# STAT 111: INTRODUCTION TO STATISTICS AND PROBABILITY I

## LECTURE 4: DESCRIPTIVE STATISTICS FOR BIVARIATE DATA

MR. BENEDICT MBEAH-BAIDEN

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, UG

February 2, 2021

# Lecture Outline I

# Lecture Outline II

4. Analyzing Quantitative Bivariate Data: Covariance and Correlation
   - Covariance
   - Correlation
   - Pearsons Product Moment Correlation Coefficient (PMCC)
   - Spearman's Rank Correlation Coefficient (SPCC)
   - Kendall's Rank Correlation Coefficient (KRCC)

# Introduction

## Learning Objectives

To introduce the student to descriptive measures that can be used to analyze bivariate data.

## Learning Outcomes

By the end of the lecture, the student should be able to:

1. Explain what bivariate data is and give examples
2. Explain the use of Contingency Tables and Contingency Coefficient
3. Compute and interpret appropriate Contingency Coefficients for different data
4. Explain the difference between Covariance and Correlation
5. Compute and interpret appropriate Correlation Coefficients for different data

# Dimensions of Data Sets

- Descriptive statistics involves describing the main features of a collection of information (data) from a group of subjects.
- The data may be Univariate, Bivariate, or Multivariate in nature.
- **Univariate Data** is obtained when measurements are made on one variable per subject.
- **Bivariate Data** is obtained when measurements are made on two variables per subject.
- **Multivariate Data** is obtained when measurements are made on many variables (three or more variables) per subject.

# Bivariate Data Analysis

- In Bivariate data analysis, data on two variables obtained from the subjects are analyzed simultaneously.
- Bivariate data involves looking at associations or relations that exist between pairs of variables, and trying to understand how these associations or relations work (i.e. in terms of direction and strength)
- The layout for a bivariate analysis is provided as follows: Let us say, there are $n$ subjects, and the two variables are Variable $X$ and Variable $Y$. Then, we have the paired points $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. or in tabular form:

| $X$ | $X_1$ | $X_2$ | $X_3$ | ... | $X_n$ |
|---|---|---|---|---|---|
| $Y$ | $Y_1$ | $Y_2$ | $Y_3$ | ... | $Y_n$ |

- For example: A collection of data consisting of the **Height** and **Weight** of 100 STAT 111 students.

# Bivariate Data Analysis

- These two variables can change and are compared to find relationships, and if there is a relationship, what is the direction and strength of the relationship.
- If one variable influences the value of the other variable then it implies the bivariate data has an **independent** and a **dependent** variable.
- **Independent Variable** is a variable whose variation or change does not depend on that of any other variable. The independent variable is also the variable under the control of the experiment.
- **Dependent Variable** is a variable whose value is influenced or controlled by that of another variable, usually the independent variable. The dependent variable is also the variable that depends on what happens in the experiment.

# Bivariate Data Analysis

- In order to effectively describe the statistical relationship between the individual variables in a bi-variate data set, numerous measures of association are employed.
- These measures provide a means of summarizing the size of the association between two variables.
- Some of the measures of association used in describing bivariate data sets include Contingency Tables, Covariance, Correlation and Regression analysis.

## Possible Types of Bivariate Data

The two pairs of data could be:

1. Two Qualitative Data (**Eg.: Sex and Residential Status**)
2. One Qualitative and One Quantitative Data (**Eg.: Sex and Exam Score**)
3. Two Quantitative Data (**Eg.: Age and Exam Score**)

# Analyzing Qualitative Bivariate Data

- When bivariate data is obtained from two qualitative (attributes or categorical) variables, the **Contingency Table** is used to arrange the data.
- A contingency table is sometimes called the **cross tabulation** or **cross classification table**

## Contingency Table

A Contingency Table is a table used to describe bivariate qualitative (categorical) data sets, showing the distribution of one variable in the rows and the other variable in the column.

