PROBABILITY & STATISTICS

BCA TU

WHAT'S INSIDE?

Theory and formulas with example of all units

info@notebahadur.com

Note Bahadur

1

Chapter-1 Introduction to statistics

⇒ statistics is concerned with scientific methods for collecting, summarizing, organizing, presenting and analyzing data as well as deriving valid conclusions and making resonable decisions on the basis of this analysis.

ⓐ <u>Mathematical statistics</u>

→ Descriptive statistics

→ Inferential statistics

<u>Descriptive statistics</u>: ~~st~~

→ statistics which deals with collection of data, presentation of data using table diagram, graph etc and summarize data using measure of central tendency, skewness, dispersion, kurtosis etc is called descriptive statistics.

<u>Inferential statistics</u>:

→ Statistics which deals with sample selection from population and statistical techniques used to draw conclusion about population on the basis of statistical measures obtained from sample.

DATE
PAGE NO. 2

* Functions of statistics:

① Condensation
→ Understanding huge mass of data by providing only few observations.

② Comparison
→ Comparing two different data from different sources.

③ forecasting:
→ To make predictions like weather, profit etc.

④ Estimation
⑤ sequential analysis
→ On the behalf of any reference to estimate something like age, height etc.

* Scope of statistics:
→ Physical and Biological science
→ Commerce, Education
→ Planning
→ Business Management
→ Information Technology etc.

3

* Limitations of statistics:
-> Statistics doesn't study individual.
-> Statistics doesn't study qualitative phenomena.
-> Statistics data do not reveal the story
-> Statistical results aren't always unquestionable.
-> Statistical laws are true on the average or in the long run

* Application of statistics in Computer
-> Data organization and coding
-> Storing the data in the computer
-> Selection of appropriate statistical measures/ techniques
-> Selection of appropriate software package
-> Execution of the computer program.

* DATA
Data is a set of collections of objects or cases. Each case has one or more attributes or qualities called variables.

① Qualitative Variable
The variable which varies in kind rather than in magnitude is called qualitative variable.
eg: Hair color, eye color, gender, smoking habits, social status etc

4

② Quantitative Variable
The variable which varies in magnitude and can be expressed in numerically is called quantitative Variable

Discrete variable
(It takes only whole number or countable numbers.
eg. No. of rooms in building, family size)

Continuous variable
(It takes all possible values i.e. whole & fractions.
eg. height, weight)

* Sources of data

① Primary Data
⇒ The data which are originally collected by investigator or researcher for the first time for the purpose of statistical enquiry is called primary data.
eg. Direct personal interview, mailed questionnaire, Indirect oral interview etc.

② Secondary Data
The data that has been already collected for a particular purpose and used for next purpose is called secondary data. eg WHO info, FNCCI info, Nepal gov etc. BBC etc.

Chapter-2

5

Descriptive statistics

* Measure of central Tendency

Objectives:

-> To get a single value that represents the characteristics of the entire data

-> To facilitate comparison

① Arithmetic Mean

ⓐ Simple Arithmetic Mean

① Individual series

ⓐ Direct method

$$\text{Mean}(\bar{x}) = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\Sigma x}{n}$$

ⓑ Deviation Method

$$\text{Mean}(\bar{x}) = A + \frac{\Sigma d}{n}$$

A = assumed mean

d = x-A

ⓒ step-Deviation Method

$$\text{Mean}(\bar{x}) = A + \frac{\Sigma d'}{n} \times h$$

A = assumed mean

$d' = \frac{x-A}{h}$

h = common factor

DATE
PAGE No. 6

② Discrete Series

(a) Direct Method

$$\text{Mean}(\bar{X}) = \frac{\Sigma fx}{n}$$

(b) Deviation Method

$$\text{Mean}(\bar{X}) = A + \frac{\Sigma fd}{n} \qquad d = X - A$$

(c) Step-Deviation Method

$$\text{Mean}(\bar{X}) = A + \frac{\Sigma fd'}{n} \times h \qquad d' = \frac{X-A}{h}$$

h = common factor

③ Continuous Series

(a) Direct Method

$$\text{Mean}(\bar{X}) = \frac{\Sigma fm}{n} \qquad m = \frac{\text{lower limit} + \text{Upper limit}}{2}$$

(b) Deviation Method

$$\text{Mean}(\bar{X}) = A + \frac{\Sigma fd}{n}$$

A = assumed mean

(c) Step-Deviation Method

$d = m - A$

$$\text{Mean}(\bar{X}) = A + \frac{\Sigma fd'}{n} \times h \qquad d' = \frac{m-A}{h}$$

DATE
PAGE No. 7

* Weighted Arithmetic Mean

Mean $(\bar{X}_w) = \frac{\Sigma wx}{\Sigma w}$

eg:

ISP	volume (w)	Price / GB (x)	wx
Vianet	1,800	80	144000
Subisu	77,00	70	5,390,000
	Σw = 980		Σwx = 5534000

* Median

ⓐ Individual series

ⓐ 24, 38, 39, 40, 52, 59, 67, 75, 77

Here, n = 9

size of Md = $\boxed{\frac{n+1}{2}}$ = $\frac{9+1}{2}$ = 5

i.e. Md = 5^{th} item i.e. 52

ⓑ Descrete series

Income (arrange in asc)	No. of Person (f)	cf
800	16	16
1000	24	40
1500	26	66
1800	30	96
2000	20	116
2500	6	122
	N = 122	

DATE
PAGE No. 8

Size of Md = $\boxed{\frac{N+1}{2}} = \frac{122+1}{2} = 61.5^{th}$ item

Since the c.f just greater than 61.5 is 66

$\therefore$ Md = 1500

© Continuous series

Height	No. of students (f)	cf
161-167	79	79
167-173	92	171
173-179	60	231
179-185	22	253
185-191	5	258
191-197	2	260
	N=260	

Here,

Size of Md = $\boxed{\frac{N}{2}} = \frac{260}{2} = 130$

$\because$ The c.f. just greater than 130 is 171 so Median lies in class (167-173)

$$Md = L + \frac{\frac{N}{2} - cf}{f} \times h$$

$$= 167 + \frac{130-79}{92} \times 6$$

$\therefore$ Md = 170.32 cm

* Partition Volues

(a) Individual series

$$Q_i = \left(\frac{n+1}{4}\right) \times i \longrightarrow i = 1, 2, 3$$

$$D_i = \left(\frac{n+1}{10}\right) \times i \longrightarrow i = 1, 2, \ldots, 9$$

$$P_i = \left(\frac{n+1}{100}\right) \times i \longrightarrow i = 1, 2, \ldots, 99$$

eg. 8, 12, 17, 20, 22, 25, 30

Note: Always order in asc or desc

(i) size of $Q_1 = \frac{n+1}{4} \times 1 = \left(\frac{7+1}{4}\right) \times 1 = 2^{nd}$ item
i.e. 12

(ii) size of $P_{60} = \left(\frac{n+1}{100}\right) \times i = \frac{7+1}{100} \times 60 = 4.8^{th}$ item

$= 4^{th} + 0.8(5^{th} - 4)$

$= 20 + 0.8(22 - 20)$

$= 21.6$

(iii) size of $D_7 = \left(\frac{n+1}{10}\right) \times i = \left(\frac{7+1}{10}\right) \times 7 = 5.6^{th}$

$= 5^{th} + 0.6(6^{th} - 5^{th})$

$= 23.8$

DATE
PAGE No. 10

(b) Descrete Series

Respiratory rate (x)	No. of Pee (f)	c.f
10	8	8
15	12	20
20	36	56
25	25	81
30	28	109
35	18	127
40	9	136
45	12	148
50	6	154
	N=154	

Here

Size of $Q_3 = \boxed{\frac{N+1}{4} \times 3} = \frac{154+1}{4} \times 3 = 116.25$

The cf just greater than 116.25 is 127

$\therefore Q_3 = 35$

IIIy,

$P_{80} = \frac{N+1}{100} \times 80 = \frac{154+1}{100} \times 80 = 124$

$\therefore P_{80} = 35$

$D_6 = \frac{N+1}{10} \times 6 = \frac{154+1}{10} \times 6 = 93$

$\therefore D_6 = 30$

DATE
PAGE No. 11

(c) Continuous series

Height (cm)	No. of person (f)	cf
161-167	79	79
167-173	92	171
173-179	60	231
179-185	22	253
185-191	5	258
191-197	2	260

$N = \Sigma f = 260$

Now,

size of $Q_1 = \frac{N}{4} = \frac{260}{4} = 65^{th}$ item

The cf just greater than 65 is 79 so Q_1 lies in class (161-167)

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$= 161 + \frac{65 - 0}{79} \times 6 = 165.93$$

lly, All ~~fs~~ are same.

$$P_{20} = L + \frac{\frac{N}{100} - cf}{f} \times h$$

$$D_7 = L + \frac{\frac{N}{10} - cf}{f} \times h$$

* Mode:

Maximum occured value in the series of observation is called mode.

