Prof. Kanchan Dhuri.

VIT Vidyalankar Institute of Technology
(Accredited A+ by NAAC)
(Autonomous Institute Affiliated to University of Mumbai)

Mid Semester Examination

Branch	Date	Sem.	Roll No. / Exam Seat No.	Subject	Student's Signature	Junior Supervisor's Name and Sign
AI	1/4/24	6		SL for DS		

Question No.	A	B	C	D	E	F	G	H	Total	Total out of (20 /30 / 40)
1										
2										
3										
4										

Examiners Signature	Student's Sign (After receiving the assessed answer sheet)

Q1. What are the types of statistics? Give examples. Also explain classification of data & what are the different scales of measurement for data?

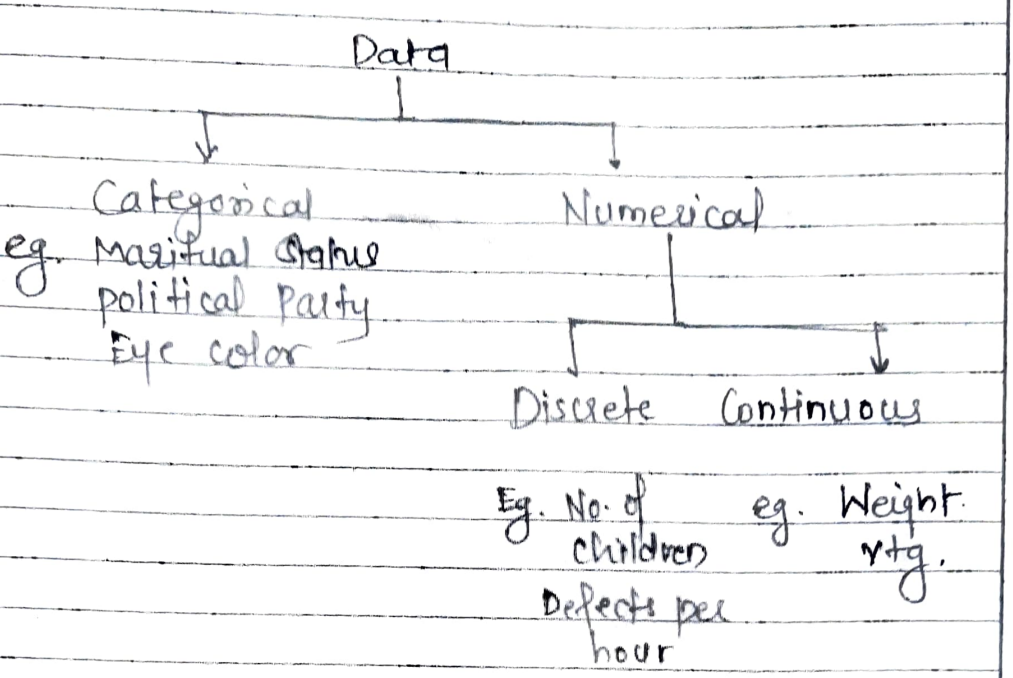
→ A branch of mathematics taking & transforming numbers into useful information for decision makers.

Types of statistics.

1) Descriptive statistics :- Collecting, summarizing & describing data. called descriptive statistics.

2) Inferential Statistics :- Drawing conclusions and/or making decisions concerning a population based only on sample data.

Classification of Data



Scales of Measurement :-

- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization & statistical analyses that are most appropriate.

Scales of measurement include :

Nominal	Interval
ordinal	Ratio.

1) **Nominal :-**
Data are labels or names used to identify an attribute of the element
A non-numeric label or numeric code may be used.
eg. Student of a university are classified by the school in which they are enrolled using a non-numeric label

such as Business, Humanities, Education
1. Business 2 Humanities 3 Education

2) Ordinal :-
The data have the properties of nominal data & the order or rank of the data is meaningful.
A non numeric label or numeric code may be used.

3) Interval :- The data have the properties of ordinal data & the interval between observation is expressed in terms of a fixed unit of measure
- Interval data are always numeric.
eg Melissa has an SAT score of ~~to~~ 1205 while Kelvin has an SAT score of 1090. Melissa scored 115 points more than ~~Kelv~~ Kevin.

4) Ratio :-
The data have all the properties of interval data the ratio of two values is meaningful
- Variables such as distance, height, weight & time use the ratio scale.
- This scale must contain a zero value that indicates that nothing exists for the variable at the zero point.
eg. Melissa's college record shows 96 credit hours. (earned, while Kevin's record shows 72 credit hours earned. Kevin has twice as many credit hours earned as Melissa.

