আমার বিজ্ঞকথা

টপিকঃ

ডেটা সায়েন্স

০০১-০১৩ Numerical Data (পর্ব-৫)

০০১-০১৩ Numerical Data (পর্ব-৫)

আমরা আগে Discrete Numerical Data এর ক্ষেত্রে Central Tendency, Dispersion, Mean Absolute Deviation এবং Median Absolute Deviation দেখেছি।

এবার আমরা Continuous Numerical Data এর ক্ষেত্রে Central Tendency, Dispersion, Mean Absolute Deviation এবং Median Absolute Deviation এগুলো দেখাবো। এক্ষেত্রে, আমরা আগে থেকে জানি যে, ডেটা সায়েন্সে Continuous Numerical Data এর ক্ষেত্রে Exclusive Interval সবচেয়ে ভালো পদ্ধতি। তাই আমরা এখানে শুধু Exclusive Interval দিয়েই Continuous Numerical Data এর কাজ করবো।

Continuous Numerical Data এর ক্ষেত্রে সেন্ট্রাল টেন্ডেন্সি:

দশম শ্রেণীর "ঘ" শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বর নিচে দেওয়া হলো —

53, 97, 87, 46, 46, 53, 91, 66, 49, 49, 91, 51, 49, 81, 67, 56, 54, 98, 95, 48, 65, 56, 48, 80, 88, 84, 50, 63, 60, 67, 66, 78, 98, 52, 56, 79, 90, 67, 61, 95, 86, 45, 48, 99, 68, 71, 85, 87, 58, 59, 93, 57, 93, 97, 82, 48, 50, 80, 90, 78

[আমরা আগেই "০০১-০১০ Numerical Data (পর্ব-২)" এ বলেছিলাম যে - দশম শ্রেণীর "ঘ" শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের ডেটাসেট হলো Population]

এক্ষেত্রে আমাদের ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন টেবিল বানানোর নিয়ম হলো, আগে sorting করা।

45, 46, 46, 48, 48, 48, 48, 49, 49, 49, 50, 50, 51, 52, 53, 53, 54, 56, 56, 56, 57, 58, 59, 60, 61, 63, 65, 66, 66, 67, 67, 67, 68, 71, 78, 78, 79, 80, 80, 81, 82, 84, 85, 86, 87, 87, 88, 90, 90, 91, 91, 93, 93, 95, 95, 97, 97, 98, 98, 99

আমরা এক্ষেত্রে লক্ষ্য করতে পাচ্ছি যে maximum value = 99 এবং minimum value = 45।

অতএব, পরিসর (Range) = minimum value – minimum value = 99 – 45 = 54

এরপর আমরা স্ট্রাজেসের ফর্মুলা (Sturges' Formula) দিয়ে class number বের করবো। এখানে "k" হলো class number এবং N হলো উপাত্ত সংখ্যা। N = 60 *[৬০ জন শিক্ষার্থী]*। এখানে n এর পরিবর্তে N দেওয়া হয়েছে, কারণ দশম শ্রেণীর "ঘ" শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের ডেটাসেট হলো Population।

$$k = 1 + 3.322\log(N) = 1 + 3.322\log(60) = 6.91 \approx 7$$

অর্থাৎ 7টি ক্লাস ইন্টার্ভাল তৈরি হবে।

এরপর আমরা class width 'j' বের করে পাই।

$$j = \frac{\text{Range}}{k} = \frac{54}{7} = 7.7 \approx 8$$

অর্থাৎ প্রত্যেক ক্লাস ইন্টার্ভালের ব্যবধান 8 হবে। এবং এক ক্লাস ইন্টার্ভালের lower limit থেকে তার পরবর্তী ক্লাস ইন্টার্ভালের lower limit এর ব্যবধান 8 হবে।

ঠিক এরকম,

এখানে 45 আর 53 এর ব্যবধান হলো 8, যেটা j এর মান

ক্লাস ইন্টার্ভাল
45 – 53
53 – 61

এখানে 45 আর 53 এর ব্যবধান হলো 8, যেটা j এর মান

এখন আমরা $k = 7, j = 8$ অনুযায়ী ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন টেবিল বানানো হলোঃ

