Descriptive Statistics
STA 111

24/11/2025

Lecturer: Mr. Nwosu Ozoemena
Topic: Introduction to ~~Statistics~~ Statistics

Definition of Statistics: It is the scientific method of data collection, summarizing, analyzing, presenting and making valid conclusion from the sample of the data collection from the population.

Branches of Statistics:
1. Descriptive Statistics
2. Inferential Statistics

* Descriptive statistics: This branch of statistics that builds data collection, summarizing, analyzing and presentation.

* Inferential Statistics: Is a branch of statistics that deal with making valid conclusion from the data collected.

Sources of data Collection:
i) Primary Sources
ii) Secondary Sources

* Primary Sources: It Involves the researcher going to field to collect the data by itself e.g Interviews, questionnaire, auction, telephone e.t.c.

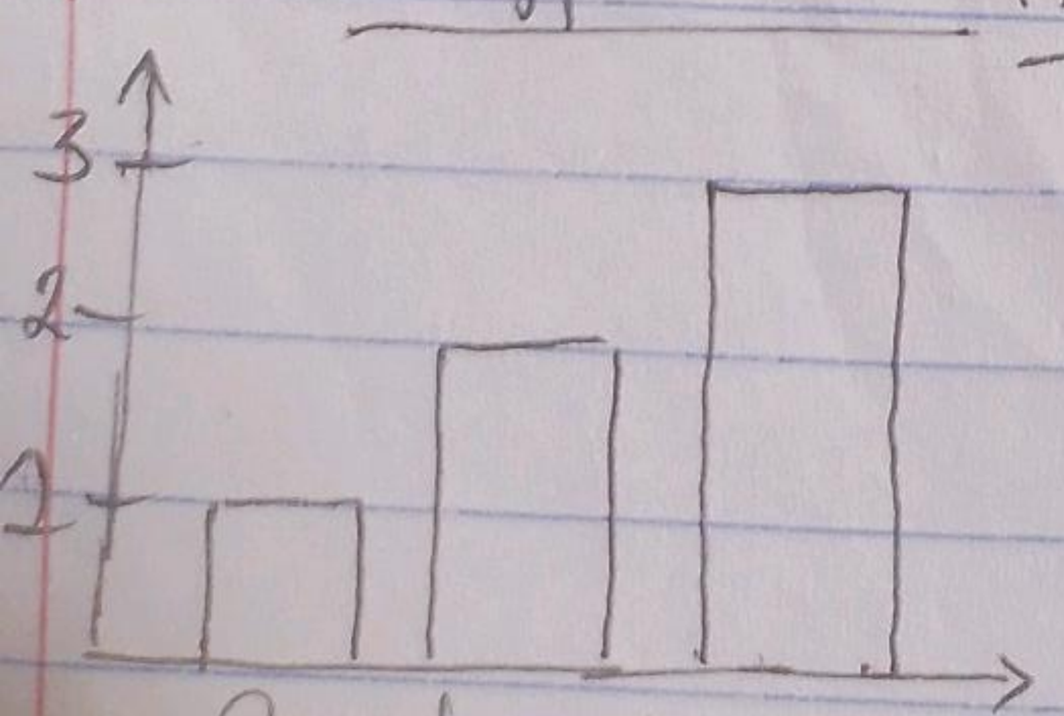
* Secondary Sources: This deals with already collected data e.g data from National Population Commission, data from Federal Road Safety Commission e.t.c.

Methods of data Collection:

