CSE 3107

06-03-2023

Applied Statistics and Queing Theory

↓

used in Machine Learning

Statistics is concearned with

- collecting → data collect করা
- Organizing → unnecessary data / Invalid data বাদ দেওয়া
- summarizing → single value যেটা পুরো distribution কে represent করবে।
- Presenting → Data present করা।
- analyzing → কি সিদ্ধান্তে পৌছাতে পারব। কি mean করে।

data

→ purpose: Make prediction

Inferential Statistics → বর্তমান data র উপর ভিত্তি করে future value predict করা। (Machine Learning এ use করা হয়)

Descriptive → data কে describe করা।

purpose: describing data → parameter

→ mean, median, std. deviation

Population

*population এর subset sample.

*population অনেক বেশি বলে করা সবসময় feasible হয় না।

Sample → mean, median, std. deviation → statistics

↳ Randomly sampling popular method.

*Data distribution যদি biased হয় তখন analyzing ভুল প্রমাণ নাও হতে পারে।

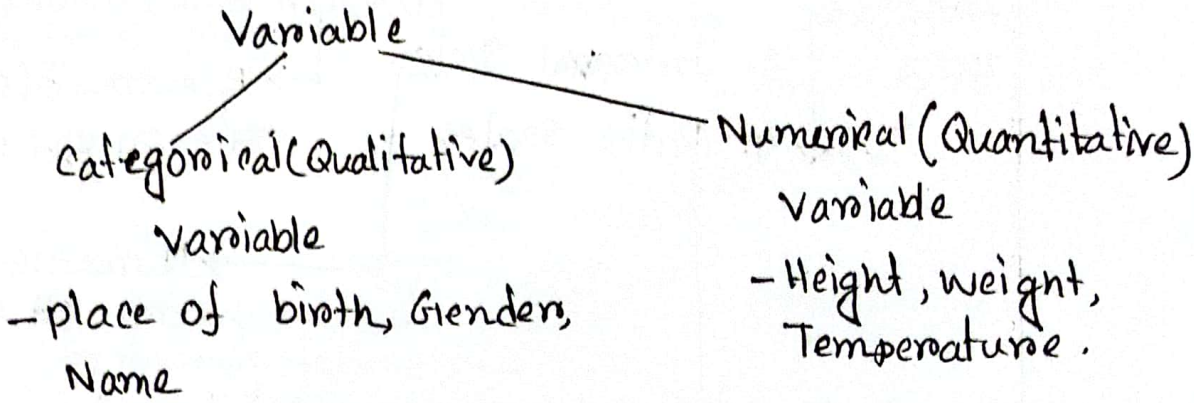
Paramater estimation : ● Sample বা statistics state থেকে parameter state এ calculate করা, ideal এর সাথে তুলনা করে।

Books : statistics by Murray R. Spiegel
Probability by Seymour Lipschutz.

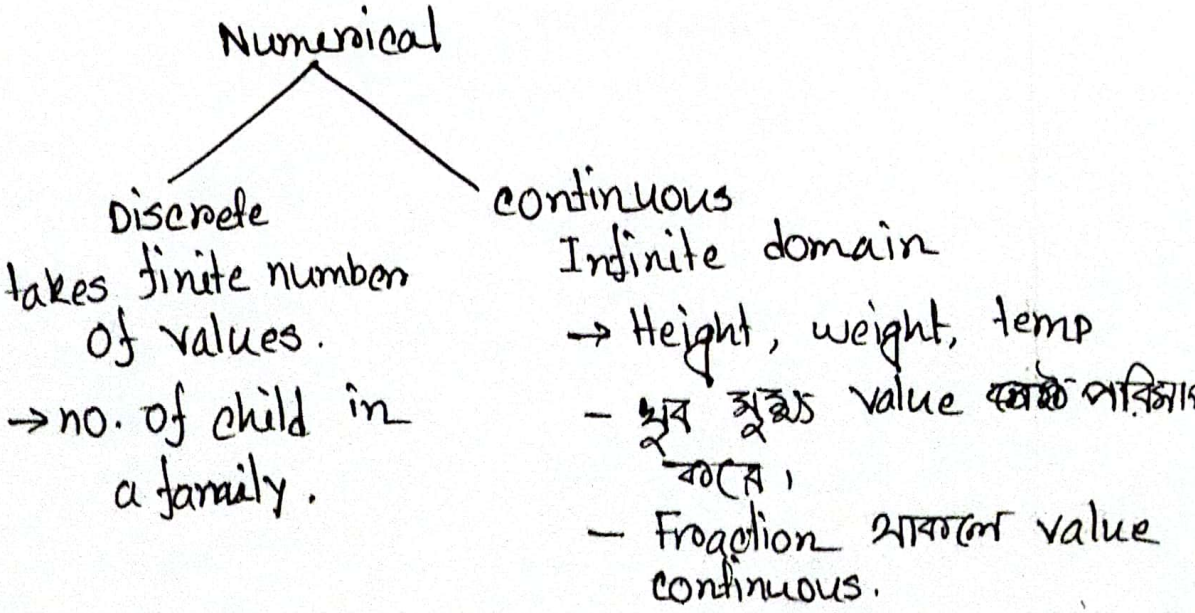
11.03.2023

Variable : A variable is an attribute that describes a person, place, thing or idea.

Variable changeable.



* Numerical থেকে categorical এ নেওয়া যায় কিন্তু কিছু data loss হয়।



*simplification এর জন্য continuous থেকে discrete এ আনা হয়।

* Level of Measurement :

* একটা variable থেকে কি পরিমান information পাবো।

Information বাড়বে

1. Nominal → Categorical, without order. কাজ info sort করা যাবে না, compare হবে না। → Nationality, Gender, Religion.

2. Ordinal → Categorical, with order. Level of order আছে। customer satisfaction (excellent, good, poor) → order আছে তাই, Agreement (agree, Neutral, disagree) সাজানো যায়।

3. Interval Scales → Numerical, NO absolute zero. zero দিয়েও কোনো value measure হয়। → zero. pH scale, Temp. in °F and °C. interval scale এ ratio নেওয়া valid না। অর্থাৎ reference নেওয়া যায় না।

4. Ratio Scales → Numerical, absolute zero exist. Temp. in K. একটা Reference value থাকে। সবাই compare করা easy.

* Univariate data : looks at only variable.
যখন 1টি data নিয়ে বা 1টি Feature নিয়ে কাজ করা হয়।

কোনো data type না।
কি পরিমান data নিয়ে কাজ হয়

* Multivariate data : looks at multiple variable.
একাধিক data or feature নিয়ে কাজ করা হয়। Classification করা যায়।

13-03-2023

Graphs

Pie chart:

* Categorical data visualization এর জন্য use করা হয়।

Bar graphs:

* Decreasing আকারে সাজানো হলে Pareto chart.

