Big Data Analytics

22 October 2024 11:27

* Empirical formulae

2 × mean + mode = 3 × median

or

3 median − 2 mean = mode

* Median → grouped data = $l + \left(\frac{n/2 - cf}{f}\right) \times h$

* mode → $= l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$

Measures of Dispersion =  → most scattered and dispersed value.

→ Dispersion in Statistics:-

It is the way of scattering the data through which we can easily find that how far does the data is from the average value, or how close.

It shows variability & consistency of data.

Measures of Dispersion is of 2 types.

Eg. - Range, variance & Standard deviation

1) Absolute measure of Dispersion.
2) Relative measure of Dispersion.

→ Absolute ✓ Range, variance, Standard Deviation, Quartile Deviation & mean Deviation.

→ Relative → coff of range, coff of variance, coff of qu. dev. coff of mean deviation

* Units of data → expressed & measured

↓

Absolute measure of Dispersion.

* Relative → 2 values vary & from there we express ... d 2 data

Also:

* Relative → 2 values vary & from that we express & measure by taking reference of 2 data & also helps in scattering of datas.

* Range → 10, 20, 15, 0, 100 } ungrouped data.

100 − 0 = (100)

eg:- 20, 24, 31, 17, 45, 39, 51, 61

Range = 61 − 17 = (44) → Range.

* Grouped data:-

	f
0 − 10	5
10 − 20	8
20 − 30	15
30 − 40	9

Range = 0 lowest
40 − highest
→ highest − lowest
→ 40 − 0 = (40)

* Mean deviation:-

$$M.D = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots |x_n - \mu|}{n}$$

** { for ungrouped data } (μ − mean)

Eg:- −5, 10, 25

⇒ $\mu = \frac{-5 + 10 + 25}{3} = (10)$ (μ = 10)

$$M.D = \frac{|-5 - 10| + |10 - 10| + |25 - 10|}{3}$$

$$\Rightarrow \frac{-15 + 0 + 15}{3} = (0) \rightarrow (M.D = 0)$$

* Eg:- mean deviation for 2, 4, 6, 8, 10

$\mu = 2 + 4 + 6 + 8 + 10 / 5 = 6$ (μ = 6) → mean

$$M.D = \frac{\Sigma|x_i - \bar{x}|}{n} = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots |x_n - \mu|}{n}$$

$$M.D = |2 - 6| + |4 - 6| + |6 - 6| + |8 - 6| + |10 - 6|$$

$$M.D = \frac{|2-6| + |4-6| + |6-6| + |8-6| + |10-6|}{5}$$

$$\rightarrow \frac{4+2+0+2+4}{5} = \frac{12}{5} = 2.4$$

M.D = 2.4

* Mean Deviation for discrete frequency distribution :-

$$M.D = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$$

7350 + 4900 + 2200 + 5200 + 4350 / 30

Eg :-

x_i wages	f_i	$\|x_i - \bar{x}\| \times f_i$
2500 - 3500	7	7350
3000 - 3500	9	4900
4000	5	2200
4500	6	5200
5000	3	4350
	30	

-1050
-550
+450
950
1450
1250

$\Sigma x_i f_i$: 7350, 24000, 20000, 27000, 13500, 106500

$$\text{mean} = \frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{106500}{30}$$

mean = 3550

4450 × 30 / 30 = 1200 × 30

M.D = 820

* Continuous → M.D

Eg :-

10 - 20	$\frac{10+20}{2} = 15$	15
20 - 30	25	10
30 - 40	35	13
40 - 50	45	12

⇒

x_i	f_i	$x_i f_i$
15	15	225
25	10	250
35	13	455
45	12	540

$$\mu = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

$$\mu = \frac{225 + 250 + 455 + 540}{50} = \frac{1470}{50} = 29.4$$

$$M.D = \frac{|15-29.4|, |25-29.4|, |35-29.4|, |45-29.4|}{50}$$

$$14.4, 4.4, 5.6, 15.6$$

M.D = |15 − 21.71|, |25 − 21.71|, |30 − ...|, ...