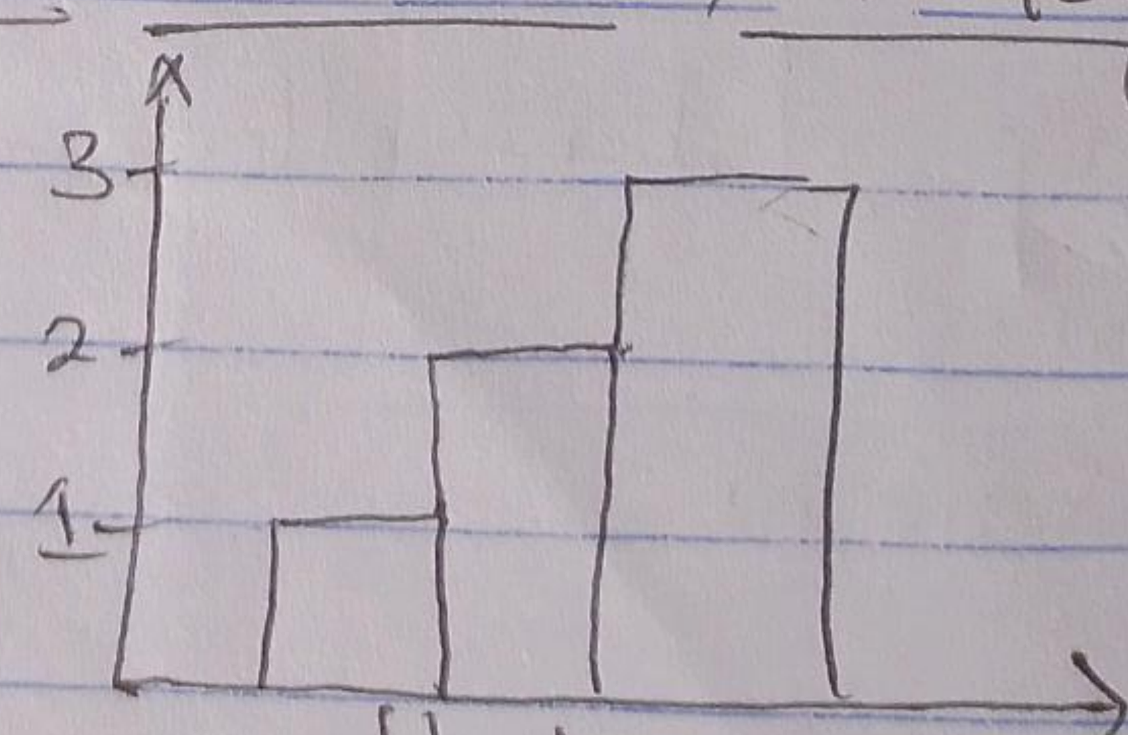
i) By Interview ⇒ Telephone
ii) By observation
iii) By Questionnaire
iv) By Graphical Representation of data → Barchart, Pie chart, Histogram, Frequency Polygon/ogive.

Note: Questionnaire is a well designed outline of question for responses from the respondent.
- Pie chart is the sectional representation of data by angle 360°.

Difference between Bar chart & Histogram



Barchart



Histogram

STA 111 01/12/2025

Topic: Measures of Central Tendency

1) Mean
2) Mode
3) Median

$\left.\begin{array}{l}\text{Mean}\\ \text{Mode}\\ \text{Median}\end{array}\right\}$ Always remember the Formulas.

Grouped data: Has frequencies attached to them.

Ungrouped data: Has no frequencies attached to them.

Mean:

Mean is the average of a number or given no.

Ungrouped mean: Suppose that we have Variables $x_1, x_2, \ldots, x_n$, then the mean ($\bar{x}$) =

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

OR $$\frac{\sum_{i=1}^{n} x}{n}$$ Where Σ means Summation or addition.

Example: Find the mean of these numbers 3, 5, 7.

Solution

$$\text{Mean}(\bar{x}) = \frac{\Sigma x}{n} = \frac{3+5+7}{3} = \frac{15}{3} = \underline{\underline{5}}$$

* **Grouped mean:** Suppose that the observations $x_1, x_2 \ldots x_n$ occurs with frequencies $f_1, f_2 \ldots f_n$. Then the mea

$$(\bar{x}) = \frac{\Sigma f x}{\Sigma f} \quad x \to \text{class mark}$$

OR

$$\text{The mean } (\bar{x}) = A + \frac{\Sigma f d}{\Sigma f}$$

Note: This Second Formular is used if and only if (iff) th Assumed mean (A) is given.

