আমার বিজ্ঞকথা

টপিকঃ

ডেটা সায়েন্স

০০১-০১০ Numerical Data (পর্ব-২)

০০১-০১০ Numerical Data (পর্ব-২)

বিঃদ্রঃ Dataset (ডেটাসেট) মানে আমি আগেই বলেছিলাম "Structured collection of data", কিন্তু আমরা যদি 1, 2, 5, 6, 3... এভাবেও ডেটা দেখতে পাই, সেটাকেও আমরা Dataset (ডেটাসেট) বলতে পারি।

সেন্ট্রাল টেন্ডেন্সি (Central Tendency)

সেন্ট্রাল টেন্ডেন্সি তিন ধরনের। যথাঃ গড় (Mean), মধ্যক (Median), প্রচুরক (Mode)।

Categorical Data এর ক্ষেত্রে গড় (Mean) পসিবল ছিলো না। কিন্তু Numerical Data এর ক্ষেত্রে গড় (Mean) পসিবল।

আমরা আগেই জেনেছি Numerical Data দুই ধরনের। যথাঃ ১) Discrete, ২) Continuous।

Discrete Numerical Data এর ক্ষেত্রে সেন্ট্রাল টেন্ডেন্সি:

গড় (Mean):

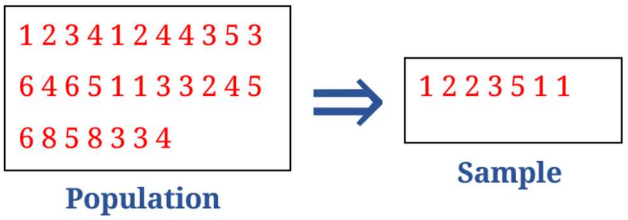
Mean এর সূত্র হলোঃ

$$\text{Mean} = \frac{\text{সকল উপাত্তের যোগফল}}{\text{উপাত্তের সংখ্যা}}$$

Mean দুই ধরনের,

১) Population Mean (μ)

২) Sample Mean ($\bar{x}$)



Population কি আর Sample কি তা আমরা আগেই দেখেছি। Population হলো সমগ্র data এর কালেকশন। Sample হলো Population এর কোনো একটা Subgroup।

তো Population Mean (μ) হলো সমগ্র data এর mean। আর Sample Mean ($\bar{x}$) হলো Population এর কোনো এক Subgroup এর mean।

Population dataset এর ক্ষেত্রে উপাত্তের সংখ্যা কে N দ্বারা প্রকাশ করা হয়। আর Sample dataset এর ক্ষেত্রে উপাত্তের সংখ্যা কে n দ্বারা প্রকাশ করা হয়।

ফর্মুলাঃ

Sample mean এর ক্ষেত্রে:-

$$\text{Sample mean, } \bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{n}$$

$$\Longrightarrow \bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + x_4 + \cdots + x_n)$$

$$\therefore \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Population mean এর ক্ষেত্রে:-

$$\text{Population mean, } \mu = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N}$$

$$\Longrightarrow \mu = \frac{1}{N}(x_1 + x_2 + x_3 + x_4 + \cdots + x_N)$$

$$\therefore \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

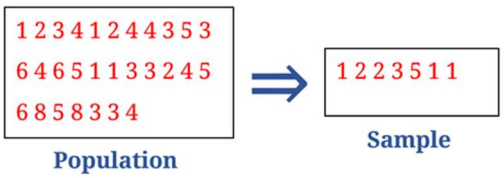
এখানে i মানে হলো ক্রম। i = 1, 2, 3,... N (Population এর ক্ষেত্রে) বা i = 1, 2, 3, ... n (Sample এর ক্ষেত্রে)। x_i মানে হলো কোনো ক্রমের উপাত্ত ভ্যালু।

যেমনঃ 93, 97, 82, 48, 50 এই Sample Dataset এ প্রথম (i=1) উপাত্ত হলো $x_1 = 93$;

দ্বিতীয় (i=2) উপাত্ত হলো $x_2 = 97$; তৃতীয় (i=3) উপাত্ত হলো $x_3 = 82$; চতুর্থ (i=4) উপাত্ত হলো $x_4 = 48$; পঞ্চম (i=5) বা সর্বশেষ (i=n) উপাত্ত হলো $x_5 = x_n = 50$। এখানে n = 5, কারণ ডেটাসেটে পাঁচটি উপাত্তই আছে।