14.7, 4.4, 5.6, 15.6

$\Sigma\lvert x_i - \bar{x}\rvert$	f_i	$\Sigma\lvert x_i - \bar{x}\rvert f_i$
14.4	15	216
4.4	10	44
5.6	13	72.8
15.6	12	187.2
	520	

$$M.D = \frac{\sum_{i=1} f_i (x_i - \bar{x})}{\sum f_i}$$

$$\Rightarrow \frac{520}{50} = 10.4$$

(*) $S.D = \sqrt{\frac{\sum_{i=1} (x_i - \bar{x})^2}{N}}$

$\sigma = \frac{f_i}{N}$

variance(σ^2) $= \frac{(x_i - \bar{x})^2}{N}$

* SD of ungrouped data:-

S.D of 'Actual mean' $= \sqrt{\frac{\sum_{i=1} (x_i - M)^2}{N}}$

Eg:- $x = \{2, 3, 4, 5, 6\}$

x_i	$x_i - M$	$(x_i - M)^2$
2	2 − 4 = −2	4
3	−1	1
4	0	0
5	1	1
6	2	4
		10

$\frac{\Sigma (x_i - M)^2}{N}$

$S.D = \sqrt{\sum_{i=1}^{n} \frac{(x_i - M)^2}{N}}$

$M = 2+3+4+5+6/5 = 4$

$\sqrt{\frac{10}{5}} = \sqrt{2} = 1.414$

* SD by assumed mean method:-

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

* Standard deviation for Discrete, grouped data.

1) Actual mean method

2) Assumed mean method.

formulae $= \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2 \times f_i}{N}}$

formulae = $\sqrt{\frac{\sum_{i=1}^{n}}{N}}$

Eg:-

x_i	f_i
10	1
4	3
6	5
8	1

Standered deviation for Discrete.

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \mu)^2}{N}}$$

$$\mu = \frac{\sum x_i f_i}{\sum f_i} = \frac{10 + 12 + 30 + 8}{10} = \frac{60}{10} = 6$$

$\mu = 6$

x_i	f_i	$(x_i - \mu)$	$(x_i - \mu)^2$	$f_i (x_i - \mu)^2$
10	1	4	16	16
4	3	-2	4	12
6	5	0	0	0
8	1	2	4	4
			24	32

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{32}{10}} = \sqrt{3.2}$$

$SD = \sqrt{3.2} = 1.788$

* for continous Grouped distribution :-

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \mu)^2}{N}}$$

Eg:-

x_i	f_i	$\bar{x}$	$f_i x_i$	$x_i - \mu$	$(x_i - \mu)^2$	$\sum f_i (x_i - \mu)^2$
0-10	2	5	10	-14	196	392
10-20	4	15	60	-4	16	64
20-30	2	25	50	6	36	72
30-40	2	35	70	16	256	512
			190			1040

$$\text{Mean} = \frac{\sum x_i f_i}{\sum f_i} = \frac{190}{10} = 19$$

$\mu = 19$

$$SD = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{1040}{10}} = 10.19$$

$SD = 10.19$

* Quantile deviation :-

Quartile deviation — lower Deviation

Quantile

Highest deviation − lowest Deviation / 2

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

Q_3 = highest deviation, Q_1 = lowest deviation.

$$Q_1 = \text{size}\left[\frac{n+1}{4}\right]^{th}$$

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

coff of $Q_D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

{Interquartile Range $= Q_3 - Q_1$}

Eg:- For ungrouped data = Discrete.

Q 150, 100, 268, 280, 195, 140, 200.

Soln:- Arrange in ascending order.

100 140 150 195 200 268 280

Find InterQuantile Range, quatile deviation & coff of Quantile Deviation.