Where d which is the deviation value

$$= d = X - A.$$

Example: Compute the mean of these data

Class	93-102	103-112	113-122	123-132
Frequency	7	29	20	4

Solution

$$\text{The mean } (\bar{x}) = \frac{\Sigma f x}{\Sigma f}$$

Class	F	X	FX
93-102	7	97.5	682.5
103-112	29	107.5	3117.5
113-122	20	117.5	2350
123-132	4	127.5	510

Note:
For $X = 93 + 102/2$
For $FX = 7 \times 97.5$

$\therefore$ Mean $(\bar{x}) = \frac{666\cancel{0}}{6\cancel{0}}$

$= 111$

Mode:

Mode is the highest occurring number or frequency given.

* Ungrouped mode: this is the highest occurring number among the set of numbers.

Example: obtain the mode of these numbers

9, 2, 7, 10, 2

Ans = 2

* Grouped mode: The formulas for finding grouped modal distribution is given by

$$Mode = L_1 + \left(\frac{D_1}{D_1 + D_2}\right)C$$

Where L_1 = Lower class boundary of modal class

D_1 = Difference b/w frequency of modal class and frequency before the modal class.

D_2 = Difference b/w frequency of modal class and frequency after the modal class.

C = class size is difference between the upper class boundary and lower class boundary.

Example: Calculate the mode of the data:

Class	93 - 102	103 - 112	113 - 122	123 - 132
Frequency	7	29	20	4

Solution

$$\text{Mode} = L_1 + \left(\frac{D_1}{D_1 + D_2}\right) C$$

The modal class = 103 – 112

$L_1 = 103 - 0.5 = 102.5$

$D_1 = 29 - 7 = 22$

$D_2 = 29 - 20 = 9$

$C = 112 + 0.5 = 112.5 - 102.5 = 10$

Note:
For whole no
103 – 0.5
For Decimals
103 – 0.05
For upper class Interval
112 + 0.5

$$\therefore \text{Mode} = L_1 + \left(\frac{D_1}{D_1 + D_2}\right) C$$

$$= 102.5 + \left(\frac{22}{22+9}\right) 10$$

$$= 102.5 + \left(\frac{22}{31}\right) 10$$

$$= 102.5 + 0.7096 \times 10$$

$$= 102.5 + 7.096$$

$$= 109.596$$

Median:

Median is the middle no. of a given data.

Ungrouped Median: This is the middle number among the set of numbers.

Example: Find the median of these numbers

2, 1, 7, 6, 5

Solution

Rearrange in ascending or descending order

$\therefore$ 1, 2, 5, 6, 7

The median is $= 5$

Example 2: Find the median of these numbers

4, 2, 6, 8

Solution

Rearrange in ascending or descending order

2, 4, 6, 8

$\therefore$ The median is $\frac{4+6}{2} = \frac{10}{2} = 5$

* Grouped Median: The formular for finding the grouped median is given by:

$$\text{Median} = L_1 + \left(\frac{\frac{N}{2} - Fb}{Fm}\right)c$$

Where L_1 = Lower class boundary of median class

N = Is the sum of the frequency

$\frac{N}{2}$ = Is the location cumulative frequency of the median class

NB: Cumulative frequency: successive addition of frequencies

Fb = Is the Sum of all the frequencies before the median class

Fm = Is the frequency of the median class

C = class Size

Example: Given the classes below:

Class:	93 - 102	103 - 112	113 - 122	123 - 132
Frequency:	7	29	20	4

Obtain the median.

Solution

$$\text{Median} = L_1 + \left(\frac{\frac{N}{2} - Fb}{Fm}\right) C$$

But $\frac{N}{2} = \frac{7+29+20+4}{2} = \frac{60}{2} = 30$

The median class is = 103 - 112

$L_1 = 103 - 0.5 = 102.5$, $Fb = 7$

$Fm = 29$, $C = 112 + 0.5 = 112.5 - 102.5 = \underline{10}$

$\therefore$ Median $102.5 + \left(\frac{30-7}{29}\right)10$

$= 102.5 + \left(\frac{23}{29}\right) 10$

$= 102.5 + 0.793 \times 10$

$= 102.5 + 7.93$

$= \underline{110.43}$

Topic 3

Measures of Partitions   9/12/2025

1) Quartiles
2) Deciles
3) Percentiles

1. Quartiles: Set of observations being divided by four equal parts.
The Formular for finding the ith Quartile is given by

$$Q_i = \left(LQ_i + \frac{\frac{iN}{4} - \Sigma FQ_i}{FQ_i} \right) c$$

Where LQ_i = Lower class boundary of the ith Quartile class

$\frac{iN}{4}$ = Location cumulative frequency of the ith Quartile class

ΣfQ_i = Sum of all the frequencies before the Ith Quartile class.

