[Day - 2]

Page No.:
Date. / /

different types of statistics

Discriptive	Inferential
- Consists of organizing and summarizing	- Collect sample data make conclusion of population data based on sample data
- includes ① Measure of central tendency i) mean ii) Median iii) Mode ② Measure of dispersion i) variance ii) Standard deviation	- includes - Hypothesis testing

college A has 1000 students.

e.g" Height = { 160, 132, 150, 140, 157 }

Discriptive - what is mean, median, mode of above data.

inferential - taking sample predict the height of 1000 students.

Page No.:
Date: / /

(1) Measure of central tendency

i) Mean

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

e.g.

Age of students = {10, 12, 15, 17, 19}

$$\mu = \frac{10+12+15+17+19}{5} = 57.8$$

ii) Median :-

Suppose Age = {10, 12, 15, 17, 19}

then μ = 57.8

If we add 87 in Age

then Age = {10, 12, 15, 17, 19, 87}

μ = 87.5

In above case 87 become outlier

to remove that we use median

∴ for even data set - arrange in asc

median = Average of two middle values

for odd data set

median = middle value.

Page No.:
Date. / /

iii) mode :-

→ the data who had maximum frequency called mode

e.g. Age = { 10, 10, 12, 15, 19 }

mode = 10

② Measure of dispersion :-

i) variance :

It tells about the spread of the data.

e.g. - Age = { 2, 2, 4, 4 } Age = { 1, 1, 5, 5 }

$\mu = 3$ $\mu = 3$

we can say about spred using above results.

but using variance we can tell.

Population (N)	Sample data (n)
$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$
x_i = data points μ = mean n = size	x_i - data points $\bar{x}$ - mean n - size

Page No.:
Date: / /

e.g. Age = {2, 2, 4, 4}

$\mu = \frac{2+2+4+4}{4}$

$\mu = 3$

x_i	μ	$(x_i - \mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1

$N = 4 \quad \Sigma (x_i - \mu)^2 = 4$

$\sigma^2 = \frac{4}{4}$

$\boxed{\sigma^2 = 1}$

∴ data is less spred

Age = {1, 1, 5, 5}

$\mu = \frac{1+1+5+5}{4}$

$\mu = 3$

x_i	μ	$(x_i - \mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4

$N = 4 \quad \Sigma (x_i - \mu)^2 = 16$

$\sigma^2 = \frac{16}{4}$

$\boxed{\sigma^2 = 4}$

∴ data has more spread

Sample variance:

$\Rightarrow s^2 = \sum_{i=1}^{n} \frac{(x - \bar{x})^2}{(n-1)}$ ⇒ Bessel's correction

population variance:-

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

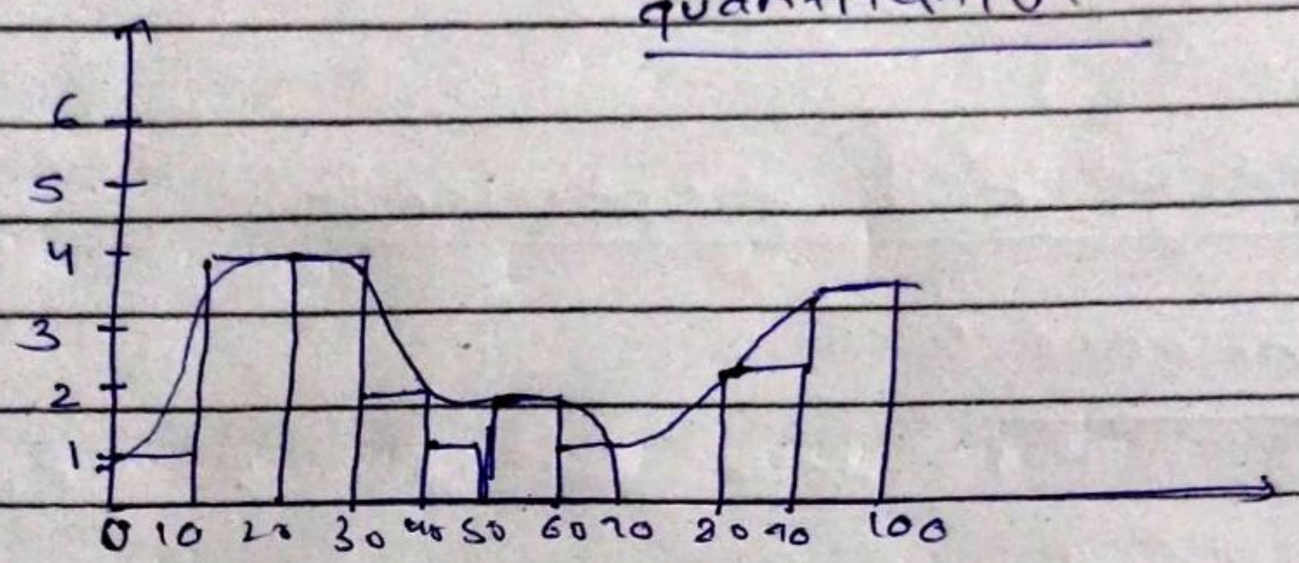
Page No.:
Date: / /

Histogram :-

Data = {10, 12, 13, 14, 20, 22, 24, 25, 26, 35, 38, 42, 55, 56, 68, 82, 84, 86, 92, 93, 94, 100 }

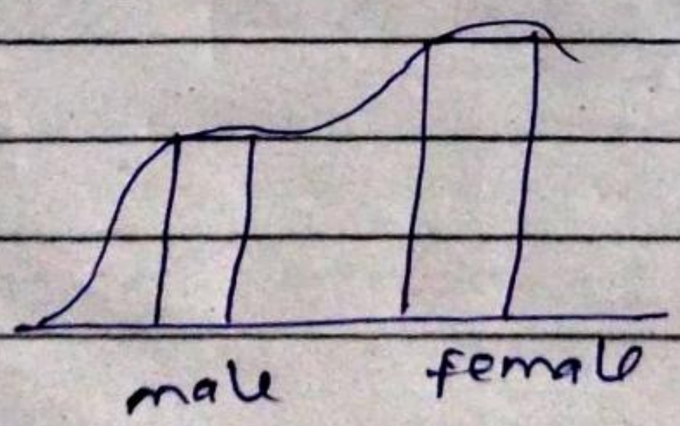
by default histogram bin = 10

quantitative



pdf :- Smoothen version of histogram, the technique is used
kernal density function

qualitative :-



Page No:
Date: / /

- five number summary / - to remove outlier

Here

① minimum
② first quartile
③ median
④ Third quartile
⑤ maximum

} using this we can create boxplot

It is used to remove outlier:-

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

Here 27 is outlier.

to find this,

[lower fence ⟷ higher fence]

$Q_1 \rightarrow 25\%$ lower fence = $Q_1 - 1.5(IQR)$

$Q_3 \rightarrow 75\%$. higher fence = $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

$Q_1 = \frac{25}{100} \times (19+1) = \frac{25 \times 20^{5}}{100_{5}} = 5^{th}$ index

$\boxed{Q_1 = 3}$

$Q_3 = \frac{75}{100} \times (19+1) = \frac{75 \times 20^{15}}{100_{5}} = 15^{th}$ index

$\boxed{Q_3 = 7}$

Page No.:
Date: / /

IQR = 7 - 3 = 4

lower fence = 3 - 1.5 (4)
= 3 - 6
lower fence = -3

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

data range [-3, 13]

∴ 27 is outlier in our data

i.e. data = {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

① min = 1 ② median = 5 ③ max = 9

④ $Q_1 = 3$ ⑤ $Q_3 = 7$

boxplot :-