Multiple Bar Graph:

* একাধিক attribute একসাথে নিলে chart করতে হয়।

Line Graph:

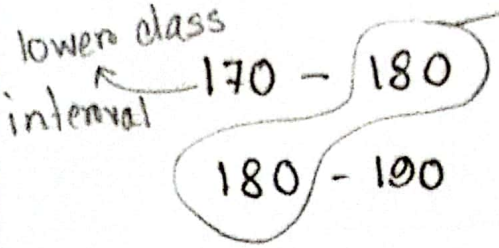
* progression show করার জন্য use করা হয়।

Frequency Distribution:

* কোনো particular value র frequency user কে দেয়,

(ক্লাস 1-10 এর বিদ্যুৎ-অবিদ্যুৎ)

Left-end convention:



a-b range এ input হবে

$a \leq x < b$;

*last number হলে include করতে হবে.

*এই convention আসছে overlap এর জন্য, ambiguity র জন্য.

class size = upper class interval − lower class interval.

*5-10 টা class interval নেওয়া উচিত.

Ogive রেখা

* Histogram and bar charts are not the same.
  ↓
  slide থেকে পড়তে হবে.

* Bin width related to class interval.

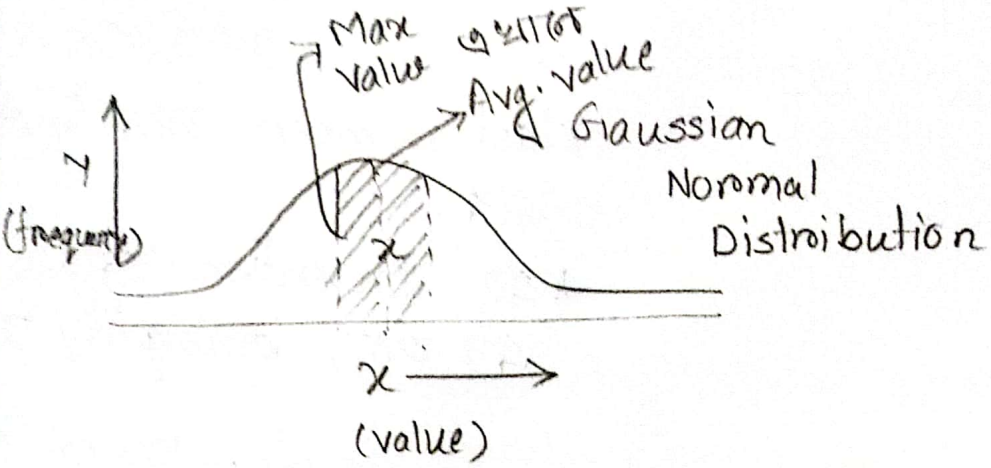
Scatterplot:

* Data distribution easily বোঝা যায়, ∞ data র মধ্যে co-relation থাকলে বুঝা যায়.

15-03-2023

Measure of Central Tendency:

* central কে কেন্দ্র করে



* Avg এর আশে পাশে maximum value কেন্দ্রীভূত থাকে।

যেমন: 60 জন student এর marks distribution

* Arithmetic mean
* Geometric "
* Harmonic "
* Quantiles "
* Median "

① Arithmetic Mean :

$x_1, x_2, x_3, \cdots x_n$

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

Tabular form এ থাকলে

$$\bar{x} = \frac{\sum f_i x_i}{n}$$

where, $n = \sum f_i$

x	f
10	2
20	3
30	1

from ungroup frequency distribution table

*Such of deviations from mean is equal to zero

$x_i \longrightarrow x_1, x_2, \cdots x_n$

$d_i \longrightarrow d_1, d_2, \cdots d_n$

$\sum d_i = 0$

$d_i = x_i - \bar{x}$

d_i = বিচ্যুতি, x_i = number, $\bar{x}$ = Avg.

Formal proof :

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

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

$\sum_{i=1}^{n} c = n$ ↓ a constant

$$= \sum_{i=1}^{n} x_i - n\bar{x}$$ ⟶ particular সংখ্যক number এর জন্য avg ($\bar{x}$) constant

$$= \sum_{i=1}^{n} x_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= 0$$

$x_1, x_2, x_3, \ldots x_n$

(A) ⟶ Assumed mean

$d_i = x_i - A$

Assumed mean actual mean এর সমান না হলে কখনই d_i 0 হবে না.

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{\sum A + \sum d_i}{n}$$

$$= \frac{nA + \sum d_i}{n}$$

$$\boxed{\bar{x} = A + \frac{\sum d_i}{n}}$$

$\frac{3+4+5}{3} = 4 \longrightarrow$ True mean

$A = 10 \longrightarrow$ Assume mean

$(3-10)+(4-10)+(5-10)$

$= -18$

$\bar{x} = 10 + \frac{-18}{3}$

$= 10 - 6 = 4$ true mean

For tabular / group data:

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i}$$

Example:

Class	Frequency, f_i	x_i	$d_i = x_i - A$	u_i
0-10	7	5	-20	-2
10-20	8	15	-10	-1
20-30	20	25	0	0
30-40	10	35	10	1
40-50	5	45	20	2

* A এর Value assume করা হয় মাঝের class এর xi এর মান

25

$n = \Sigma f = 30$

$$\bar{x} = A + \frac{\Sigma f_i d_i}{n}, \quad d_i = x_i - A$$

class width, $c = 10$ * di এখানে 10 এর সুনির্দিষ্ট এর reason class width

→ shortest Formula

$$\bar{x} = A + \frac{\sum f_i u_i}{n} \times c$$

→ class width

→ class distribution even হলে।

$$u_i = \frac{d_i}{c}$$

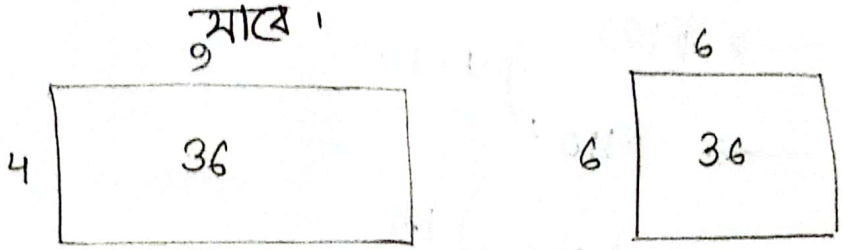
$$u_i = \frac{x_i - A}{c}$$

$$\bar{x} = 25 + \frac{(-14 - 8 + 0 + 10 + 10)}{50} \times 10$$

$$=$$

② Geometric Mean :

