আমার বিজ্ঞকথা

টপিকঃ

ডেটা সায়েন্স

০০১-০১১ Numerical Data (পর্ব-৩)

০০১-০১১ Numerical Data (পর্ব-৩)

Discrete Numerical Data এর ক্ষেত্রে Dispersion:

আমরা এর আগে Discrete Numerical Data এর ক্ষেত্রে Dispersion এর Range, Variance, Standard Deviation পড়েছি। এবার আমরা Interquartile Range নিয়ে জানবো।

*কিন্তু Interquartile Range নিয়ে পড়ার আগে আমাদের **Percentile** আর **Quartiles** নিয়ে জানা অতীব প্রয়োজন।*

- Percentile:

Percentile হলো কোনো ডেটাসেটের এমন একটা পরিমাপ যেখানে দেখানো হয় যে কোনো উপাত্তের নির্দিষ্ট শতাংশ দ্বারা বোঝা যায় যে সেই ডেটাসেটের কতগুলো উপাত্ত সেই উপাত্তের নিচে বা সমান রয়েছে, আর কতগুলো উপাত্ত সেই উপাত্তের চেয়ে বড়। আরো সহজে বলতে গেলে, Percentile হলো এমন একটি পরিমাপ যা কোনো ডেটাসেটকে ১০০টি সমান অংশে বিভক্ত করে এবং নির্দিষ্ট মানের নিচে কত শতাংশ ডেটা রয়েছে তা নির্দেশ করে।

যেমনঃ দশম শ্রেণীর "ঘ" শাখার শিক্ষার্থী **আবির** উচ্চতর গণিত পরীক্ষায় 75 percentile (P_{75}) মার্কস পাইলো। এর মানে হলো ঐ "ঘ" শাখার 75% শিক্ষার্থী উচ্চতর গণিত পরীক্ষায় **আবিরের** চেয়ে কম মার্কস বা এর সমান মান পেয়েছে। আর বাদ বাকি 25% শিক্ষার্থী উচ্চতর গণিত পরীক্ষায় **আবিরের** চেয়ে বেশি মার্কস পেয়েছে।

P_{99} (99 Percentile) দ্বারা বুঝায় যে 99% ডেটা তার থেকে ছোট বা সমান কিন্তু (100 – 99)%= 1% ডেটা তার থেকে বড় বা সমান। P_1 (1 Percentile) দ্বারা বুঝায় যে 1% ডেটা তার থেকে ছোট বা সমান কিন্তু (100 – 1)%= 99% ডেটা তার থেকে বড় বা সমান।

Median এর ক্ষেত্রে P_{50} (50 Percentile) হয়।

ডেটাসেটকে ১০০ ভাগে ভাগ করলে প্রত্যেকটা ভাগ 1% করে হবে। ঠিক এরকমঃ

1%	1%	1%	.....	1%	1%	1%
P_1	P_2	P_3		P_{98}	P_{99}	P_{100}

Percentile নির্ণয় এর আগে ডেটাসেট কে increasing order এ সাজাতে হবে। এরপর লোকেশন বের করে Percentile বের করতে হবে। P_3 (3 percentile) এখানে "3" হলো percentile rank।

লোকেশন এর সূত্রঃ $L_k = \frac{k\,(n+1)}{100}$ তম ভ্যালু

এখানে k = percentile rank (যেমনঃ 22 Percentile হলে k = 22)

দুইভাবে Percentile বের করা যায়। Interpolation Method অনুযায়ী, অথবা Midpoint Method অনুযায়ী। তবে বর্তমানে সবচেয়ে ভালো নিয়ম হলো Interpolation Method।

উদাহরণঃ মনে করি, 30, 14, 24, 16, 32, 18, 22, 29, 17, 27, 20 – এই ডেটাসেটের P_{22} (22 Percentile) বের করতে চাই। তাহলে, প্রথমে একে increasing order এ sort করে পাইঃ- 14, 16, 17, 18, 20, 22, 24, 27, 29, 30, 32।

