WEEK 4

Association Between 2 Categorical Variables

Nominal → Ordinal

→ Organize bivariate categorical data into a two way - contingency table.

→ If data is ordinal, maintain order of the variable in the table.

Gender	Own Smartphone		
	Yes	No	Row Total
Male	42	14	56
Female	34	10	44
Col. Total	76	24	100

NOMINAL
(order does not matter)

Income	Own Smartphone		Row Total
	Yes	No	
High	18	2	20
Middle	39	27	66
Low	5	9	14
Col Total	62	38	100

ORDINAL
(Order matters)

Row Relative Frequency

Divide each cell frequency in a row by its row total.

Column Relative Frequency

Divide each cell frequency in a column by its column total.

How do we know two categorical variables are associated with each other?

→ We use the notion of 'Relative Frequencies' be it row or column.

(i) If the row/column relative frequencies are the same for all rows/columns, then we say that two variables are not associated with each other.

(ii) If the row/column relative frequencies are different for some rows/columns, then we say that the two variables are associated with each other.

Stacked Bar Chart

- It represents the counts for a particular category. Each bar is further broken down into smaller segments, with each segment representing the freq. of that particular category within the segment.
- A 100% stacked bar chart is useful to part-to-whole relationships.

ASSOCIATION BETWEEN TWO NUMERICAL VARIABLES

Scatter plot

A scatter plot is a graph that displays pairs of values as points on a two dimentional plane.
x-variable → explanatory ; y-variable → response

Visual Test for Association & Describing it.

1. Direction → up or down ?
2. Curvature → linear or curved ?
3. Variation → tightly clustered or variable ?
4. Outliers → some unexpected plots on graph.

Measures of Association

(1) COVARIANCE → quantifies the strength of the linear association b/w 2 numerical variables.

Key points : (1) When large (small) values of x tend to be associated with large (small) values of y → the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also be tend to be same.

(2) When large (small) values of x tend to be associated with small (large) values of y → the signs of deviations, $(x_i - \bar{x})$ & $(y_i - \bar{y})$ will also tend to be different.

x	y	sign of dev.
large	small	different
small	large	different
large	large	same
small	small	same

Pop. Covariance : $Cov(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$

Sample Covariance : $Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Units of covariance are those of x-variable times those of y-variable

(2) CORRELATION → more easily interpretted measure of linear association.

→ derived from covariance.

→ ~~not~~ no unit

→ correlation measure always lie b/w -1 & +1.

The pearson correlation coeff, r, between x & y is :

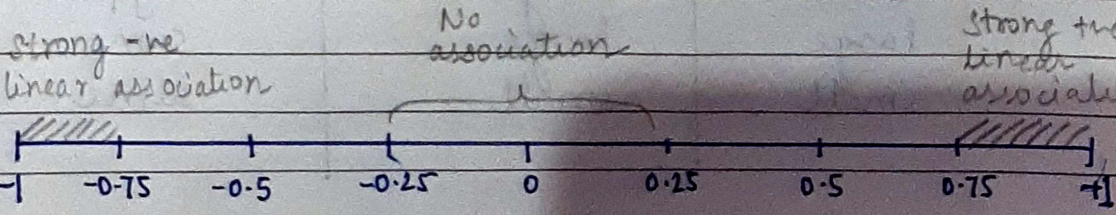
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{cov(x,y)}{S_x \cdot S_y}$$

Important Observations

COVARIANCE : → If $cov(x,y)$ is +ve, it means both x, y are going up or both are going down (never happens)

→ If $cov(x,y)$ is -ve, it means one is going up & the other down

CORRELATION:



Fitting A Line

→ The linear association's strength between the variables was measured using the measures of covariance and correlation.

→ The linear association can be described using the ~~measures of covariance~~ equation of line.

Equation of a line : $Y = mx + C$

→ We have r, i.e, correlation coeff
$(-1 \leq r \leq +1)$

from 'r' we get R^2, i.e, goodness of fit measure
$(0 \leq R^2 \leq 1)$ $\quad R^2 = r \times r$

→ If R^2 is closer to 0, it means there is no association b/w variables.

→ If R^2 is closer to 1, it means there is association b/w variables.

→ from slope also we can identify association; it is same as 'r'.
Slope is closer to +1 or −1, there is strong association.
Slope is closer to 0, no association.

Point Bi-Serial Correlation Coefficient

→ Association blw categorical & numerical variable

→ The point bi-serial correlation coeff. is :

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{S_x}\right)\sqrt{p_0 p_1}$$

where Y_0 → mean values of group associated with 0
Y_1 → mean values of group associated with 1.

p_0 → $\frac{\text{no. of obs. in '0' gp}}{\text{total observation}}$ (same for p_1)

S_x → standard deviation of numerical variable

→ If r_{pb} is closer to -1, +1 → strongly associated variables

→ If r_{pb} is closer to 0, no association bhw variables