# Basic Statistics

Dr. Sudipta Das

Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

# Outline I

# *Chapter 5: Bivariate Data and Correlation*

# Bivariate Data I

- Simultaneous analysis of two variables (attributes).
- It explores the concept of relationship between two variables
  - Is there any association between two variables
    - If yes, then what is the strength of this association,
  - or Is there any difference between two variables
    - If yes, then what is the significance of that difference
  - or Is there any affect of one variable on another variable
    - If yes, then what will be change in the affected variable if we change the other variable

# Types of Bivariate Analysis I

- Correlation Analysis
  - Used to study the strength of a relationship between two variables (e.g. height and weight).
- Regression Analysis
  - Used to study the effect of one variable (independent/predictor) on other (dependent/response) (e.g. advertisement expenditure on revenue)

# Correlation Analysis I

- Measure dependence between two variables whenever there is paired data, i.e.,

$$(x_1, y_1), \ldots, (x_n, y_n)$$

  - For example
    - Height and weight of individual
    - Rainfall and crop yield
    - Smoking and lung cancer
    - Study hours and marks obtained

# Correlation Analysis II

- Observations from Figure 1a and Figure 1b
  - Figure 1a shows
    - Taller trees generally have higher volume
    - We can also see that there are examples of relatively taller trees having relatively smaller volume.
    - Even though such exceptions are there, the overall pattern of greater volume for taller trees cannot be denied
  - Figure 1b shows
    - A similar conclusion can be made from the plot on the right side

# Correlation Analysis III

- Comparison between two scatter plots
  - The scatter plot on the right side shows a clearer relation.
  - The scatter of points on that plot resembles a straight line.
  - This conclusion is somewhat qualitative.
  - Question
    - Can it be made more precise?
    - Can we measure dependence by a number?

# Correlation Analysis IV

- In simplest way, the association between two variables can measured by calculating the following quantity which is called covariance between two variables
- For paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the covariance between the two variables is calculated by the formula

$$cov(X, Y) = s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

# Correlation Analysis V

- Some results on covariance
  - Alternate formula

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

  - If $U = \frac{X-a}{b}$ and $V = \frac{Y-c}{d}$, then

$$s_{UV} = \frac{1}{bd} s_{XY}$$

  - Covariance is distributive in nature, i.e.,

$$cov(a_1 + b_1 X_1 + b_2 X_2, a_1' + b_1' Y_1 + b_2' Y_2)$$

$$= b_1 b_1' cov(X_1, Y_1) + b_1 b_2' cov(X_1, Y_2)$$

$$+ b_2 b_1' cov(X_2, Y_1) + b_2 b_2' cov(X_2, Y_2)$$

# Product Moment Correlation I

- Product Moment Correlation: Unit less Measurement
- For paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the product moment correlation coefficient between the two variables $X$ and $Y$ is calculated by the formula

$$cor(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \times \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

- Pearson's Product Moment Correlation

# Product Moment Correlation II

- Correlation is expressed in a scale of -1 to 1.
  - The value 1 occurs when the all the points in the scatter plot lie on a single straight line (indicating a very strong linear dependence between the variables).
  - The value -1 also occurs when the all the points in the scatter plot lie on a single straight line, but that line has a negative slope. In general, negative values of correlation indicate negative linear dependence, i.e., higher value of one variable generally corresponds to lower value of the other.
  - Zero correlation corresponds to no linear dependence at all. Correlation close to zero may occur when the scatter of points is not elongated in any direction, i.e., the scatter shows no tilt along a straight line with positive or negative slope.

Figure 2

# Product Moment Correlation III

- Other than -1, 1, 0
- Positive values of correlation represent **positive** linear dependence
  - stronger positive dependence produces values closer to 1.
- Negative values of correlation represent **negative** linear dependence
  - stronger negative dependence produces values closer to -1.
- Numerical -> Correlation.xlsx -> Example 1

# Product Moment Correlation IV

- Some result on Pearson's correlation
  - If $U = \frac{X-a}{b}$ and $V = \frac{Y-c}{d}$, then

$$r_{UV} = \frac{|b||d|}{bd} r_{XY}$$

    - $U = \frac{X-\bar{X}}{s_X}$ and $V = \frac{Y-\bar{Y}}{s_Y} \rightarrow r_{UV} = r_{XY}$
  - For any set of $n$ pairs of values,

$$-1 \leq r_{XY} \leq 1$$

[Hint: $Var(U \pm V) = 1 + 1 \pm 2r_{XY} \geq 0$]

# Product Moment Correlation V

- Limitation of Pearson's correlation coefficient
  - Measures linear dependence:
    - Pearson's correlation between $x = \{0.01, .02, \ldots 1\}$ and $y = x^4$ is 0.8649
  - Sensitive to outlier:
    - When the $y = x$ and if last value of variable $y$ (i.e., y[101]) changes from 1 to 10 Pearson's correlation changes from 1 to 0.4516

# Rank Correlation I

- Instead of using the exact values of data points, we use the rank of the data points
- We measure the degree of similarity between two ranking
- Two common types of Rank correlation:
  - Spearman's correlation
  - Kendall's correlation