FQ_i = frequency of the ith Quartile class

C = Class size

Example: Obtain the 2nd Quartile of these classes

Classes	93-102	103-112	113-122	123-132
Frequency	7	29	20	4

Solution

$$Q_2 = \left(\frac{LQ_2 + \frac{2N}{4} - \Sigma f Q_2}{FQ_2} \right) c$$

But $\frac{2N}{4} = \frac{2 \times 60}{4} = \frac{120}{4} = 30$

The 2nd Quartile class $= 103 - 112$

$LQ_2 = 103 - 0.5 = 102.5$

$\Sigma f Q_2 = 7$

$FQ_2 = 29$

$C = 10$

$\therefore Q_2 = 102.5 + \left(\frac{30 - 7}{29} \right) 10$

$= 102.5 + \left(\frac{23}{29} \right) 10$

$= 102.5 + 0.793 \times 10$

$= 102.5 + 7.93$

$= 110.43$

2. **Deciles:** These are the set of observations been divided by 10 equal parts.

The formular for finding the ith Decile is given by

$$D_i = \left(LD_i + \frac{\frac{iN}{10} - \Sigma f D_i}{FD_i} \right) C$$

Where D_i = ith Decile

LD_i = Lower class boundary of the ith Decile class

$\frac{iN}{10}$ = Location Cumulative frequency of the ith Decile class

$\Sigma f D_i$ = Sum of all the frequencies before the ith Decile class.

FD_i = freqency of the ith Decile class

C = Class Size

Example: Find the 6th Decile of the classes below:

Mass (Kg)	60-62	63-65	66-68	69-71	72-74
No. of students	5	18	42	27	8

Solution

$$D_6 = \left(LD_6 + \frac{\frac{6N}{10} - \Sigma f D_6}{FD_6} \right) C$$

But $\frac{6N}{10} = \frac{6 \times 100}{10} = 60$

The 6th Decile class $= 66 - 68$

$LD_6 = 65.5$

$\Sigma f D_6 = 23$

$FD_6 = 42$

$C = 68.5 - 65.5 = 3$

$\therefore D_6 = 65.5 + \frac{(60 - 23)3}{42}$

$= 65.5 + \left(\frac{37}{42}\right)3$

$= 65.5 + 0.88 \times 3$

$= 65.5 + 2.64$

$D_6 = \underline{\underline{68.14}}$

3. Percentiles: The set of observations or data been divided by 100 equal parts.

The Formular for finding the Kth Percentile is given by:

$$P_K = \left(LP_K + \frac{\frac{KN}{200} - \Sigma FP_K}{FP_K} \right) C$$

Where P_K = The Kth Percentile

LP_K = Lower class boundary of the Kth Percentile class.

$\frac{KN}{100}$ = location Cumulative frequency of the Kth Percentile class.

$\Sigma f P_K$ = Sum of all the frequencies before the Kth Percentile class

FP_K = frequency of the Kth Percentile class

C = class size.

Example: find the 70th Percentile of the class below:

Mass (kg)	60-62	63-65	66-68	69-71	72-74
No. of students	5	18	42	27	8

Solution

$$P_{70} = LP_{70} + \left(\frac{\frac{70N}{100} - \Sigma f P_{70}}{fP_{100}} \right) C$$

But $\frac{70N}{100} = \frac{70 \times \cancel{100}}{\cancel{100}} = 70$

$\therefore$ The 70th Percentile class = 69-71

$LP_{70} = 68.5$

$\Sigma f P_{70} = 65$

$F_{P_{70}} = 27$

$C = 3$

$$\therefore P_{70} = 68.5 + \left(\frac{70-65}{27}\right)3$$

$$= 68.5 + \left(\frac{5}{27}\right)3$$

$$= 68.5 + 0.185 \times 3$$

$$= 68.5 + 0.555$$

$$P_{70} = 69.06$$

NB: You can use either ith, Kth or any other alphabet.