(a) Individual series

10, 27, 24, 12, 27, 20, 18, 15, 30, 27

∴ Mode = 27

Note: If the data is bi-modal (2 mode) or multi-modal (more than 2 mode) then

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

(b) Descrete series

x	28	29	30	31	32	33	34
f	10	20	40	65	50	15	10

Here the value 31 has max frequency i.e. 65 so Mode = 31.

(c) Continuous series

x	10-20	20-30	30-40	40-50	50-60	60-70
f	9	18	31	17	16	9

By inspection Mode lies in class 30-40

$L = 30$, $f_1 = 31$, $f_0 = 18$, $f_2 = 17$, $h = 10$

$\therefore$ Mode = $\boxed{L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h}$

$= 30 + \frac{31-18}{62-18-17} \times 10$

$= 34.81$

* Measure Dispersion

Methods of studying Dispersion:

(a) Range

→ The difference between maximum and minimum values in a given data is called Range.

i.e. Range = L − S

Hy,

coefficient of range = $\frac{L-S}{L+S}$

Imp Note:

Age (yrs)	No. of person
16-20	10
21-25	15
26-30	17
31-35	8

If we convert the given inclusive classes into exclusive classes we get the first class as 15.5 - 20.5 and the last class as 30.5 - 35.5

so,

DATE
PAGE No. 14

largest limit (L) = 35.5
lowest limit (S) = 15.5
∴ Range (R) = $L - S$ = $35.5 - 15.5$
= 20 years

Coeff. of Range = $\frac{L-S}{L+S} = \frac{20}{51} = 0.39$

Note: We convert inclusive classes into exclusive classes by using the correction factor:

$$cf = \frac{\text{Lower limit of 2nd class} - \text{Upper limit of 1st class}}{2}$$

$$= \frac{21-20}{2} = 0.5$$

Now Add 0.5 to upper limits and subtract 0.5 to lower limits

* Quartile Deviation: (Semi-Inter Q.Range)

Inter-quartile Range = $Q_3 - Q_1$

Quartile deviation (QD) = $\frac{Q_3 - Q_1}{2}$

coefficient of QD = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

DATE
PAGE No. 15

Individual series

730, 1150, 1450, 1850, 2680, 2800, 3150, 3670, 6190, 8700

Q_1 = size of $\left(\frac{n+1}{4}\right)^{th}$ item

= size of $\left(\frac{10+1}{4}\right)^{th}$ item = 2.75^{th} item

= 2^{nd} + 0.75 (3^{rd} − 2^{nd}) item

= 1375

Q_3 = size of $3\left(\frac{n+1}{4}\right)^{th}$ item

= ~~1462.5~~ 4300

∴ $QD = \frac{Q_3 - Q_1}{2} = \frac{4300 - 1375}{2} = 1462.5$

coeff. of $QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.5$

Discrete Series

size	f	cf
2	3	3
4	5	8
6	10	18
8	12	20
10	6	26
12	4	30
	N=30	

DATE
PAGE No. 16

Q_1 = size of $\left(\frac{N+1}{4}\right)^{th}$ item

= size of $\left(\frac{30+1}{4}\right)^{th}$ item

= size of 30.75^{th} item

= ~~30 10~~ 6

Q_3 = size of $3\left(\frac{N+1}{4}\right)^{th}$ item

= size of $3\left(\frac{30+1}{4}\right)^{th}$ item

= 10

$\therefore QD = \frac{Q_3 - Q_1}{2} = \frac{10-6}{2} = 2$

coeff of $QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{10-6}{10+6} = 0.25$

* ~~step-Deviati~~ Continuous Series

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times h$$

Remaing same as above ↑

17

* Mean Deviation or Average Deviation

In Individual Observations

MD from mean $= \frac{1}{n} \Sigma|x-\bar{x}|$

MD from median $= \frac{1}{n} \Sigma|x-M_d|$

MD from mode $= \frac{1}{n} \Sigma|x-M_o|$

In Discrete or continuous ~~ab~~ f distribution

MD from mean $= \frac{1}{N} \Sigma f|x-\bar{x}|$

MD from median $= \frac{1}{N} \Sigma f|x-M_d|$

MD from mode $= \frac{1}{N} \Sigma f|x-M_o|$

coeffients:

coeff. of MD from mean $= \frac{\text{M.D from mean}}{\text{mean}}$

coeff. of MD from median $= \frac{\text{MD from median}}{\text{median}}$

Coeff. of MD from mode $= \frac{\text{MD from mode}}{\text{mode}}$

DATE
PAGE No. 18

* Standard Deviation:

Individual series

$$\text{standard Deviation } SD = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

Continuous & Discrete series

$$SD = \sqrt{\frac{\Sigma f x^2}{N} - \left(\frac{\Sigma f x}{N}\right)^2}$$

*Individual Series

Direct Method

$$SD(\sigma) = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \text{ or } \sqrt{\frac{1}{n}\Sigma (x-\bar{x})^2}$$

Deviation method

$$SD(\sigma) = \sqrt{\frac{\Sigma f x^2}{f}\ \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

where,

$d = X - A$ (A is assumed mean)

* Discrete and Continuous series

Direct Method

$$SD(\sigma) = \sqrt{\frac{\Sigma f x^2}{N} - \left(\frac{\Sigma f x}{N}\right)^2} \text{ or}$$

where

$N = \Sigma f$

$$\sqrt{\frac{1}{N}\Sigma f(X-\bar{X})^2}$$

19

Deviation Method

$$SD(\sigma) = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \quad (d = X - A)$$

Step Deviation Method

$$SD(\sigma) = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \quad \left(d' = \frac{X-A}{h}\right)$$

EXAMPLE OF INDIVIDUAL SERIES

Bacteria No. (x)	$x - A = d$	d^2	$x - \bar{x}$	$(x-\bar{x})^2$
120	-5	25	-2	4
110	-15	225	-12	144
115	-10	100	-7	49
122	-3	9	0	0
126	1	1	4	16
140	15	225	18	324
125	0	0	3	9
121	-4	16	-1	1
110	-15	225	-12	144
131	6	36	9	81
$\Sigma x = 1220$	$\Sigma d = -30$	$\Sigma d^2 = 862$		$\Sigma(x-\bar{x})^2 = 772$

Here

$A = 125$

$$\text{Mean } (\bar{x}) = \frac{\Sigma x}{n} = \frac{1220}{10} = 122$$

DATE
PAGE No. 20

(i) SD By Direct Method

$$\sigma = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2} = \sqrt{\frac{772}{10}} = 8.786$$

(ii) SD by deviation method

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{862}{10} - \left(\frac{-30}{10}\right)^2}$$

$$= 8.786$$

* Continuous Series

class	Mid value	f	$d' = (X-A)/h$	fd'	fd'^2
-10 to 0	-5	19	-3	-57	171
0 to 10	5	24	-2	-48	96
10 to 20	15	49	-1	-49	49
20 to 30	25	87	0	0	0
30 to 40	35	31	1	31	31
40 to 50	45	27	2	54	108
		N = 237		$\Sigma fd' = -69$	$\Sigma fd'^2 = 455$

Now

$A = 25$

$$\text{Mean}(\bar{x}) = A + \frac{\Sigma fd'}{N} \times h$$

$$= 25 + \frac{(-69)}{237} \times 10$$

$$= 22.09$$

$$\therefore SD(\sigma) = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

$$= \sqrt{\frac{455}{237} - \left(\frac{-69}{237}\right)^2}$$

$$= 13.56$$

* Combined Standard Deviation:

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where, $d_1 = \bar{X}_1 - \bar{X}_{12}$ & $\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$

$d_2 = \bar{X}_2 - \bar{X}_{12}$

* Variance (Individual series)

Varience = $(SD)^2$

Here

$$SD(S) = \sqrt{\frac{1}{n-1}\Sigma(x-\bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - n\bar{x}^2\right)}$$

g

X	x^2
12	144
7	49
4	16
9	81
0	0
7	49
3	9
$\Sigma x = 42$	$\Sigma x^2 = 348$

Ans

$$\bar{X} = \frac{\Sigma x}{n} = \frac{42}{7} = 6$$

$$S = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - n\Sigma x^2\right)}$$

$$= 4$$

$$\therefore \text{varience} = s^2 = 4^2 = 16$$

DATE
PAGE No. 22

Coefficient of Variation:

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

Less CV less variable or more consistant or more homogeneous or more stable

More CV more variable or less consistant or less homogeneous or less stable

Example

$n = 200$

$\bar{X} = 60$

$\sigma = 20$

if two items 3 & 67 are wrongly taken instead of 13 & 17 then find mean & standard Deviation.

$$\bar{X} = \frac{\Sigma X}{n} \Rightarrow \Sigma X = n\bar{X} = 200 \times 60 = 12000$$

Also

$$\sigma^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

$$(20)^2 = \frac{\Sigma X^2}{200} - \left(\frac{12000}{200}\right)^2$$

$$\Rightarrow \Sigma X^2 = 800000$$

PAGE No. 23

If the wrong items 3 & 67 are replaced by the correct values 13 & 17 respectively, we get

correct $\Sigma X = 12000 - 3 - 67 + 13 + 17 = 11960$

correct $\Sigma X^2 = 800000 - 3^2 - 67^2 + 13^2 + 17^2$

$= 795960$

$\therefore$ Corrected mean $\bar{x} = \frac{\text{Correct } \Sigma X}{n} = 59.8$

Correct sd $(\sigma) = \sqrt{\frac{\text{Correct } \Sigma X^2}{n} - \left(\frac{\text{Correct } \Sigma X}{n}\right)^2}$

$= \sqrt{\frac{795960}{200} - (59.8)^2}$

$= 20.09$

$\therefore$ correct CV $= \frac{\text{correct } \sigma}{\text{correct } \bar{X}} \times 100\% = \frac{20.09}{59.8} \times 100\%$

$= 33.60\%$

DATE
PAGE No. 29

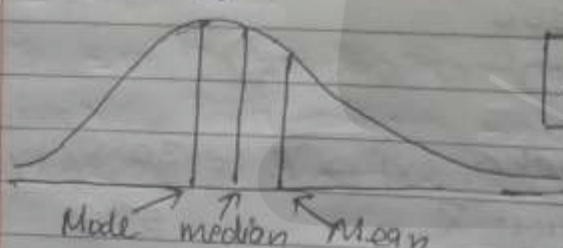
* SKEWNESS

→ If in a distribution mean = median = mode, then it is known as symmetrical distribution. otherwise it is known as non-symmetrical distribution i.e. mean ≠ median ≠ mode.

(a) Symmetrical Distribution

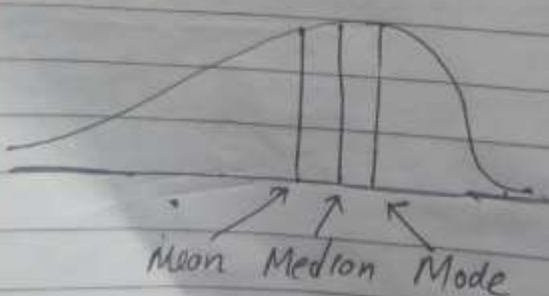
mean = median = mode

(b) Positively skewed distribution



$\boxed{\bar{X} > M_d > M_o}$

(c) Negatively skewed distribution



$\boxed{M_o > M_d > \bar{X}}$

Measures of skewness

(a) Karl-Pearson's coefficient of skewness

$$S_{kp} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \boxed{\frac{\bar{x} - M_o}{\sigma}}$$

or,

$$\boxed{S_{kp} = \frac{3(\bar{x} - M_d)}{\sigma} \text{ if mode is ill i.e. no frequency}}$$

Interpretation

- $S_{kp} = 0$ distribution is symmetrical (not skewed)
- $S_{kp} > 0$, Distribution is positively skewed (right skewed)
- $S_{kp} < 0$, Distribution is negatively skewed (left skewed)

Example:

x	x^2
2	4
3	9
6	36
9	81
15	225
$\Sigma x = 35$	$\Sigma x^2 = 355$

Now $\Sigma x = 35$, $\Sigma x^2 = 355$, $n = 5$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{35}{5} = 7$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{355}{5} - \left(\frac{35}{5}\right)^2}$$

$$= \sqrt{32} = 5.656$$

Here, Mode is ill defined so,

$$M_d = \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$= \left(\frac{5+1}{2}\right)^{th} \text{ item} = 3^{rd} \text{ item i.e. } 6$$

DATE
PAGE No. 26

$\therefore S_{kp} = \frac{3(\bar{x} - M_d)}{\sigma}$

$= \frac{3(7-6)}{5.656}$

$= 0.53$

So, $0.53 > 0$ So the distribution is positively skewed.

BOWLEY's Measure of Skewness

$$S_{kB} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Interpretation

$S_{kB} = 0$, Distribution is symmetrical (Not skewed)

$S_{kB} > 0$, Distribution is +vely skewed

$S_{kB} < 0$, Distribution is -vely skewed

steps:

find Q_1

find Q_3

find M_d

Find S_{kB}

#

kurtosis

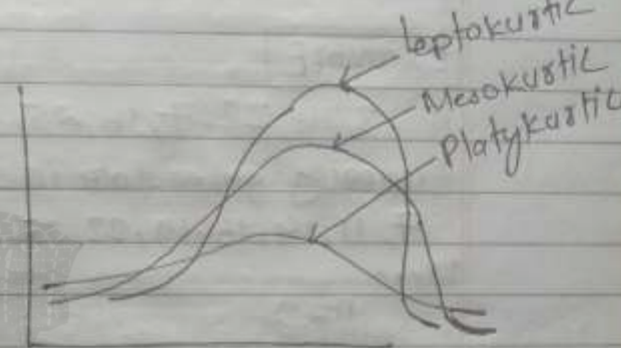
It is the measure of peakedness or curve or flatness of given distribution.

Types of kurtosis

(i) Leptokurtic curve

(ii) Mesokurtic curve

(iii) Platykurtic curve



(a) Percentile Coeff. of kurtosis (kelly's coeff. of kurtosis)

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \quad \text{or} \quad K = \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})}$$

steps:

find Q_1

find Q_3

find P_{90}

find P_{10}

find K

steps:

find P_{75}

find P_{25}

find P_{90}

find P_{10}

find K

DATE
PAGE No. 28

If $K = 0.263$, Distribution is mesokurtic (normal)
If $K > 0.263$, Distribution is leptokurtic
If $K < 0.263$, Distribution is platykurtic

Example

Q) 3, 14, 22, 7, 16, 25, 11, 19, 27
Arranging given data in ascending order
7, 11, 14, 16, 19, 22, 25, 27
Here
$n = 9$

$P_{10} = 10(n+1)/100^{th}$ item $= 10(9+1)/100^{th}$ item $= 1^{st}$ item
$= 3$

$P_{90} = 90(n+1)/100 = 90(9+1)/100^{th}$ item $= 9^{th}$ item
$= 27$

$Q_1 = (n+1)/4^{th}$ item $= (9+1)/4 = 2.5^{th}$ item
$= \frac{2^{nd} \text{ item} + 3^{rd} \text{ item}}{2} = 9$

$Q_3 = 3(n+1)/4 = 3(9+1)/4^{th}$ item $= 7.5^{th}$ item
$= \frac{7^{th} \text{ item} + 8^{th} \text{ item}}{2} = \frac{22+25}{2} = 23.5$

Now
$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{23.5 - 9}{2(27-3)} = 0.302$

$\therefore K = 0.302 > 0.236$ so the distribution is leptokurtic.

Correlation and Regression Analysis
Correlation and Regression Analysis

Univariate Data: when data are collected according to a single variable is called univariate data. eg age of individuals.

Bivariate Data: When data are collected according to two variables is called ~~univ~~ Bivariate data. eg. data on height & weight of students.

* CORRELATION

If two or more variables are so related that the change of one variable brings ~~se~~ change in the value of other variable, then the variables are said to be correlated. This relationship between the variables is called Correlation.

Types of Correlation.

① Positive and Negative correlation

→ If both the variables move in the same direction ie. if the increase or decrease in the value of one variable results the increase or decrease in the value of other variable, then the two variables are said to be positive correlated.

If both variables move in oppositive direction, then correlation between the two variables

PAGE No. 30

is called Negative Correlation.

(2) Linear & Non-linear Correlation:
-> The correlation between two variables is said to be linear if corresponding to a unit change in the other variable over the entire range of the value.

If corresponding to a unit change in one variable, there is no constant change in other variable then the correlation is said to be Non linear.

(3) Simple, Multiple and partial correlation
Simple: The relationship between two (ie. one dependent and one independent) variable is called simple correlation.

Multiple: The study of relationship among 3 or more variables simultaneously is known as multiple correlation.

partial: The study of relationship between 2 variables keeping the effect of all remaining variables constant when there are 3 or more variables is called partial correlation.

