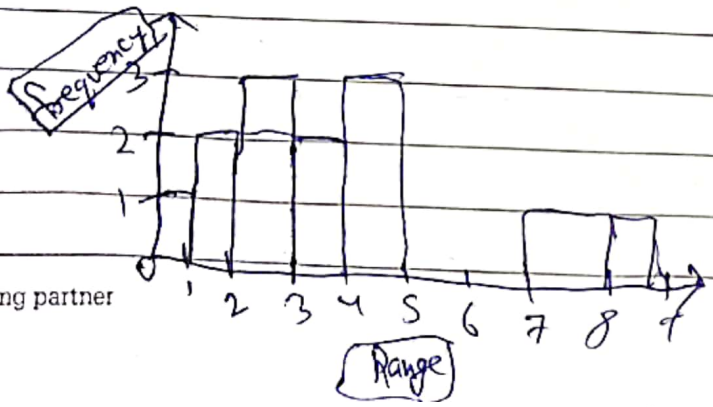
bins = 10

bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$



(b) Discrete Continuous:-

Eg:- No. of Bank Accounts = {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}



Probability Mass formula

your writing partner

PDF ⇒ Probability Density function → Continous
PMF ⇒ Probability Mass function → Discrete

Subject _____ MON TUE WED THR FRI SAT SUN

(2) Measure of Central Tendency

(1) Mean
(2) Medium
(3) Mode

→ Measure of Central Tendency is a single value that attempt to describe a set of data identifying the central position.

(2.1) Mean

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

$$\text{Average / Mean} = \frac{1+2+3+4+5}{5} \Rightarrow \frac{15}{5} \Rightarrow 3$$

Sample (n)

$N >> n$

Population (N) ✓ Sample (n) ✓

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

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

MON TUE WED THR FRI SAT SUN

Subject

(1) Population mean (μ) = $\frac{\sum_{i=1}^{N} X_i}{N}$

Eg:- Population Age = {24, 23, 2, 1, 28, 27}

$N = 6$

Population mean(μ) = $\frac{24+23+2+1+28+27}{6}$

$\mu = 17.5$

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

Eg:- Sample Age = {24, 2, 1, 27}

$n = 4$

Sample mean ($\bar{x}$) = $\frac{24+2+1+27}{4} = \frac{54}{4}$

$\bar{x} = 13.5$

np.nan ⇒ null value

Subject _____ MON TUE WED THR FRI SAT SUN

Eg: **Practical Application [Feature Engineer]**

Age	Salary	Family Size
—	—	—
—	—	—
NAN	—	—
—	—	—
—	NAN	—
—	NAN	NAN
NAN	—	—

Loss of Information

Note: [Where there are NAN value, we will replace with mean of that Age & put its =

Eg:

Age	Salary
24	45
28	50
29	NAN
NAN	60
31	75
36	80
NAN	NAN
outliers 80	200

Age = 29.6 (mean) ⇓ 38

Salary = 62 ⇓ 85

mean of Age = $\frac{24+28+29+31+36}{5} = \frac{148}{5} \Rightarrow 29.6$

your writing partner

Subject

MON TUE WED THR FRI SAT SUN

2.2 (b) Median

(a) $\{1,2,3,4,5\}$ ⇒

$\bar{x} = 3$

(b) $\{1,2,3,4,5,100\}$ ← outlier

$\bar{x} = 19.16$

Step to find out Median:-

1. Sort the Number
2. Find the central Number.
   - (a) If no No. of elements are even we find the average of central element
   - (b) If no No. of elements are odd we find the central elements

Eg: $X = \{1,2,3,4,5,6,7,8,100,120\}$ ⇒ Even

⇒ median $= \frac{5+6}{2} = 5.5$

Eg: $X_1 = \{0,1,2,3,4,5,6,7,8,100,1200\}$

mean ⇒ 25.6
median = 5

your writing partner

Subject

MON TUE WED THR FRI SAT SUN

(A) No outliers → Mean
(A) With outliers → Median

(2)(3) **Mode :-** ⇒ { Most frequent occuring element }

Eg ⇒ { 1, 2, 2, 3, 3, 3, 4, 5 }

Mode = 3 //

Eg ⇒ { 1, 2, 2, 2, 3, 3, 3, 4, 5 }

Mode = 2, 3 //

Eg:- **Dataset**

Type of Flowers { Categorical variable }

Lily
Sunflower
Rose
NAN ← Rose
Rose
Sunflower
Rose
NAN ← Rose

your writing partner

Note :- In Categorical Variable, we follow Mode for NAN value

Q. Why Sample Variance is divisible by n-1 ??

Subject _____ MON TUE WED THR FRI SAT SUN

Measure of Dispersion

1. Variance (σ^2) ← [Spread of Data]
2. Standard Deviation (σ)

Variance

Population Variance (σ^2) :-	Sample Variance (S^2)
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