Soln:- InterQuantile Range $= Q_3 - Q_1$

$$Q_1 = \left[\frac{n+1}{4}\right]^{th} \quad n=7 \quad \left[\frac{7+1}{4}\right]^{th} = \left\{\frac{8}{4}\right\} = 2^{nd}$$

$Q_1 = 2^{nd} =$ 140

$Q_3 = 3 \wedge 2 = 6^{th} =$ 268

InterQuartile Range = 268 − 140 = 128

Quantile Deviation = $\frac{128}{2} =$ 64 ; $\left(\frac{Q_3 - Q_1}{2}\right)$

coff of Quantile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{128}{408}$

$= 0.31$

Q

x_i	f_i	cf
60	25	25
62	21	46

InterQuantile :- $Q_3 - Q_1$

$$Q_1 = \left[\frac{n+1}{4}\right]^{th} \quad n = 199 \quad = \frac{200}{4} = 50^{th}$$

60	25	25
62	21	46
68	29	75
70	18	92
75	24	112
80	20	136
88	24	160
96	17	177
97	22	199

$Q_1 = \left[\frac{n+1}{4}\right]$ th $= \frac{n}{4}$

$Q_1 = 68$

$Q_3 = 3 \times 50 = 150 = 88$

Inter $= Q_3 - Q_1$

$88 - 68 = 20$

Quartile deviation $= \frac{Q_3 - Q_1}{2} = \frac{20}{2} = 10$

Coff of Quartile Deviation $= \frac{20}{154} = 0.128$

* Quartile Deviation for continous frequency (Grouped)

Size	f	cf
0-10	9	9
10-20	15	24
20-30	29	53
30-40	24	77
40-50	3	80
50-60	20	100
	100	

$Q_1 = \left(\frac{N}{4}\right)^{th}$, $Q_3 = 3\left[\frac{N}{4}\right]^{th}$.

$Q_1 = l + \frac{(n/4 - (cf)) \times i}{f}$

$Q_3 = l + 3 \times \frac{(n/4 - (cf) \times i}{f}$

$Q_1 = \left(\frac{N}{4}\right)^{th} = 25^{th}$

$l = 20$, $cf = 24$, $f = 29$, $i = 10$

$Q_1 = 20 + \left(\frac{25 - 24}{29}\right) \times 10$

$20 + \frac{1}{4} \times 10$

$Q_1 = 20.34$

$Q_3 = 3\left(\frac{N}{4}\right) = 75^{th}$ $l = 30$, $cf = 53$, $f = 24$

$Q_3 = 30 + 3 \times \left(\frac{75 - 53}{24}\right) \times 10$

$= 39.16$

$Q_3 = 39.16$

Inter Quartile Range $= Q_3 - Q_1 = 39.16 - 20.34 = 18.82$

Quartile Deviation $= \frac{Q_3 - Q_1}{2} = \frac{18.82}{2} = 9.41$

Coff $= \frac{18.82}{59.5} = 0.31$

$Q_1 = \left(\frac{N}{4}\right)^{th} = 50^{th}$

$Q_3 = 150^{th}$

		cf
20-25		100

Q

Size	f	cf
0-5	6	6
5-10	18	24
10-15	30	54
15-20	46	100
20-25	60	160
25-30	40	200

$Q_1 = \left(\frac{N}{4}\right)^{th} = 50^{th}$

10-15

$l = 10,\ cf = 24,\ f = 30$

$Q_1 = 10 + \left(\frac{50-24}{30}\right) \times 5 = 14.3$

20-25

$l = 20,\ cf = 100$

$f = 60$ ×5

$Q_3 = 20 + 3\left(\frac{150-100}{60}\right)$

$Q_3 = 20 + 12.5$

$= 32.5$

$Q_1 = 14.3,\ Q_3 = 32.5$

Inter Quantile Range $= Q_3 - Q_1$

$32.5 - 14.3 = 18.2$

Quantile Deviation $= \frac{18.2}{2} = 9.1$

Coff $= \frac{18.2}{46.8} = 0.38$

* Coff of Range :- $\left\{\frac{\text{Largest} - \text{Smallest}}{\text{Largest} + \text{Smallest}}\right\}$