এখন একে তম হিসেবে দেখে নেই।

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

Midpoint Method অনুযায়ীঃ-

22 Percentile এর লোকেশন, $L_{22} = \frac{22(11+1)}{100} = 2.64$ তম

2.64 তম মানে 2 তম আর 3 তম এর মাঝামাঝি। 2 তম ভ্যালু হলো $x_2 = 16$ এবং 3 তম ভ্যালু হলো $x_3 = 17$

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

তাই,

$$P_{22} = \frac{16 + 17}{2} = 16.5$$

Interpolation Method অনুযায়ীঃ-

22 Percentile এর লোকেশন, $L_{22} = \frac{22(11+1)}{100} = 2.64$ তম

2.64 তম মানে 2 তম আর 3 তম এর মাঝামাঝি। 2 তম ভ্যালু হলো $x_2 = 16$ এবং 3 তম ভ্যালু হলো $x_3 = 17$

এখানে, $L_{22} = 2.64$ তম $= 2 + 0.64 = L_{22(\text{integer})} + L_{22(\text{fraction})}$

$L_{22(\text{integer})}$ হলো L_{22} এর পূর্ণসংখ্যা (integer) এর মান = 2

$L_{22(\text{fraction})}$ হলো L_{22} এর দশমিক সংখ্যা (fraction) এর মান = 0.64

এক্ষেত্রে আমরা $L_{22(\text{integer})}$ তম ভ্যালু ও $[L_{22(\text{integer})} + 1]$ তম ভ্যালু নিয়ে 22 percentile বের করবো। এক্ষেত্রে $L_{22(\text{integer})}$ তম ভ্যালু হলো $x_{L_{22(\text{integer})}}$; এবং $[L_{22(\text{integer})} + 1]$ তম ভ্যালু হলো $x_{[L_{22(\text{integer})}+1]}$।

$L_{22(\text{integer})} = 2$; এইজন্যেই 2 তম ভ্যালু হলো $x_{L_{22(\text{integer})}} = x_2 = 16$ আর (2+1) = 3 তম ভ্যালু $x_{[L_{22(\text{integer})}+1]} = x_{[2+1]} = x_3 = 17$ নিয়ে 22 percentile বের করবো।

তাই, $P_{22} = x_{L_{22(integer)}} + \{ L_{22(fraction)} \left(x_{[L_{22(integer)}+1]} - x_{L_{22(integer)}} \right) \}$

Or, $P_{22} = x_2 + \{0.64 \times (x_3 - x_2)\}$

Or, $P_{22} = 16 + \{0.64 \times (17 - 16)\}$

$\therefore P_{22} = 16.64$

এখানে, $L_{22(integer)} = 2$

$[L_{22(integer)} + 1] = 2 + 1 = 3$

সুতরাং $x_{L_{22(integer)}} = x_2 = 16$

$x_{[L_{22(integer)}+1]} = x_3 = 17$

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

দেখা যাচ্ছে Midpoint Method অনুযায়ী $P_{22} = 16.5$ কিন্তু Interpolation Method অনুযায়ী $P_{22} = 16.64$

সঠিক হলো $P_{22} = 16.64$, যেটা Interpolation Method অনুযায়ী বের করা হয়েছে। কারণ Midpoint Method খুব পুরোনো নিয়ম, কিন্তু Interpolation Method হলো আধুনিক নিয়ম।

percentile rank "k" হলে, ডেটাসেটের ডেটা সংখ্যা "n" হলে, Interpolation Method অনুযায়ী percentile বের করার সূত্রঃ-

$$P_k = x_{L_{k(integer)}} + \{ L_{k(fraction)} \left(x_{[L_{k(integer)}+1]} - x_{L_{k(integer)}} \right) \}$$

$$\text{যেখানে লোকেশন, } L_k = \frac{k\,(n+1)}{100} = L_{k(integer)} + L_{k(fraction)}$$