Topic: Measures of Dispersion/Variation

i) Range
ii) Mean Deviation
iii) Variance
iv) Standard Deviation
v) Co-efficient of Variation

* Range: Is the difference between the highest and lower number.

For grouped data:

E.g find the Range of these numbers; 1, 2, 5, 8, 6.

Solution

Range = 8 − 1 = 7

⇒ Inter Quartile Range: Is the difference between the higher Quartile and lower Quartile $= Q_3 - Q_1$

⇒ Semi Inter Quartile Range: Is the difference between the higher Quartile and lower Quartile divided by 2.

$$\frac{Q_3 - Q_1}{2}$$

Assignment 1 [For yourself]

Use the table to solve for Interquartile and Semi-Interquartile Range.

Mean Deviation: Suppose the variables $X_1, X_2, \dots, X_n$ occur with frequencies $F_1, F_2, \dots, F_n$, Then the mean deviation is given as

$$M.D = \frac{\Sigma f |x - \bar{x}|}{\Sigma f}$$

Variance: Suppose the variables $X_1, X_2, \dots, X_n$ occur with Frequencies $F_1, F_2 \dots F_n$. Then the Variance is:

$$\text{Variance} = \frac{\Sigma f (x - \bar{x})^2}{\Sigma f}$$

* Standard Deviation. $= \sqrt{\text{Variance}}$

* Co. efficient of variation is given by

$$C.V = \frac{S.D}{\bar{X}} \times 100\%$$

Example: Given the data below

Classes	60-62	63-65	66-68	69-71	72-74
F	5	18	42	27	8

Find: i) The mean Deviation

ii) The Variance

iii) The Standard deviation

iv) The Co-efficient of variation

Solution

Classes	F	X	Fx	$\lvert X-\bar{X}$	$F\lvert X-\bar{X}\rvert$	$(X-\bar{X})^2$	$F(X-\bar{X})^2$
60-62	5	61	305	6.45	32.25	41.6025	208.0125
63-65	18	64	1152	3.45	62.1	11.9025	214.245
66-68	42	67	2814	0.45	18.9	0.2025	8.505
69-71	27	70	1890	2.55	68.85	6.5025	175.5675
72-74	8	73	584	5.55	44.4	30.8025	246.42
	100		6745		226.5		852.75

Note: $\lvert\bar{X}$ absolute changes negative numbers to positive

But $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{6745}{100} = 67.45$

$M.D = \frac{226.5}{100} = 2.265$

Variance $= \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{852.75}{100} = 8.5275$

Standard deviation $= \sqrt{\text{Variance}} = \sqrt{8.5275} = 2.92$

Co-efficient of Variation

$= C.V = \frac{S.D}{\bar{x}} \times 100\%$

$= \frac{2.92 \times 100}{67.45}$

$= 0.04329 \times 100$

$= 4.329\%$

Assignment 2: (Calculate on 20 leaves)

Given the data below

Classes	93-102	103-112	113-122	123-134
F	7	29	20	4

Find i) The mean deviation
ii) The Variance
iii) The standard deviation
iv) The Co-efficient of Variation.

~~Regression Analysis~~

~~Regression Analysis is used to investigate two or more values~~

STA 111 15/12/2025

~~Topic 3~~

Regression Analysis

This is the statistic analysis used to investigate the relationship between two (2) or more variables (Quantitative) such that one can predict from the others or after.

For Example:

The yield of maize can be caused by the amount of fertilizer applied. The variable which is the caused is called the independent variable or predator variable or explanatory variable. While the variable which is the effect is the dependent variable or explained variable or Response variable.

