PL03 : Descriptive Statistics II

Video1 : Understanding Z-scores

- Z-Score and Standard Deviation
- Z-score is a measure of DISTANCE from the MEAN.

$\bar{x} = 50$ & $\sigma = 5$; If value = 55, then Z-score = +1.0
If value = 40, then Z-score = -2.0

$$z = \frac{\text{data point} - \text{mean value}}{\text{standard deviation}}$$

$$\boxed{z = \frac{x - \bar{x}}{s}} \approx \boxed{z = \frac{x - \mu}{\sigma}}$$

(sample) (population)

Video 2: Standard Deviation

- Greater variability means the values are more spread out.

- How far is each data point from the mean ? (distance)
  ↳ This is the question, Variance & Standard Deviation will help us answer.

- The standard deviation is just the positive square root of the variance.

Mean = $\bar{x}$; Standard deviation = σ ; Variance = σ^2

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$ ← Formula

- e.g class 1 , $\sigma = 7.9$, data points are spread away, farther
  class 2 , $\sigma = 4.74$, means data points are closer to the mean.

- COEFFICIENT OF VARIATION

↳ $$\left(\frac{\text{Standard deviation} \times 100}{\text{Mean}} \right) \%$$

- The coefficient of variation is a relative measure of variability.
- Usually expressed as Percentage
- It measures the standard deviation relative to the mean
  ⇒ How large is σ relative to μ ?

percentage

- Since it is a ratio, it is helpful to compare data having different means & standard deviations.

- By the same property, it is also UNIT-INDEPENDENT.

e.g Coefficient of variation, case 1 = 9.29%
" " " , case 2 = 5.58%

The C.O.V is much smaller relative to the mean in class 2.

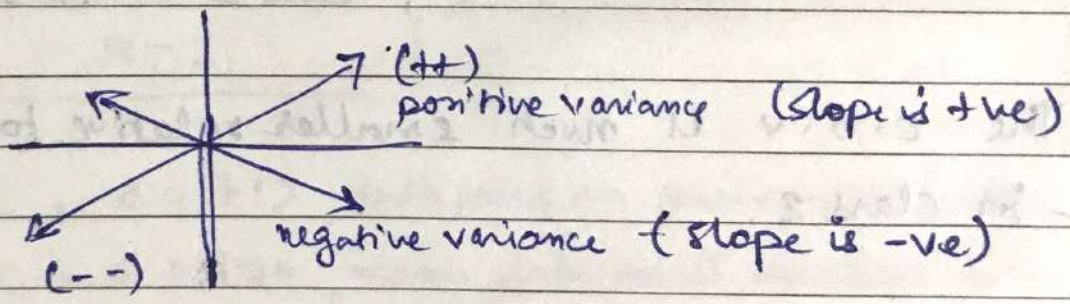
Video 3 : NFL field Goals - I
Video 4 : NFL field Goals - II

Video 5: Bivariate Relationships : COVARIANCE

- COVARIANCE – simplified CO-vary
  i.e, vary together
  – one variable up, the other is also up.

- LINEAR RELATIONSHIP – 'One change, the other change'
- How do variables behave as a PAIR.
- COVARIANCE – A descriptive measure of the linear association between two variable.
  - Positive value, Increasing linear relationship
  - Negative value, decreasing linear relationship

- Graph:



- FORMULA:

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample Covariance | Population Covariance

- COVARIANCE - A descriptive measure of the linear association between two variables

↘ - Direction = Sign + ↗

Video 6 : Bivariate Relationships : The COVARIANCE MATRIX

- COVARIANCE does not tell about STRENGTH of relationship it tells, about Increasing / Decreasing linear relationship.
- eg 4 variables ; x_1, x_2, x_3, x_4
  Sample ; $n = 20$

Statistics of Interest : Mean | Variance | Standard Deviation

- MATRIX of SCATTER PLOTS
  - A figure that plots each variable against every other variable

	x_1	x_2	x_3	x_4
x_1	$Var(x_1)$	$Cov(x_1, x_2)$	$Cov(x_1, x_3)$	$Cov(x_1, x_4)$
x_2		$Var(x_2)$	$Cov(x_2, x_3)$	$Cov(x_2, x_4)$
x_3			$Var(x_3)$	$Cov(x_3, x_4)$
x_4				$Var(x_4)$

- The Diagonal of a Covariance Matrix provides the variance of each variable
- we can get σ, by finding root of S^2

- Either side of Diagonal is Duplicated.

Video 7: BIVARIATE RELATIONSHIPS : Understanding CORRELATION

COVARIANCE vs CORRELATION.

- Provides the DIRECTION of the linear Relationship between two variables. (positive, negative, near zero)
- Provides DIRECTION as well STRENGTH

- Has NO UPPER BOUND or LOWER BOUND, and its SIZE is dependent on the SCALE of the variables
- Correlation is ALWAYS BETWEEN -1 & +1, and its scale is INDEPENDENT of the scale of the variables themselves.