* Multiplicative in nature এরকম data র জন্য।
* এমন একটা value নিবো যেন একবার multiplication করলে তাকে একটা সংখ্যার square root করা যাবে।



$x_1, x_2, \cdots, x_n$

$$G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

$$= \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

$$\log G = \frac{1}{n} \log(x_1 x_2 \cdots x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log_0 x_n)$$

$$\boxed{\log G = \frac{\Sigma \log x_i}{n}}$$

$$\boxed{\log G = \frac{\Sigma f_i \log x_i}{\Sigma f_i}} \rightarrow \text{Group / Tabular data}$$

$$\therefore G = \text{anti log}\left(\frac{\Sigma \log x_i}{n}\right)$$

18-03-2023

* Microsoft stock gained 10% in year 1, 20% in year 2 and 30% in year 3. What is the average yearly rate of return?

Year 0 ⟶ $100
Year 1 ⟶ $110 ⟶ 1.1
Year 2 ⟶ $132 ⟶ 1.2
Year 3 ⟶ $171.60 ⟶ 1.3

→ consider 3 sides of cube

$G = \sqrt[3]{1.1 \times 1.2 \times 1.3} = 1.1972$ → Geometric mean of 3 side of equal

$100 \times 1.1972 = \$119.72$
$\$119.72 \times 1.1972 = \$143.33$
$\$143.33 \times 1.1972 = \$171.59$

$\$100 \times (1.1972)^3 \longrightarrow 19.72\%$

$1.1972 = (1 + 0.1972)$ → $100 पर

Arithmetic mean calculate करने पर $100 \times (1.2)^3 = \$172.8$ which is quite accurate but geometric mean gives the actual rate.

Harmonic Mean :

$$\frac{1}{H} = \frac{1}{N}\sum\frac{1}{x}$$

$$\boxed{H = \frac{N}{\sum\frac{1}{x}}}$$

2, 3, 4 → 3টা সংখ্যা

$$\therefore H = \frac{3}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 2.77$$

Harmonic series : $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5} \ldots$

* কোনো particular frequency play বন্ধ হলে সেটা একা play করা হয় না কিন্তু oven tune play হয় সেখানে একটা value তার previous ও next value র Harmonic mean. ও Harmonic series থেকে analysis করে Harmonic mean বের করা হয়।

For group / Tabular data :

$$H = \frac{\sum f}{\sum\frac{f}{x}}$$

* A car can travel 25 miles at 25 mph, 25 miles at 50 mph and 25 miles at 75 mph. What is the average velocity?

$$\text{Average velocity} = \frac{\text{Total distance}}{\text{total time}}$$

$t = \frac{s}{v} \longrightarrow = \frac{75}{\frac{25}{25} + \frac{25}{50} + \frac{25}{75}}$

$$= \boxed{\frac{3}{\frac{1}{25} + \frac{1}{50} + \frac{1}{75}}}$$ → Harmonic Mean এর formula.

$= 40.9$ mph

Arithmetic mean : $\frac{25+50+75}{3} = 50$ mph

↓

অনেক বেশি diff আসে.

* যখন উপরেরটা fixed (এখানে travel distance) এবং নিচেরটা variable (sub unit এ variable) হলে Harmonic mean use করা যাবে.

* swim 1 minute of freestyle at 3 km/h and
1 minute of butterfly at 2 km/h speed

↓

এখানে time fixed but distance variable ($\frac{\text{distance}}{\text{time}}$)
তাই arithmetic mean দিয়ে বের করা যাবে.

$$\frac{3+2}{2} = 2.5 \text{ km/h}$$

* swim one lap of free style at 3 km/h
and one lap of butterfly at 2 km/h.

↓

fixed distance

$$H = \frac{2}{\frac{1}{3}+\frac{1}{2}} = 2.4 \text{ km/h}$$

22 . 03 . 2023

Median :

Odd number : $\frac{N+1}{2}$ th

Even number : Avg. $\frac{N}{2}$ th and $(\frac{N}{2}+1)$th value

* at most 50% data median এর চেয়ে বেশি হবে.

2, 4, (5), 10, 12

Median এর চেয়ে বড় Number আছে $= \frac{2}{5} \times 100$

$= 40\%$

" " " ছোট " " $= \frac{2}{5} \times 100$

$= 40\%$.

* Avg outline দ্বারা sensitive . but median না .

2, 4, 5, 10, 1000

Median = 5

avg. = অনেক huge .

For even number

2, 4, (5, 10) 12, 25

↓

Avg নিয়ে হয়

Median = 7.5

Mode :

* frequency distribution এর mode বলা হয় সেই value যার frequency বেশি।

2 2 2 3 3 5

x	f
2	3
3	2
5	1

→ unimodal

mode = 2

2 2 2 3 3 3 → uniform (সবগুলোর frequency same)

2, 2, 2, 3, 3, 3, 5, 7, 10 → Bimodal.

x	f
2	3
3	3
5	1
7	1
10	1

→ Bimodal.

Median for group data:

weight	frequency	cummulative frequency
30-40	18	18
40-50	37	(55) p.c.f
median class → 50-60	45 → f_m	100
60-70	27	127
70-80	15	142
80-90	8	150

$N = \Sigma f = 150$

Median হবে $\frac{150}{2}$ বা 75 no. value র class এ আছে

$$\text{Median} = L + \frac{\frac{N}{2} - p.c.f}{f_m} \times c$$

L = lower class boundary of median class

f_m = frequency of median class

p.c.f = preceeding cummulative frequency / cummulative freq. of class before median

$(\Sigma f)_1$

c = class width

Lower class boundary (যেটাই true lower boundary)
49.5 ছোট value 40-49
" বড় " 50-59
Lower class interval

lower boundary	Weight
30-39	30-40
40-49	40-50
50-59	50-60

$\frac{50+49}{2} = 49.5$ এটা হলো (50-60) এর true lower boundary

$$\text{Median} = 49.5 + \frac{\frac{150}{2} - 55}{45} \times 10$$

$$= 53.494$$

*Median class খুঁজতে হলে প্রথম class এর frequency অনেক বেশি হয়।

Mode for group data:

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

- L = lower class boundary of modal class.

