Page No.
Date

Intermediate Stats.

1. Measure of Central Tendency
2. measure of Dispersion
3. Gaussian Distribution
4. Percentiles and Quartiles
5. 5 Number Summary (Box Plots)

* Measure of Central Tendency

1. mean
2. median
3. mode.

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

(μ) mean
($\bar{x}$) x Ba

(i) mean (Average) — Distribution

x_1, x_5

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

Population (N) | Sample (n)

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

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

$= \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = \boxed{3}$

Average = 3 | Distribution of data = 3

* Population vs Sample Example.

Page No.
Date

* Population data {24, 23, 2, 1, 28, 27}
(N)

* Population mean

$$\mu = \frac{24 + 23 + 2 + 1 + 28 + 27}{6}$$

Central distribution = 17.5

* Sample data (n) = {23, 2, 28, 27}
* Sample mean

$$\bar{x} = \frac{23 + 2 + 28 + 27}{4}$$ (Random Pic Value)

Central distribution = 20

* Examples of An mean, Median and mode

medn

DATASET

Age	Salary	family size
—	—	—
—	NAN	—
NAN	—	NAN
—	—	NAN
—	NAN	—
—	—	—
NAN	—	NAN

NAN = (Empty Value).

As, here we have an dataset of an age, salary, and family size, where as there are some empty value (NAN) so here we have to fullfill the data set.

As, if we delete the empty Colourm it may cause an loss of Information.

So, we will fill it by Avg Colaumn.

Eg :- Age = {24, 26, NAN, 21, 20, 18}

$$\text{Average}(m) = \frac{24 + 26 + 21 + 20 + 18}{5}$$

$$= 21.8$$

As, here we can replace the NAN Colaumn with value of 21.8, So the dataset will an {24, 26, 21.8, 21, 20, 18}

Median :- The middle number in Sorted list of Number more descriptive of data set than Aver

As, here if we have an dataset {1, 2, 3, 4

$$(m) \text{ Average} = \frac{1+2+3+4+5}{5} = 3$$

with same data set If we add 100 {1, 2, 3, 4, 5, 100}.

$$(m_s) = \frac{1+2+3+4+5+100}{6} = \frac{115}{6}$$

As Now {1, 2, 3, 4, 5} ↔ {1, 2, 3, 4, 5, 1

we can see an difference between dataset.

Range = (5 - 1) = 4

[1, 2, 3, 4, 5]

Sorted Order

Range (1 - 100) = 99

[1, 2, 3, 4, 5, 100]

Outlier

(By Calculating Range and seen huge range

Page No.
Date

difference over both dataset Identify that there is an Outlier.

* Defination of Outlier.

Outlier is an number that is Complete different than entire distribution.

If the data set has an Outlier we can replace the value using median.

* Step to find median. {1, 2, 3, 4, 5, 100}

(1) Sort the Numbers.
(2) Find the Central Number.
   (i) If the no. of element are even we find Average.
   (ii) If the no. of element are odd we find Central number.

Even no.
{1, 2, [3, 4], 5, 100}

$\frac{3+4}{2} = 3.5$

median 3.5

Odd no.
{1, 2 [4], 5, 100}

median 4

MODE : The most frequent Occuring element.

Dataset. { Categorical Variable

Types of flower

1. Lily
2. Sunflower
3. Rose
4. Lily
5. NAN ← Replace with something
6. Rose
7. Rose
8. Sunflower

As, here most frequent Occuring elem is an Rose. (So, NAN Replace with Ro

Uses :-

- mean (Avg)
- median → (outliers)
- mode → (Categorical replacement)

Page No.
Date

* Assignment

Find mean, meddin and mode of following distribution.

$X = \{24, 25, 26, 27, 28, 90, 100, 1000, 1200, 1400, 1400, 1400\}$

(i) Mean (Average)

$$(M) = \frac{24+25+26+27+28+90+100+1000+1200+1400+1400+1400}{12}$$

$$= \frac{6720}{12} = 560.$$

mean = 560.

(ii) Median

$\{24, 25, 26, 27, 28, \boxed{90, 100}, 1000, 1200, 1400, 1400, 1400\}$

$$\frac{90+100}{2} = \frac{190}{2} = 95$$

median = 95

(iii) Mode = 1400

(2) Measure of Dispersion

(1) Variance (σ^2) : Sigma^2.

(2) Standard Deviation (σ) : Sigma

Dispersion (Distribution of Data).

1. Variance

Variance is used to understand proper distribution between data.

Example :- If we have Two data As

$X = \{1, 1, 2, 2, 4\}$ $\quad$ $Y = \{0, 2, 2, 3, 3\}$

$m = \underline{2}$ $\quad$ $m = \underline{2}$

As, here we can see that the mean of both data set is equal, so it not consider that both data set are same

* Population Variance (σ^2) Sigma^2

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x^i - m)^2}{N} \rightarrow \text{Population mean}$$

(Spread of Distribution)

As example.

A = {1, 2, 2, 3, 4, 5, 10} (Range : 10 - 1 = 9)

⇵ Range difference.

B = {2, 2, 4, 6} (Range : 6 - 2 = 4)

As here with A Dataset we can see that the spread of data distribution is high; with B Dataset distribution is low.

If Variance is Become high the spread of distribution is high.

If Variance is Become low Spread of distribution is lesser.

* Sample Variance.

Bessels Correction

Degree of freedom

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

$\bar{x}$ → Sample mean.