যদি x_i কে সহজ ভাষায় x_⟨i⟩ ধরি, তাহলে,

$$P_k = x\_\langle L_{k(integer)} \rangle + \{ L_{k(fraction)} (x\_\langle L_{k(integer)} + 1 \rangle - x\_\langle L_{k(integer)} \rangle) \}$$

এখন percentiles কেমনে বের করতে হয়, তা সহজেই শিখে গেছেন।

এবার চলুন আমরা সেই আগের ডেটাসেট 30, 14, 24, 16, 32, 18, 22, 29, 17, 27, 20 এর P_{15}, P_{25} , P_{50}, P_{75}, এবং P_{85} বের করি। (ডেটাসেট এর নাম দিলাম "ক")

প্রথমে এই "ক" ডেটাসেট কে increasing order এ sort করে পাইঃ- 14, 16, 17, 18, 20, 22, 24, 27, 29, 30, 32। n = 11।

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

15 Percentile এর লোকেশন, $L_{15} = \frac{15(11+1)}{100} = 1.8$ তম = 1 + 0.8

এক্ষেত্রে 1 তম ভ্যালু x_1 = 14; এবং (1+1) = 2 তম ভ্যালু x_2 = 16

সুতরাং 15 Percentile, $P_{15} = x_1 + \{0.8 \times (x_2 - x_1)\}$

$\therefore P_{15} = 14 + \{0.8 \times (16 - 14)\} = 15.6$

25 Percentile এর লোকেশন, $L_{25} = \frac{25(11+1)}{100} = 3$ তম = 3 + 0.0

এক্ষেত্রে 3 তম ভ্যালু x_3 = 17। সুতরাং 25 Percentile, $P_{25} = x_3 = 17$

যেহেতু পুরোপুরি আমরা তম পেয়েছি, তাই যেই তম পেয়েছি, সেই তম ভ্যালু 25 percentile হিসেবে গণ্য হয়েছে।

50 Percentile এর লোকেশন, $L_{50} = \frac{50(11+1)}{100} = 6$ তম = 6 + 0.0

এক্ষেত্রে 6 তম ভ্যালু x_6 = 22। সুতরাং 50 Percentile, $P_{50} = x_6 = 22$

50 Percentile (P_{50}) কে Median বলা হয়, তাই Median = 22

75 Percentile এর লোকেশন, $L_{75} = \frac{75(11+1)}{100} = 9$ তম = 9+ 0.0

এক্ষেত্রে 9 তম ভ্যালু x_9 = 29। সুতরাং 75 Percentile, $P_{75} = x_9 = 29$

85 Percentile এর লোকেশন, $L_{85} = \frac{85(11+1)}{100} = 10.2$ তম = 10 + 0.2

এক্ষেত্রে 10 তম ভ্যালু x_{10} = 30; এবং (10+1) = 11 তম ভ্যালু x_{11} = 32

সুতরাং 85 Percentile, $P_{85} = x_{10} + \{0.2 \times (x_{11} - x_{10})\}$

$\therefore P_{85} = 30 + \{0.2 \times (32 - 30)\} = 30.4$

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

Interpolation Issue in Percentile Estimation

এবার চলুন, আমরা সেই "ক" ডেটাসেট 30, 14, 24, 16, 32, 18, 22, 29, 17, 27, 20 এর P_1 এবং P_{99} নির্ণয় করি।

প্রথমে এই "ক" ডেটাসেট কে increasing order এ sort করে পাইঃ- 14, 16, 17, 18, 20, 22, 24, 27, 29, 30, 32। n = 11।

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

1 Percentile এর লোকেশন, $L_1 = \frac{1(11+1)}{100} = 0.12$ তম = 0 + 0.12

আমরা আগেই জেনেছি যে $L_k = \frac{k\,(n+1)}{100} = L_{k(integer)} + L_{k(fraction)}$ যেখানে $L_{k(integer)}$ হলো L_k এর পূর্ণসংখ্যা (integer) এর মান, এবং এই $L_{k(integer)}$ ধরেই $L_{k(integer)}$ তম ভ্যালু এবং $[L_{k(integer)} + 1]$ তম ভ্যালু নিয়ে percentile বের করবো। কিন্তু $L_{1(integer)} = 0$। এখন 0 তম ভ্যালু (x_0) তো বাস্তবে নেই, এখন কিভাবে আমরা 1 Percentile বের করবো?