$k = 7$
এখানে 7 টি ক্লাস ইন্টার্ভাল রয়েছে

ক্লাস ইন্টার্ভাল	ট্যালি মার্ক	ফ্রিকুয়েন্সি	কিউমুলেটিভ ফ্রিকুয়েন্সি
45 – 53	𝍸 𝍸 ||||	14	14
53 – 61	𝍸 𝍸	10	14 + 10 = 24
61 – 69	𝍸 ||||	9	24 + 9 = 33
69 – 77	|	1	33 + 1 = 34
77 – 85	𝍸 |||	8	34 + 8 = 42
85 – 93	𝍸 ||||	9	42 + 9 = 51
93 – 101	𝍸 ||||	9	51 + 9 = ***60***
Total		***60***	

বিঃ দ্রঃ পরীক্ষার নাম্বার 101 হয় না, কিন্তু এখানে ক্লাস ইন্টার্ভাল width এর কারণে 101 হয়ে গেছে।

এখন, এই ডেটা টেবিলের মধ্যমান (x_i) আর ফ্রিকুয়েন্সি (f_i) বের করে পাইঃ

ক্লাস ইন্টার্ভাল	মধ্যমান (x_i)	ফ্রিকুয়েন্সি (f_i)	$f_i x_i$
45 – 53	49	14	686
53 – 61	57	10	570
61 – 69	65	9	585
69 – 77	73	1	73
77 – 85	81	8	648
85 – 93	89	9	801
93 – 101	97	9	873
Total	$\sum_{i=1}^{k} x_i = 511$	$\sum_{i=1}^{k} f_i = 60$	$\sum_{i=1}^{k} f_i x_i = 4236$

মধ্যমান বের করার নিয়ম

$$\text{মধ্যমান} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

উদহারণঃ

45 – 53 এর মধ্যমান $= \frac{45+53}{2} = 49$

53 – 61 এর মধ্যমান $= \frac{53+61}{2} = 57$

গড় (Mean):

Mean দুই ধরনের,

১) Population Mean (μ)

২) Sample Mean $(\bar{x})$

1 2 3 4 1 2 4 4 3 5 3
6 4 6 5 1 1 3 3 2 4 5
6 8 5 8 3 3 4

Population

$\Rightarrow$

1 2 2 3 5 1 1

Sample

উভয়ক্ষেত্রে সূত্র একইঃ-

$$\text{Sample mean}, \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \qquad \text{Population mean}, \mu = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

(কোনো ডেটাসেটের ক্ষেত্রে Population বা Sample উল্লেখ না থাকলে আমরা সাধারণভাবে Sample mean বের করি। কিন্তু আমরা আগেই বলেছি যে দশম শ্রেণীর “ঘ” শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের ডেটাসেট হলো Population। তাই এক্ষেত্রে আমরা Population mean বের করবো।)

সুতরাং, দশম শ্রেণীর “ঘ” শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের Mean:

$$\mu = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{4236}{60} = 70.6$$

আমরা আগেই “০০১-০১০ Numerical Data (পর্ব-২)” এ বলেছিলাম যে - দশম শ্রেণীর “ঘ” শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের mean, $\mu = \frac{53+97+87+46+46+53+91+66+49+\cdots+50+80+90+78}{60} = 70.07$ । কিন্তু এখানে ফ্রিকুয়েন্সি টেবিল বানিয়ে mean বের করার ক্ষেত্রে দেখাচ্ছে $\mu = 70.6$। এর কারণ হলো, ফ্রিকুয়েন্সি টেবিল ছাড়া mean বের করা হচ্ছে arithmetic calculation। আর ফ্রিকুয়েন্সি টেবিল দিয়ে mean বের করা হচ্ছে শুধুমাত্র মধ্যমান আর ফ্রিকুয়েন্সির গুণফলগুলোকে সমষ্টি করে mean বের করা, এবং এটা হলো approximation। এক্ষেত্রে ডেটাসেটের সকল উপাত্তগুলোকে মধ্যমান আকারে ধরা হচ্ছে, যার কারণে দশমিকে পার্থক্য দেখা যাচ্ছে।

মধ্যক (Median):

ক্লাস ইন্টার্ভাল	ফ্রিকুয়েন্সি (f_i)	কিউমুলেটিভ ফ্রিকুয়েন্সি (cf_i)
45 – 53	14	14
53 – 61	10	14 + 10 = 24
61 – 69	9	24 + 9 = 33
69 – 77	1	33 + 1 = 34
77 – 85	8	34 + 8 = 42
85 – 93	9	42 + 9 = 51
93 – 101	9	51 + 9 = *60*
Total	$\sum_{i=1}^{k} f_i = 60$	