Δ_1 = Difference between modal class and pre modal class

$= f_m - f_1$

Δ_2 = difference between modal class and post modal class

$= f_m - f_2$

Weight	Frequency
30 - 40	18
40 - 50	37 → f_1
50 - 60	45 → f_m
60 - 70	27 → f_2
70 - 80	15
80 - 90	8

$\Delta_1 = f_m - f_1$
$= 45 - 37 = 8$

$\Delta_2 = f_m - f_2$
$= 45 - 27$
$= 18$

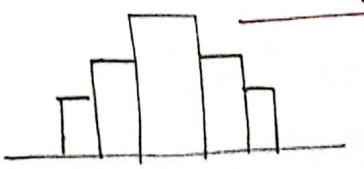
$$\text{Mode} = 50 + \frac{8}{8+18} \times 10$$

$$= 52.58$$

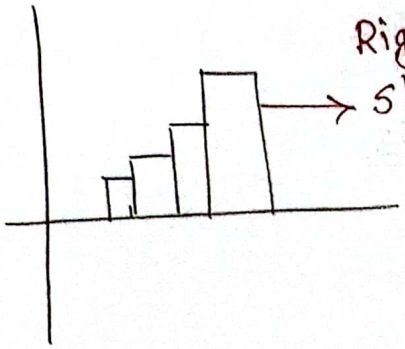
Empirical Relation of mode, Median and Mean :
For moderately skewed distribution

$$\text{Mode} = 3\,\text{Median} - 2\,\text{mean}$$

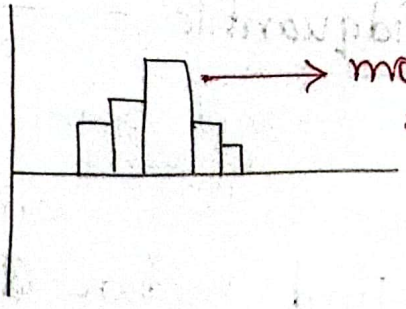
→ proof করা যাচ্ছে না, only observation করে পাওয়া



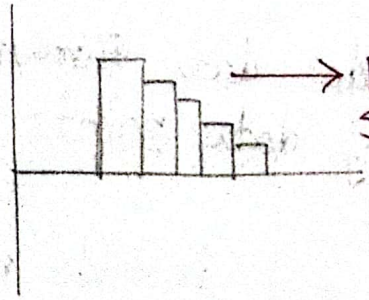
→ symmetrical distribution



Right (+ve) skewed distribution (এক side এ বেশি)



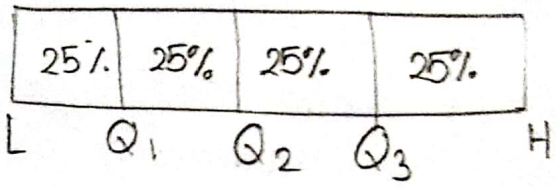
moderately skewed distribution



→ Left (-ve) skewed distribution

29.04.2023

Quantiles : Quantiles are values that divide total frequency into 4 parts



Q_1 ⟶ At most 25% data are smaller than Q_1 and at most 75% data are larger.

Q_3 ⟶ At most 75% data are smaller than Q_3 and ~~25~~ at most 25% data are larger

Q_2 ⟶ 50 - 50

$$IQR = Q_3 - Q_1$$

$$\text{Midquantile} = \frac{Q_1 + Q_3}{2}$$

↳ mid point যেখানে এটা আসে Q_2 না

Steps to find Quartiles:

* Find median Q_2
* Use median to divide data Further.
  - ⟶ For odd no of data points includes median in both halves.
  - ⟶ For even, do not include median in either.

2, 3, (5) → Q_1, 7, (8) → Q_2, 9, (10) → Q_3, 12, 15 → Sorted data

10, 12, 14, 15, 14, 16, 17, 18, 10, 19, 17, 17

Sorted: 10, 10, (12, 14) → $Q_1 = 13$, 14, (15, 16) → $Q_2 = 15.5$, 17, (17, 17) → $Q_3 = 17$, 18, 19

Grouped Data :

$$Q_i = L + \frac{\frac{iN}{4} - p.c.f}{f} \times h$$

L = Lower class of boundary of quartiles class.

$$Q_2 = L + \frac{\frac{N}{2} - p.c.f}{f} \times h$$

* Quartile median এর generalization

f = frequency of quartile class

*quartile class is identified by $\frac{iN}{4}$ th observation.

Profit	No. of Companies	c.f
20-30	4	4
30-40	8	12
40-50	18	30
50-60	30	60
60-70	15	75
70-80	10	85
80-90	8	93
90-100	7	100

$N = 100$

$$Q_2 = 49.5 + \frac{\frac{2 \times 100}{4} - 30}{30} \times 10$$

$$= 56.17$$

$$Q_1 = 39.5 + \frac{\frac{1 \times 100}{4} - 12}{18} \times 10$$

$$= 46.72$$

$$Q_3 = 59.5 + \frac{\frac{3 \times 100}{4} - 60}{15} \times 10$$

$$= 69.5$$

Percentiles :

$$P_i = L + \frac{\frac{i \times N}{100} - p.c.f}{f} \times h$$

$\frac{i \times N}{100}$ th ~~or~~ observation

$P_{75} = Q_3$

$P_{50} = Q_2$

$P_{25} = Q_1$

Deciles :

$$D_i = L + \frac{\frac{i \times N}{10} - p.c.f}{f} \times h$$

$i = 1, 2, 3, \ldots . 9$

$D_2 = P_{20}$

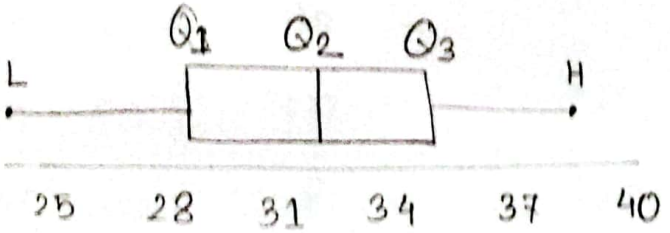
06 - 05 - 2023

Box Whisken plot / 5 number Summary :