এক্ষেত্রে আমরা লোকেশন এর সাথে ১ যোগ করবো। যখন লোকেশন ১ এর কম হবে, তখন ১ যোগ করতে হবে।

1 Percentile এর লোকেশন, $L_1 = \frac{1(11+1)}{100} + 1 = 0.12$ তম + 1 = 1.12 তম = 1 + 0.12

এক্ষেত্রে 1 তম ভ্যালু x_1 = 14; এবং (1+1) = 2 তম ভ্যালু x_2 = 16

সুতরাং 1 Percentile, $P_1 = x_1 + \{0.12 \times (x_2 - x_1)\}$

$P_1 = 14 + \{0.12 \times (16 - 14)\} = 14.24$

এবার আমরা চেষ্টা করি P_{99} নির্ণয় করার।

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

99 Percentile এর লোকেশন, $L_{99} = \frac{99(11+1)}{100} = 11.88$ তম = 11 + 0.88

আমরা আগেই জেনেছি যে $L_k = \frac{k\,(n+1)}{100} = L_{k(integer)} + L_{k(fraction)}$ যেখানে $L_{k(integer)}$ হলো L_k এর পূর্ণসংখ্যা (integer) এর মান, এবং এই $L_{k(integer)}$ ধরেই $L_{k(integer)}$ তম ভ্যালু এবং $[L_{k(integer)} + 1]$ তম ভ্যালু নিয়ে percentile বের করবো। কিন্তু $L_{1(integer)} = 11$ এবং $[L_{k(integer)} + 1] = 11 + 1 = 12$ । এখন 12 তম ভ্যালু (x_{12}) তো বাস্তবে নেই, এখন কিভাবে আমরা 99 Percentile বের করবো?

এক্ষেত্রে আমরা লোকেশন এর সাথে ১ বিয়োগ করবো। যখন লোকেশন n এর বেশি হবে, তখন ১ বিয়োগ করতে হবে।

অতএব, 99 Percentile এর লোকেশন,

$L_{99} = \frac{99(11+1)}{100} - 1 = 11.88$ তম $- 1 = 10.88$ তম = 10 + 0.88

এক্ষেত্রে 10 তম ভ্যালু x_{10} = 30; এবং (10+1) = 11 তম ভ্যালু x_{11} = 32

সুতরাং 99 Percentile, $P_{99} = x_{10} + \{0.88 \times (x_{11} - x_{10})\}$

$P_{99} = 30 + \{0.88 \times (32 - 30)\} = 31.76$

P_{100} এর মানঃ এবার চলুন সেই "ক" ডেটাসেট 14, 16, 17, 18, 20, 22, 24, 27, 29, 30, 32 এর P_{100} নির্ণয় করি। (Increasing order sorting দুইবার দেখিয়ে দিয়েছি, তাই আরেকবার সেই sorting না দেখিয়ে টেবিল টাই দেখালাম)

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
তম (i)	1	2	3	4	5	6	7	8	9	10	11
উপাত্ত	14	16	17	18	20	22	24	27	29	30	32

100 Percentile এর লোকেশন, $L_{100} = \frac{100(11+1)}{100} = 12$ তম। এই 12 তম থেকে 1 বিয়োগ দিলে হবে 12 - 1 = 11 তম।

এক্ষেত্রে 11 তম ভ্যালু x_{11} = 32। সুতরাং 100 Percentile, $P_{100} = x_{11} = 32$