Qualitative & Quantitative data.

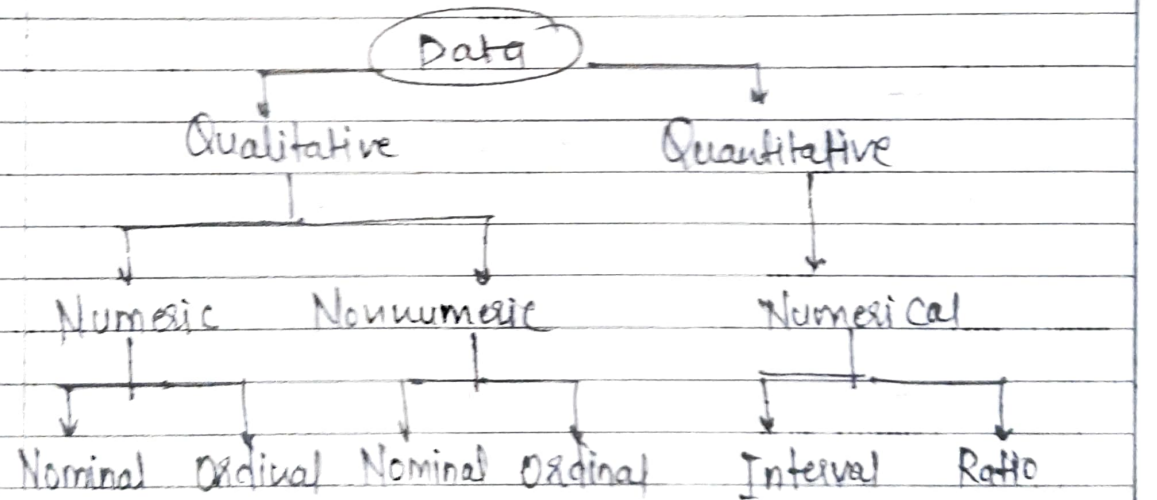
1) Data can be further classified as being qualitative or quantitative

Qualitative Data :-

- Labels or names used to identify an attribute of each element.
- Often referred to as categorical data.
- Use either the nominal or ordinal scale of measurement
- Can be either numeric or nonnumeric.

Quantitative Data :-

Discrete, if measuring how many
Continuous, if measuring how much
Quantitative data are always numeric.
Ordinary arithmetic operations are meaningful for quantitative Data.

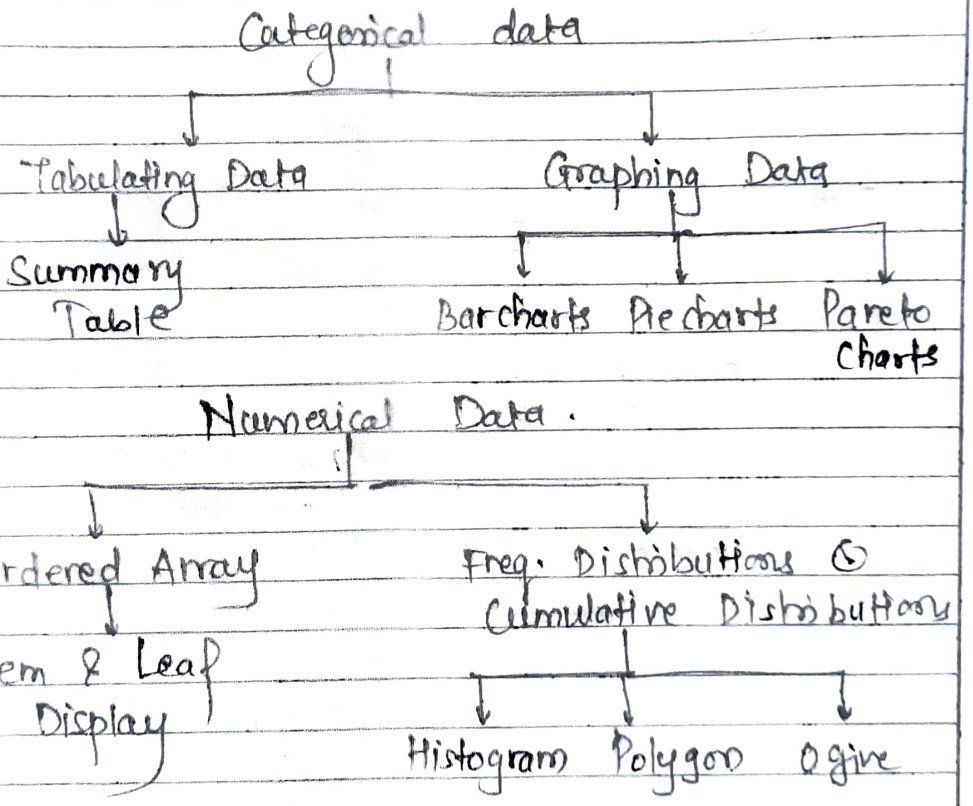


Q2) What are the various ways to present summary of Categorical data & Numerical data.

