Statistical Methods Using R

Unit - 01

Measures of Central Tendency

The central tendency is stated as the statistical measure that represents the single value of entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

Types: Mean, Mode, Median

- Mean Value: The average value of dataset.

→ $\bar{x} = \frac{\text{Sum of all observation}}{\text{no. of observations}}$ (Individual) by Observations

→ $\bar{x} = \frac{\Sigma f_i x_i}{N}$ (Discrete) → observations + frequency

→ $\bar{x} = \frac{\sum_{i=0}^{n} f_i x_i}{N}$ where x_i is middle value of $l_i - h_i$ (Continous) Interval obs + freq.

Que. Find the mean height of the students from the following frequency table

Height (inch)	No. of St. (f_i)	$x_i f_i$
64	1	64
65	6	390
66	10	660
67	22	1474
68	21	1428
69	17	1173
70	14	980
71	5	355
72	3	216
73	1	73
	100	6813

Mean $= \frac{\Sigma f_i x_i}{N}$

$\bar{x} = \frac{6813}{100}$

$\bar{x} = 68.13$ inch

Que. Calculate the airthmatic mean of marks from following table distribution

Marks	No. of St.	x_i	$x_i f_i$
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
	100		2800

$\bar{X} = \frac{\Sigma f_i x_i}{N}$

$\bar{X} = \frac{2800}{100}$

$\bar{X} = 28$

Que. Mean of 1st five odd integers

Sol

$x \rightarrow 1, 3, 5, 7, 9$ Sum $= 1+3+5+7+9 = 25$

$N = 5$

$\therefore$ Mean $= \frac{\text{Sum}}{N} = \frac{25}{5} = 5$

Que. Class consist of 4 boys & 3 girls having avg. marks 20 & 30 respectively calculate avg. marks of class

Sol

Avg. of Boys = 20 No. of boys = 4

Avg. of girls = 30 No. of girls = 3

$\therefore$ Total marks of boys = Avg. x No of boys $= 20 \times 4 = 80$

Total Marks of girls = Avg. x No. of girls $= 30 \times 3 = 90$

Total Marks of Class = Marks B + Marks G = 80 + 90 = 170

Total students in class = No. of B + No. of G = 4 + 3 = 7

Avg. Marks of Class = $\frac{\text{Total Marks}}{\text{Students}} = \frac{170}{7} = 24.28$

(Ascending order)

- Median Value : Middle value of given data

→ Odd = $\left(\frac{N+1}{2}\right)^{th}$ term

→ Even : $\frac{\left(\frac{N}{2}\right)^{th} + \left(\frac{N}{2}+1\right)^{th}}{2}$

→ Continous : $l + \frac{\frac{N}{2} - f_0}{f} \times i$

Find the median for following distribution

x	f	Cumlative freq
0-10	22	22
10-20	38	60
20-30	46	106
30-40	35	141
40-50	20	161
	161	

here $N = 161$

$\therefore \frac{N}{2} = \frac{161}{2} = 80.5$

$\therefore$ cf just greater than $\frac{N}{2}$ is 106

thus Median Class is 20-30

Thus $N = 161$ $f_0 = 60$ (previous cf)

$f = 46$ (fr. of Med. class)

$l = 20$ (lower limit of Med. class)

$\therefore$ Median $= l + \frac{\frac{N}{2} - f_0}{f} \times i$

$= 20 + \frac{\frac{161}{2} - 60}{46} \times 10$

$= 20 + \frac{20.5}{46} \times 10$

$= 20 + 4.45 = 24.23$

Qw. Find the median of data

Marks	student	cf
0-10	2	2
10-20	18	20
20-30	30	50
30-40	45	95
40-50	35	130
50-60	22	152
60-70	6	158
70-80	3	161

$\frac{N}{2} = \frac{161}{2} = 80.5$

$\therefore$ Median $= 30 + \frac{80.5 - 50}{45} \times 10$

$= 30 + \frac{305}{45}$

$= 30 + 6.77$

Median $= 36.77$

Camlin Page
Date / /

Mode Value : Most Occurring Value in data

Most frequent

conti. Mode $= l + \frac{f - f_1}{2f - f_{-1} - f_1} \times i$

Find the Mode of following distribution

CI	freq.
0-7	19
7-14	25
14-21	36
21-28	72
28-35	51
35-42	43
42-49	28

$l = 21 \quad i = 7 \quad f = 72 \quad f_1 = 51 \quad f_{-1} = 36$

$$\text{Mode} = l + \frac{f - f_1}{2f - f_{-1} - f_1} \times i$$

$$= 21 + \frac{72 - 36}{2 \times 72 - 36 - 51} \times 7$$

$$= 21 + \frac{36}{57} \times 7$$

$$= 21 + 4.42 = 25.42$$

Measure of Dispersion:

Dispersion is the state of getting spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of data.