DATE
PAGE No. 37

* KARL PEARSON'S CORRELATION COEFFICIENT

It is denoted by 'r'

formulas-

$$r = \frac{Cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Direct Method

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2} - \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

Deviation Method

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} - \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$u = X - A$

$v = Y - B$

Step Deviation Method

$$r = \frac{n\Sigma u'v' - \Sigma u' \Sigma v'}{\sqrt{n\Sigma u'^2 - (\Sigma u')^2} - \sqrt{n\Sigma v'^2 - (\Sigma v')^2}}$$

$u' = X - A$

$v' = Y - B$

A = Assumed mean of X-series

B = Assumed mean of y-series

Correlation lies between -1 and 1

high	Moderate	Low		low	Moderate	high	
-1	-0.7	-0.5	0	0.5	0.7	1	

← Negative Correlation → | ← positive correlation →

No correlation

* PROBABLE ERROR

$$PE(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

If $r < P.E.(r)$ then r is not significant

If $r > P.E.(r)$ then r is significant

r = correlation of karl's pearson

n = no. of obs.

* Estimation of coefficients using Least Square Method

Deviation Method

$$\Sigma u = na + b\Sigma v$$
$$\Sigma uv = a\Sigma v + b\Sigma v^2$$

[eq^n u on v]

where $u = X - A$ (A = assumed mean in X-series)

$V = Y - B$ (B = assumed mean in Y-series)

DATE
PAGE No. 33

Step - Deviation Method

$$\Sigma u' = na + b\Sigma v'$$
$$\Sigma u'v' = a\Sigma v' + b\Sigma v'^2$$

when, $u' = \frac{x-A}{h}$, $v' = \frac{y-B}{K}$

Example:

Q

X	12	15	13	7	10	5	22	9	13	7
Y	77	85	48	59	75	41	94	65	79	70

$\Sigma x = 103$, $\Sigma y = 693$, $\Sigma x^2 = 1335$, $\Sigma y^2 = 50447$, $\Sigma xy = 7881$

fitting a simple linear regression model y on x.

Soln

To fit $y = a + bx$

$$\Sigma y = na + b\Sigma x$$

or $693 = 10a + 103b$ —— (i)

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

or, $7881 = 103a + 1335b$ —— (ii)

Solving (i) & (ii)

coeff. of a	coeff. of b	constant
10	103	693
103	1335	7881

$$D = \begin{vmatrix} 10 & 103 \\ 103 & 1335 \end{vmatrix} = 13350 - 10609 = 2741$$

$$D_1 = \begin{vmatrix} 693 & 103 \\ 7881 & 1335 \end{vmatrix} = 925155 - 811743 = 113412$$

$$D_2 = \begin{vmatrix} 10 & 693 \\ 103 & 7881 \end{vmatrix} = 78810 - 71379 = 7431$$

now

$$a = \frac{D_1}{D} = \frac{113412}{2741} = 41.37$$

$$b = \frac{D_2}{D} = \frac{7431}{2741} = 2.711$$

Regression eqn of y on x

$$y = a + bx = 41.37 + 2.711x$$

now

$b = 2.711$ it means y changes by 2.711 per unit change in x.

REGRESSION :

It is a statistical tool used to determine the nature of relationship that exists among two or more variables.

Dependent Variable: The unknown variable which ~~bot~~ we are going to determine is called dependent variable.

DATE
PAGE No. 35

Independent Variable: The known variable during determination of ~~inde~~ dependent variable is called independent variable.

* Measures of Variation

Total sum of square (TSS) = Sum of square due to regression (SSR) + Sum of square due to error (SSE)

i.e. TSS = SSE + SSR

for the regression model $y = a + bx$, where y is dependent variable and x is independent variable.

$TSS = \Sigma(y-\bar{y})^2 = \Sigma y^2 - n\bar{y}^2$ (Total variation)

$SSE = (y-\hat{y})^2$

Also given by
$SSE = \Sigma y^2 - a\Sigma y - b\Sigma xy$ (Unexpected variation)

$SSR = TSS - SSE$ (Explained variation)

standard error of the Estimate.

$$S_e = \sqrt{\frac{SSE}{n-k-1}}$$

SSE = Sum of square due to error
k = no. of independent variables in regression model
n = no. of observations.

DATE
PAGE No. 36

When $s_e = 0$, there is no variation of observed data around regression line.

Coefficient of Determination

~~Regression~~ Coeff of Determination $(R) = \frac{SSR}{TSS}$

for regression eqn of y on x

$TSS = \Sigma(y-\bar{y})^2 = \Sigma y^2 - n\bar{y}^2$

$SSE = \Sigma(y-\hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$

$SSR = TSS - SSE$

Note: coeff of determination can be obtained by squaring coeff. of correlation i.e.

$R^2 = r^2$

Example:

If $\Sigma x = 16$, $\Sigma y = 42$, $\Sigma xy = 160$, $\Sigma y^2 = 382$ and $\Sigma x^2 = 74$. find R^2.

Here

$TSS = \Sigma(y-\bar{y})^2 = \Sigma y^2 - n\bar{y}^2$

$= 382 - 5\left(\frac{42}{5}\right)^2$

$= 382 - 352.8$

$= 29.2$

DATE
PAGE No. 37

To find a and b

$\Sigma y = na + b\Sigma x$

$42 = 5a + 16b$ —— (i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$160 = 16a + 74b$ —— (ii)

Coeff. of a	Coeff. of b	constant
5	16	42
16	74	160

$$D = \begin{vmatrix} 5 & 16 \\ 16 & 74 \end{vmatrix} = 370 - 256 = 114$$

$$D_1 = \begin{vmatrix} 42 & 16 \\ 160 & 74 \end{vmatrix} = 3108 - 2560 = 548$$

$$D_2 = \begin{vmatrix} 5 & 42 \\ 16 & 160 \end{vmatrix} = 800 - 672 = 128$$

Now

$$a = \frac{D_1}{D} = \frac{548}{114} = 4.802$$

$$b = \frac{D_2}{D} = \frac{128}{114} = 1.122$$

DATE
PAGE No. 38

$\therefore SSE = \Sigma(y-\hat{y})^2$

$= \Sigma y^2 - a\Sigma y - b\Sigma xy$

$= 382 - 4.807 \times 42 - 1.122 \times 160$

$= 382 - 201.894$

$= 0.586$

$SSR = TSS - SSE$

$= 29.2 - 0.586$

$= 28.614$

Now

$R^2 = \dfrac{SSR}{TSS}$

$= \dfrac{28.614}{29.2}$

$= 0.979$

PROBABILITY

DATE
PAGE No. 39

Most of the phenomena exists in our daily life where the results cannot be predicted with certainly are known as unpredictable phenomeena. In this case what we assume is called probability.
It ranges from 0 to 1.
0 for the event which is not possible and 1 is for the event which is sure.

Same Terminologies:

Experiment: It is any procedure which is conducted under certain situation.
eg. tossing a coin

Deterministic Experiment: Experiment in which outcome can be predicted with certainty.
eg - Selecting white ball from a bag contains white balls.

Random Experiments: Repeatedly conducted experiment under essentially homogeneous conditions and it reasults any one of the various possible outcomes but not unique.
eg. You developed an app and tested for lunch the apps may have coding error or it may be correct. So the experiment is testing of app.

Trial and Event: Performing a random experiment is called trial and outcome or combination of outcome are called events.
eg. If a coin is tossed repeatedly, the result is not unique. We get any of the two faces head or tail. So tossing a coin is a trial of a random experiment and getting one of head and tail is event.

Exhaustive Cases (N)
The total number of possible outcomes of a random experiment is called the exhaustive cases for the experiment.

Favourable cases or Events (m)
Total number of outcomes of a random experiment which result in the happening of an event.
eg. If we toss two coins ~~total no of possible outcomes HH, TT,~~ HT, TH are favourable cases for getting 'H' or 'T'.

~~Man~~ **Mutually Exclusive Events or Cases**
→ Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiment. eg. If we toss a coin if 'head' comes ~~to~~ 'T' can't come at the same time.

Sample Space
The set of all possible outcomes of a random experiment is called sample space.

Equally Likely Cases
The outcomes are said to be equally likely or equally probable if none of them is expected to occur in preference to other.
eg. If we toss a coin the probability of getting 'H' and 'T' ~~at~~ is equal.

Impossible event
An event which cannot happen in random experiment is called impossible event eg. getting face numbered 7 on rolling a dice.

Sure Event
An event which is certain to happen in random experiment is called sure event.
eg. Getting 'H' or 'T' on tossing a coin.

Independent Events
Events are said to be independent of each other if happening of any one of them is not affected by and doesn't affect the happening of any one of others. eg. In tossing of a dice ~~getting~~ repeatedly, ~~gett~~ the event of getting '5'

DATE
PAGE No. 42

in first throw is independent of getting '5' in second, third and subsequent throws.