class.	class mid pt	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2
~~60 but less than~~	~~55~~.	

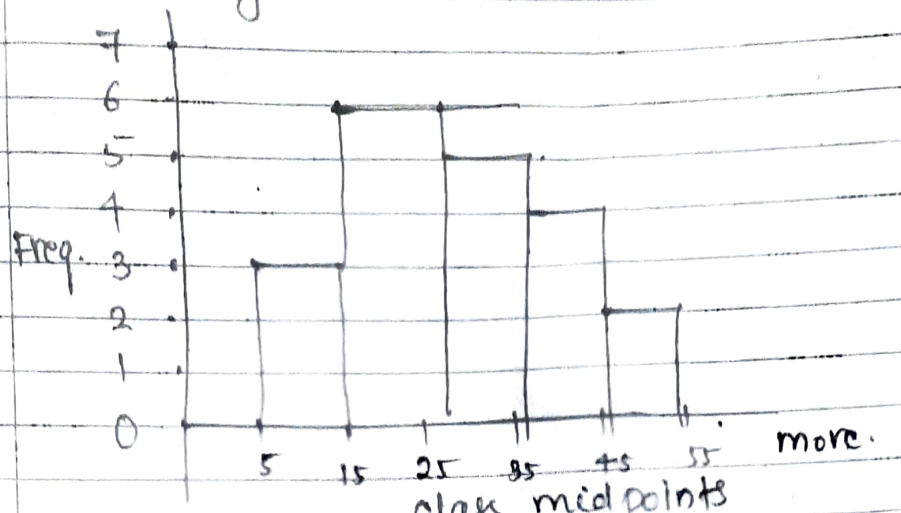
Plot polygon, histogram & ogive graph for above mentioned data.

⇒

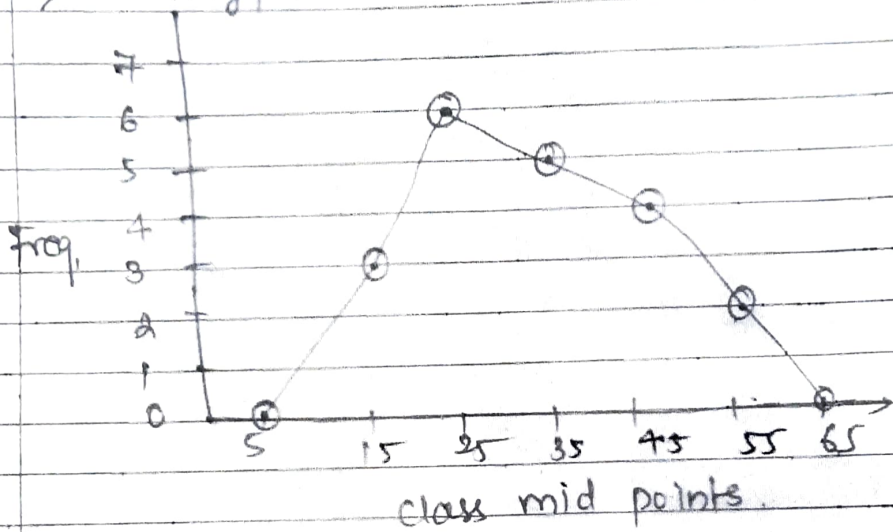


class.	Freq.	mid pt	Relative Freq.	%	Cum. Freq	cum %
10 ≧ 20	3	15	0.15	15	3	15
20 ≧ 30	6	25	0.30	30	9	45
30 ≧ 40	5	35	0.25	25	14	70
40 ≧ 50	4	45	0.20	20	18	90
50 ≧ 60	2	55	0.10	10	20	100
Total	20	—	1.00	100		