→ Since Correlations is Independent, the comparison of variables which are different regardless of their UNITS
eg. Temperature vs Energy Consumption.

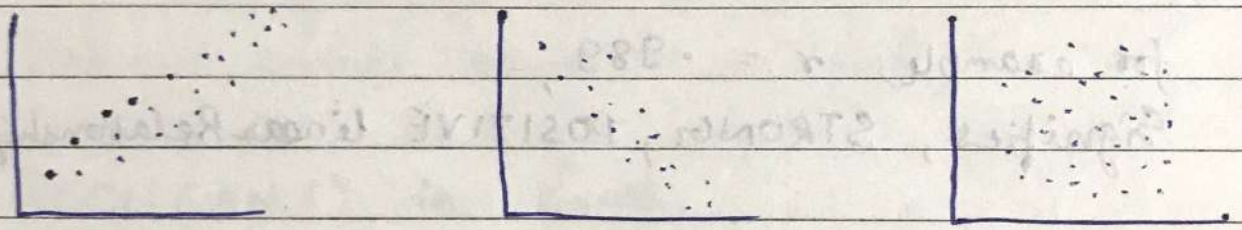
- Covariance is NOT STANDARDIZED. (in other words)
- It is STADARDIZED (like Z-score is a standarized measure of Variation)

→ like Standardization allows to Compare Variables that are measured using different scales.

- CORRELATION CAVEATS

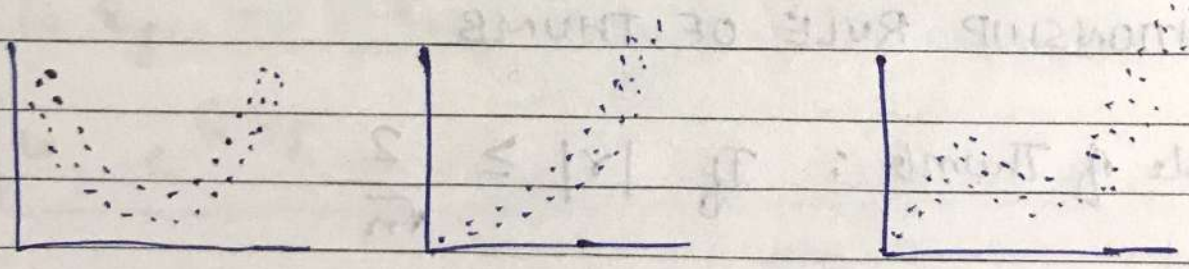
(i) Before computing Correlation, always look at Scatter plot
(ii) Correlation is only applicable to LINEAR RELATIONSHIPS.
(iii) Correlation is NOT CAUSATION. (dog barking vs moon please)
(iv) Correlation Strength does not mean, it is statistically significant.

- General Correlation Patterns



near +1 (Slope is +ve) | near −1 (Slope is −ve) | near 0 (No slope)

- Non-linear Relationships



Quadratic | Exponential | Polynomial

∴ Always look at Scatter Plot of the Data.

- CORRELATION FORMULA :

r is called the (Pearson) correlation coefficient.

$$r = \frac{\text{Covariance}(x,y)}{\text{Standard Deviation}(x) \times \text{Standard Deviation}(y)}$$

$$r = \frac{\text{Cov}(x,y)}{S_x S_y} \quad \& \quad \text{Cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- e.g for example, $r = .989$,
  Signifies, STRONG, POSITIVE linear Relationship.

We can say, that no. of Workers causes the no. of tables produced.
In this case, we can say, it is CAUSATION, because, the variables are real/practical.

- RELATIONSHIP RULE OF THUMB

- Rule of Thumb : If $|r| \geq \frac{2}{\sqrt{n}}$

then a Relationship exists.

Video 8 : Geometric Mean & Standard Deviation.

- The sample mean is only suitable for additive processes.
- The geometric mean is suitable for multiplicative proces

- All values must be positive for Geometric Mean
- Geometric mean is often used for financial growth, growth in biology, agriculture, medicine, etc.

- Any rate of change over sequential periods of any length
- Unequal periods should not be used.

- GEOMEAN() in Excel.

- NATURAL LOGARITHM METHOD

→ Geometric mean formula : $\bar{x}_g = \sqrt[n]{(x_1)(x_2)(x_3)\ldots(x_n)}$

→ $\ln \bar{x}_g = \dfrac{\ln x_1 + \ln x_2 + \ln x_3 \cdots \ln x_n}{n}$

↳ $e^{\ln \bar{x}_g} = \bar{x}_g$

e.g length = 4 ; 6 ; 9

ln(x) = 1.386 , 1.791 , 2.19722

Averge = 1[illegible] (↓) /3

$e^{Av} = \bar{x}_g$