Types : Range, Standard Deviation, Variance, Quartile deviation

Range - It is defined as the difference b/w largest value & smallest value in distribution

Range = H - L Range co-efficient = $\frac{H-L}{H+L}$

Quartiles - It divides the entire set into four equal parts. There are three quartiles Q_1, Q_2, Q_3. where Q_1 is median of 1st half. Q_2 is median of complete data. Q_3 is median of 2nd half of data.

Lowest Value Q_1 Q_2 Q_3 Highest Value

Inter-quartile Range = $Q_3 - Q_1$

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

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

$Q_1 = l + \frac{\frac{N}{4} - f_0}{f} \times i$

$Q_3 = l + \frac{\frac{3N}{4} - f_0}{f} \times i$

here class is decided on basis of $\frac{N}{4}$ & $\frac{3N}{4}$ value ∴ cf which is just greater than $\frac{N}{4}$ & $\frac{3N}{4}$

Standard Deviation :- It is measured used in descriptive statistics to understand how the data points in a set are spread out from the average value.

$$\sigma = SD = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n}} \qquad \sqrt{\frac{\Sigma f_i (x_i - \mu)^2}{n}}$$

Variance :-

for individual distribution

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

$\bar{x}$ = Mean

N = no. of obsv

for discrete distribution

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{N}$$

$\bar{x}$ = Mean = $\frac{\Sigma f_i x_i}{N}$

$N = \Sigma f_i$ = Sum of freq.

for continous distribution

here $x_i = \frac{h_i + l_i}{2}$ $\bar{x}$ = Mean = $\frac{\Sigma f_i x_i}{N}$

$N = \Sigma f_i$ = Sum of freq.

Camlin Page
Date / /

Find the quartile deviation.

CI	Freq	CF
0-10	28	28
10-20	31	59
20-30	18	77
30-40	21	98
40-50	36	134
50-60	49	183
60-70	20	203
	$\Sigma f_i = 203$	

$\frac{N}{4} = \frac{203}{4} = 50.75$

$l = 10 \quad i = 10 \quad f_0 = 28 \quad f = 31$

$Q_1 = 10 + \frac{50.75 - 28}{31} \times 10$

$Q_1 = 17.33$

for Q_3 $3\frac{N}{4}^{th} = 152.25$

$l = 50 \quad i = 10 \quad f_0 = 134 \quad f = 49$

$Q_3 = 50 + \frac{152.25 - 134}{49} \times 10 = 53.72$

$Q_d = \frac{Q_3 - Q_1}{2} = \frac{38.39}{2} = 18.2$

Find Quartile deviation of following distribution

0-10	5	5
10-20	3	8
20-30	4	12
30-40	3	15
40-50	3	18
50-60	4	22
60-70	7	29
70-80	9	38
80-90	7	45
90-100	8	53

$\frac{N}{4} = \frac{53}{4} = 13.25 \qquad \frac{3N}{4} = \frac{3 \times 53}{4} = 39.75$

$Q_1 = l + \frac{\frac{N}{4} - f_0}{f} \times i \qquad Q_3 = l + \frac{\frac{3N}{4} - f_0}{f} \times i$

$Q_1 = 30 + \frac{13.25 - 12}{3} \times 10 \qquad Q_3 = 80 + \frac{39.75 - 38}{7} \times 10$

$Q_1 = 30 + 4.16 = 34.16 \qquad Q_3 = 82.5$

$Q_d = \frac{82.5 - 34.16}{2} = 24.17$

23/01/24

Qu Calculate the variance & standard deviation

x	2	4	6	8	10	
f	3	5	9	5	3	25
$x_i f_i$	6	20	54	40	30	
$x_i - \bar{x}$	-4	-2	0	2	4	
$(x_i - \bar{x})^2$	16	4	0	4	16	
$f_i(x_i - \bar{x})^2$	48	20	0	20	48	

Mean $= \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{150}{25} = 6$

Var $= \sqrt{\frac{136}{25}} =$

Sd $= \sqrt{\text{Var}} = \sqrt{\frac{136}{25}} = 2.33$

Var $= \frac{136}{25} = 5.44$

Qu Calculate the standard deviation for following.