Q

1575 = largest

1380 = smallest

Range $= 1575 - 1380 = 195$

Coff of Range $= \frac{195}{2955} = 0.065 = \frac{\text{Largest} - \text{Smallest}}{\text{Largest} + \text{Smallest}}$

* Coff of Mean deviation :-

$$MD = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{N}$$

Coff $= \frac{MD}{\bar{x}}$ $\qquad \frac{MD \times}{\bar{x}}$

Eg :- 2, 4, 6, 8, 10

x_i	$x_i - \bar{x}$	$f(x_i - \bar{x})$
2	2 - 6 = 4	
4	4 - 6 = 2	
6	6 - 6 = 0	
8	8 - 6 = 2	
10	10 - 6 = 4	

$\bar{x} = \frac{\sum x_i}{N} = 30/5 = 6$ $\qquad \mu = 6$

$\frac{\sum x_i f_i}{\sum f_i} = \mu$

$MD = \frac{\sum f(x_i - \bar{x})}{N}$

f	12	12000

Q

x_i	f	(12)
2500	7	17500
3000	9	27000
4000	5	36000
4500	6	27000
5000	3	15000

$\frac{\Sigma x_i f_i}{\Sigma f_i} = M$

$= \frac{116500}{30} = 3883.3$

$= \frac{N}{820}$

$\text{coff} = \frac{MD}{M} = \frac{820}{3883} = 0.23$

* coff of mean deviation in continous (grouped) :-

x_i	x_i	f
15	10-20	15
25	20-30	10
35	30-40	13
45	40-50	12

$\frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{225 + 250 + 455 + 540}{40}$

$M = 29.4$

$x_i - \bar{x}$	$f(x_i - \bar{x})$
15 - 29.4	14.4
25 - 29.4	4.4
35 - 29.4	5.6
45 - 29.4	15.6

216
44
72.8
187.2
520

$MD = \frac{520}{50} = 10.4$

$MD = 10.4$

$\text{coff} = \frac{10.4}{29.4} = 0.35$

* coffecient of variance :-

$= \frac{\text{Standered Deviation}}{\text{mean}} \times 100$

$SD = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{N}}$

Q

$2 + 3 + 4 + 5 + 6 / 5 = 4$ $\quad M = 4$

$2-4, \; 3-4, \; 4-4, \; 5-4, \; 6-4$

$\rightarrow 2 \quad 1 \quad 0 \quad 1 \quad 2$

$4 \quad 1 \quad 0 \quad 1 \quad 4$

$SD = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.414$

$\Rightarrow SD/\text{mean} = \frac{1.414}{4} = 0.3535$

Q

x_i	f_i		
10-12	1	2	16
4-6	3		0
6-8	5		16

$\frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{10 + 12 + 30 + 8}{10}$

$M = 4$

$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - M)^2}{N}}$

$\sqrt{\frac{32}{4}} = \sqrt{8} = 2\sqrt{2}$

4	3	0	0
6-4	5	4	4
8-6	4		

$\sqrt{\frac{32}{4}} = \sqrt{8} = 2\sqrt{2}$

Q

	f_i	$x_i f_i$
5 = 0 - 10	2	10
15 = 10 - 20	4	60
25 = 20 - 30	2	50
35 = 30 - 40	2	70
	10	190