The dependent variable is represented by y while the Independent variable is represented by x

Simple Linear Regression: Is a relationship Involving only two variables y and x. The model for simple linear Regression is represented Mathematically as $Y_1 = \beta_0 + \beta_1 X_i + e_i$

Where

Y_1 = Dependent variable

X_1 = Independent Variable

β_0 = Interception of y-axis

β_1 = Slope or gradient of the line

e_i = Error Component Superimposed on the linear relation.

β_0 and β_1 are the regression Parameters or Co-efficient of the simple linear regression Which is derived by least square method.

For $\beta_1 = \dfrac{N \Sigma X_1 Y_1 - (\Sigma X_1)(\Sigma Y_1)}{N \Sigma X_1^2 - (\Sigma X_1)^2}$

For $\beta_0 = \frac{1}{N}(\Sigma Y_i - \beta_1 \Sigma X_i)$ OR $\bar{Y} - \beta_1 \bar{X}$

Example: In a Factory, the relationship between working hours

X and the average productivity of labour per worker Y has been determined and the following data obtained

(hrs) X	1	2	3	4	5
(kg) Y	20	18	18	17	15

i) Obtain the least Square regression equation of Y on X.
ii) Estimate the value of y when $x = 6$.

Solution

i)
$$Y_1 = \beta_0 + \beta_1 X_1 + e_1$$

Regression equation of $y = x$

$$X_1 = \beta_0 + \beta_1 Y_1 + Z_1$$

First find β_1

$$\beta_1 = \frac{N\Sigma X_1 Y_1 - \Sigma X_1 \Sigma Y_1}{N\Sigma X_1^2 - (\Sigma X_1)^2}$$

X_i	Y_i	XY	X^2
1	20	20	1
2	18	36	4
3	18	54	9
4	17	68	16
5	15	75	25
15	88	253	55

$$\Rightarrow \beta_1 = \frac{5 \times 253 - (15)(88)}{5 \times 55 - 225}$$

$$= \frac{1265 - 1320}{275 - 225}$$

$$= \frac{-55}{50}$$

$$\beta_1 = -1.1$$

Note:

$\bar{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$

$\frac{\sum Y_i}{N} = \frac{88}{5} = 17.6$

For B_0

$B_0 = \bar{Y} - B_1\bar{X}$

$= 17.6 - (-1.1)3$

$= 17.6 - 3.3$

$= 20.9$

(e) $\therefore Y = B_0 + B_1X_i + e_i$

$Y = 20.9 - 1.1X_i + e_i$

Estimate the value of y when $x = 6$

$Y = 20.9 - 1.1 \times 6$

$= 20.9 - 6.6$

$Y = 14.3$

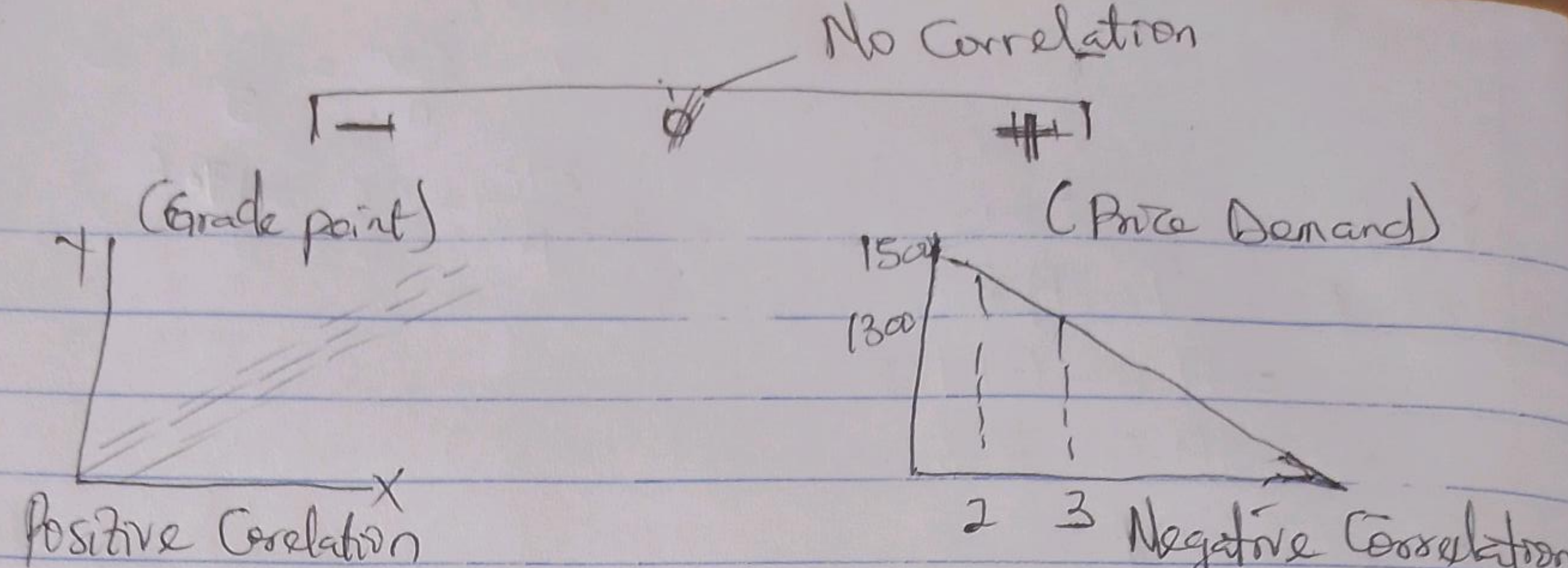
Correlation Analysis

Correlation Analysis shows us the strength of the relationship whether strong, weak, positive, negative, linear etc.

It is the strength of measure or degree of relationship between two random variables. The value of the Correlation Co-efficient lines between -1 and $+1$; This means that $-1 \leq r \leq +1$

Where r is the Correlation Co-efficient.

When it takes up 0 (zero value) it means that there is no Correlation between the two variables or that the variables are linearly uncorrelated.



Two methods of finding Correlation Co-efficient:

1) **Pearson's product moment correlation Co-efficient**

It is given as $r = \dfrac{n\Sigma X_iY_i - (\Sigma X)(\Sigma Y_i)}{\sqrt{(n\Sigma X_i^2 - (\Sigma X)^2)(n\Sigma Y_i^2 - (\Sigma Y_i)^2}}$

Example:

Student	Studying (x)	Exam Score (y)	XY	X^2	Y^2
A	2	50	100	4	2500
B	3	55	165	9	3025
C	5	75	375	25	5625
D	6	80	480	36	6400
E	8	85	680	64	7225
	24	345	1800	138	24775

~~Calculate~~ Given the table above. Calculate the Pearson product moment correlation between X and Y variables

Solution

$$r = \frac{n\Sigma X_iY_i - (\Sigma X)(\Sigma Y_i)}{\sqrt{(n\Sigma X_i^2 - (\Sigma X)^2(n\Sigma Y_i^2 - (\Sigma Y_i)^2}}$$

$$r = \frac{5 \times 1800 - 24 \times 345}{\sqrt{((5 \times 138) - 576)(5 \times 24775 - (345)^2)}}$$

$$= \frac{9000 - 8280}{\sqrt{(690 - 576)(123875 - 119025)}}$$

$$= \frac{720}{\sqrt{(114)(4850)}} = \frac{720}{\sqrt{552900}} = \frac{720}{743.57}$$

$$= 0.9683 \simeq \underline{\underline{0.97}} \Rightarrow \text{strong positive correlation}$$

ii) **Spearman's Rank Correlation**

It is given as $r = 1 - \dfrac{6 \sum d_i^2}{n(n^2 - 1)}$

Where d_i = difference in rank

Example:

X	Y	R_x	R_y	$d_i = R_x - R_y$	d_i^2
1	20	1	5	-4	16
2	18	2	3	-1	1
3	18	3	3	0	0
4	17	4	2	2	4
5	15	5	1	4	16
					37

$$r = \frac{1 - 6\Sigma d_i^2}{n(n^2-1)}$$

$$r = \frac{1 - 6 \times 37}{5(25-1)}$$

$$r = 1 - \frac{222}{120}$$

$$= 1 - 1.85$$

$$= -0.85 \Rightarrow \text{Strong Negative Correlation.}$$