আমার বিজ্ঞকথা

টপিকঃ

ডেটা সায়েন্স

০০১-০০৯ Numerical Data (পর্ব-১)

০০১-০০৯ Numerical Data (পর্ব-১)

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।

Numerical Data মানে যে ডেটা Number আকারে থাকে। একে দুই ভাগে ভাগ করা হয়। যথা: ১) Discrete, ২) Continuous

Discrete: যে ডেটাকে গণনা করা যায়, তাকে Discrete বলে। যেমনঃ বক্সে ১২ টা পেন্সিল আছে। এখানে ১২ হলো Discrete কারণ পেন্সিলগুলো গুণে গুণে হিসাব করা করা হয়েছে।

Continuous: যে ডেটা গণনা করা যায় না, তাকে Continuous বলে। যেমন: কারো উচ্চতা গণনা করা যায় না। যেমনঃ একজন ব্যক্তির উচ্চতা ৫ ফুট, এটা গণনা করা যায় না। Continuous এ আমরা গণনা করতে পারি না, কিন্তু সেটা measure করতে পারি।

Numerica Data এর জন্য ফ্রিকুয়েন্সি ডিস্ট্রিবিউশনঃ

Discrete:

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।

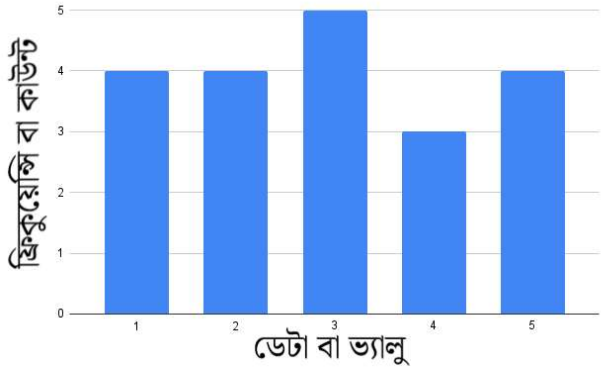
1, 3, 5, 2, 2, 5, 3, 1, 3, 5, 2, 3, 1, 1, 4, 4, 5, 2, 4, 3

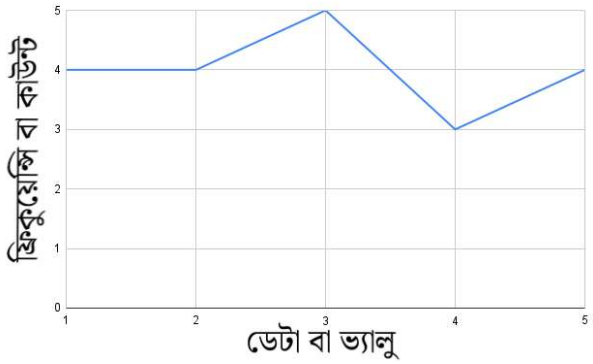
এখানে প্রত্যেক পল্লী বিদ্যুৎ কেন্দ্রে ভালো এক্সপেরিয়েন্স থাকা কর্মীদের সংখ্যা হয় 1 জন আছে, অথবা 2 জন আছে, অথবা 3 জন আছে, অথবা 4 জন আছে, অথবা 5 জন আছে। এখানে 1, 2, 3, 4, 5 হলো distinct values।

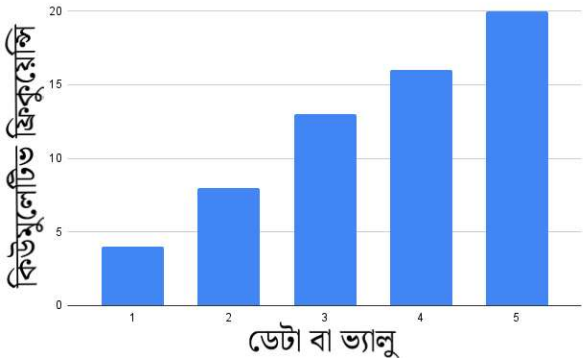
এক্ষেত্রে নিম্নে এর ফ্রিকুয়েন্সি টেবিল বানানো হলোঃ

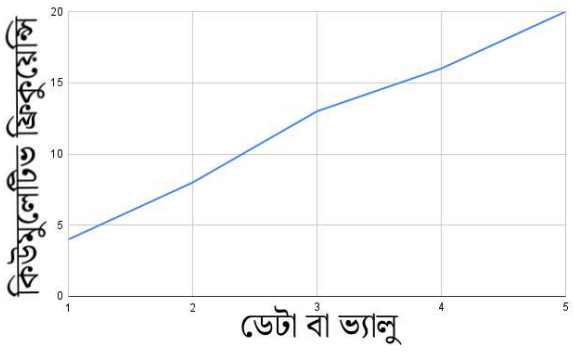
ভ্যালু	ট্যালি মার্ক	ফ্রিকুয়েন্সি	রিলেটিভ ফ্রিকুয়েন্সি	কিউমুলেটিভ ফ্রিকুয়েন্সি
1	||||	4	4/20 = 0.2	4
2	||||	4	4/20 = 0.2	4 + 4 = 8
3	|||| |	5	5/20 = 0.25	8 + 5 = 13
4	|||	3	3/20 = 0.15	13 + 3 = 16
5	||||	4	4/20= 0.2	16 + 4 = ***20***
Total		***20***	***1***	

আমরা এটা কে গ্রাফিক্যালি এভাবে দেখাইঃ (Courtesy: This visualization was generated using Google Sheets.)









