আমার বিজ্ঞকথা

টপিকঃ

ডেটা সায়েন্স

০০১-০১২ Numerical Data (পর্ব-৪)

০০১-০১২ Numerical Data (পর্ব-৪)

আমরা Discrete Numerical Data এর Central Tendency, Dispersion এগুলা শেষ করেছি। এরপর আমরা Discrete Numerical Data এর দুইটি জিনিস শিখবো। Mean Absolute Deviation এবং Median Absolute Deviation। এদের দুইজনকেই MAD দ্বারা প্রকাশ করা হয়, যা confusion তৈরি করে। তাই আমি এদেরকে দুই ধরনের রূপ দিচ্ছি। Mean Absolute Deviation কে $\bar{x}_{ad}$, μ_{ad} দ্বারা দেখানো হলো এবং Median Absolute Deviation $\tilde{x}_{ad}$ কে দ্বারা প্রকাশ করা হলো। Sample এর ক্ষেত্রে Mean Absolute Deviation কে $\bar{x}_{ad}$ দ্বারা প্রকাশ করানো হয়েছে, আর Population এর ক্ষেত্রে Mean Absolute Deviation কে μ_{ad} দ্বারা প্রকাশ করানো হয়েছে।

বিঃদ্রঃ Median কে $\tilde{x}$ দ্বারা প্রকাশ করা হয়, কিন্তু আমরা এর আগে Median কে সেই চিহ্নে দেখিনি। কিন্তু আমরা এখন এটুকু জেনেছি যে Median কে $\tilde{x}$ দ্বারা প্রকাশ করা হয়।

Discrete Numerical Data এর ক্ষেত্রে Mean Absolute Deviation:

মনে করি, আমাদের কাছে দুইটি 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। আমরা জানি যে এক্ষেত্রে Dispersion দ্বারা তাদের ভিন্নতা প্রকাশ করা যায়, যেটা Mean, median, mode এ তাদের ভিন্নতা সম্ভব নয়। কিন্তু তাদের ভিন্নতা Mean Absolute Deviation আর Median Absolute Deviation দ্বারা প্রকাশ করা হয় *(যদিও*

Mean Absolute Deviation আর Median Absolute Deviation – এরা দুজনেই Dispersion এর অংশ, কিন্তু সাধারণভাবে আমরা Dispersion কে চার প্রকার দেখিয়েছিলাম।)

- <u>Mean Absolute Deviation এর ফর্মুলাঃ</u>

Sample এর ক্ষেত্রে Mean Absolute Deviation এর সূত্র –

$$\bar{x}_{ad} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

এখানে n হলো Sample dataset এর ক্ষেত্রে উপাত্তের সংখ্যা; $\bar{x}$ হলো Sample dataset এর Sample mean আর x_i হলো i ক্রমানুযায়ী উপাত্তসমূহ [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।]

Population এর ক্ষেত্রে Mean Absolute Deviation এর সূত্র –

$$\mu_{ad} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$$

এখানে N হলো Population dataset এর ক্ষেত্রে উপাত্তের সংখ্যা; μ হলো Population dataset এর Population mean আর x_i হলো i ক্রমানুযায়ী উপাত্তসমূহ [উদাহরণ আগেও দেখিয়ে দিয়েছি, যদিও সেটা Sample dataset এর ক্ষেত্রে ছিলো, কিন্তু একই মিথড Population Dataset এর ক্ষেত্রেও হবে, শুধুমাত্র সর্বশেষ (i=N) উপাত্তকে x_N দ্বারা প্রকাশ করা হবে]

তবে কোনো ডেটাসেটের ক্ষেত্রে Population বা Sample উল্লেখ না থাকলে আমরা সাধারণভাবে Sample mean বের করি, এবং Sample Mean Absolute Deviation বের করি।

এখন Dataset 1: 2, 3, 3, 4, 5, 5, 5, 6, 7, 8 এবং Dataset 2: 1, 4, 4, 4, 5, 5, 5, 6, 6, 9এই দুটো ডেটাসেটের mean, median, mode যেহেতু একই, এবং population নাকি sample তা

উল্লেখ করা নেই, সেই হিসেবে তাদের Sample Mean Absolute Deviation ($\bar{x}_{ad}$) বের করবো। উভয় ডেটাসেটের ক্ষেত্রে, n = 10।

এখানে Dataset 1: 2, 3, 3, 4, 5, 5, 5, 6, 7, 8 এর Mean $\bar{x} = 4.8$।

Dataset 1 এর Mean Absolute Deviation ($\bar{x}_{ad}$) হলো,

$$\bar{x}_{ad} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\Rightarrow \bar{x}_{ad} = \frac{1}{10}\sum_{i=1}^{10}|x_i - 4.8|$$

$$\Rightarrow \bar{x}_{ad} = \frac{|x_1 - 4.8| + |x_2 - 4.8| + \ldots + |x_{10} - 4.8|}{10}$$

$$\Rightarrow \bar{x}_{ad} = \frac{|2 - 4.8| + |3 - 4.8| + |3 - 4.8| + |4 - 4.8| + |5 - 4.8|}{10}$$

$$\frac{+|5 - 4.8| + |5 - 4.8| + |6 - 4.8| + |7 - 4.8| + |8 - 4.8|}{10}$$

$$\Rightarrow \bar{x}_{ad} = 1.44$$

Dataset 1 এর Mean Absolute Deviation 1.44

আবার Dataset 2: 1, 4, 4, 4, 5, 5, 5, 6, 6, 9 এর Mean $\bar{x} = 4.8$।

Dataset 2 এর Mean Absolute Deviation ($\bar{x}_{ad}$) হলো,

$$\bar{x}_{ad} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\Rightarrow \bar{x}_{ad} = \frac{1}{10}\sum_{i=1}^{10}|x_i - 4.8|$$

