Date

Python Data Analytics

→ Analysis means evaluating the past data to figure out why/ how it happened whereas analytic is exploring possible future by using past.

Descriptive analysis → What happened → eg. company yearly report

Diagnostic analysis → Why it happened

Predictive analytics → What will happen

Prescriptive analysis → What to do for better results.

- Level of Data measurement

1. Nominal Data → classifies into distinct categories, no ranking implied
   eg. gender, marital status

2. Ordinal scale → Distint categories in which ranking is implied.
   eg. student grades (A, B, C, D ..).

3 Interval scale → ordered scale in which difference b/w measurements have meaningful value but

Spiral

Date _____

measurements dont have true zero pt.
eg. tempreture, potential etc.

4 Ration scale → difference b/w measurement is meaningful quantity and they have true zero point
eg. weight, age, salary.

Usage potential max → ~~nominal~~ Ratio
min → nominal

~~Data~~ ~~visuals~~ Visualisation

1. → Ogive curve
→ joining the mid-points of cumulative freq ~~histogram~~ bar graph.

2. Frequency polygon
→ join mid-points of histogram to observe data trend.

3. Relative ogive
→ normalised ogive curve

4. Pareto Chart
→ Used to identify importance of a variable

5. Scatter plot
→ To analyse trend of a variable wrt another variable

Spiral

Date _____

Central tendencies → Arithmatic mean
→ Weighted mean
→ Median
→ Percentile

Dispersion → Skewness
→ Kurtosis
→ Range
→ Interquartile range
→ Variance
→ Standard score
→ Coefficient of variation

1. Central Tendencies

→ mean of grouped data = $\frac{\Sigma f_i M_i}{\Sigma f_i} = \mu$

M_i = mid-point of class interval

→ eg. interval ⇒ 20 - 30

$M_i = 25$

→ Median of grouped data

$$= \frac{N/2 - cf_p}{f_{med}} (w) + L$$

L → Lower limit of median class
cf_p → cumulative freq of prev. class
f_{med} → freq. of median class.
w → width of med class
N → total of frequencies

median class is one containing $N/2$ cumulative freq.

Spiral

Date _____

→ mode of grouped = $L_{mo} + \left[\frac{d_1}{d_1+d_2}\right] w$

modal class has highest freq.

d_1 → diff. b/w modal class freq. and prev class freq.

d_2 → diff. b/w modal and next class freq.

w → class width.

Note

→ If data Left / Right skewed use median but if symmetric use whatever u want.

→ $i = \frac{P(n)}{100}$ pth percentile location.

2. Dispersion.

- Quartiles eg (106, 109, 114, 116, 121, 122, 125, 129)

$i_1 = \frac{25}{100}(n)$ ⇒ [redacted]

$Q_1 = \frac{109+114}{2} = 111.5$

$i_2 = \frac{50}{100}(n) = \frac{50}{100}(8) = 4$; $Q_2 = \frac{116+121}{2} = 118.5$

$i_3 = \frac{75}{100}(8) = 6$; $Q_3 = \frac{122+125}{2} = 123.5$

- interquartile range

$= Q_3 - Q_1$

- Popular variance

$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$

X → i^{th} sample [redacted]

μ → mean

N → no. of samples

→ $\sqrt{\sigma^2} = \sigma$ = std. deviation

Spiral

Date _____

- Sample varience

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

s = sample standard deviation.

~~Varience~~ Note

→ Varience ∝ Risk

→ 68% value lie within 1 std. deviation of mean in bell shaped curve.

- Chebyshev's theoram

→ [illegible] ; [for $k > 1$] fraction

of all data will lie

→ The fraction of data lying with the k std deviations of the mean is at least

$$1 - \frac{1}{k^2} \quad \text{for } k > 1$$

→ This is the lower bound

- Coeff of varience

→ relative dispersion

$$C.V. = \frac{\sigma}{\mu} \times 100\ \%$$

→ lower CV → better option

Spiral

Date _____

- Varience std for grouped data

1 Population

$$\sigma^2 = \frac{\sum f(M-\mu)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

2 Sample

$$s^2 = \frac{\sum f(M-\bar{x})}{n-1}$$

$$s = \sqrt{s^2}$$

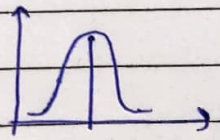
Note ~~ig so~~

→ In sample we divide by n-1 to provide more unbiased estimate of population varience (unbiased to sample)

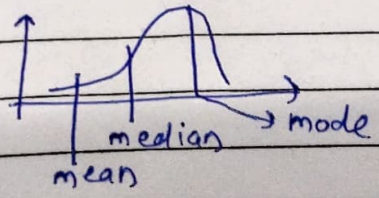
→ Above is called bessel correction

- Skewness

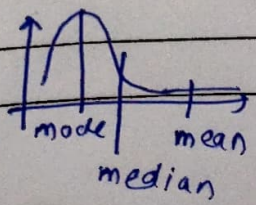
1. Symmetrical



2. Left skewed (-ve skewed)



3. Right skewed



Spiral

Date _____

- coefficient of skewness

$$S = \frac{3(\mu - Md)}{\sigma}$$

$S < 0$ → -ve skew ; $S = 0$ → symmetric

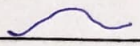
$S > 0$ +ve skewed

- Kurtosis

Leptokurtic → high thin

Meso kurtic → normal in shape

Platy kurtic → flat and spredout



NOTE

Median = Q_2