এভাবেই আমরা ফ্রিকুয়েন্সি আর কিউমুলেটিভ ফ্রিকুয়েন্সি এর গ্রাফ বানাতে পারি।

Continuous:

Continuous এর ক্ষেত্রে ক্লাস ইন্টার্ভাল থাকে। এই ক্লাস ইন্টার্ভালে দুটি অংশ থাকে, তা হলো lower limit আর upper limit। যেমনঃ 10 – 15 (এখানে 10 হলো lower limit, আর 15 হলো upper limit).

ইন্টার্ভাল দুই ধরনের –

১) Exclusive Interval

২) Inclusive Interval

Exclusive Interval: এই ইন্টার্ভাল পদ্ধতিতে একটি ক্লাস ইন্টার্ভালের upper limit এ যে সংখ্যা থাকে, পরের ক্লাস ইন্টার্ভালের lower limit এ সেই সংখ্যাই থাকে। যেমনঃ

ক্লাস ইন্টার্ভাল	ফ্রিকুয়েন্সি
10 – 20	15
20 – 30	13
30 – 40	24
40 – 50	30
50 – 60	12

এখানে প্রথম ক্লাস ইন্টার্ভাল হলো 10 – 20, আর এর পরের ক্লাস ইন্টার্ভাল হলো 20 – 30। লক্ষ্য করে দেখুন, এখানে প্রথম ক্লাস ইন্টার্ভালের upper limit হলো 20 আর এর পরের ক্লাস ইন্টার্ভালের lower limit হলো 20.

এক্ষেত্রে মনে রাখতে হবে যে, Exclusive Interval এর সময় একটি ক্লাস ইন্টার্ভালের upper limit এ যে ভ্যালু থাকবে, সেই ভ্যালু কখনোই ঐ ইন্টার্ভালে অন্তর্ভুক্ত হবে না। যেমনঃ 10 – 20 এই ইন্টার্ভালে 20 ভ্যালু অন্তর্ভুক্ত হতে পারবে না, বরং 20 – 30 তে 20 ভ্যালু অন্তর্ভুক্ত হতে পারবে । অর্থাৎ 10 – 20 হলে 10 থেকে 19 পর্যন্ত ভ্যালু অন্তর্ভুক্ত হতে পারবে।

Inclusive interval: এই ইন্টার্ভাল পদ্ধতিতে একটি ক্লাস ইন্টার্ভালের upper limit এ যে সংখ্যা থাকে, পরের ক্লাস ইন্টার্ভালের lower limit এ সেই সংখ্যা থাকে না। যেমনঃ

ক্লাস ইন্টার্ভাল	ফ্রিকুয়েন্সি
11 – 15	11
16 – 20	6
21 – 25	9
26 – 30	10
31 – 35	12

এখানে প্রথম ক্লাস ইন্টার্ভাল হলো 11 – 15, আর এর পরের ক্লাস ইন্টার্ভাল হলো 16 – 20। লক্ষ্য করে দেখুন, এখানে প্রথম ক্লাস ইন্টার্ভালের upper limit হলো 15 আর এর পরের ক্লাস ইন্টার্ভালের lower limit হলো 16.

ডেটা সায়েন্সের ক্ষেত্রে Exclusive Interval সবচেয়ে ভালো পদ্ধতি।

Class width হলো lower limit আর upper limit এর ডিফারেন্স।

Class mark হলো কোনো ক্লাসের lower limit আর upper limit এর মধ্যে গড়।

Continuous Numerical ফ্রিকুয়েন্সি ডিস্ট্রিবিউশনের উদাহরণঃ

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

এই ডেটাসেট হচ্ছে Population Dataset।

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]*। Population dataset এর ক্ষেত্রে N হবে, Sample dataset এর ক্ষেত্রে n হবে।