9	12	-3
0	15	0
9	18	3
25	10	-5
1	14	-1
36	21	6
1	14	-1
1	16	1

$\frac{12+15+18+10+14+21+14+16}{8}$

Mean $= \frac{120}{8} = 15$

Var $= \frac{82}{8} = 10.25$

Sd $= \sqrt{10.25} = 3.20$

Qu Calculate the sd of following

x	f	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
14	4	56	-2.77	7.67	59.03
15	5	75	-1.77	3.13	15.65
16	10	160	-0.77	0.59	5.9
17	8	136	0.23	0.05	0.4
18	6	108	1.23	1.51	9.06
19	4	76	2.23	4.97	19.88
20	3	60	3.23	10.43	31.29
	40	671			

Mean $= \frac{671}{40} = 16.77$

Var $= \frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i} = \frac{141.21}{40}$

Var $= 3.53$

Sd $= 1.87$

Calculate S.d.

Mark	freq	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$	x_i	
0-10	4	-18.12	328.33	1313.32	5	20
10-20	8	-8.12	65.93	527.44	15	120
20-30	12	1.88	3.53	42.36	25	300
30-40	6	11.88	141.13	846.78	35	210
40-50	2	21.88	478.73	957.46	45	90
	32			~~3754.56~~ 3687.6		740

$$\text{Mean} = \frac{740}{32} = 23.12$$

$$\text{Var} = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{3687.6}{32} = 115.23$$

$$\text{Sd} = \sqrt{\text{Var}} = 10.73$$

A garden contains 39 plants. The following plants were chosen at random & height were recorded in cm. 38, 51, 46, 79, 57. Calculate their height sd.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
38	-16.2	262.44
51	-3.2	10.24
46	-8.2	67.24
79	24.8	615.04
57	2.8	7.84
271		962.8

$$\text{Mean} = \frac{271}{5} = 54.2$$

$$\text{Var} = \frac{962.8}{5} = 192.56$$

$$\text{Sd} = \sqrt{192.56} = 13.87$$

Sample

$$\text{Var} = \frac{962.8}{4} = 240.7$$

$$\text{Sd} = \sqrt{240.7} = 15.5$$

Find the Mean, Median, Mode of

Qu. Find Mean Mode Median

C.I	f	cf
0-20	6	6
20-40	8	14
40-60	10	24
60-80	12	36
80-100	6	42
100-120	5	47
120-140	3	50
	50	

$\frac{N}{2} = \frac{50}{2} = 25^{th}$

Median $= l + \frac{\frac{N}{2} - f_0}{f} \times i$

Median $= 60 + \frac{25-24}{12} \times 20$

Median $= 60 + 1.66 = 61.66$

Mode

Mode $= l + \frac{f - f_0}{2f - f_0 - f_1} \times i$

$= 60 + \frac{12-10}{24-10-6} \times 20 = 60 + \frac{40}{8}$

Mode = 65

CI	f	x_i	$f_i x_i$
0-20	6	10	60
20-40	8	30	240
40-60	10	50	500
60-80	12	70	840
80-100	6	90	540
100-120	5	110	550
120-140	3	130	390
			3120

Mean $= \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{3120}{50} = 62.4$

Camlin Page
Date / /

Probability

Theory of probability is the branch of mathematics which has been developed to deal with situations involving uncertainity.

The probability of an event 'E' can be written as

$$P(E) = \frac{\text{No. of favarouble outcome of event E}}{\text{No. of all possible outcome of event.}}$$

eg Tossing a coin $P(H) : P(T) = \frac{1}{2}$

Laws of Probability –

1) If A & B are two events (subsets of sample space) & are not disjoin then probability

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

eg An urn contain 10 Black & 10 white balls. Find the probability of drawing two balls of same color.

Sol Black = 10 white = 10 Total = 20

$$P = \frac{^{10}C_2}{^{20}C_2} + \frac{^{10}C_2}{^{20}C_2} = \frac{9}{2\times19} + \frac{9}{2\times19} = \frac{9}{19}$$

2) Conditional Probability

It is a probability measure that describe the likely hood of an event occurring, given that another event has already occured.