As here we can see that their is some difference in formulas $(n-1)$

Page No.
Date

Population Standard deviation. [Variance]

Formula. $\sigma = \sqrt{\sigma^2}$

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

Sample Standard deviation [Variance]

Formula $s = \sqrt{s^2}$

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

Dataset : $x = \{23, 21, 20, 19, 24, 27, 28\}$

x	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
23	23.14	-0.14	0.196
21	23.14	2.14	4.5796
20	23.14	3.14	9.8596
19	23.14	4.14	17.1396
24	23.14	-0.86	0.7396
27	23.14	3.86	14.8996
28	23.14	4.86	23.6196
			71.0336

Formula

$$\sigma^2 = \frac{71.0336}{6} = 11.8$$

$$\sigma = \sqrt{11.8} = 3.44$$

Variance = 11.8

Standard deviation = 3.44

(3) Percentiles and Quartiles.

To understand the percentiles lets get understand percentage.
Basics of example.

Calculating the Percentage of even and odd numbers.

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

∴ Percentage of even no = $\frac{\text{No. of even number}}{\text{Total no. of numbers}}$

⇓

50 % = $\frac{4}{8} = \frac{1}{2} = 0.5$

∴ Percentage of odd no = $\frac{\text{no. of odd number}}{8}$

⇓

50% = $\frac{4}{8} = \frac{1}{2} = 0.5$

Percentiles.

As, In many of exam we dee their is an percentiles such as, SAT, CAT Or GRE the result where display as an percentile 99%, 85% Or 84%

Page No.
Date

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

If A person got an 99 percentile in exam. It means that the person has got better marks than 99% of the entire students.

Examples :-

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

Q. What is the percentile of 10 ?

formula : Percentile Rank of x = number of values below x / n (sample) x 100

$$= \frac{16}{20} \times 100 = 80$$

Percentile = 80

As, here we can say that 10 is greater than 80 percentage of entire distribution

Page No.
Date

Q. What is the percentile of 11 ?

Percentile Rank of x = $\frac{\text{no. of value below } x}{n} \times 100$

$= \frac{17}{20} \times 100$

$= \underline{85}$ Percentile.

Ans, If we want to calculate or find the percentile ..

Q. Find the 25% percentile ?

formula : value = $\frac{\text{Percentile}}{100} \times n+1$

$= \frac{25}{100} \times 20+1$

$= 5.25$ (Index).

(4) 5 number Summary (Box Plot)

(i) Minimum
(ii) First Quartile (25%) (Q1)
(iii) Median
(iv) Third Quartile (75%) (Q3)
(v) Maximum.

} Removing Outlier.

This five number Summary help to remove an Outliers.

As we understand the Concept with an Example.

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

>-3
<13

Outlier

As here we can see that an 27 is an Outlier but also Consider that 1 is also an Outlier. As In this Condition we Create an fence to understant / determine the outlier.

Standard formula based On data distribution.

lower fence = Q1 - 1.5 (IQR)
higher fence = Q3 + 1.5 (IQR)

IQR = (Inter Quartile range)

As, first find Q1 and Q3 of dataset.

Lower fence

$Q1 = (25 \text{ Percentile})$

$= \frac{25}{100} \times (n+1)$

$\therefore \frac{25}{100} \times (20) = \underline{5}$

So the 5th Index of dataset is

Higher fence

$Q3 = (75 \text{ Percentile})$

$= \frac{75}{100} \times (n+1)$

$\therefore \frac{75}{100} \times (20) = \underline{15}$

So the 15th Index of dataset is 7

$IQR = Q3 - Q1$

$= 7 - 3$

$= \underline{4}$ (IQR)

Now we will Calculate an Lower fence and higher fence.

Page No.
Date

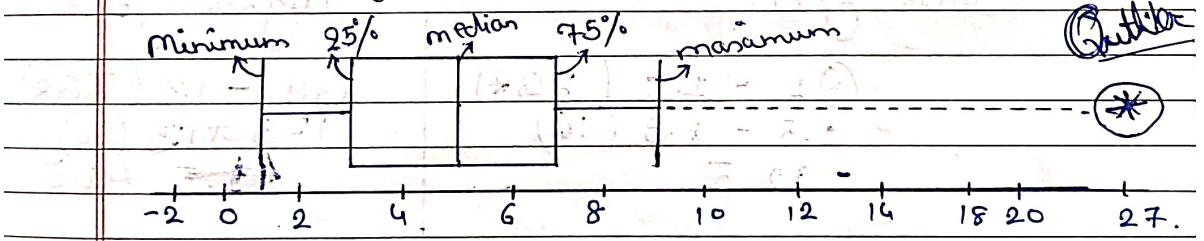
Lower fence = Q1 - 1.5 (IQR)
= 3 - 1.5 × 4
= -3

higher fence = Q3 + 1.5 (IQR)
= 7 + 1.5 × 4
= 13

As after getting an every value Now Just put the value using 5 number summary

minimum = 1
~~Maximum~~ = 1
first Quartile (Q1) = 3
median = 5
Third Quartile = 7
maximum = 9

Box Plot Identify Outlier



Assignment

* Find minimum, maximum, Q1, Q3 median of following dataset

{-8, 1, 2, 4, 5, 6, 8, 15, 20, 120}
Outlier

$Q1 = \frac{25}{100} \times 11 = 2.75$ (Index)

$Q3 = \frac{75}{100} \times 11 = 8.25$ (Index)

* Q1 = 1.5
Q3 = 17.5

$\frac{1+2}{2} = \frac{3}{2} = 1.5$

$\frac{15+20}{2} = \frac{35}{2} = 17.5$

* IQR = Q3 - Q1
= 17.5 - 1.5
= 16

Lower fence
Q1 - 1.5 (IQR)
1.5 - 1.5 (16)
-22.5

Higher fence
Q3 + 1.5 (IQR)
17.5 + 1.5 (16)
41.5

{-8, 1, 2, 4, 5, 6, 8, 15, 20, 120}

minimum = -8

Q1 = 1.5

median = 5.5

Q3 = 17.5

maximum = 20