L, H, Q_1, Q_2, Q_3 → Quantiles

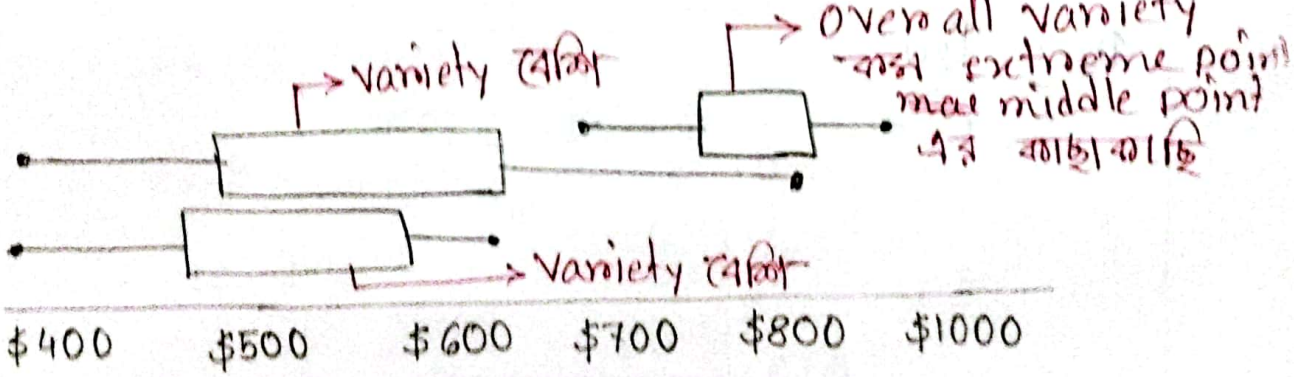
Data : 25, 28, 29, 29, 30, 34, 35, 35, 37, 38

$L = 25$, $H = 38$, $Q_2 = 32$, $Q_1 = 29$, $Q_3 = 35$



* Multiple box whisken plot কে একে অপরের against এ plot ~~করা~~ করে

* Data টা central point থেকে যত দূর ~~আছে~~ থাকবে data র variation তত বেশি।



* box যত বড় variation তত বেশি

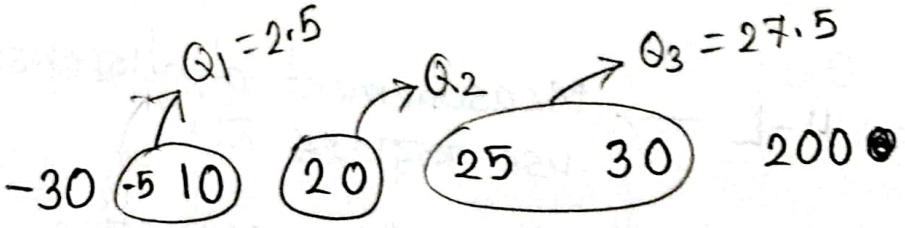
* ~~Odd happen~~ Outlier ~~Die~~ Detection using IQR

* Outlier ⟶ Particular value distribution follow না করলে তা outlier value.
Outlier highest value থেকে খানিক বের করা যাবে না, comparison লাগবে.

*Outlier বের করতে Inter Quantile Range use করা হয়.

$$IQR = Q_3 - Q_1$$

$x > Q_3 + 1.5 \times IQR$
$x < Q_1 - 1.5 \times IQR$
⟶ এই condition মিললে data কে outlier বলা হয়.



$Q_1 = 2.5$

$Q_2 = 20$

$Q_3 = 27.5$

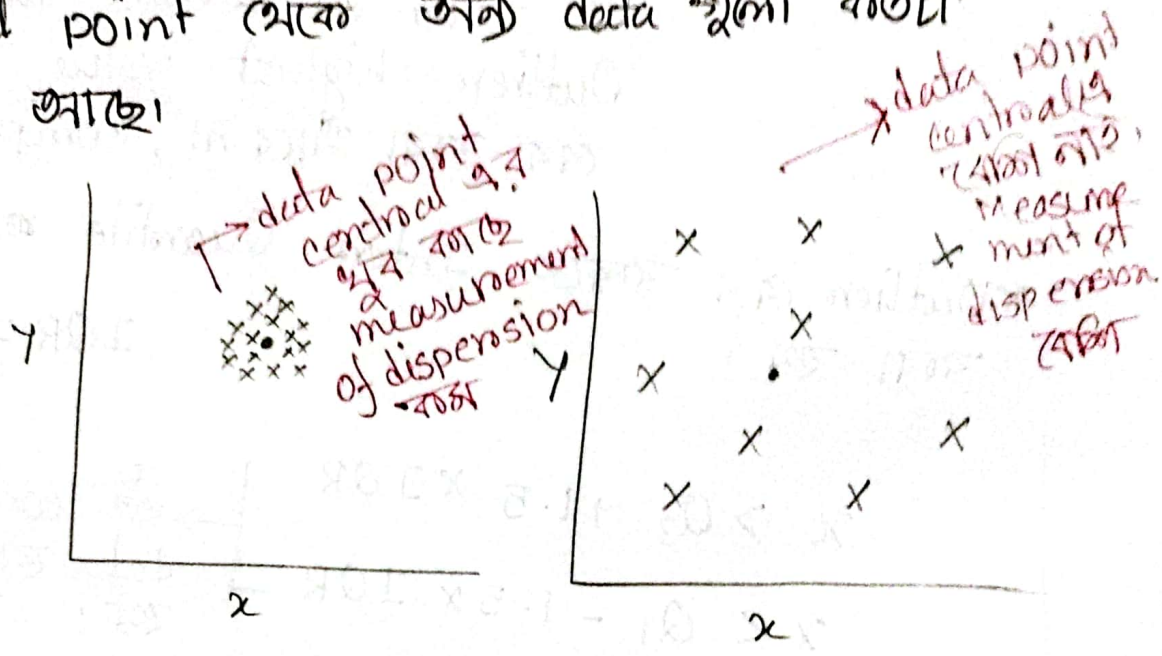
$IQR = 27.5 - 2.5 = 25$

$Q_3 + 1.5 \times IQR = 65$
$Q_1 - 1.5 \times IQR = -35$
⟶ 65 এর চেয়ে বড় ও -35 এর চেয়ে ছোট value outlier হবে.

∴ 200 এখানে outlier.