- <u>Population Mean আর Sample Mean এর উদহারণঃ</u>

এখানে Population উপাত্ত দেওয়া আছে



উপাত্তের সংখ্যা N = 29।

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

এর Population mean হলো $\mu = \frac{1+2+3+4+1+2+4+4+3+\cdots+8+3+3+4}{29} = 3.76$

এই Population থেকে একটা Sample হলো 1 2 2 3 5 1 1। উপাত্তের সংখ্যা হলো 7।

Sample mean হলো $\bar{x} = \frac{1+2+2+3+5+1+1}{7} = 2.14$

আরো ভালো উদাহরণ দিলে বুঝতে পারবেন। দশম শ্রেণীর "ঘ" শাখাতে মোট ৬০ জন শিক্ষার্থী আছে। সেই 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

এখন আমি দশম শ্রেণীর "ঘ" শাখার মাত্র ১০ জন শিক্ষার্থীর পদার্থবিজ্ঞান পরীক্ষার নম্বর নিলাম –

85, 87, 58, 59, 93, 57, 93, 97, 82, 48

এখানে দশম শ্রেণীর "ঘ" শাখাটা হলো population আর সেই দশম শ্রেণীর "ঘ" শাখা থেকে আমার নেওয়া ১০ জন শিক্ষার্থী হলো sample

সুতরাং, এক্ষেত্রে Population mean,

$$\mu = \frac{53+97+87+46+46+53+91+66+49+\cdots+50+80+90+78}{60} = 70.07$$

আর Sample mean,

$$\bar{x} = \frac{85+87+58+59+93+57+93+97+82+48}{10} = 75.9$$

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

- Discrete Numerical Data এর ক্ষেত্রে ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন তৈরির মাধ্যমে সহজে mean বের করা:

ধরে নিন, একটা ডেটাসেটে রিপোর্ট দেওয়া আছে যে কতগুলো পল্লী বিদ্যুৎ কেন্দ্রে ভালো এক্সপেরিয়েন্স থাকা কর্মী আছে। এক্ষেত্রে ২০টি পল্লী বিদ্যুৎ কেন্দ্র থেকে এই ডেটা বা উপাত্তের উত্তর পাওয়া গেল।

1, 3, 5, 2, 2, 5, 3, 1, 3, 5, 2, 3, 1, 1, 4, 4, 5, 2, 4, 3

এখানে প্রত্যেক পল্লী বিদ্যুৎ কেন্দ্রে ভালো এক্সপেরিয়েন্স থাকা কর্মীদের সংখ্যা হয় 1 জন আছে, অথবা 2 জন আছে, অথবা 3 জন আছে, অথবা 4 জন আছে, অথবা 5 জন আছে। এখানে 1, 2, 3, 4, 5 হলো distinct values।

এখন এর mean বের করতে হলে আমরা এভাবে mean বের করবো

$$\bar{x} = \frac{1+3+5+2+2+5+3+1+3+5+2+3+1+1+4+4+5+2+4+3}{20} = 2.95$$

এক্ষেত্রে দেখুন কতটা লম্বা আকারে যোগ করে mean বের করা লেগেছে। সেই কাজটা কে লম্বা না করার জন্য আমরা ফ্রিকুয়েন্সির সাহায্যে সহজে mean বের করবো।

ভ্যালু (x_i)	ফ্রিকুয়েন্সি (f_i)	$f_i x_i$
1	4	4*1 = 4
2	4	4*2 = 8
3	5	5*3 = 15
4	3	3*4 = 12
5	4	4*5 = 20
Total	$\sum_{i=1}^{r} f_i = 20$	$\sum_{i=1}^{r} f_i x_i = 59$

এবার একটু মনোযোগ দিয়ে লক্ষ্য করুণ।

এখানে i মানে হলো row এর ক্রম। r হলো মোট row এর সংখ্যা। এই টেবিল অনুযায়ী r = 5।

i = 1, 2, ... r

যেমনঃ প্রথম রো হলো i=1। প্রথম রো i=1 এর ভ্যালু x_1 হলো 1। প্রথম রো i=1 এর ফ্রিকুয়েন্সি f_1 হলো 4।