# Analyzing Qualitative Bivariate Data

- Contingency tables are named based on the **size of the table**. That is, Contingency tables are named based on the **number of rows and columns** the table has.
- . If $X$ and $Y$ are two categorical data sets with $m$ and $n$ distinct categories respectively, then the contingency table for both variables will be an $mxn$ contingency table as shown on the table below:

Table: An *mxn* Contingency Table for variables $X$ and $Y$

| | $Y = y_1$ | $Y = y_2$ | $Y = y_3$ | ... | $Y = y_n$ | *Total* |
|---|---|---|---|---|---|---|
| $X = x_1$ | $O_{11}$ | $O_{12}$ | $O_{13}$ | ... | $O_{1n}$ | $R_1$ |
| $X = x_2$ | $O_{21}$ | $O_{22}$ | $O_{23}$ | ... | $O_{2n}$ | $R_2$ |
| $X = x_3$ | $O_{31}$ | $O_{32}$ | $O_{33}$ | ... | $O_{3n}$ | $R_3$ |
| ... | ... | ... | ... | ... | ... | ... |
| $X = x_m$ | $O_{m1}$ | $O_{m2}$ | $O_{m3}$ | ... | $O_{mn}$ | $R_m$ |
| *Total* | $C_1$ | $C_2$ | $C_3$ | ... | $C_n$ | $N$ |

# Explanation of Notations

- The quantity $O_{ij}$ is the number of observations with $X = x_i$ and $Y = y_j$. That is the number of observations with characteristic $x_i$ of variable $X$ and $y_j$ of variable $Y$.
- $R_i = \sum_{j=1}^{n} \emptyset_{ij}$ is the total of observations of the $i - th$ row (i.e. the total number of observations with the characteristic $x_i$ of variable $X$ across all characteristics of variable $Y$).
- $C_j = \sum_{i=1}^{m} \emptyset_{ij}$ is the total of observations of the $j - th$ column (i.e. the total number of observations with the characteristic $y_j$ of variable $Y$ across all characteristics of variable $X$).
- $N = \sum_{i=1}^{m} \sum_{j=1}^{n} \emptyset_{ij}$ is the total number of observations under consideration (i.e. the sample size).

# Contingency Coefficients

## Contingency Coefficients

**A Contingency Coefficient** is a measure of the extent of association or dependence between two categorical variables or attributes.

- A contingency coefficient is computed from a contingency table.
- A number of contingency coefficients are available to assess the degree of association between variables on the contingency table.
- They include:
  1. Phi Contingency Coefficient
  2. Cramer's V Contingency Coefficient
  3. Kappa Contingency Coefficient
  4. Tschuprow's T Contingency Coefficient
  5. Pearson's Contingency Coefficient

# Phi Contingency Coefficient

- The most common measure of magnitude of effect for two binary variables is the Phi Coefficient Coefficient (PCC).
- The PCC is a measure of association which is based on a measure called the Chi-square ($\chi^2$).
- The Phi coefficient adjusts the chi-square statistic by the sample size.
- The symbol for Phi coefficient is the Greek letter phi, written as $\phi$.

# Phi Contingency Coefficient

- The Phi coefficient is defined as:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{1}$$

- where $N$ is the total number of observations and $\chi^2$ is computed as the Pearson chi square test statistic computed as:

$$\chi^2 = \frac{\sum(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

- and $E_{ij}$ is calculated as

$$E_{ij} = \frac{R_i * C_j}{N} \tag{3}$$

# Phi Contingency Coefficient

- The Phi coefficient can also be defined in terms of a shortcut notation for the frequencies in the four cells, using the **Fourfold table Notation** or the **Azen and Walker Notation** as shown below:

Table: Azen and Walker Notation

| | $Y = y_1$ | $Y = y_2$ | *Total* |
|---|---|---|---|
| $X = x_1$ | $O_{11}$ | $O_{12}$ | $R_1$ |
| $X = x_2$ | $O_{21}$ | $O_{22}$ | $R_2$ |
| *Total* | $C_1$ | $C_2$ | $N$ |

$$\phi = \frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})}} \tag{4}$$

OR alternatively

Table: Fourfold Notation

| | $Y = y_1$ | $Y = y_2$ | *Total* |
|---|---|---|---|
| $X = x_1$ | $A$ | $B$ | $R_1$ |
| $X = x_2$ | $C$ | $D$ | $R_2$ |
| *Total* | $C_1$ | $C_2$ | $N$ |

$$\phi = \frac{AD - BC}{\sqrt{(R_1)(R_2)(C_1)(C_2)}} \tag{5}$$

# Phi Contingency Coefficient: Its Properties

- The Phi Coefficient varies from 0 (corresponding to no association between variables) to 1 (corresponding to complete association between variables).
- It cannot have a value less than zero since the minimum value for $\chi^2$ is zero.
- If there is no association between the variables so that $\chi^2$ equals zero, then $\phi$ also equals zero. If $\chi^2$ value is small, $\phi$ will also be relatively small and vice versa.
- The chi square ($\chi^2$)value indicates the strength of the relationship between the two variables.
- $\phi$ is appropriate for a $2x2$ contingency table. For contingency tables with dimensions other than $2x2$ , the other contingency coefficients are employed.