① $P(A/B) = \frac{P(A \cap B)}{P(B)}$

② $P(B/A) = \frac{P(A \cap B)}{P(A)}$

Q A pair coin is tossed twice, such that E = event of having both head & tail. F = event of having at most one tail. Find the probability of E & F and E/F.

Sol $S = \{HH, HT, TH, TT\}$ $E = \{HT, TH\}$ $F = \{HH, HT, TH\}$

$P(E) = \frac{2}{4} = \frac{1}{2}$ $P(F) = \frac{3}{4}$ $P(E \cap F) = \frac{2}{4}$

$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{2/4}{3/4}$

$P(E/F) = \frac{2}{3}$

Q In a class 40% of students like Maths, 25% like physics, & 15% like both the subjects. One student select at random. Find probability that he likes physics if it is known that he like maths.

Sol $P(M) = \frac{40}{100} = \frac{4}{10}$ $P(P) = \frac{25}{100} = \frac{1}{4}$

$P(M \cap P) = \frac{15}{100}$

$\therefore P(P/M) = \frac{15/100}{40/100} = \frac{3}{8} = 0.33$

Q4. A box contains 6 Red, 4 white & 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour. (0.512)

Sol 6R, 4W, 5B

2, 1, 1

1, 2, 1

1, 1, 2

$\frac{{}^6C_2 \cdot {}^4C_1 \cdot {}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1 \cdot {}^4C_2 \cdot {}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1 \cdot {}^4C_1 \cdot {}^5C_2}{{}^{15}C_4}$

$= \frac{6 \times 5 \times 4 \times 5}{2 \times 15 \times 14 \times 13 \times 12} + \frac{3 \times 6 \times 4 \times 5 \times 4}{15 \times 14 \times 13 \times 12} + \frac{6 \times 4 \times 5 \times 4}{15 \times 14}$

$\frac{20 + 12 + 16}{7} = \frac{48}{7} =$

$$\frac{{}^6C_2\,{}^4C_1\,{}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1\,{}^4C_1\,{}^5C_2}{{}^{15}C_4} + \frac{{}^6C_1\,{}^4C_2\,{}^5C_1}{{}^{15}C_4}$$

$$= \frac{\overset{3}{\cancel{6}}\times 5\times 4\times 5\times \overset{2}{\cancel{4}}\times 3\times 2}{2\times \overset{3}{\cancel{15}}\times \overset{7}{\cancel{14}}\times 13\times \overset{2}{\cancel{12}}} + \frac{\overset{3}{\cancel{6}}\times 4\times \overset{2}{\cancel{5}}\times 4\times \overset{2}{\cancel{4}}\times 3\times 2}{2\times \overset{3}{\cancel{15}}\times \overset{7}{\cancel{14}}\times 13\times \overset{2}{\cancel{12}}} + \frac{\overset{3}{\cancel{6}}\times 4\times 3\times 5\times \overset{2}{\cancel{4}}\times 3\times 2}{2\times \overset{3}{\cancel{15}}\times \overset{7}{\cancel{14}}\times 13\times \overset{2}{\cancel{12}}}$$

$$= \frac{\overset{20}{\cancel{40}}}{2\times 91} + \frac{\overset{16}{\cancel{32}}}{2\times 91} + \frac{\overset{12}{\cancel{24}}}{2\times 91} = \frac{48}{91} = 0.527$$

Three persons A, B & C have applied for a job in a private company & the chance of their selection is ratio 1:2:4 The probability is that A, B & C can introduce changes to improve to the company are 0.8, 0.5 & 0.3 respectively If the change doesn't take place. Find the probability that it is due to the appointment of C.

Three urns are there containing white & black balls 1st urn has 3W & 2B balls, 2nd has 2W & 3B balls & 3rd urn has 4W & 1B ball. Without any biasing one urn is chosen from that 1 ball is chosen randomly which was white. What is the probablity that it came from 3rd urn.

An urn contains 9 balls two of which are red, 3 Blue & 4 black. Three balls are drawn at random. What is the probability that 3B are of different colours.

$$\frac{{}^2C_1 \cdot {}^3C_1 \cdot {}^4C_1}{{}^9C_3} = \frac{2\times 3\times 4\times 3\times 2}{9\times 8\times 7} = \frac{2}{7}$$