Dependent Events

Events are said to be dependent of each other if happening of any one of them is affected by happening of others.

eg. A card is drawn from a pack of cards and without replacing drawn card if next card is drawn then second draw is affected by first draw.

Counting:

Permutation	Combination
→ It is the arrangement of an object.	It is the selection of an object.
→ It is denoted by ^nP_r.	It is denoted by ^nC_r or $C(n,r)$
→ $P(n,r) = \frac{n!}{(n-r)!}$; $r \le n$	$C(n,r) = \frac{n!}{(n-r)!\,r!}$
→ for repeated case	And
$P(n,r) = \frac{n!}{p!\,q!\,r!}$	$C(n,r) = C(n, n-r)$

DATE
PAGE No. 43

Example of Permutation

Q. In how many ways can 3 persons be seated in 5 chairs?

Soln $n=5, r=3$

$$P(n,r) = \frac{n!}{(n-r)!} = \frac{5!}{(5-3)!} = 60 \text{ Ans}$$

In case of repeated case.

g. 'STATISTICS'

Here

$n=10$

$p =$ no. of s $= 3$

$q =$ no. of t $= 3$

$r =$ no. of i $= 2$

$$\therefore \text{Permutation} = \frac{n!}{p!q!r!} = \frac{10!}{3!3!2!} = 50400 \text{ Ans}$$

Ex. of Combination

Q. In how many ways 2 computers can be selected out of 10 computers in lab.

Here $n=10, r=2$

$$\therefore C(n,r) = \frac{n!}{(n-r)!\,r!} = \frac{10!}{(10-2)!\,2!} = \frac{10!}{8! \times 2!} = 45 \#$$

Imp Question:

Q. What is the probability that a leap year selected at random contains 53 sunday?

Soln

A leap year has 366 days

A year has 52 weeks i.e. $52 \times 7 = 364$ days

Hence, remaing days $= 366 - 364 = 2$

Two remaining days may be [Sun-Mon, Mon-Tue, Tue-Wed, Wed-Thu, Thu-Fri, Fri-Sat, Sat-Sun]

$\therefore$ Total no. of cases $(N) = 7$

favourable number of cases for sunday $(m) = 2$

$\therefore$ The Probability of 53 Sundays $(P) = \frac{m}{N}$

$= 2/7$

* LAWS OF PROBABILITY

For any two events A & B

→ probability of occurance of atleast one event $P(A \text{ or } B) = P(A \cup B) = 1 - P(\overline{A \cup B})$

→ probability of occurance of event A only $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

→ probability of occurance of event B only $P(\bar{A} \cap B) = P(B) - P(A \cap B)$

→ probability of occurance of only one event =

$P(A \cup B) - P(A \cap B)$

Here A and B should be independent of each other.

→ If A & B are not mutually exclusive events then probability of happening to at least one of the event is given by

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Also, for the event A, B, C

$P(A \cup B \cup C) = P(A) + P(B) + P(C) + P(A \cap B \cap C) - P(A \cap B) - P(B \cap C) - P(A \cap C)$

Also

If A & B are mutually exclusive events then

$P(A \cup B) = P(A) + P(B)$

In General form

$$P(A_1 \text{ or } A_2 \text{ or} \ldots \text{ or } A_n) = \sum_{i=1}^{n} P(A_i)$$

DATE
PAGE No. 46

* Mutiplicative law of probability

① The probability of simultaneous happening two independent events A and B is given by product of their individual probabilities.

i.e.

$$P(A \cap B) = P(A \text{ and } B) = P(A)\,P(B)$$

for three independent events A, B & C

$$P(A \cap B \cap C) = P(A \text{ and } B \text{ and } C) = P(A)\,P(B)\,P(C)$$

② The probability of simultaneous happening of two dependent events A & B is given by the product of probability of event A and conditional probability of event B given that A has already happened or the product of probability of event B and conditional probability of event A given that B has already happened.

$$P(A \cap B) = P(A \text{ and } B) = P(A)\,P(B/A)$$
$$= P(B)\,P(A/B)$$

for three dependent events A, B & C

$$P(A \cap B \cap C) = P(A \text{ and } B \text{ and } C) = P(A)\,P(B/A)\,P(C/A \cap B)$$

Conditional Probability

For any two dependent events the probability of an event given that other has already happened is called conditional probability.

The conditional probability of event A given that B has already happened is denoted by P(A/B) and given by

$P(A/B) = P(A \cap B) / P(B) \quad \because P(B) > 0$

The conditional probability of event B given that A has already happened is denoted by ~~P(A~~ P(B/A) and given by

$P(B/A) = P(A \cap B) / P(A) \quad \because P(A) > 0$

Ex.

A card is drawn from a pack of 52 cards. What is the probability that card drawn is king given that card was spade.

Soln

let card is king = K & card is spade = S

then $N = 52$

$M(K) = 4$

$M(S) = 13$

$M(K \cap S) = 1$

$P(K \cap S) = m(K \cap S) / N$

$= \frac{1}{52}$

$P(S) = \frac{m(S)}{N} = 1/4$

PAGE No. 48

$$\therefore P(K|S) = P(K \cap S) / P(S)$$
$$= \frac{1/52}{1/4}$$
$$= \frac{1}{13}$$

* RANDOM VARIABLE

→ It is a rule which assigns one and only real value to each outcome of a random experiment.

Discrete Random variable

→ A random variable is called discrete if it takes integer values.
eg. no. of files in folder.

Continuous Random Variable

→ A random variable is called continuous if it takes all possible values within a certain intervals.
eg. amount of rainfall in rainy season.

DATE
PAGE No. 49

* PROBABILITY DISTRIBUTION

→ It is a formula that specifies the probabilities associated with the random variables.

A probability distribution must satisfy two conditions:

(i) Probability associated with each random variable should be greater than or equal to 0.

(ii) Total of all the probability values associated with random variables should be 1.

(1) NORMAL DISTRIBUTION

Probability mass function of Bionomial Distribution

$$p(x) = C(n, x)\, p^{x} q^{n-x}$$

Here,

n and p are parameters of the distribution

A random variable x following Binomial distribution is denoted by $X \sim B(n, p)$.

Ex

Ten unbiased coins are tossed simultaneously. Find the probability of obtaining

(i) Exactly 6 heads (ii) At least 8 heads (iii) No heads

Soln

let X = number of heads.

PAGE No. 50

probability of obtaining head $(p) = \frac{1}{2}$ then

$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$

It's pmf is $p(x) = C(n, x)\, p^x q^{n-x}$

(i) $P(\text{exactly 6 heads}) = C(10, 6)\left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{10-6}$

$= \frac{210}{1024}$

$= 0.205$

(ii) $P(\text{at least 8 heads}) = P(x \geq 8)$

$= P(x=8) + P(x=9) + P(x=10)$

$= C(10,8)\left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^{10-8} + C(10,9)\left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^{10-9}$

$+ C(10,10)\left(\frac{1}{2}\right)^{10} \cdot \left(\frac{1}{2}\right)^{10-10}$

$= \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024}$

$= 0.054$

(iii) $P(\text{no heads}) = P(x=0) = C(10,0)\left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10-0}$

$= \frac{1}{1024}$

$= 0.0009$

51

* Mean and Variance of Binomial Distribution
let x be a random variable following Bionomial distribution with parameter n and p i.e. $x \sim B(n,p)$ then
Mean $(E(x)) = np$
Variance $V(x) = npq$

Q. The mean and varience of Bionomial distribution are 3 and 2 respectively. Find the number of cases out of 10000 repetitions of the experiments have less than or equal to 2 success.