Eg: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

$$\mu = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} \Rightarrow 5.5$$

Eg: $X_1 \Rightarrow \{1, 2, 3, 4, 5\}$

$\mu = 3$

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

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

Eg: $X_2 = \{1, 2, 3, 4, 5, 6, 80\}$

$\mu = 14.4$

$$\sigma^2 \Rightarrow \frac{(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2}{7}$$

=

Subject _____

MON TUE WED THR FRI SAT SUN

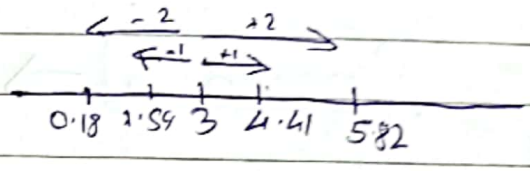
(8) Standard Deviation ($\sqrt{\sigma^2}$) $\Rightarrow \sigma$

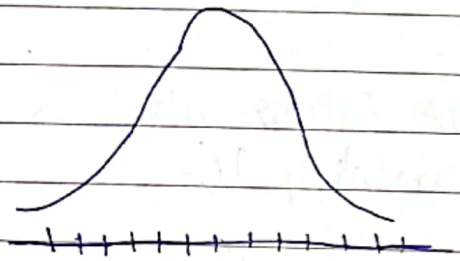
Eg:- $\{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma^2 = 2$

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





(9) ~~Box~~

~~Percentiles~~

Subject

MON TUE WED THR FRI SAT SUN

(A) Percentiles And Quartiles

(B) Eg: Percentiles ⇒ GMAT, LAT, JEE, GRE, NEET

Percentile

→ Define it:

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

99 percentile :— It means the person has got better marks than, 99% of the entire mark.

(Eg:) Dataset

Ascending order

⇒ { 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12 }

(Q) What is a percentile Rank of 10?

⇒ Percentile Rank of x = $\frac{\text{\# No. of value below } x}{n}$

$= \frac{16}{20} = \frac{4}{5}$ ⇒ 80 percentile,

Rank 6 ⇒ $\frac{7}{20}$ ⇒ 35%.

your writing partner

Scanned with CamScanner

Q) Percentile Rank of 8 ⇒ $\frac{9}{20}$ ⇒ 55%

#) What is the value that exists at 25 percentile??

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

= $\frac{25}{100} \cdot \times 20 = 5^{Th}$ Index

Output value is 5 //

Q) 95% value = $\frac{95}{100} \times 21 = 19.95$

Subject _____ MON TUE WED THR FRI SAT SUN

⭐ Five Number of Summary (Box Plot)

1. Minimum
2. first Quartile (25 percentile) Q_1
3. ~~Second Quartile~~ (50 percentile) Median
4. Third Quartile (75 percentile) (Q_3)
5. Maximum

$\frac{3+3}{2} = 3$

Eg:- {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 = $Q_3 - 1.5(IQR)$

$IQR = Q_3 - Q_1$ (75 − 25)

⇓

Inter Quantile Range (IQR)

25% $Q_1 = \frac{25}{100} * (n+1) = \frac{25}{100} \times 21 = 5.25 \Rightarrow$ Index $\Rightarrow 3$

75% $Q_3 = \frac{75}{100} \times (n+1) \Rightarrow \frac{75}{100} \times 21 \Rightarrow 15.75$ Index $\Rightarrow \frac{8+7}{2} \Rightarrow 7.5$

① Lower fence $\Rightarrow 3 - (1.5)(4.5) = \boxed{-3.65}$

② Higher fence $\Rightarrow 7.5 - (1.5)(4.5) = \boxed{14.25}$

Subject _____

MON TUE WED THR FRI SAT SUN

⑧ Five Number Summary for Box-Plot (previous Page Ques)

1. Minimum = 1
2. Q_1 = 3
3. Median = 5
4. Q_3 = 7.5
5. Maximum = 9

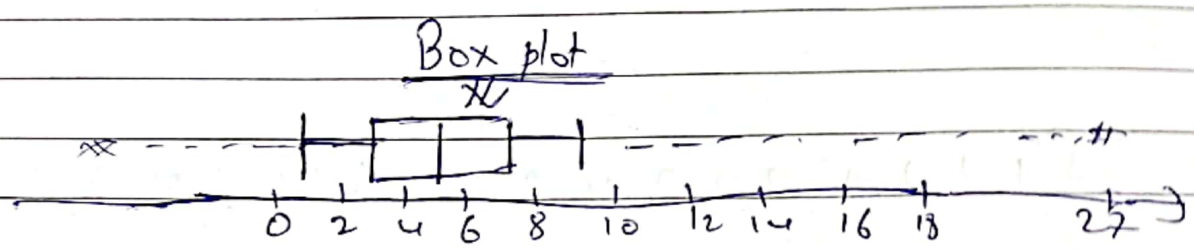


fig:- Plotting the Box-Plot with above five Number summary Information