Measure of Dispersion

* Central point থেকে অন্য data গুলো কতটা ছড়িয়ে আছে।



* Range = H-L ⟶ Measurement of dispersion এ use করা হয় না।

* Measure of dispersion data র variation বোঝা যাবে। Range এর ক্ষেত্রে always এমন হয় না।

* ~~Quartile Deviation~~, Q.D = $\frac{Q_3 - Q_1}{2} = \frac{IQR}{2}$

A: 6, 46, 46, 46, 46

B: 6, 6, 6, 6, 46

C: 6, 10, 15, 35, 46

R = 46-6 = 40

A, B, C এর range same হলেও A তে B তে data variation বেশি

* Quantile Deviation, Q.D = $\frac{Q_3 - Q_1}{2}$ = $\frac{10.8}{2}$ h

Profit (In lakh)	# Of company	cf
20 - 30	4	4
30 - 40	8	12
40 - 50	18	30
50 - 60	30	60
60 - 70	15	75
70 - 80	10	85
80 - 90	8	93
90 - 100	7	100

$$Q_i = L + \frac{\frac{i \times N}{4} - p.c.f}{f} \times h$$

$$Q_2 = 49.5 + \frac{\frac{2 \times 100}{4} - 30}{10} \times 10 = 56.17 \text{ lakh}$$

$Q_1 = 46.72$ lakh

$Q_3 = 69.50$ lakh

$Q.D = 11.39$ lakh

Mean Deviation :

$$M.D = \frac{\sum_{i=1}^{N} |(x_i - \bar{x})|}{N}$$

For grouped data, $M.D = \frac{\sum f_i |x_i - \bar{x}|}{N}$

	f	x	$\lvert x_i - \bar{x} \rvert$
20-30			
30-40			

10 68 90 40

$$\bar{x} = \frac{208}{4} = 52$$

$$MD = \frac{|10-52| + |68-52| + |90-52| + |40-52|}{4}$$

$$= 27$$

Variance,

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$ → MD এর কাছাকাছি just square করা হয়।

Standard Deviation,

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

[Group data র ক্ষেত্রে নিতে হবে]

* Standrand deviation এ unit কম্পে যায়। যা consistent তাই use করা হয় বিভিন্ন operation. Analysis করার কাজে use হয়।
* বিভিন্ন problem এর distribution এ calculation এর জন্য use করা হয় variance.

Variance এর shortcut method for table,

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$= \frac{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\sum x_i^2}{n} - \frac{2\bar{x}\sum x_i}{n} + \frac{\sum \bar{x}^2}{n}$$

$$= \frac{\sum x_i^2}{n} - 2\bar{x}^2 + \frac{n\bar{x}^2}{n}$$

$$= \frac{\sum x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$\sigma^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2$$

→ Assumed mean এর instance

$$\sigma^2 = \frac{\Sigma f_d^2}{n} - \left(\frac{\Sigma f_d}{n}\right)^2 \quad \text{Variance.}$$

$$d_i = x_i - A$$

$$u_i = \frac{x_i - A}{c}$$

$$= \left[\frac{\Sigma f_u^2}{n} - \left(\frac{\Sigma f_u}{n}\right)^2\right] \times c^2$$

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

→ Bessel's correction

* ~~Bast~~ Biased estimation : (n-1) দিয়ে ভাগ দিলে biased হয় না.

08-05-2023

Coefficient of variation:
↳ unitless.

Absolute measurement: একক থাকে (unit)

Relative " : Comparison type করে & Different elements এর

Group 1 (lbs) : 120, 130, 140, 150, 160
$\mu_1 = 140$ $\sigma_1 = 14.14$

Group 2 (kg) : 60, 65, 70, 75, 80 $\mu_2 = 70$ $\sigma_2 = 7.07$

কোন group এ variation বেশি?

dispersion same হবে, unit আলাদা বলে দেখে মনে হচ্ছে double. এখানে dispersion দিয়ে measure করলে correct comparison আসবে না as unit আলাদা. এজন্য আলাদা unit এর data র জন্য coefficient of variation use করা হয়.

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

$$CV_1 = \frac{14.14}{100} \times 100\% = 10.10\%$$

$$CV_2 = \frac{7.07}{70} \times 100\% = 10.10\%$$

→ দুইটা group এর data র dispersion same.

68-95-99.7 Rule, Empirical Rule for Normal distribution!

Z-score: standardization/standard score:

$$z = \frac{x - \mu}{\sigma}$$

* একটা particular value mean এর কত standard deviation এগিয়ে বা পিছিয়ে সেটা measure করবে, z-score এটা relative measurement

Q. Suppose mean of an exam is 80 and std. deviation 6.3. A student achieved 92.5. Student এর standard score কতো?

$$z = \frac{92.5 - 80}{6.3} = \frac{1.97}{\downarrow}$$

Avg এর চেয়ে কতো পরিমান std deviation এগিয়ে

$\mu = 80$

$\sigma = 5$

student A → 87

$$z_A = \frac{x - \mu}{\sigma} = \frac{87 - 80}{5} = 1.50$$

$\mu = 73$

$\sigma = 6$

student B → 82

$$z_B = \frac{x - \mu}{\sigma} = \frac{82 - 73}{6} = 1.50$$

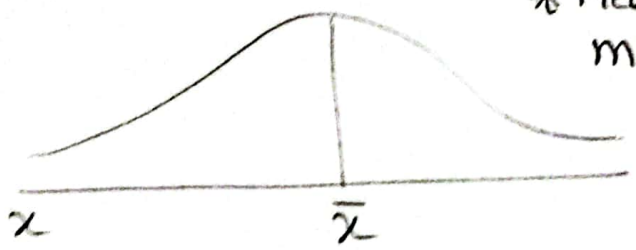
* comperitively student B result ভালো করেছে।

$x_i \longrightarrow$ 5, 7, 10, 12, 15, 16, 20, 25 $\quad \bar{x} = 13.75$

$\sigma = 6.24$

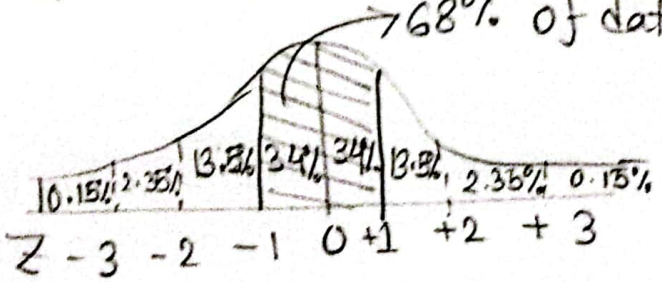
$z \longrightarrow$ -1.40, -1.08, -0.60, -0.28, 0.28, 0.36, 1.00, 1.80

Normal distribution!



* Mean, median and mode সমান

যদি normal distribution কে z scale এ রাখা হয় তাহলে standardized normal distribution



* কিছু data যদি এমন হয় যে normal distribution এর pattern follow করে তখন mean থেকে 1 standard deviation data (-1 to +1) র 68% অর্থাৎ 68% data থাকবে।

* -2 থেকে +2 range এ 95% data থাকবে.