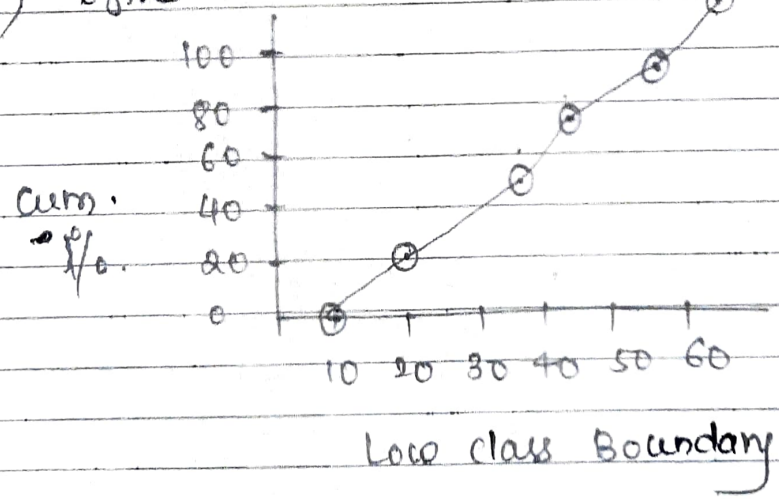
a) Histogram :-



b) Polygon.



c) Ogive



Q3) Define discrete prob. distribution function. Explain Binomial & Poisson's Distribution in details (eg. Formula for expected mean, Variance & Standard deviation)

⇒ Discrete Prob. distributions:-
The prob. distribution for a random variable describes how probabilities are distributed over the values of the random variable.

- We can describe a discrete prob. distribution with a table, graph. or equation.
- The prob. distribution is defined by a prob. function, denoted by $f(x)$, which provides the prob. for each value of the random variable.
- The required conditions for a discrete probability functions are:

$$f(x) \geq 0$$
$$\Sigma f(x) = 1.$$

a] Binomial Distribution:- Describes discrete, not continues, data resulting from an experiment known as Bernoulli Process.

Four properties of a Binomial Exp:
1. The experiment consists of a sequence of n identical trials.
2. Two outcomes, success & failure are possible on each trial.
3. The prob. of a success, denoted by p, does not change from trial to trial.
4. The trials are independent.

- The number of successes occurring in the n- trials.
- We let x denote the no of successes occurring in the n trials.

Binomial prob. functions :-

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

where :
f(x) = the prob. of x successes in n trials
n = the no. of trials.
p = the prob. of success on any one trial

Expected value :- $E(x) = \mu = np$

Variance (x) = $\sigma^2 = np(1-p)$

Standard Deviation (σ) = $\sqrt{np(1-p)}$.

2] Poisson Distribution :-

- The poisson distribution & the binomial distribution have some similarities but also several differences.
- The binomial distribution describes a distribution of two possible outcomes designated as successes & failure from a given number of trials.
- The poisson distribution focuses only on the number of ~~trials~~. discrete occurrences over some interval or continuum.
- The poisson distribution describes the

occurrence of rare events. In fact, the Poisson formula has been referred to as the law of improbable events.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$f(x)$ = prob. of x occurrences in an interval
μ = mean number of occurrences in an interval
$e = 2.71828$.

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

A property of the poisson distribution is that the mean & variance are equal.

Q2)

A)

Class	Freq (f)	mid pt (x)	$f \times x$	$x_i - \bar{x}$
10.0 - 10.9	1	10.5	10.5	-4.71
11.0 - 11.9	4	11.5	46.0	-3.71
12.0 - 12.9	6	12.5	75.0	-2.71
13.0 - 13.9	8	13.5	108.0	-1.71
14.0 - 14.9	12	14.5	174.0	-0.71
15.0 - 15.9	11	15.5	170.5	0.29
16.0 - 16.9	8	16.5	132.0	1.29
17.0 - 17.9	7	17.5	122.5	2.29
18.0 - 18.9	6	18.5	111.0	3.29
19.0 - 19.9	2	19.5	39.0	4.29
Total	65		988.5	

a) $\bar{x} = \frac{\Sigma(f \times x)}{n} = \frac{988.5}{65} = 15.2077$ pounds.

b) Stand. dev $= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \frac{82.941}{9} = \sqrt{9.215} = 3.035$

c) Variance $= s^2 = 9.215 = 9.215$

Q2 B) Find outlier ?