Soln

Mean $E(x) = 3$
Variance $V(x) = 2$
then $\frac{npq}{np} = \frac{2}{3} \Rightarrow q = \frac{2}{3} \Rightarrow p = 1 - \frac{2}{3} = \frac{1}{3}$

$, np = 3 \Rightarrow n = 9$

Then, $x \sim B(9, \frac{1}{3})$ and $N = 10{,}000$
It's pmf is $p(x) = c(n,x)\, p^x q^{n-x}$

Then the expected Number of cases with number of success less than or equals to 2
$= N \times P(X \le 2)$
$= 10{,}000 \times \{P(0) + P(1) + P(2)\}$
$= 1000 \times \{C(9,0)(\frac{1}{3})^0(\frac{2}{3})^{9-0} + {}^9C_1(\frac{1}{3})^1(\frac{2}{3})^{9-1} + {}^9C_2(\frac{1}{3})^2(\frac{2}{3})^{9-2}$
$= 9586.5$ //

* Fitting a Binomial Distribution

The random variable following Bionomial Distribution can be used to find the expected frequencies of each values of the random variable using relation $N \times p(X=x)$, where N is the total frequency of the observed data.

eg/ Fit the bionomial distribution to the data given below.

x	0	1	2	3	4
f	28	62	46	10	4

soln

let X be the random variable following Binomial distribution with parameter $n=4$ and p

x	f	fx
0	28	~~28~~ 0
1	62	62
2	46	92
3	10	30
4	4	16
	$N = \Sigma f = 150$	$\Sigma fx = 200$

Here

$N = 150$

The mean of observed data $(\bar{x}) = \frac{\Sigma fx}{N} = \frac{200}{150}$

$= 4/3$

Then, $np = 4/3$

$\Rightarrow 4 \times p = 4/3$

$\Rightarrow p = 1/3$

$\Rightarrow q = 1 - 1/3 = 2/3$

pmf $p(x) = C(4, x)(1/3)^{x}(2/3)^{4-x}$

Then expected frequency is obtained by substituting the value of $x = 0, 1, 2, 3 \ldots$

x	$p(x) = C(4,x)(1/3)^{x}(2/3)^{4-x}$	Expected frequency $= N \times p(x)$	
0	0.1979	29.689	30
1	0.395	59.289	59
2	0.296	44.4	44
3	0.0985	14.78	15
4	0.0123	1.845	2
total	1	150	150

PAGE No. 54

* POISSON DISTRIBUTION

The probability mass function of poisson distribution is

$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x = 0, 1, 2, \cdots$

Here,

λ is the parameter of poisson distribution and a random variable and it is ~~doned~~ denoted by $X \sim P(\lambda)$

Ex: Network breakdowns are unexpected rare events that occur every 3 weeks on the average, compute the probability that more than 4 breakdowns during a 21-week period.

Soln

Let x = number of network breakdown

Average no. of breakdown every 3 weeks = 1

Average no. of network breakdown during 21 weeks $= \frac{1}{3} \times 21 \Rightarrow 7$

Hence,

$\lambda = 7$

its pmf $p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$

$p(x>4) = 1 - p(x \leq 4)$

$= 1 - \{p(0) + p(1) + p(2) + p(3) + p(4)\}$

$= 1 - \left\{ \frac{e^{-7} 7^0}{0!} + \frac{e^{-7} 7^1}{1!} + \frac{e^{-7} 7^2}{2!} + \frac{e^{-7} 7^3}{3!} + \frac{e^{-7} 7^4}{4!} \right\}$

$\therefore p(x>4) = 1 - e^{-7} \times 189.791$
$= 0.826$

$\therefore$ The probability of more than 4 networks breakdown in 21 weeks is 0.826

* Mean and variance of Poisson distribution
let x be a random variable following Poisson distribution with parameter
i.e. $x \sim P(\lambda)$ then
Mean $E(x) = \lambda$
variance $V(x) = \sigma^2$

Eg. If a poisson variate x is such that $p(x=1) = 2p(x=2)$. find $p(x=0)$. mean and variance.

Soln

$x \sim P(\lambda)$

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

According to question

$p(x=1) = 2p(x=2)$

or, $\frac{e^{-\lambda}\lambda^1}{1!} = 2\frac{e^{-\lambda}\lambda^2}{2!} \Rightarrow \lambda = 1$

$\therefore p(x=0) = \frac{e^{-1}1^0}{0!} = 0.367$

DATE
PAGE No. 56

Mean $E(x) = \lambda = 1$
Variance $V(x) = \lambda = 1$

* Fitting of poisson Distribution

→ The random variable following Poisson distribution can be used to find the expected frequencies of each values of the random variable using relation $N \times p(X=x)$, where N is the total frequency of the observed data.

Example

x	0	1	2	3	4
f	123	59	14	3	1

Sol,

X be the random variable following Poisson variable with parameter λ.
Then the probability mass function of random variable X is

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, \ldots$$

for poisson distribution

$$\text{mean}(\bar{X}) = \frac{\Sigma fx}{N} = \frac{100}{200} = 0.5$$

X	0	1	2	3	4	Total	
f	123	59	14	3	1	200	$= N$
fx	0	59	28	9	4	100	$= \Sigma fx$

PAGE No. 57

∴ pmf $p(x) = \frac{e^{-0.5} (0.5)^{x}}{x!}$

The expected frequency is given by

$f(x) = N \times p(x)$ where $x = 0, 1, 2, \ldots$

To calculate the expected frequency.

X	$p(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$	Expected frequency $= N\,p(x)$ $= 200 \times p(x)$
0	0.6065	121.3 ≈ 121
1	0.30325	60.65 ≈ 61
2	0.0758125	15.325 ≈ 15
3	0.00505417	2.552 ≈ 3
4	0.00015794	0.32 ≈ 0
Total		200

PAGE No. 38

* CONTINUOUS DISTRIBUTION
-> Probability distribution of continuous random variable having probability density function is called continuous distribution.
its types
-> Normal Distribution
-> Exponential Distribution etc.

(a) Normal Distribution
-> It is one of the most important continuous theoritical distributions in statistics. Most of the data related to economic and business statistics or even in social and physical sciences conform to this distribution.

(1) Probability density function of normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where
$-\infty < x < \infty$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$
with parameters μ & σ^2. It is written as $X \sim N(\mu, \sigma^2)$

PAGE No. 59

* Standard Normal Distribution

=) A continuous random variable $z = \frac{x-\mu}{\sigma}$ is said to ~~be~~ follow standard normal distribution if its probability density function is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2} \qquad -\infty < z < \infty$$

Note

Some properties of Normal distribution

1. Mean = Median = Mode = μ
2. Max. amplitude of curve is $\frac{1}{\sqrt{2\pi}}$ & occurs at $x = \mu$.
3. $QD = \frac{2}{3}\sigma$
4. $MD = \sqrt{\frac{2}{\pi}}\,\sigma \simeq \sigma$.
5. $QD : MD : SD = 10 : 12 : 15$
6. The curve is asymptotic to x axis
7. The area under normal curve is unity

etc.

DATE
PAGE No. 60

6) Number of files in folder of a computer programmer is normally distributed with average number of files in a folder is 78 with standard deviation of 10. Find the probability that number of files in a folder lies between 61 and 94.

Soln

let X = number of files in folder.

$X \sim N(\mu, \sigma^2)$

$X \sim N(78, 100)$

let, $z = \frac{x-\mu}{\sigma}$ be the standard normal variate

$p(61 < x < 94) = ?$

when, $x = 61$, $z = \frac{61-78}{10} = -1.7$

when $x = 94$, $z = \frac{94-78}{10} = 1.6$

Then

$p(61 < x < 94) = p(-1.7 < z < 1.6)$

$= p(-1.7 < z < 0) + p(0 < z < 1.6)$

$= p(0 < z < 1.7) + p(0 < z < 1.6)$

$= 0.4554 + 0.4452$

$= 0.9$

(Note: values are taken from table)

DATE
PAGE No. 61

* Fitting of Normal Distribution

In normal distribution pdf if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Q. 1

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
frequency	3	15	37	52	61	30	14	8

Soln

class	Frequency	Mid value(x)	fx	fx^2
10-20	3	15	45	675
20-30	15	25	375	9375
30-40	37	35	1295	69825
40-50	52	45	2340	105300
50-60	61	55	3355	184525
60-70	30	65	1950	126750
70-80	14	75	1050	78750
80-90	8	85	680 ~~608~~	57800
	N = 220		$\Sigma fx = 11090$	$\Sigma fx^2 = 633000$

Ans
$$\mu = \bar{x} = \frac{\Sigma fx}{N} = \frac{11090}{220} = 50.409$$

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2} = \sqrt{\frac{633000}{220} - \left(\frac{11090}{220}\right)^2} = 18.335$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{18.335\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50.409}{18.335}\right)^2}$$

Page No. 62

class ~~Below 10~~	lower limit X_L	$z_i = \frac{x_i - \mu}{\sigma}$	$\varphi(z) = \frac{1}{2\pi}\int_{-\infty}^{z} e^{-\frac{z^2}{2}}$	$\Delta\varphi(z) = \varphi(z_{i+1}) - \varphi(z_i)$	expected frequency $= N\Delta\varphi(z)$
Below 10	$-\infty$	$-\infty$	0	0.0139	3.058 ≈ 3
10-20	10	-2.203	0.0139	0.0356	7.83 ≈ 8
20-30	20	-1.658	0.0495	0.114	25.08 ≈ 25
30-40	30	-1.113	0.1635	0.1242	27.32 ≈ 27
40-50	40	-0.567	0.2877	0.2003	44.06 ≈ 44
50-60	50	-0.022	0.488	0.2139	47.05 ≈ 47
60-70	60	0.523	0.7019	0.11535	33.77 = 34
70-80	70	1.068	0.8554	0.0909	19.99 ≈ 20
80-90	80	1.613	0.9463	0.0379	8.33 ≈ 8
90 & over	90	2.159	0.9842		

CH-5 : SAMPLE SURVEY

63

Population:

→ A group or aggregate of all objects/units/ individuals/items etc. related to the topic of the study is called a population.