# Phi Contingency Coefficient: Examples

## Example 1

The Dean of Students randomly sampled 70 students in the University of Ghana and asked then whether they are Fresh Students or Continuing Students as well as whether they are Residential or Non-Residential students as shown in the table below. Calculate the phi coefficient.

Table: Table for Example 1

| | Residential | Non-Residential | Total |
|---|---|---|---|
| Fresh Students | 21 | 6 | 27 |
| Continuing Students | 30 | 13 | 43 |
| Total | 51 | 19 | 70 |

# Phi Contingency Coefficient: Solution to Example 1

- The variables here are the Level of Study (i.e. fresh student or continue students) and Residential Status (Resident or Non-Resident). Hence $i = 1, 2$ and $j = 1, 2$ . Total number of observations is 70 (i.e. $N = 70$ )
- Step 1: Calculate $\chi^2$ since the phi coefficient is dependent on it.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6}$$

- and $E_{ij}$ is calculated as

$$E_{ij} = \frac{R_i * C_j}{N} \tag{7}$$

$$E_{11} = \frac{R_1 * C_1}{N} = \frac{27 * 51}{70} = \frac{1377}{70} = 19.6714 \tag{8}$$

# Phi Contingency Coefficient: Solution to Example 1 Contd.

$$E_{12} = \frac{R_1 * C_2}{N} = \frac{27 * 19}{70} = \frac{513}{70} = 7.3286 \quad (9)$$

$$E_{21} = \frac{R_2 * C_1}{N} = \frac{43 * 51}{70} = \frac{2193}{70} = 31.3286 \quad (10)$$

$$E_{22} = \frac{R_2 * C_2}{N} = \frac{43 * 19}{70} = \frac{817}{70} = 11.6714 \quad (11)$$

- and $\chi^2$ is then computed as

$$\chi^2 = \frac{(21 - 19.6714)^2}{19.6714} + \frac{(6 - 7.3286)^2}{7.3286} + \frac{(30 - 31.3286)^2}{31.3286} + \\ + \frac{(13 - 11.6714)^2}{11.6714} = 0.5382$$

# Phi Contingency Coefficient: Solution to Example 1 Contd.

- Step 2: Calculate $\phi$.

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{12}$$

$$\phi = \sqrt{\frac{0.5382}{70}} = 0.0877 \tag{13}$$

# Phi Contingency Coefficient: Solution to Example 1

Using Fourfold table Notation

$$\phi = \frac{AD - BC}{\sqrt{(R_1)(R_2)(C_1)(C_2)}} \tag{14}$$

$$\phi = \frac{(21)(13) - (6)(30)}{\sqrt{(27)(43)(51)(19)}} \tag{15}$$

$$\phi = \frac{273 - 180}{\sqrt{(1125009}} \tag{16}$$

$$\phi = \frac{93}{1060.6644} = 0.0876 \tag{17}$$

# Phi Contingency Coefficient: Solution to Example 1

Using Azen and Walker Notation

$$\phi = \frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})}} \tag{18}$$

$$\phi = \frac{(21)(13) - (6)(30)}{\sqrt{(27)(43)(51)(19)}} \tag{19}$$

$$\phi = \frac{273 - 180}{\sqrt{(1125009}} \tag{20}$$

$$\phi = \frac{93}{1060.6644} = 0.0876 \tag{21}$$

# Cramer's V Contingency Coefficient (CVCC)

- This measure of association is also based on Pearsons chi square test statistic and it is appropriate for contingency tables that have dimensions larger than 2x2 .
- The PCC is a measure of association which is based on a measure called the Chi-square ($\chi^2$).
- The Cramers V is sometimes referred to as Cramers phi and is denoted $V$ or $\phi_c$.
- It is defined as

$$V = \phi_c = \sqrt{\frac{\phi^2}{t}} = \sqrt{\frac{\chi^2}{Nt}} \tag{22}$$

where $\chi^2$ is computed as the Pearson chi square test statistic, $N$ is the total number of observations, $t = (k-1)$ and is $k = min(r, c)$ (i.e. the lesser of the number of rows and columns).