$$k = 1 + 3.322 \log N$$

$$k = 1 + 3.322 \log (60)$$

$$k = 6.91 \approx 7$$

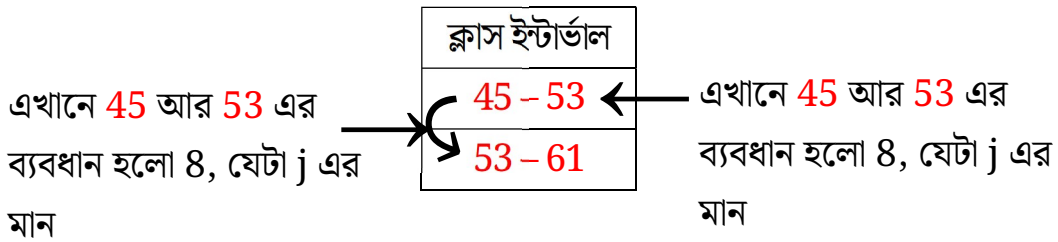
অর্থাৎ 7টি ক্লাস ইন্টার্ভাল তৈরি হবে।

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,



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

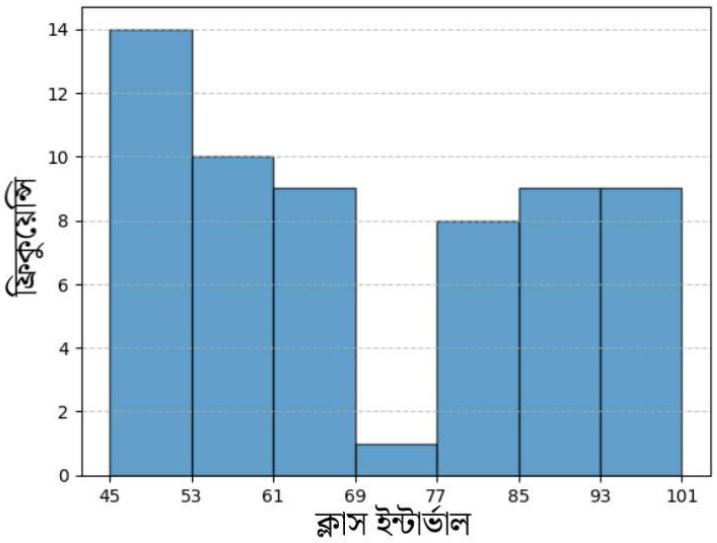
k = 7
এখানে 7 টি
ক্লাস ইন্টার্ভাল
রয়েছে

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।

আমরা এই Exclusive interval দেওয়া ফ্রিকুয়েন্সি ডিস্ট্রিবিউশন কে হিস্টোগ্রাম করে দেখাইঃ

(Courtesy: This visualization was generated using Python.)

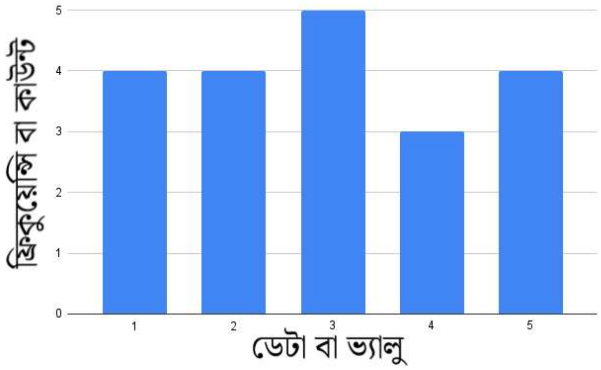


এভাবে আমরা Exclusive interval দেওয়া continuous ফ্রিকুয়েন্সি ডিস্ট্রিবিউশনের গ্রাফিক্যাল রিপ্রেজেন্টেশন বা ডেটা ভিজুয়ালাইজেশন করতে পারি।

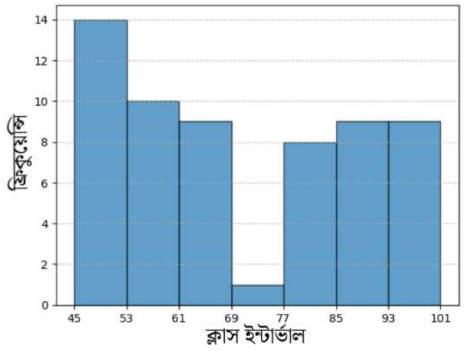
বার চার্ট আর হিস্টোগ্রাম, পার্থক্য কী?

বার চার্ট ব্যবহৃত হয় Categorical data, Discrete numerical data আর, Continuous numerical data in inclusive interval এর ক্ষেত্রে। হিস্টোগ্রাম ব্যবহৃত হয় Continuous numerical data in exclusive interval এর ক্ষেত্রে।

বার চার্ট এ দুটি স্তম্ভ (bin) এর মাঝে ফাঁকা থাকে। কিন্তু হিস্টোগ্রামে তা থাকে না।



(১)



(২)

এখানে (১) নম্বর চিত্র হলো বার চার্ট আর (২) নম্বর চার্ট হলো হিস্টোগ্রাম।