Finite Population:

A population is said to be a finite population if the number of units in the population is countable in certain time.

eg. Total no. of students in TU

Infinite Population:

⇒ A population is said to be infinite if it is not countable in certain time.

eg. Total no. of stars in the sky.

Sample

Some units selected from population which gives information about the population is called sample.

Random Sample

On selecting units from population each time unit selected has equal chance to appear in the sample is called random sample

PAGE NO. 64

Non-random sample
On selecting units from population each time unit selected has no equal chance to appear in the sample is called non-random sample

Census survey
A study conducted on all units of population is called census.
eg. Population census survey

Sample survey.
The survey carried out by selecting representative sample of the study population is known as sample survey.
eg. family planning survey

Sampling.
When one by one study of all units of a population is not possible due to some factors like time, cost, manpower etc we take a small representative part from the population of the study and this small representative part selected for the study from population is called sample

eg.
selection of word during speech recognition with maximum probability

Factors Affecting sample size
→ Population parameter to be estimated
→ Cost of sampling
→ The information known before sampling
→ Variability of the population
→ Practicality of the data collection (how tough to collect)
→ Precision required for the research

Advantages of sampling
→ Reduce Cost
→ Greater speed
→ Greater Scope
→ Greater Accuracy

CH-6

SAMPLE SURVEY METHODS

DATE
PAGE No. 66

The sampling method or technique usually depends upon the nature of the population under investigation. The sampling techniques that are commonly used can be classified as:

1. Random or Probability Sampling Technique
2. Non-random or Non-Probability Sampling Technique

(i) Random or Probability Sampling Technique

→ It is defined as the method of sampling techniques in which each unit of the population has some fixed ~~set~~ probability of being selected in the sample.

Random sampling techniques

- → Sample Random Sampling
- → Stratified Random Sampling
- → Systematic Random Sampling
- → Cluster Random Sampling
- → Multi-stage Random sampling

(ii) Non-random or Non-Probability Sampling Techniques

→ It is defined as the method of sampling technique in which each unit in a sample is

PAGE No. 67

selected on the basis of personal judgment. These are several non-random sampling methods for selecting samples from a population

- → Judgment or Purprosive Sampling
- → Convenience Sampling
- → Quota Sampling
- → Snowball sampling

CH-7 DESIGN OF EXPERIMENT

DATE
PAGE No. 68

ANALYSIS OF VARIENCE (ANOVA)

The ~~total~~ systemetic proress for achieving the variation is called the analysis of variance.

F statistic and its distribution:

F statistic is defined as the ratio of two independent chi square variates divided by their respective degrees of freedom.

let $X \sim \chi^2_m$ and $Y = \chi^2_n$ the F statistic is given by

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}}$$

it follows Snedecor's F distribution with (m,n) degrees of freedom. Its probability density function is given by

$$f(F) = \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}-1}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \cdot \frac{F^{\frac{m}{2}-1}}{\left[1+\frac{m}{n}F\right]^{\frac{m+n}{2}}} ; 0 \le F < \infty$$

Also,

$$\boxed{F = \frac{S_1^2}{S_2^2}} , \quad S_1^2 > S_2^2$$

Here S_1 & S_2 are population variance.

DATE
PAGE No. 69

Q. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Find F.

Soln

$n_1 = 8$, $n_2 = 10$, $\Sigma(x-\bar{x})^2 = 94.5$, $\Sigma(y-\bar{y})^2 = 101.7$, $\alpha = 5\%$.

$$S_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1 - 1} = \frac{94}{7} = 13.5$$

$$S_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2 - 1} = \frac{101.7}{9} = 11.3$$

$$\therefore F = \frac{S_1^2}{S_2^2} = \frac{13.5}{11.3} = 1.19$$

* ANALYSIS OF ONE-WAY CLASSIFIED DATA

→ The total variation present in any set of numerical data is classified according to a factor is called one-way classification.

The main objective of the ANOVA is to examine if there is significant difference between the class means.

DATE
PAGE No. 70

Degree of freedom

(a) Degree of freedom for total sum of square = n-1
here one degree of freedom is lost ~~dot~~ due to linear constraint $\sum_{i=1}^{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}}_{..}) = 0$

(b) Degree of freedom for sum of square due to class = k-1

(c) Degree of freedom for sum of square due to error = n-k

(d) Degree of freedom for TSS = Degree of freedom for SSC + Degree of freedom for SSE
i.e. ~~n-1~~ = (k-1)+(n-k)
= n-1

Mean sum of Square (MSS)

→ The sum of square divided by the corresponding degree of freedom gives the respective mean sum of square or variance.

~~Mean~~ Mss due to class (MSC) = SSC / k-1
Mss due to error (MSE) = SSE / n-k

Test statistic:

If H_0 is true, $E(MSC) = E(MSE)$

$F = \frac{MSC}{MSE}$ gives the value 1 for null hypothesis.

If H_1 is true, $E(MSC) > E(MSE)$

$F = \frac{MSC}{MSE}$ gives value greater than 1 for alternative hypothesis

Critical value

- let α% be the level of significance, then critical value at α% level of significance for $(k-1, n-k)$ degree of freedom is $F_\alpha(k-1, n-k)$

Decision

Reject H_0 at α% level of significance if $F > F_\alpha(k-1, n-k)$, accept otherwise.

Analysis of variance (ANOVA) table for one-way classified data:

Source of variation	d.f	S.S	MS	F_{cal}	F_{tab}
class	$k-1$	SSC	$MSC = \frac{SSC}{k-1}$	$F_c = \frac{MSC}{MSE}$	$F_\alpha(k-1, n-k)$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	TSS			

DATE
PAGE No. 72

Relation to calculate TSS, SSC and SSE

We have

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{n}$$, G is grand total of observation

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - CF$$ (G^2/n is called correction factor)

Also

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} \frac{(T_i)^2}{n_i}$$

SSC = TSS - SSE (Since, TSS = SSC + SSE)

$$= \sum_{i=1}^{k} \frac{(T_i)^2}{n_i} - CF$$

Example:

The weight in grams of a number of copper wires each of length 1m were obtained. These are shown below classified according to die from which they come.

DATE
PAGE No. 73

Die No

I	II	III	IV	V
2	0	4	3	2
4	7	1	1	4
8	5	3	5	2
7	6	0	3	5
4		5	4	
9		2		

Are the weights of copper wire for different die are unequal?

								T_i	T_i^2
	I	2	4	8	7	4	9	34	1156
Die	II	0	7	5	6			18	324
No	III	4	1	3	0	5	2	15	225
	IV	3	1	5	3	4		16	256
	V	2	4	2	5			13	169
						Total		96	2130

Here

$n_1 = 6, n_2 = 4, n_3 = 6, n_4 = 5, n_5 = 4, \Sigma T_i = 96, \Sigma T_i^2 = 2130,$

$\Sigma n_i = 25$

Problem to test.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1: At least one μ_i is different, $i = 1, 2, 3, 4, 5$

now

$G = \Sigma\Sigma y_{ij} = \Sigma T_i = 96,$

$\frac{G^2}{N} = \frac{96^2}{25} = 368.64,$

DATE
PAGE No. 74

$$\sum\sum y_{ij}^2 = 2^2+4^2+8^2+7^2+4^2+9^2+0^2+7^2+5^2+6^2+4^2+1^2+3^2+0^2+5^2+2^2+3^2+1^2+5^2+3^2+4^2+2^2+4^2+2^2+5^2$$
$$= 504$$

$$\sum \frac{T_i^2}{n_i} = \frac{1156}{6} + \frac{324}{4} + \frac{225}{6} + \frac{256}{5} + \frac{169}{4}$$
$$= 404.69$$

$$TSS = \sum\sum y_{ij}^2 - G^2/N = 504 - 368.64 = 135.36$$

$$SSR = \sum \frac{T_i^2}{n_i} - G^2/N = 404.61 - 368.64 = 35.97$$

$$SSE = TSS - SSR$$
$$= 135.36 - 35.97$$
$$= 99.38$$

ANOVA table.