একিভাবে, দ্বিতীয় রো হলো i=2। দ্বিতীয় রো i=2 এর ভ্যালু x_2 হলো 2। দ্বিতীয় রো i=2 এর ফ্রিকুয়েন্সি f_2 হলো 4।

তাহলে Mean হবে,

$$\bar{x} = \frac{\sum_{i=1}^{r} f_i x_i}{\sum_{i=1}^{r} f_i} = \frac{59}{20} = 2.95$$

এভাবে আমরা Discrete Numerical Data এর ক্ষেত্রে ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন তৈরির মাধ্যমে সহজেই mean বের করতে পারি।

Population mean এর ক্ষেত্রে ফর্মুলাঃ

$$\mu = \frac{\sum_{i=1}^{r} f_i x_i}{\sum_{i=1}^{r} f_i}$$

মধ্যক (Median):

Median হলো সেই সংখ্যা, যেই সংখ্যা ডেটাসেটের উপাত্তগুলিকে কে দুই অংশে ভাগ করে। এক্ষেত্রে উপাত্তের সংখ্যা n হলে,

n এর মান বিজোড় হলে, Median = $\frac{n+1}{2}$ তম উপাত্ত।

n এর মান জোড় হলে, Median = $\frac{n}{2}$ তম উপাত্ত এবং $\left(\frac{n}{2}+1\right)$ তম উপাত্ত।

অনেক ক্ষেত্রে n এর মান জোড় হলে Median = $\frac{\frac{n}{2}\text{ তম উপাত্ত} + \left(\frac{n}{2}+1\right)\text{ তম উপাত্ত}}{2}$ ধরা হয়, যদিও সেটা অনেক ক্ষেত্রে দশমিক মানে আসে এবং সেই সংখ্যা হলো $\frac{n}{2}$ তম উপাত্ত এবং $\left(\frac{n}{2}+1\right)$ তম উপাত্ত – এই দুটি উপাত্তের গড়।

যেমনঃ অষ্টম শ্রেণীর "ক" শাখায় ১০ জন শিক্ষার্থীর গণিত পরীক্ষার নম্বর দেখানো হলো

85, 87, 58, 59, 93, 57, 94, 97, 82, 48

এর median বের করার আগে increasing order এ sort করতে হবে।

48, 57, 58, 59, 82, 85, 87, 93, 94, 97

এখন n = 10 (জোড়)।

তাহলে median = $\frac{10}{2}$ তম উপাত্ত এবং $\left(\frac{10}{2}+1\right)$ তম উপাত্ত = 5 তম উপাত্ত এবং 6 তম উপাত্ত।

Value (x_i)	48	57	58	59	82	85	87	93	94	97
তম	1	2	3	4	5	6	7	8	9	10

5 তম উপাত্ত হলো 82 এবং 6 তম উপাত্ত হলো 85

তাই median = 82 এবং 85

অথবা, median = $\frac{\frac{10}{2}\text{ তম উপাত্ত} + \left(\frac{11}{2}+1\right)\text{ তম উপাত্ত}}{2} = \frac{82+85}{2} = 83.5$ [যদিও এই মান ডেটা তে নেই]।

যদি ডেটাসেটে ৯ জন শিক্ষার্থী হতো, যেমনঃ ১০ জন শিক্ষার্থীর গণিত পরীক্ষার নম্বর 85, 87, 58, 59, 93, 57, 94, 97, 82, 48 থেকে 48 বাদ দিয়ে ৯ জন শিক্ষার্থীর গণিত পরীক্ষার নম্বর দেখানো হতো তাহলেঃ- 85, 87, 58, 59, 93, 57, 94, 97, 82

median বের করার আগে increasing order এ sort করতে হবে।

57, 58, 59, 82, 85, 87, 93, 94, 97

এখন n = 9 (বিজোড়)।

তাহলে median = $\frac{9+1}{2}$ তম উপাত্ত = 5 তম উপাত্ত

Value (x_i)	57	58	59	82	85	87	93	94	97
তম	1	2	3	4	5	6	7	8	9

5 তম উপাত্ত হলো 85। তাই median = 85

প্রচুরক (Mode): Discrete Numerical data এর ক্ষেত্রে যার ফ্রিকুয়েন্সি বেশি বা যে সংখ্যাটার occurrence বেশি, সেই শ্রেণী হলো mode।