* -3 থেকে +3 range এ 99.7% data বা area পাড়বে.

* এটি 68-95-99.7 Rule

* $z > +3$, $z < -3$ হলে সে data extreme outlier.

10-05-2023

data standardized করার পর (Z-score এ convert করার পর)

mean $\longrightarrow 0$

Variance / SD $= 1$

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ & σ $\longrightarrow$ parameters

$P(x)$ = Probability density function

standardized Normal distribution,

$$N(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-x/2}$$

$\int_{-\infty}^{0} N(x; 0, 1) = 0.5 \longrightarrow$ ∞ to 0 পর্যন্ত area under curve.

Q: Mean grade on final exam is 72 and standard deviation is 9. The top 10% student will get A+. What is the minimum mark required to get A+?

$$Z = \frac{x - \mu}{\sigma}$$

$\therefore x = \mu + Z\sigma$

$= 72 + Z \times 9$

$= 72 + (1.28 \times 9)$ [$Z = 1.28$ from Z-score table]

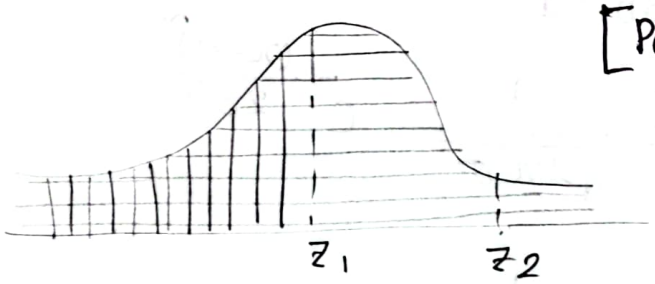
↓

Left sided Z-score table

row $\rightarrow$ 1.2
column $\rightarrow$ 0.08 } $\Rightarrow Z = 90\%$ almost

H.W:

range of marks in central 50%



[Percentile থেকে z-score বের করা calculator এ নাই] বাকিগুলো আছে

20.05.2023

Left sided z-value ⟶ percentile [Left, side এ যতগুলো area আছে]

↓

calculator এ করা যায়

1.0 ⟶ 84%

1.1

1.2

⋮

increment করলে area ও বাড়ে।

Chebyshev's Theorem

The proportion of any distribution that lies within K standard deviation of the mean is at least

$1 - \frac{1}{K^2}$, $K > 1$

↳ lower bound

K	1.7	2	2.5	3
$1 - \frac{1}{K^2}$	0.65	0.75	0.84	0.89

Normal distribution এ 2 এর জন্য 95% হলেও, এখানে বলেছে at least 75% এর মধ্যেই থাকবে। তাছাড়া সব ধরনের distribution এর জন্য empirial rule নাই। সেক্ষেত্রে এই theorem কাজে লাগে।

Moments :

$x_1, x_2, x_3, \ldots, x_n$

$$\overline{x_r} = \frac{\sum_{i=1}^{N} x_i^r}{N}$$

Central moment $\mu_r = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^r}{N}$

$r = 0, \mu_0 = 1$

$r = 1 \ldots \mu \longrightarrow$ more important

$r = 1, \mu_1 = 0 \longrightarrow$ sum of dispersim

$r = 2, \mu_2 = \sigma^2 \longrightarrow$ variance

$r = 3, \longrightarrow$ data এর symmetry / Assymtry measure করে।

$r = 4 \longrightarrow$ data এর tale এবং peak এর ratio ~~meas~~ measure করে।

Raw moment, $\mu'_r = \dfrac{\sum_{i=1}^{N} (x_i - A)^r}{N}$

A = Assumed mean

A reduces calculation

for grouped data

$$\mu'_r = \frac{\sum f_i (x_i - A)^r}{N}$$

$$\mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N}$$

$\mu_1 = 0$

$$\mu_1' = \frac{\Sigma(x_i - A)}{N} = \frac{\Sigma x_i}{N} - \frac{\Sigma A}{N}$$

$$= \bar{x} - \frac{NA}{A}$$

$$= \bar{x} - A$$

$$\mu_2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

$$= \frac{\Sigma\{(x_i - A) - (\bar{x}_i - A)\}^2}{N}$$

$$= \frac{\Sigma\{(x_i - A)^2 - 2(x_i - A)(\bar{x} - A) + (\bar{x} - A)^2\}}{N}$$

$$= \frac{\Sigma(x_i - A)^2}{N} - 2(\bar{x} - A)\frac{\Sigma x_i - A}{N} + \frac{\Sigma(\bar{x} - A)^2}{N}$$

$$= \mu_2' - 2\mu_1'\,\mu_1' + \frac{\Sigma\mu_1'^2}{N} \quad \text{(} \Sigma\mu_1'^2 \text{ is } N\mu_1'^2 \text{)}$$

$$= \mu_2' - 2\mu_1'^2 + \mu_1'^2$$

$$= \mu_2' - \mu_1'^2$$

$\therefore$ $\boxed{\mu_2 = \mu_2' - \mu_1'^2}$ → Relation bet^n raw and central moment

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$\mu_1' = \frac{\sum f_i(x_i - A)}{N}$$

$$= \frac{\sum (f_i u_i - h)}{N}$$

$$= \frac{\sum f_i u_i}{N} \times h$$

$$\therefore \mu_2' = \frac{\sum f_i u_i^2}{N} \times h^2$$

$$\mu_3' = \frac{\sum f_i u_i^3 \times h^3}{N}$$

$$\therefore \boxed{\mu_r' = \frac{\sum f_i u_i^r}{N} \times h^r}$$

$u_i = \dfrac{x_i - A}{h}$

$\Rightarrow x_i - A = u_i h$

Skewness, moments and kurtosis:

$\frac{\sum f_i u_i}{N} \times h$, $\frac{\sum f_i u_i^2}{N} \times h$; $u_i = \frac{x_i - A}{h}$

Assumed mean = raw moment

about the mean = central moment

আলাদা assumed mean বসলে আলাদা raw moment but আলাদা central moment same থাকবে.

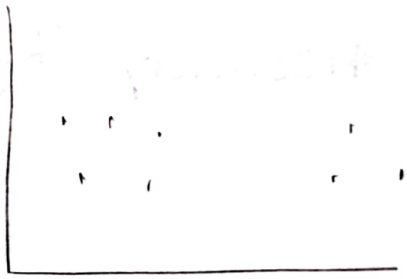
1st central moment = 0 ⟶ Always.