850	875	4700	4900	5300	5700	6700	7300	7700	8100
8300	8400	8700	8700	8900	9300	9500	9500	9700	1100
10300	10500	10700	10800	11000	11300	11300	11800	12700	129
13100	13500	13800	14900	16300	17200	18500	20300	21310	213

$Q_1 = 8100.$

$Q_1 = \frac{1}{4} \times n = 10^{th} =$

$= \frac{40}{4} \Rightarrow 10 = 8100$

$Q_3 = \frac{3}{4} \times n = \frac{3}{4} \times 40 = 30$

$[Q_3 = 12900.]$

$IQR = 12900 - 8100 = 4800$

$1.5 * IQR = 7200.$

$1.5 \times 4800 = 7200$

Outliers = $Q_1 - 7200 = 8100 - 7200 = 900$

Outliers = $Q_3 + 7200 = 12900 + 7200 = 20100$

Any data point below 900 & above 20100 are outliers.

→ Outlier :- An outlier is a single data point that goes far outside the average value of a group of statistics. outlier may be exceptions that stand outside individual samples of populations as well.

locating quartiles :-
find a quartile by determining the value on

the appropriate position in the ranked data.

Q_1 = first quartile position = $Q_1 = \frac{(n+1)}{4}$ ranked value

$Q_2 = \frac{n+1}{2}$

$Q_3 = \frac{3(n+1)}{4}$ where n = number of observed value

Interquartile range = $Q_3 - Q_1$

Five number summary

X smallest
First quartile (Q_1)
Median (Q_2)
Third quartile Q_3
X largest.

Q2c)

Gender	First time offender	Repeat offender
Male	60	70
Female	44	76
Total	104	146

a) The prob that the shoplifter is male

$P(M) = \frac{60+70}{250} = 0.520$

b) The prob. that the shoplifter is first time offender, given that the shoplifter is male

$P(F/m) = \frac{P(F \cap m)}{P(m)} = \frac{60/250}{130/250} = 0.465$

c) ~~P(w/~~ The prob. that the shoplifter is female, given that the shoplifter is a repeat offender.

$$P(W/R) = \frac{P(W \cap R)}{P(R)} = \left(\frac{76}{250}\right) / \left(\frac{146}{250}\right)$$

$$= 0.521.$$

d) The prob. that the shoplifter is female, given that the shoplifter is a first time offender.

$$P(W/F) = \frac{P(W \cap F)}{P(F)} = \left(\frac{44}{250}\right) / \left(\frac{104}{250}\right) = 0.423$$

e) The prob. that the shoplifter is both male & a repeat offender.

$$P(M \text{ and } R) = 70/250 = 0.280.$$

Q3 A] Three types of probabilities

1] Classical approach
2] Relative frequency approach
3] Subjective approach

1] classical approach:-

prob. of an event = (no. of outcomes where the event occurs) / (total no. of possible outcomes)

$$P(H) = \frac{1}{(1+1)} = \frac{1}{2}.$$

Prob of getting head when coin is tossed.

VIT Vidyalankar Institute of Technology
(Accredited A+ by NAAC)
(Autonomous Institute Affiliated to University of Mumbai)

Mid Semester Examination

Branch	Date	Sem.	Roll No. / Exam Seat No.	Subject	Student's Signature	Junior Supervisor's Name and Sign

Question No.	A	B	C	D	E	F	G	H	Total	Total out of (20 /30 / 40)
1										
2										
3										
4										

Examiners Signature	Student's Sign (After receiving the assessed answer sheet)

$P(5) = 1/6$ for the dice rolling

Classical prob is also called as a priory prob. because we don't need to perform experiments.

2] Relative Frequency (RF):-
Live upto 85 years, plant near river will substantially kill the fish.
We need experiment to answer these.
This method uses the relative frequencies of past occurrences as probabilities.
- How often something has happened in past - we predict future.
- more trails greater accuracy.

3] Subjective probability :- Based on belief, experience, when event has occurred once

or few times.
Because most higher-level & social & managerial decisions are concerned with specific, unique, stations, rather than with a long series of identical situation, decisions makers use this prob.

Probability under conditions of statistical independence :-

→ Statistical independence : The occurence of one event has no effect on the prob. of occurence of any other event.

1. Marginal prob. under statistical independence :-

Tossing of fair coin, outcome of second toss is independent of outcome of first toss.

2. Joint prob. under statistical independence :-

The prob. of two or more independent events occurring together or in succession is the product of their marginal probabilities.

Joint prob. of two independent events :

$P(AB) = P(A) \times P(B)$

$P(AB)$ = prob. of events A & B occurring together, this is known as a joint prob.