যেমনঃ 1, 3, 5, 2, 2, 5, 3, 1, 3, 5, 2, 3, 1, 1, 4, 4, 5, 2, 4, 3

ভ্যালু	ফ্রিকুয়েন্সি
1	4
2	4
3	5
4	3
5	4

এখানে 3 সবার চেয়ে বেশি occur করেছে। 3 এর ফ্রিকুয়েন্সি বেশি। তাই mode হলো 3

যদি কোনো ডেটাসেটে সর্বাধিক occurance বা সর্বাধিক ফ্রিকুয়েন্সি না থাকে, তাহলে সেই ডেটা তে mode নেই। যেমনঃ 1, 2, 3, 4, 5 – এই ডেটা তে mode নেই, কারণ প্রত্যেকের occurance বা ফ্রিকুয়েন্সি একবার করে।

Dispersion:

আমাদের কেন Dispersion এর প্রয়োজন? দুই বা ততোধিক ডেটাসেটে যখন mean, median বা mode একই মানের হয়ে যায়, তখন আমাদের Dispersion পরিমাপ করা লাগে।

যেমনঃ মনে করি দুটি ডেটাসেট দেওয়া আছে

Dataset 1: 2,3,3,4,5,5,5,6,7,8 Dataset 2: 1,4,4,4,5,5,5,6,6,9

Dataset 1 এর mean = 4.8; median = 5; mode = 5

একইভাবে, Dataset 2 এর mean = 4.8; median = 5; mode = 5

কিন্তু তারা তো আলাদা ডেটাসেট। তাই এক্ষেত্রে আমাদের Dispersion লাগে।

Dispersion চার প্রকার। যথাঃ ১) Range, ২) Variance, ৩) Standard Deviation, ৪) Interquartile Range (IQR)

Discrete Numerical Data এর ক্ষেত্রে Dispersion:

Range:

সর্বোচ্চ আর সর্বনিম্ন ভ্যালুর তফাৎ হলো Range। সূত্রঃ Range = Max – Min।

Dataset 1: 2,3,3,4,5,5,5,6,7,8 এবং Dataset 2: 1,4,4,4,5,5,5,6,6,9 – এই দুইটি ডেটাসেটের mean, median বা mode একই মানের। কিন্তু তাদের Range ভিন্ন। যেমনঃ Dataset 1 এর range = 8 – 2 = 6 কিন্তু Dataset 2 এর range = 9 – 1 = 8।

Variance: Variance দুই প্রকার।

যথাঃ ১) Population Variance (σ^2); এবং ২) Sample Variance (s^2)

ফর্মুলাঃ

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

$$\text{Sample Variation, } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

এখানে,
μ = Population mean
$\bar{x}$ = Sample mean
N = Population dataset এর উপাত্ত সংখ্যা
n = Sample dataset এর উপাত্ত সংখ্যা

যেমনঃ 2,3,3,4,5 – এই ডেটাসেটের variance বের করতে হবে। (শুধু variance বলা থাকলে, population–sample উল্লেখ না থাকলে sample variance বের করতে হবে।)

এখনঃ

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

এখানে,
n = 5

$$\bar{x} = \frac{2+3+3+4+5}{5} = 3.4$$

$$\Rightarrow s^2 = \frac{1}{5-1}\sum_{i=1}^{5}(x_i - 3.4)^2$$

$$\Rightarrow s^2 = \frac{(2-3.4)^2 + (3-3.4)^2 + (3-3.4)^2 + (4-3.4)^2 + (5-3.4)^2}{4}$$

$$\Rightarrow s^2 = 1.3$$

- <u>Discrete Numerical Data এর ক্ষেত্রে ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন তৈরির মাধ্যমে সহজে variance বের করা:</u>

