



আমার বিজ্ঞকথা



টপিকঃ

ডেটা সায়েন্স

০০১-০১৪ Numerical
Data (পর্ব-৬)





০০১-০১৪ Numerical Data (পর্ব-৬)

আমরা Continuous Numerical Data এর ক্ষেত্রে Central Tendency শিখেছি, এবার Continuous Numerical Data এর ক্ষেত্রে Dispersion শিখবো।

Continuous Numerical Data এর ক্ষেত্রে Dispersion:

Range:

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) = maximum value – minimum value = 99 – 45 = 54

range থেকে ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন টেবিল বানানোঃ

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]

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

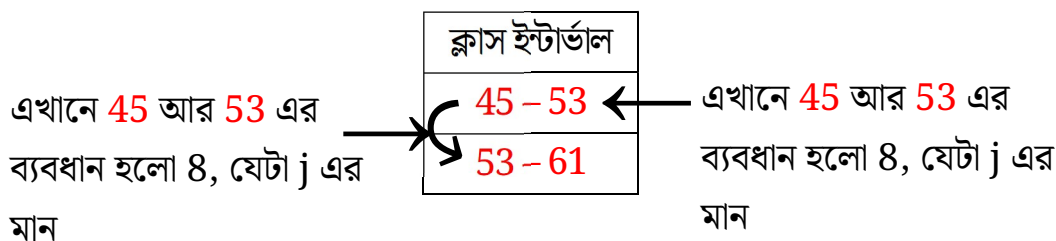
অর্থাৎ ৭টি ক্লাস ইন্টারভাল তৈরি হবে।

এরপর আমরা class width ‘j’ বের করে পাই।

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

অর্থাৎ প্রত্যেক ক্লাস ইন্টারভালের ব্যবধান ৮ হবে। এবং এক ক্লাস ইন্টারভালের lower limit থেকে তার পরবর্তী ক্লাস ইন্টারভালের lower limit এর ব্যবধান ৮ হবে।

ঠিক এরকম,



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

ক্লাস ইন্টারভাল	ট্যালি মার্ক	ফ্রিকুয়েন্সি	কিউমুলেটিভ ফ্রিকুয়েন্সি
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		60	

$k = 7$
এখানে ৭ টি ক্লাস ইন্টারভাল রয়েছে

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

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

ক্লাস ইন্টারভাল	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	$\sum_{i=1}^k x_i = 511$	$\sum_{i=1}^k f_i = 60$	$\sum_{i=1}^k f_i x_i = 4236$

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

উদাহরণঃ

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

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

Variance:

এই ডেটাসেটের Population mean আমরা আগেই বের করেছি।

$$\text{Population mean, } \mu = 70.6$$

Continuous Numerical Data এর ক্ষেত্রে Variance এর সূত্রঃ

$$\text{Sample variance, } s^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\left(\sum_{i=1}^k f_i \right) - 1}$$

$$\text{Population variance, } \sigma^2 = \frac{\sum_{i=1}^k f_i (x_i - \mu)^2}{\sum_{i=1}^k f_i}$$

কোনো ডেটাসেটের ক্ষেত্রে Population বা Sample উল্লেখ না থাকলে আমরা সাধারণভাবে Sample variance বের করি।

কিন্তু আমরা আগেই বলেছি যে দশম শ্রেণীর “ঘ” শাখার ৬০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বরের ডেটাসেট হলো

Population। তাই এক্ষেত্রে আমরা

Population variance (σ^2) বের করবো।

এখন,

ক্লাস ইন্টারভাল	মধ্যমান (x_i)	$(x_i - \mu)^2$	ফ্রিকুয়েন্সি (f_i)	$f_i(x_i - \mu)^2$
45 – 53	49	466.56	14	6531.84
53 – 61	57	184.96	10	1849.6
61 – 69	65	31.36	9	282.24
69 – 77	73	5.76	1	5.76
77 – 85	81	108.16	8	865.28
85 – 93	89	338.56	9	3047.04
93 – 101	97	696.96	9	6272.64
Total	$\sum_{i=1}^k x_i = 511$		$\sum_{i=1}^k f_i = 60$	$\sum_{i=1}^k f_i(x_i - \mu)^2 = 18854.4$

$$\text{অতএব, population variance, } \sigma^2 = \frac{\sum_{i=1}^k f_i(x_i - \mu)^2}{\sum_{i=1}^k f_i} = \frac{18854.4}{60} = 314.24$$

Standard Deviation:

Standard Deviation দুই প্রকার। যথাঃ

১) Population Standard Deviation (σ)

২) Sample Standard Deviation (s)

$$\text{Sample standard deviation, } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^k f_i(x_i - \bar{x})^2}{\left(\sum_{i=1}^k f_i\right) - 1}}$$

$$\text{Population standard deviation, } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \mu)^2}{\sum_{i=1}^k f_i}}$$

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

$$\text{অতএব, Population standard deviation, } \sigma = \sqrt{\sigma^2} = \sqrt{314.24} = 17.73$$

তো আমরা এই পর্যন্তই *Continuous Numerical Data* এর ক্ষেত্রে *Dispersion* এর *range*, *variance*, *standard deviation* পর্যন্ত শিখেছি। পরবর্তীতে আমরা *Continuous Numerical Data* এর ক্ষেত্রে *Dispersion* এর *Interquartile Range* শিখবো।