এখানে, Median বের করার জন্য $\frac{1}{2}\sum_{i=1}^{k} f_i$ নির্ণয় করতে হবে। আর সেই $\frac{1}{2}\sum_{i=1}^{k} f_i$ যদি কোনো ক্লাসের কিউমুলেটিভ ফ্রিকুয়েন্সির চেয়ে কম হয় এবং তার পূর্ববর্তী ক্লাসের কিউমুলেটিভ ফ্রিকুয়েন্সির চেয়ে বেশি হয়, তাহলে $\frac{1}{2}\sum_{i=1}^{k} f_i$ যে ক্লাসের কিউমুলেটিভ ফ্রিকুয়েন্সির চেয়ে কম, সেই ক্লাস হলো "median class"।

এখানে, $\frac{1}{2}\sum_{i=1}^{k} f_i = \frac{60}{2} = 30$। এই 30 হলো 61 – 69 ক্লাসের কিউমুলেটিভ ফ্রিকুয়েন্সি 33 এর চেয়ে কম; আবার এই 30 হলো 61 – 69 ক্লাসের পূর্ববর্তী ক্লাস 53 – 61 এর কিউমুলেটিভ ফ্রিকুয়েন্সি 24 এর চেয়ে বেশি। সুতরাং 61 – 69 ক্লাস হলো "Median Class"।

ক্লাস ইন্টার্ভাল	ফ্রিকুয়েন্সি (f_i)	কিউমুলেটিভ ফ্রিকুয়েন্সি (cf_i)
45 – 53	14	14
53 – 61	10	14 + 10 = 24
61 – 69	9	24 + 9 = 33
69 – 77	1	33 + 1 = 34
77 – 85	8	34 + 8 = 42
85 – 93	9	42 + 9 = 51
93 – 101	9	51 + 9 = *60*
Total	$\sum_{i=1}^{k} f_i = 60$	

এখানে,

Median Class এর Lower Limit, $L = 61$

Median Class এর আগের Class এর কিউমুলেটিভ ফ্রিকুয়েন্সি, $B = 24$

Median Class এর ফ্রিকুয়েন্সি, $G = 9$

Median Class এর Width,

$W = (Max - Min) + 1 = (69 - 61) + 1 = 9$

$\frac{1}{2}\sum_{i=1}^{k} f_i = \frac{60}{2} = 30$

অতএব,

$$\text{median, } \tilde{x} = L + \frac{\frac{1}{2}\sum_{i=1}^{k} f_i - B}{G} \times W$$

$$= 61 + \frac{30 - 24}{9} \times 9$$

$$= 67$$

প্রচুরক (Mode):

ক্লাস ইন্টার্ভাল	ফ্রিকুয়েন্সি (f_i)
45 – 53	14
53 – 61	10
61 – 69	9
69 – 77	1
77 – 85	8
85 – 93	9
93 – 101	9
Total	$\sum_{i=1}^{k} f_i = 60$

এখানে সর্বোচ্চ ফ্রিকুয়েন্সি হলো 14 যা 45 – 53 ক্লাস ইন্টার্ভালে আছে।

তাহলে modal class হলো 45 – 53।

Modal class এর Lower Limit, L = 45

Modal class এর পূর্ববর্তী class এর ফ্রিকুয়েন্সি, $f_0 = 0$

;[কারণ Modal class এর আগে কোনো ক্লাস নেই, তাই f_0 এর মান 0]

Modal class এর ফ্রিকুয়েন্সি, $f_1 = 14$

Modal class এর পরবর্তী class এর ফ্রিকুয়েন্সি, $f_2 = 10$

h = Upper Limit – Lower Limit = 53 – 45 = 8

সুতরাং,

$$\text{mode, } \; Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 45 + \left(\frac{14 - 0}{(2 \times 14) - 0 - 10}\right) \times 8$$

$$= 51.22$$

তো আমরা এই পর্যন্ত Continuous Numerical Data এর ক্ষেত্রে সেন্ট্রাল টেন্ডেন্সি (Mean, Median, Mode) শিখে গেছি।