# Rank Correlation II

- Spearman's correlation (or Spearman's rho).
  - If the numerical values of each variable is replaced by its rank (i.e., position from the lower end after the data values are arranged in order of values),
  - then Pearson's correlation between these modified variables (ranks) is Spearman's correlation.

$$\rho_{XY} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

    - where $d_i = u_i - v_i$ differences in the rank
    - $u_i$ is the rank of $x_i$ in $X$
    - $v_i$ is the rank of $y_i$ in $Y$
- $\rho_{XY} = r_{UV}$
    - Hint: $d_i = (u_i - \bar{u}) - (v_i - \bar{v}) \rightarrow \frac{1}{n} \sum_{i=1}^{n} d_i^2 = ?$
- Numerical -> Correlation.xlsx -> Example 2

# Rank Correlation III

- Kendall's correlation (or Kendall's tau).
  - This correlation is computed by considering two pairs of data values at a time, checking whether the order of values of the first variable is in concordance or discordance with the order of values of the second variable,
  - and then contrasting the counts of all concordances and discordances, with suitable scaling.

$$\tau_{XY} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} sgn(x_i - x_j) sgn(y_i - y_j) = \frac{C-D}{C+D}$$

    - where C = Number of concordant pairs and D = Number of discordant pairs
  - Numerical -> Correlation.xlsx -> Example 3

# Rank Correlation IV

- Like Spearman's rho Kendall's tau also depends on the ranks of the data points, rather the values themselves.
  - For this reason, both of these are called rank correlations.
- Remain the same with any change in the unit of measurement.
- Similarities between Pearson's Correlation and Rank Correlation
  - Values in -1 to +1
    - +1 in the case of perfect positive dependence
    - -1 in the case of perfectly negative dependence, and
    - 0 in the case of independence.

# Rank Correlation V

- Overcoming limitation of Pearson's correlation coefficient
  - Measures linear dependence
    - Pearson's correlation between $x = \{0, 1, .01\}$ $y = x^4$ is 0.8649
    - Both the Spearman's correlation and Kendall's correlation are 1
  - Sensitive to outlier
    - if $y[101]$ changes from 1 to 10 Pearson's correlation changes from 1 to 0.4516
    - Both the Spearman's correlation and Kendall's correlation are 1

# Rank Correlation VI

- Rank Correlation with Tie
- First:- Spearman's correlation (or Spearman's rho) with tie
  - It is *Pearson's* correlation between adjusted rank series

$$\rho_{XY} = \frac{\frac{n(n^2-1)}{6} - \left(\sum_{i=1}^{n} d_i^2 + m_2 \frac{2^3-2}{12} + m_3 \frac{3^3-3}{12} + \cdots\right)}{\sqrt{\left[\frac{n(n^2-1)}{6} - 2\left(m_{x_2}\frac{2^3-2}{12} + m_{x_3}\frac{3^3-3}{12} + \cdots\right)\right]\left[\frac{n(n^2-1)}{6} - 2\left(m_{y_2}\frac{2^3-2}{12} + m_{y_3}\frac{3^3-3}{12} + \cdots\right)\right]}}$$

  - $d_i =$ differences in the rank
  - $m_{x_2}, m_{x_3}, \ldots$ are the frequencies of two ties, three ties and so on from the series $X$.
  - $m_{y_2}, m_{y_3}, \ldots$ are the frequencies of two ties, three ties and so on from the series $Y$.
  - $m_i = m_{x_i} + m_{y_i}$

# Rank Correlation VII

- An approximate formula for Spearman's rho with tie

$$\rho_{XY} = 1 - \frac{6}{n(n^2-1)}\left[\sum_{i=1}^{n} d_i^2 + m_2\frac{2^3-2}{12} + m_3\frac{3^3-3}{12} + \cdots\right]$$

  - where $d_i$ = differences in the rank
  - $m_2, m_3, \ldots$ are the frequencies of two ties, three ties and so on.
- Numerical -> Correlation.xlsx -> Example 4

# Rank Correlation VIII

- Second:- Kendall's correlation (or Kendall's tau) with tie

$$\tau_{XY} = \frac{C - D}{\sqrt{\left[C + D - u_2 \frac{2(2-1)}{2} - u_3 \frac{3(3-1)}{2} - \cdots\right]\left[C + D - v_2 \frac{2(2-1)}{2} - v_3 \frac{3(3-1)}{2} - \cdots\right]}}$$

  - $u_2, u_3, \ldots$ are the frequencies of two ties, three ties and so on from the series $X$
  - $v_2, v_3, \ldots$ are the frequencies of two ties, three ties and so on from the series $Y$
- Numerical -> Correlation.xlsx -> Example 5

# Rank Correlation IX

- Comparison between Spearman's rho and Kendall's tau
  - Kendall's Tau:
    - Usually smaller values than Spearman's rho correlation.
    - Robust against error, since the calculation is based on concordant and discordant pairs.
    - $p-$values are more accurate with smaller sample sizes.
  - Spearman's rho:
    - Usually have larger values than Kendall's tau.
    - Much more sensitive to error and discrepancies in the data, since the calculation is based on deviations.