# Cramer's V Contingency Coefficient: Its Properties

- The Cramer V Coefficient varies from 0 (corresponding to no association between variables) to 1 (corresponding to complete association between variables).
- For $2x2$ contingency tables, $(k-1) = (2-1) = 1$ which makes

$$V = \phi_c = \sqrt{\frac{\chi^2}{Nt}} = \sqrt{\frac{\chi^2}{N}} = \phi \tag{23}$$

- Hence Cramers V is the same as the phi coefficient for 2x2 contingency tables.

# Cramer V Contingency Coefficient: Examples

## Example 2

A researcher is interested in finding out if the choice of university is dependent on the area of origin of students. The data the researcher obtained from students is summarized below. Calculate Cramers V coefficient for the data.

Table: Table for Example 2

| | UG | KNUST | UCC | Others | Total |
|---|---|---|---|---|---|
| Northern | 56 | 14 | 12 | 22 | 104 |
| Middle Belt | 15 | 27 | 8 | 24 | 74 |
| Southern | 19 | 2 | 26 | 47 | 94 |
| Total | 90 | 43 | 46 | 93 | 272 |

# Cramer V Contingency Coefficient: Solution to Example 2

- The variables here are area of origin (i.e. northern, middle belt or southern) and Choice of University (UG, KNUST, UCC or others). Hence $i = 1, 2, 3$ and $j = 1, 2, 3, 4$ and $k = min(r, c) = min(3, 4) = 3$. Total number of observations is 272 (i.e. $N = 272$)
- Step 1: Calculate $\chi^2$ since the Cramer's V coefficient is dependent on it.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{24}$$

- and $E_{ij}$ is calculated as

$$E_{ij} = \frac{R_i * C_j}{N} \tag{25}$$

# Cramer's V Contingency Coefficient: Solution to Example 2 Contd.

$$E_{11} = \frac{R_1 * C_1}{N} = \frac{90 * 104}{272} = \frac{9360}{272} = 34.4118 \tag{26}$$

- The other $E'_{ij}s$ are computed similarly:
- and $\chi^2$ is then computed as $\chi^2 = 75.3662$
- Step 2: Calculate Cramer's V Contingency Coefficient (CVCC)

$$V = \phi_c = \sqrt{\frac{75.3662}{(272)(2)}} = \sqrt{\frac{75.3662}{(544)}} = \sqrt{0.1385} = 0.3722 \tag{27}$$

# Kappa Contingency Coefficient

- Kappa Contingency Coefficient, denoted $K$ , is a measure of association (correlation or reliability) between two measurements on the same individual when the measurements are categorical.
- It tests if the counts along the diagonal are significantly large.
- Because Kappa is used when the same variable is measured twice, it is only appropriate for square tables where the row and column categories are the same.
- Kappa is often used to study the agreement of two raters (inter-rater reliability) such as judges or doctors, where each rater classifies each individual into one of $k$ categories.

# Kappa Contingency Coefficient Contd.

- Inter-rater reliability occurs when data raters (or collectors) give the same score to the same data item.
- $K$ can also be calculated when one rater rates two trials on each sample.
- $K$ varies from 0 (representing agreement equivalent to chance) to 1 (representing a perfect agreement).
- $K$ is defined as

$$K = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e} \tag{28}$$

where $P_o$ is the relative observed agreement among the raters; $P_e$ is the hypothetical probability of chance (expected) agreement between the raters.

# Interpreting the Kappa Contingency Coefficient

- The following rule of thumb is used in interpreting the Kappa Contingency Coefficient (K):
- $0.00 \leq K \leq 0.20$: implying poor agreement or association between the two raters .
- $0.21 \leq K \leq 0.40$: implying fair agreement or association between the two raters.
- $0.41 \leq K \leq 0.60$: implying moderate agreement or association between the two raters.
- $0.61 \leq K \leq 0.80$: implying good agreement or association between the two raters.
- $0.81 \leq K \leq 0.99$: implying very good agreement or association between the two raters.