$\mu = 19$

$SD = \sqrt{\frac{\sum f(x_i - \mu)^2}{N}}$

$SD = 10.198$

$coff = SD/mean = 10.198/19 = 0.52$

Q

$SD = 8.5$

$\mu = 14.5$

$coff = \frac{8.5}{14.5} \times 100 = 58.6\%$

Q

$SD = 1.4$

$\sigma = 26.5$

$coff = \frac{SD}{mean}$

$\frac{1.4}{26.5} = 0.0528 \times 100 = 52.8\%$

Q

$\mu = 13$

$coff = 38$

$coff = \frac{SD}{m} \times 100$

$38 = \frac{SD}{13} \times 100$

$m = \frac{13 \times 100}{38}$

$\frac{38 \times 13}{100} = SD$

$SD = 4.94$

Q

$\frac{SD}{m} = coff$

$\frac{10}{65} = 15\%$ → maths

$\frac{12}{60} = 20\%$ economics

$\frac{12}{60} = 20\%$

$\frac{14}{57} = 24\%$ → Economics

---

* Null hypothesis:- agreement condition

H_0 → $\mu_1 = \mu_2$

* Alternative hypo:- $\mu_1 \neq \mu_2$ → H_a

* one tailed test → left → H_0 $\mu \geq 50$, $H_a = \mu < 50$

→ right → H_0: $\mu \leq 50$, $H_a = \mu > 50$

* Two tailed test:- In which we include both hypothesis & assuming them

* Type I & II errors:-

i) Type I → null hypothesis ✓ but we rejected (×)

ii) Type II → null hypothesis × but we accepted (β)

* z testing:-

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = sample mean
μ = population mean
σ = standard deviation
n = size

Q Data $n = 25$, $\mu = 200$, $\sigma = 5 mg/dc$

Soln:- firstly we accept null hypothesis.

H_0: Average cholesterol is 200

H_a: A· —— is not 200

Significance = 0.05

$$z = \frac{203.8 - 200}{\frac{5}{5}} = 3.8$$

rejected null hypothesis

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

z tabulated = 0.05

−1.96 to 1.96 +

$z = 2.98$ (more) compare.

* Standard deviation value when not given in + test.

t = $\mu > 30$

$$S = \left\{ \frac{1}{n-1} \left[\Sigma \bar{x}^2 - \frac{(\Sigma x)^2}{n} \right] \right\}$$ → standard deviation

Now t test formulae:-

$$t_{critical} = \left.\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}\right\} \rightarrow \text{when } s \text{ is not given}$$

or

$$\left.\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}\right\} \text{ when } s \text{ is already given.}$$

122.1 , 118.2

$n \rightarrow x_{after} - x_{before}$

$\Rightarrow 118.2 - 122.1 = -3.9$

$s = 1.37$

$n = -3.9$

$$t_{critical} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad 3.9 / \frac{1.37}{\sqrt{10}}$$

$$s = \frac{1}{n-1}\left[\Sigma \bar{x}^2 - \frac{(\Sigma x)^2}{n}\right]$$

(s)

For 2 values $$s = \frac{1}{n_A + n_B - 2}\left[\Sigma \bar{x}^2 - \frac{(\Sigma \bar{x})^2}{n_A} + \Sigma x_B^2 - \frac{\Sigma x_B^2}{n_B}\right]$$

* chi square test:-

$$\chi_c^2 = \frac{\Sigma (O_i - E_i)^2}{E_i}$$

Q

Water/Actors	Y	N	total
m	140	44	184
f	178	38	216

$E_i = \frac{RT \times CT}{GT}$

$M \rightarrow Y$

$\frac{184 \times 318}{400} = 61.4$

F		178	38	216
		318 (Oi)	82 (Ei)	400
M	Y	140	146	36
M	N	44	38	36
F	Y	178	172	36
F	N	38	44	36

$\frac{184 \times 318}{400} =$

M → N

$\frac{82 \times 184}{400} = 38$

F → Y $\frac{318 \times 216}{400} = 172$

F → N $\frac{82 \times 216}{400} = 44$

$(O_i - E_i)^2 / E_i$

* f test hypothesis:-

$$F = \frac{S_1^2}{S_2^2}$$

$$\left\{ S_1^2 = \frac{\Sigma (x_A - \bar{x})^2}{n_A - 1} \right\}$$

$$\left\{ S_2^2 = \frac{\Sigma (x_B - \bar{x})^2}{n_B - 1} \right\}$$

Q

x_1	x_1
20	
16	
26	
24	
23	
22	

when $(\bar{x})$ is not given

$$S_1^2 = \frac{n_1 (S_1^2)}{n_1 - 1} \quad , S_2^2 = \frac{n_2}{n_2 - 1} (S_2)^2$$