Source of variation	Degree of freedom	Sum of Squares	Mean Squares	F_{cal}	F_{tab}
Row (Die No)	4	SSR = 35.99	MSR = 8.99	$F_R = 1.81$	$f_{0.05}(4,20) = 2.87$
Error	20	SSE = 99.38	MSE = 4.96		
Total	24	TSS = 135.36			

Test statistic

$$f_R = \frac{MSR}{MSE} = 1.81$$

PAGE No. 75

Critical value
Let 5% be the level of significance. Critical or tabulated value of F at 5% level of significance with 4 and 20 degree of freedom is $F_{0.05(4,20)} = 2.87$

Decision:
$F_R = 1.81 < F_{0.05(4,20)} = 2.87$, accept H_0 at 5% level of significance.

Conclusion:
The weights of copper wire for different die are equal.

DATE
PAGE No. 76

* ANALYSIS OF TWO-WAY CLASSIFIED DATA

⇒ The total variation present in any set of numerical data is classified according to two factors is called two-way classification.

Problem to set test:

H_{0R} : $\mu_{1.} = \mu_{2.} = \mu_{3.} = \cdots = \mu_{m.}$ (Population means of all m rows are equal)

H_{1R} : At least one $\mu_{i.}$ is different
$i = 1, 2, 3, \ldots, m$
(Population mean of at least one ~~row~~ of m row is different)

H_{0C} : $\mu_{.1} = \mu_{.2} = \cdots = \mu_{.n}$
(population means of all n columns are equal)

H_{1C} : At least one $\mu_{.j}$ is different. $j = 1, 2, 3, \ldots, n$)
(population mean of at least one column of n column is different)

~~TSS~~ Total sum of square (TSS) = Sum of square due to row (SSR) + sum of square due to column (SSC) + sum of square due to error (SSE)

$\therefore$ TSS = SSR + SSC + SSE

DATE
PAGE No. 78

Degree of freedom

1. Degree of freedom for TSS $= N-1$
2. Degree of freedom for SSR $= m-1$
3. Degree of freedom for SSC $= n-1$
4. Degree of freedom for SSE $= N-1-(m-1)-(n-1)$

$$= mn - \cancel{1} - m + \cancel{1} - n + 1 \quad (N = mn)$$
$$= mn - m - n - 1$$
$$= m(n-1) - 1(n-1)$$
$$= (m-1)(n-1)$$

Mean Sum of Square (MSS)

The sum of square divided by the corresponding degree of freedom gives the respective mean sum of square or variance.

MSS for Row (MSR) $= SSR / m-1$

MSS for Column (MSC) $= SSC / n-1$

MSS for Error (MSE) $= SSE / (m-1)(n-1)$

Test statistic

If H_{0R} is true, $E(MSR) = E(MSE)$

$F_R = \frac{MSR}{MSE}$ gives the value 1 for null hypothesis.

If H_{1R} is true, $E(MSR) > E(MSE)$

$F_R = \frac{MSR}{MSE}$ gives value greater than 1 for alternative hypothesis.

PAGE No. 79

Similar for column as well.

Critical value:
let $\alpha\%$ be the level of significance, then critical value at $\alpha\%$ level of significance for $\{m-1, (m-1)(n-1)\}$ degree of fredom is

$F_{\alpha}\{m-1, (m-1)(n-1)\}$

And critical value at $\alpha\%$ level ~~is~~ of significance for $\{n-1, (m-1)(n-1)\}$ degree of freedom is

$F_{\alpha}\{n-1, (m-1)(n-1)\}$

Decision:
Reject H_{0R} at $\alpha\%$ level of significance if $F_R > F_{\alpha}(m-1, (m-1)(n-1))$, accept otherwise.

Reject H_{0C} at $\alpha\%$ level of significance if $F_C > F_{\alpha}\,\cancel{(n-1)}\,(n-1, (m-1)(n-1))$, accept otherwise.

PAGE No. 80

ANOVA TABLE

Source of variation	D.f	SS	MS	F_{cal}	F_{tab}
Rows	$m-1$	SSR	$MSR = \frac{SSR}{m-1}$	$F_R = \frac{MSR}{MSE}$	$F_\alpha(m-1, (m-1)(n-1))$
Columns	$n-1$	SSC	$MSC = \frac{SSC}{n-1}$	$F_C = \frac{MSC}{MSE}$	$F_\alpha(n-1, (m-1)(n-1))$
Error	$(m-1)(n-1)$	SSE	$MSE = \frac{SSE}{(m-1)(n-1)}$		
Total	$mn-1$	TSS			

Relation to calculate TSS, SSR, SSC and SSE

$$TSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{mn},$$ (G is the grand total of the observations)

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2 - CF$$ (G^2/mn is called correction factor)

Also

$$SSR = n \sum_{i=1}^{m} (y_{ij} - \bar{y}_{..})^2$$

$$= \frac{1}{n} \sum_{i=1}^{m} T_i^2 - \frac{G^2}{mn}$$

$$= \frac{1}{n} \sum_{i=1}^{m} q_i^2 - CF$$

81

Again,

$$SSC = m\sum_{j=1}^{n} (\bar{y}_{.j} - \bar{y}_{..})^2$$

$$= \frac{1}{m}\sum_{j=1}^{n} T_{.j}^2 - \frac{G^2}{mn}$$

$$= \frac{1}{m}\sum_{j=1}^{n} T_{.j}^2 - CF$$

$SSE = TSS - SSR - SSC$ (since $TSS = SSR + SSC + SSE$)

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{m} T_{i.}^2 - \frac{1}{m}\sum_{j=1}^{n} T_{.j}^2 + CF$$

Example

An experiment was conducted to determine the effects of different dates of planting and different methods of planting on the yield of sugarcane.

		Date of planting			
		Oct	Nov	Feb	Mar
Methods of planting	I	7	4	5	2
	II	10	5	5	3
	III	8	4	5	2

Does the method of planting effect mean yield and date of planting affect mean yield.

Solution:

Problem to test

$H_{0R}: \mu_{I.} = \mu_{II.} = \mu_{III}$

H_{1R}: At least one $\mu_{i.}$ is different, i= I,II,III

$H_{0R}: \mu_{.1} = \mu_{.2} = \mu_{.3}$

H_{1R}: At least one $\mu_{.j}$ is different, j= 1(Oct), 2(Nov), 3(Feb), 4(Mar)

Here

	Date of planting					$T_{i.}$	$T_{i.}^2$
Method of planting		1	2	3	4		
	I	7	4	5	2	18	324
	II	10	5	5	3	23	529
	III	8	4	5	2	19	361
	$T_{.j}$	25	13	15	7	G= 60	$\Sigma T_{i.}^2$=1214
	$T_{.j}^2$	625	169	225	49	$\Sigma T_{.j}^2 = 1068$	

$N = m \times n = 3 \times 4 = 12$, $G = \Sigma T_{i.} = \Sigma T_{.j} = 60$, $G^2/N = \frac{60^2}{12} = 300$

$TSS = \Sigma\Sigma x_{ij}^2 - G^2/N = (7^2 + 4^2 + 5^2 + 2^2 + 10^2 + 5^2 + 5^2 + 3^2 + 8^2 + 4^2 + 5^2 + 2^2) - 300 = 362 - 300 = 62$

$SSR = \Sigma\frac{T_{i.}^2}{n} - \frac{G^2}{N} = \frac{1214}{4} - 300 = 303.5 - 300 = 3.5$

$SSC = \Sigma\frac{T_{.j}^2}{m} - \frac{G^2}{N} = \frac{1068}{3} - 300 = 356 - 300 = 56$

$SSE = TSS - SSR - SSC = 62 - 3.5 - 56 = 2.5$

ANOVA table

Source of variation	Degree of freedom	Sum of squares	Mean squares	$F_{calculated}$	$F_{tabulated}$
Row (Method of planting)	2	SSR = 3.5	MSR =1.75	F_R = 4.206	$F_{0.05(2,6)}$ =5.14
Column (Date of planting)	3	SSC = 56	MSC =18.666	F_C= 44.87	$F_{0.05(3,6)}$ = 4.76
Error	6	SSE = 2.5	MSE = 0.416		
Total	11	TSS =62			

Test statistic

$F_R = \frac{MSR}{MSE} = 4.206$, $F_C = \frac{MSC}{MSE} = 44.87$

Critical value

Let 5% be the level of significance, then critical value at 5% level of significance with 2 and 6 degree of freedom is $F_{0.05\,(2,6)} = 5.14$ and critical value at 5% level of significance with 3 and 6 degree of freedom is $F_{0.05\,(3,6)} = 4.76$

Decision

$F_R = 4.206 < F_{0.05\,(2,6)} = 5.14$, accept H_{0R} at 5% level of significance.

$F_C = 44.87 > F_{0.05\,(3,6)} = 4.76$, reject H_{0C} at 5% level of significance.

Conclusion

Methods of planting do not affect the mean yield of sugarcane but dates of planting affect the mean yields of sugarcane.