# Data Layout for calculating the Kappa Contingency Coefficient

- The data layout for calculating a Kappa Contingency coefficient is given below:

| | | Rater 1 | | |
|---|---|---|---|---|
| | | *Yes* | *No* | *Total* |
| *Rater2* | *Yes* | $a$ | $b$ | $m_1$ |
| *Rater2* | *No* | $c$ | $d$ | $m_2$ |
| | *Total* | $n_1$ | $n_2$ | $N$ |

- $a$ and $d$ represent the number of times the two raters agree while $b$ and $c$ represent the number of times the two raters disagree.

# Data Layout for calculating the Kappa Contingency Coefficient Contd.

- If there are no disagreements, $b$ and $c$ would be zero, and the observed agreement $P_o$ is 1.
- If there are no agreements, $a$ and $d$ would be zero, and the observed agreement $P_o$ is 0.
- The observed agreement$P_o$ is calculated as:

$$P_o = \frac{(a+d)}{N} \tag{29}$$

- The expected agreement$P_e$ is calculated as:

$$P_e = [(\frac{n_1}{N}) * (\frac{m_1}{N})] + [(\frac{n_2}{n}) * (\frac{m_2}{N})] \tag{30}$$

# Kappa Contingency Coefficient: Example

**Example 5**

A lecturer wanted to measure the effectiveness of his mode of delivery during lectures. He picked two students from his class and asked them to evaluate the usefulness of a series of 100 lectures delivered by him. Student 1 and Student 2 agree that the lectures are useful 15% of the time and not useful 70% of the time. The data is presented in the table below. Calculate the degree of agreement between the two students using the Kappa Contingency Coefficient, $K$.

| | | *Student*1 | | |
|---|---|---|---|---|
| | | *Yes* | *No* | *Total* |
| *Student*2 | *Yes* | 15 | 5 | 20 |
| *Student*2 | *No* | 10 | 70 | 80 |
| | *Total* | 25 | 75 | 100 |

# Kappa Contingency Coefficient: Solution to Example 5

- The observed agreement $P_0$ is calculated as:

$$P_o = \frac{(15 + 70)}{100} = 0.85 \tag{31}$$

- The expected agreement $P_e$ is calculated as:

$$P_e = [(\frac{25}{100}) * (\frac{20}{100})] + [(\frac{75}{100}) * (\frac{80}{100})] = 0.65 \tag{32}$$

- Hence the Kappa Contingency Coefficient is given as:

$$K = \frac{P_o - P_e}{1 - P_e} = \frac{0.85 - 0.65}{1 - 0.65} = 0.57 \tag{33}$$

# Kappa Contingency Coefficient: Trial Example

**Example 6**

The table below shows the results of a study assessing the prevalence of depression among 200 patients treated in a primary care setting using two methods to determine the presence of depression; one based on information provided by the individual (i.e., proband) and the other based on information provided by another informant (e.g., the subjects family member or close friend) about the proband. Calculate the degree of agreement between the two approaches using the Kappa Contingency Coefficient, $K$.

| | | *Informant* | | |
|---|---|---|---|---|
| | | *Depressed* | *NotDepressed* | *Total* |
| *Proband* | *Depressed* | 65 | 50 | 115 |
| *Proband* | *NotDepressed* | 19 | 66 | 85 |
| | *Total* | 84 | 116 | 200 |

# Analyzing Quantitative Bivariate Data: Covariance and Correlation

- An example of a quantitative bivariate data is spousal ages of a sample of married couples as shown below:

| Husband | 36 | 72 | 37 | 36 | 51 | 50 | 47 | 37 | 41 |
|---|---|---|---|---|---|---|---|---|---|
| Wife | 35 | 67 | 33 | 35 | 50 | 46 | 47 | 36 | 41 |

## Types of Relationships

There are different ways in which two variables may be related:

1. absence of relationship
2. negative linear relationship
3. positive linear relationship
4. nonlinear relationship