$P(A)$ = marginal prob. of event A occurring

$P(B)$ = marginal prob. of event B occurring

$P(H_1 H_2) = 0.5 * 0.5 = 0.25$.
This is the prob. of heads in two succession tosses)

3) Conditional probabilities under statistical independence :-
It is written as $P(B/A)$. The prob. of event B given A has occurred.

$$P(B/A) = P(B)$$

What is the prob. that the second toss of a fair coin will result in heads, given that heads resulted in first toss.
$P(H_2/H_1)$, we know that independence means the first toss's result could not affect the result of second toss.
$P(H_2/H_1) = 0.5$.

Summary

Prob. under statistical independence	Types of prob	Symbol	Formula
	marginal	P(A)	P(A)
	Joint	P(AB)	P(A) × P(B)
	Conditional	P(B/A)	P(B)

Probabilities under conditions of Statistical Dependence :

- When prob. of some event is dependent on or affected by the ocurrence of some other event.

1) Conditional Prob. under statistical Dependence:
Assume a box has 10 balls as follows:

Event	Prob. of event	
1	0.1	colored & dotted
2	0.1	
3	0.1	
4	0.1	colored & stripped
5	0.1	Gray & dotted
6	0.1	~~Gray & stripped~~
7	0.1	Gray & striped
8	0.1	
9	0.1	
10	0.1	

eg. If a colored ball is drawn.

1. what is the prob. that it is dotted
$P(D/C) = 0.3/0.4$

2. What is the prob. that it is striped
$P(S/C) = 0.1/0.4$

conditiondl probabilities for statistical dependent events.

2. Joint prob. under statistical dependence:
We know that conditional prob. under statistical dependence:
$P(B/A) = P(BA) / P(A)$

$P(BA) = P(B/A) * P(A)$

- $P(BA)$: Joint prob of events B & A
- $P(B/A)$: Prob. of event B given that A has happened
- $P(A)$: prob. of events A.

What is $P(D/G) = P(DG) / P(G)$

$= 0.2 / 0.6$

$= 1/3.$

3) Marginal probabilities under statistical dependence: Are computed by summing up the prob. of all the joint events in which the simple event occurs.

$P(c) = P(cD) + P(cs) = 0.3 + 0.1 = 0.4$

Q3) b) Explain covariance & significance of coefficient of correlation method.
for following data a) Compute covariance.
b) Compute coefficient of correlation
c) Which is valuable in expressing relationship?
d) What conclusion can you reach about relationship?

Soln → Relationship between two numerical variable
1) Covariance.
2) Coefficient of correlation.

The covariance measures the strength of the linear relationship between two numerical variables (X & Y)

The sample covariance:

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- only concerned with the strength of the relationship
- No casual effect is implied

Interpreting Covariance :-
Covariance between two variables :

$cov(x,y) > 0 \rightarrow$ x & y tend to move in the same direction

$cov(x,y) < 0 \rightarrow$ x & y tend to move in opposite direction.

$cov(x,y) = 0 \rightarrow$ x & y are independent

The covariance has a major flaw:
It is not possible to determine the relative strength of the relationship from the size of the covariance.

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$r = \frac{cov(x,y)}{S_x S_y}$$

Where,

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

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

$$S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

- The population coefficient of correlation is referred as ρ.
- The sample coefficient of correlation is referred to as r.
- Either ρ or r have the following features:

Ranges between -1 & 1

- The closer to -1, the stronger the negative linear relationship
- The closer to 1 the stronger the positive linear relationship.
- The closer to 0, the weaker the linear relationship.

	x (calories)	y (Fat)	a $x_i - \bar{x}$	b. $y_i - \bar{y}$	axb. axb
1	240	8	-140	-7.78	1089.2
2	260	3.5	-120	-12.28	1473.6
3	350	22	-30	6.22	-186.6
4	350	20	-30	4.22	-126.6
5	420	16	40	0.22	8.8
6	510	22	130	6.22	808.6
7	530	19	150	3.22	483

$\bar{x} = 380$

$\bar{y} = 15.78.$

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\boxed{Cov(x,y) = 591.66}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = 113.14$$

$$Sy = \sqrt{\sum_{i=1}^{7} \frac{(Yi - \bar{Y})^2}{n-1}}$$

$$= 7.26_{11}$$

$$r = \frac{COV(X,Y)}{Sx\ Sy}$$

$$= \frac{591.66}{113.14 \times 7.26}$$

$$[r = 0.72]$$

c) Which is valuable in expressing relationship
r = Correlation

d) r = 0.72 is more closer to '+1' hence
It is strong positive relationship