আচ্ছা, 100 Percentile (P_{100}) এর মান সর্বশেষ ভ্যালু 32 হলে, 1 Percentile (P_1) কেন সর্বপ্রথম ভ্যালু 14 হলো না? কেন 1 Percentile (P_1) এর ভ্যালু 14.24 হলো?
এর কারণ হলো 100 Percentile (P_{100}) দ্বারা exactly সর্বোচ্চ ভ্যালু বোঝায়,
কিন্তু 1 Percentile (P_1) দ্বারা exactly সর্বনিম্ন ভ্যালু বোঝায় না।
তাই $P_{100} = 32$ এবং $P_1 \neq 14$

- Quartiles:

25 Percentile (P_{25}) কে first quartile (Q_1) বলা হয়। 50 Percentile (P_{50}) কে Second Quartile (Q_2) বলা হয়। 75 Percentile (P_{75}) কে Third Quartile (Q_3) বলা হয়।

যেহেতু 50 Percentile (P_{50}) কে Median বলা হয়, সে হিসেবে Second Quartile (Q_2) কেও Median বলা হয়।

100 Percentile (P_{100}) কে fourth quartile () বলা হয়, যদিও এর ব্যবহার এখানে করার প্রয়োজন নেই।

Quartile এর ক্ষেত্রে ডেটাসেটকে ৪ ভাগে ভাগ করলে প্রত্যেকটা ভাগ 25% করে হবে। ঠিক এরকমঃ

25%	25%	25%	25%
	P_{25}	P_{50}	P_{75}

আমরা কিছুক্ষণ আগেই সেই "ক" ডেটাসেট 30, 14, 24, 16, 32, 18, 22, 29, 17, 27, 20 এর P_{25}, P_{50}, এবং P_{75} বের করেছি। $P_{25} = 17$; $P_{50} = 22$; $P_{75} = 29$

সুতরাং "ক" ডেটাসেটের First Quartile, $Q_1 = P_{25} = 17$;
Second Quartile, $Q_2 = P_{50} = 22$; Third Quartile, $Q_3 = P_{75} = 29$

Decile ডেটাসেট কে 10 ভাগে বিভক্ত করে।
এক্ষেত্রে $D_1 = P_{10}$; $D_2 = P_{20}$; $D_3 = P_{30}$; $D_4 = P_{40}$; $D_5 = P_{10}$; $D_6 = P_{60}$; $D_7 = P_{70}$; $D_8 = P_{80}$; $D_9 = P_{90}$; $D_{10} = P_{100}$।
কিন্তু আমাদের Interquartile Range বের করার জন্য Decile এর প্রয়োজন নেই। যেহেতু percentile বের করা শিখেছি, তাহলে Quartile এর মতো Decile বের করাও অনেক সহজ।

Interquartile Range:

Interquartile Range কে সংক্ষেপে IQR বলা হয়। Interquartile Range হলো First Quartile (Q_1) আর Third Quartile (Q_3) এর মধ্যে পার্থক্য। Interquartile Range এর সুত্রঃ

$$IQR = Q_3 - Q_1$$

আমরা কিছুক্ষণ আগেই সেই "ক" ডেটাসেট 30, 14, 24, 16, 32, 18, 22, 29, 17, 27, 20 এর P_{25}, P_{50}, এবং P_{75} বের করেছি। $P_{25} = 17$; $P_{50} = 22$; $P_{75} = 29$

"ক" ডেটাসেটের First Quartile, $Q_1 = P_{25} = 17$;
Second Quartile, $Q_2 = P_{50} = 22$; Third Quartile, $Q_3 = P_{75} = 29$

সুতরাং "ক" ডেটাসেটের $IQR = Q_3 - Q_1 = 29 - 17 = 12$

অর্থাৎ, কিভাবে Interquartile Range নির্ণয় করা হয়, তা আমরা শিখে গেছি।

তো আমরা Discrete Numerical Data এর Interquartile Range বের করা শেখার মাধ্যমে আমরা পুরোপুরিভাবে Discrete Numerical Data এর Dispersion শেখা শেষ করেছি।
ধন্যবাদ।