22-05-2023

Standardized moment:

* Transition Invariant:

  * value র সাথে constant terom যোগ করলে value shift হবে but dispersion (ছড়িয়ে যাবে না) হবেনা, standard deviation change হবেনা।



z score, $z = \frac{x - \bar{x}}{\sigma}$

std. deviation দিয়ে ভাগ

~~* Scale Variant~~

* standard deviation দিয়ে অনেক সময় data কে ভাগ করা যায়।

* distribution এর ক্ষেত্রে shape এর কিছু property follow করে ~~কোনো~~ moment, skewness

~~* St~~ * standard deviation দিয়ে কোনো data ভাগ দিলে unitless হবে। which will make the comparison easy and Dimentionless.

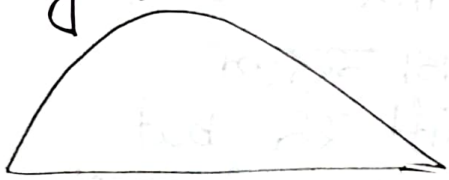
standardized moment:

$\frac{\mu}{\sigma}, \frac{\mu^2}{\sigma^2}, \frac{\mu}{\sigma^3}$ → std deviation দিয়ে ভাগ করা হলো।

4.hi

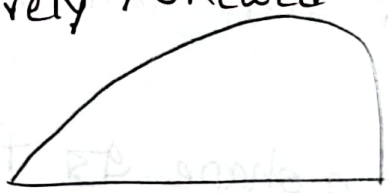
Skewness: distribution কোন দিকে data asymmetric থাকবে।

*Positively / Right Distribution:



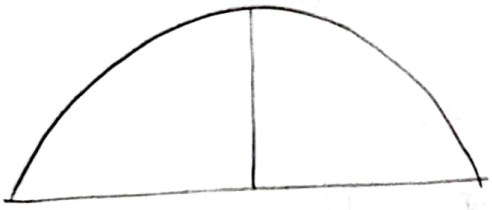
* Mean সবচেয়ে বড় হয় এরপর median and mode

*Negitively / Skewed Distribution:



* Mean সবচেয়ে ছোট mode বেশি।

*Symmetrical Distribution



* mean, mode, median same.

Coefficient of Skewness:

$$Skp = \frac{Mean - Mode}{\sigma}$$

→ Positive distribution (+)
→ Negative " (-)

↳ 1st skewness coefficient

* Skewness Factor same থাকতে হবে, size ছোট বড় করলেও, shape same থাকবে just scale up বা down হবে।

$$Skp = \frac{3(mean - median)}{\sigma}$$

↳ 2nd coefficient of skewness

* ~~Quantile দিয়েও~~
* Quantile and Percentile দিয়ে calculate করা যায়

skewness range $-\infty$ to ∞, Ideally value (-3 to 3) থাকে because of std. deviation

* Purely symmetric value হলে Skp = 0

3rd central moment → (+) হলে +ve skewness
→ (-) " (+) " "

$$\frac{\mu_3}{\sigma^3}$$

Karl Pearson skewness :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$\mu_2 \rightarrow$ Variance, কখনো positive না

β_1 always (+)ve হবে, refers to magnitude, direction refer করতে পারে না.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sigma^2)^{3/2}} = \frac{\mu_3}{\mu^{3/2}} = \sqrt{\beta_1}$$

↓ Karl Pearson moment coefficient of skewness

↙ 2nd central moment & = Variance

$\gamma_1 \rightarrow$ (+) হলে (+)ve skewed

$\gamma_1 \rightarrow$ (−) " (−)ve "

Kurtosis ⟶ 4th standard central distribution.

23·05·2023

$$\gamma_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}} = \frac{\mu_3}{\sigma^3}$$

* raw moment দিয়ে বের করা easy

Kurtosis:

* Skewness distribution এর দুইটা জিনিস নিয়ে concern

① Peak.

② Tail

Peak and Tail এর উপর Kurtosis depend করে।

Kurtosis distribution

① Leptokurtic:
Sharpen than normal distribution, Fat tail than normal distribution

② Mesokurtic:
Normal distribution এর মতো peak ও tail

③ Platykurtic:
Tail thinner than normal distribution,
Peak Flatter than normal distribution

* moment এর square power দিয়ে
* " " qubic " " Skewness বুঝায়
* " " forth " " Kurtosis (Peak & tail) বুঝায়

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum (x_i - \bar{x})^2\right]^2}$$

mesokurtic ⟶ $\beta_2 = 3$
leptokurtic ⟶ $\beta_2 > 3$
platykurtic ⟶ $\beta_2 < 3$

* moment generatic function ⟶ polynomical এর power থাকে।

Excess kurtosis, $\gamma_2 = \beta_2 - 3$

↳ platy to (-) হবে
lepto (+) ,,

↳ kurtosis বের করতে হলে এটা হবে (using 4th and 2nd central moment)

Corelation

* একটা বৃদ্ধি করলে আরেকটায় কি change করে।

Positive Negative:

Simple & multiple:

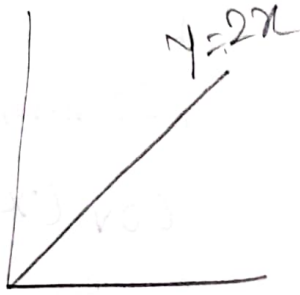
(i) simple → দুইটা variable এর মধ্যে

(ii) Multiple → কয়েকটার মধ্যে

Linear: Linear সম্পর্ক আছে কিনা। একটা change করলে আরেকটা ও scale এই change হচ্ছে কিনা

Method of estimation corelation:

* perfect positive corelation:
  perfectly একটা line এ fit করবে তাল negative



* perfect negative corelation:
  perfectly একটা line এ fit করবে তবে তাল (-) হবে

* Highly degree of positive

* High degree of Negative

* No corelation !

কোনো সম্পর্ক নাই একটা বাড়লে আরেকটা কমবে না বাড়বে ঠিক নাই ।

* corelation হলেই যে cause and effect হবে এমন না ।

* আবার Causaion হলেই corelation হবে ।

Covariance :

$$Cov(x,y) = \sigma_{x,y} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

$$\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{N}$$

↓ Variance

* দুইটা variable এর জন্য calculate করা হয় ।

* Covariance দুইটা ~~Van~~ ~~value~~ variable এর corelation নির্নয় করে ।

Covariance

(+) positively corelated

0 No corelation

(-) Negatively corelated