$$\Rightarrow \bar{x}_{ad} = \frac{|x_1 - 4.8| + |x_2 - 4.8| + \ldots + |x_{10} - 4.8|}{10}$$

$$\Rightarrow \bar{x}_{ad} = \frac{|1 - 4.8| + |4 - 4.8| + |4 - 4.8| + |4 - 4.8| + |5 - 4.8|}{10}$$

$$\frac{+|5 - 4.8| + |5 - 4.8| + |6 - 4.8| + |6 - 4.8| + |9 - 4.8|}{10}$$

$$\Rightarrow \bar{x}_{ad} = 1.34$$

Dataset 2 এর Mean Absolute Deviation 1.34

লক্ষ্য করে দেখুন, Dataset 1 আর Dataset 2 এর mean, median, mode অভিন্ন হলেও Dataset 1 এর Mean Absolute Deviation 1.44 এবং Dataset 2 এর Mean Absolute Deviation 1.34।

তাই আমরা বলতে পারি, দুই বা ততোধিক ভিন্ন Dataset এর mean, median, mode অভিন্ন হলে তাদের Mean Absolute Deviation ভিন্ন।

Discrete Numerical Data এর ক্ষেত্রে Median Absolute Deviation:

আবার সেই Dataset 1 আর Dataset 2 দ্বারা আমরা Median Absolute Deviation নির্ণয় করবো।

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

Median Absolute Deviation এর সুত্রঃ

$$\tilde{x}_{ad} = \text{median}(|x_i - \tilde{x}|)$$

$$\tilde{x}_{ad} = \text{median}(|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \ldots, |x_n - \tilde{x}|) \text{ [For sample]}$$

$$\tilde{x}_{ad} = \text{median}(|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \ldots, |x_N - \tilde{x}|) \text{ [For population]}$$

তবে কোনো ডেটাসেটের ক্ষেত্রে Population বা Sample উল্লেখ না থাকলে আমরা সাধারণভাবে Sample Median Absolute Deviation বের করি।

Dataset 1 আর Dataset 2 এর ক্ষেত্রে Population বা Sample উল্লেখ করা নেই। এদের দুইজনের median, $\tilde{x} = 5$। উভয় ডেটাসেটের ক্ষেত্রে, $n = 10$।

Dataset 1 এর Median Absolute Deviation $\tilde{x}_{ad(1)}$ হলে

$$\tilde{x}_{ad(1)} = \text{median}(|2 - \tilde{x}|, |3 - \tilde{x}|, |3 - \tilde{x}|, |4 - \tilde{x}|, |5 - \tilde{x}|, |5 - \tilde{x}|, |5 - \tilde{x}|, |6 - \tilde{x}|, |7 - \tilde{x}|, |8 - \tilde{x}|)$$

$$\Rightarrow \tilde{x}_{ad(1)} = \text{median}(|2 - 5|, |3 - 5|, |3 - 5|, |4 - 5|, |5 - 5|, |5 - 5|, |5 - 5|, |6 - 5|, |7 - 5|, |8 - 5|)$$

$$\Rightarrow \tilde{x}_{ad(1)} = \text{median}(3, 2, 2, 1, 0, 0, 0, 1, 2, 3)$$

$$\Rightarrow \tilde{x}_{ad(1)} = \text{median}(0, 0, 0, 1, 1, 2, 2, 2, 3, 3) \text{ [Sorting the data]}$$

এখন 0, 0, 0, 1, 1, 2, 2, 2, 3, 3 - এর median হলো 5 তম ও 6 তম ডেটার গড় (median এর সূত্র অনুযায়ী) $\frac{1+2}{2} = 1.5$

$$\Rightarrow \tilde{x}_{ad(1)} = 1.5$$

Dataset 1 এর Median Absolute Deviation হলো 1.5

অনুরূপভাবে Dataset 2 এর Median Absolute Deviation $\tilde{x}_{ad(2)}$ হলে

$$\tilde{x}_{ad(2)} = \text{median}(|1 - \tilde{x}|, |4 - \tilde{x}|, |4 - \tilde{x}|, |4 - \tilde{x}|, |5 - \tilde{x}|, |5 - \tilde{x}|, |5 - \tilde{x}|, |6 - \tilde{x}|, |6 - \tilde{x}|, |9 - \tilde{x}|)$$

$$\Rightarrow \tilde{x}_{ad(2)} = \text{median}(|1 - 5|, |4 - 5|, |4 - 5|, |4 - 5|, |5 - 5|, |5 - 5|, |5 - 5|, |6 - 5|, |6 - 5|, |9 - 5|)$$

$$\Rightarrow \tilde{x}_{ad(2)} = \text{median}(4, 1, 1, 1, 0, 0, 0, 1, 1, 4)$$

$$\Rightarrow \tilde{x}_{ad(2)} = \text{median}(0, 0, 0, 1, 1, 1, 1, 1, 4, 4) \text{ [Sorting the data]}$$

এখন 0, 0, 0, 1, 1, 1, 1, 1, 4, 4 - এর median হলো 5 তম ও 6 তম ডেটার গড় (median এর সূত্র অনুযায়ী) $\frac{1+1}{2} = 1$

$$\Rightarrow \tilde{x}_{ad(2)} = 1$$

Dataset 2 এর Median Absolute Deviation হলো 1

দেখাই যাচ্ছে, Dataset 1 আর Dataset 2 এর Median অভিন্ন হলেও Median Absolute Deviation আলাদা।