1, 3, 5, 2, 2, 5, 3, 1, 3, 5, 2, 3, 1, 1, 4, 4, 5, 2, 4, 3 – এই dataset এর mean বের করা যেমন কঠিন তেমনি এর variance বের করাও কঠিন। এক্ষেত্রে আমরা ফ্রিকুয়েন্সি

ডিস্ট্রিবিউশন তৈরির মাধ্যমে সহজে variance বের করছি। ডিস্ট্রিবিউশনে row এর সংখ্যা r=5

ভ্যালু (x_i)	ফ্রিকুয়েন্সি (f_i)	$f_i x_i$	$f_i(x_i - \bar{x})^2$
1	4	4*1 = 4	$4 * (1 - 2.95)^2 = 15.21$
2	4	4*2 = 8	$4 * (2 - 2.95)^2 = 3.61$
3	5	5*3 = 15	$5 * (3 - 2.95)^2 = 0.0125$
4	3	3*4 = 12	$3 * (4 - 2.95)^2 = 3.3075$
5	4	4*5 = 20	$4 * (5 - 2.95)^2 = 16.81$
Total	$\sum_{i=1}^{r} f_i = 20$	$\sum_{i=1}^{r} f_i x_i = 59$	$\sum_{i=1}^{r} f_i(x_i - \bar{x})^2 = 38.95$

$$\text{Variance}, s^2 = \frac{\sum_{i=1}^{r} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{r} f_i - 1}$$

$$\Rightarrow s^2 = \frac{38.95}{59-1} = 0.672$$

$$\text{Mean}, \bar{x} = \frac{\sum_{i=1}^{r} f_i x_i}{\sum_{i=1}^{r} f_i} = \frac{59}{20} = 2.95$$

এভাবে আমরা Discrete Numerical Data এর ক্ষেত্রে ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন তৈরির মাধ্যমে সহজেই variance বের করতে পারি।

ফর্মুলাঃ

$$\text{Sample Variance}, s^2 = \frac{\sum_{i=1}^{r} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{r} f_i - 1}$$

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

Standard Deviation: Standard Deviation দুই প্রকার। যথাঃ

১) Population Standard Deviation (σ)

২) Sample Standard Deviation (s)

ফর্মুলাঃ

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

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

অথবা,

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

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

অর্থাৎ Variance এর বর্গমূল হলো Standard Deviation

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

আমরা কিছুক্ষণ আগেই 1, 3, 5, 2, 2, 5, 3, 1, 3, 5, 2, 3, 1, 1, 4, 4, 5, 2, 4, 3 – এই dataset এর Variance বের করেছিলাম, এবং এর Variance এর মান 0.672 ।

এখন এই ডেটাসেট এর Standard deviation = $\sqrt{0.672}$ = 0.82

(শুধু Standard deviation বলা থাকলে, population–sample উল্লেখ না থাকলে sample standard deviation বের করতে হবে।)

<u>কেন আমরা Sample Variance বা Sample Standard Deviation এর ক্ষেত্রে $n-1$ বা $\sum_{i=1}^{r} f_i - 1$ ব্যবহার করি?</u>

উত্তরঃ

এই $n-1$ বা $\sum_{i=1}^{r} f_i - 1$ ব্যবহার করা কে বলা হয় Bessel's correction। এই Bessel's correction করার কারণ হলো, যখন আমরা sample mean ব্যবহার করি, তখন sample variance একটু কম অনুমান করে। sample mean নিজেই sample-এর ভিত্তিতে গণনা করা হয়, তাই এটি "population mean" থেকে সামান্য দূরে থাকতে পারে। যদি আমরা সরাসরি n দিয়ে ভাগ করি, তাহলে আমরা প্রকৃত variance কে একটু কম গণনা করবো, যা bias তৈরি করবে। Bessel's correction প্রয়োগ করে sample variance-কে কম অনুমান করার সমস্যা সমাধান করা হয়।

তো আমরা এই পর্যন্তই Discrete Numerical Data এর ক্ষেত্রে সেন্ট্রাল টেন্ডেন্সি আর range, variance, standard deviation পর্যন্ত দেখেছি। পরবর্তীতে আমরা Discrete Numerical Data এর Interquartile Range নিয়ে দেখবো।