# Analyzing Quantitative Bivariate Data: Covariance and Correlation Contd.

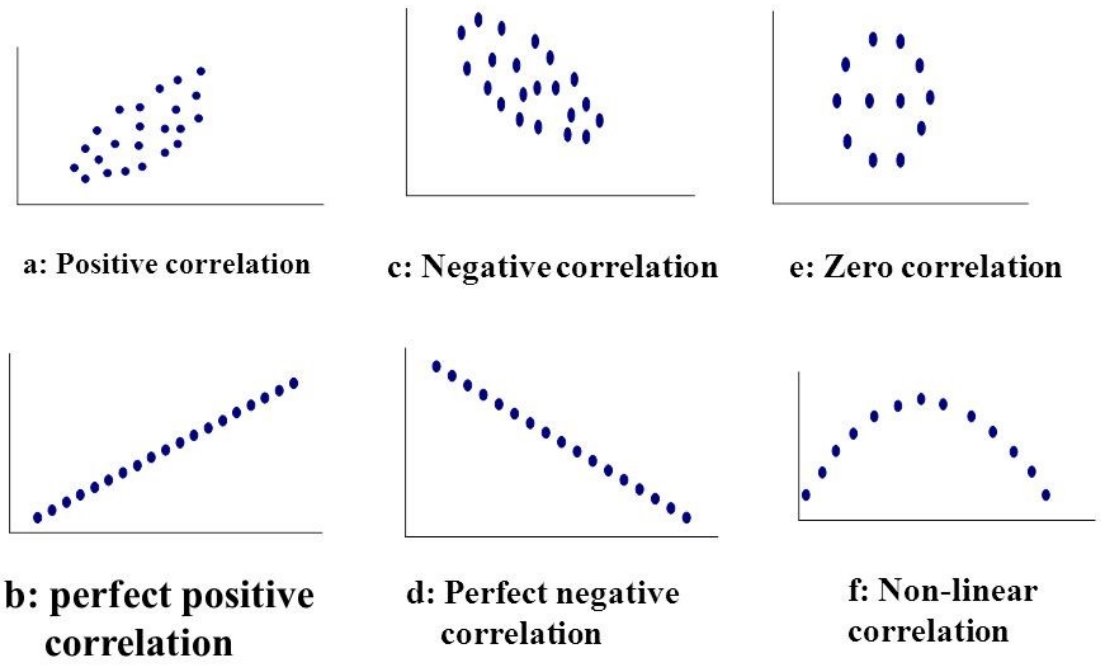


Figure: Types of relationships between two variables

# Measures of Linear Association: Covariance

- When comparing data from two quantitative variables, two of the most popular measures of **linear association** are **covariance** and **correlation**.
- **Covariance** is a measure of the joint variability of two variables.
- Let us say we have $N$ individuals and the two quantitative variables are $X$ and $Y$. Then we have $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$.
- Intuitively, the **sum of cross products** measures the direction and strength of association between variables $X$ and $Y$.

# Measures of Linear Association: Covariance Contd.

- The sum of cross products between variables $X$ and $Y$ is given by

$$\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \tag{34}$$

- The value of the sum of cross products $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$ will be large in magnitude either positively or negatively when there is an association or correlation between X and Y .
- Therefore the sum of cross products $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$ form the basics of correlation analysis.
- When we average out the sum of cross products $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$ we obtain the measure for the co-variance between X and Y.

# Measures of Linear Association: Covariance Contd.

- For two random variables $X$ and $Y$ , the **population covariance** is defined as:

$$Cov(X, Y) = S_{XY} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N} \tag{35}$$

- For two random variables $X$ and $Y$ , the **sample covariance** is defined as:

$$Cov(X, Y) = S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{36}$$

# Computational Formulars for Covariance

- Computationally the **population covariance** between two random variables $X$ and $Y$ is defined as:

$$Cov(X,Y) = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(n)} \tag{37}$$

- and the **sample covariance** between two random variables $X$ and $Y$ is defined as:

$$Cov(X,Y) = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(n-1)} \tag{38}$$

# Properties of Covariance

- The **sign** of the covariance shows the **tendency or direction** in the linear relationship between the variables.

- If larger values of one variable mainly correspond to larger values of the other, and vice versa, the covariance is **positive**.
- Conversely if larger values of one variable correspond to smaller value of the other variable, and vice versa, the covariance is **negative**.
- If the covariance is "much larger than 0", it implies there exists a positive linear relationship between the variables.
- If the covariance is "much smaller than 0", it implies there exists a negative linear relationship
- Random variables whose covariance is **zero** are called **uncorrelated variables**.

# Drawbacks of Covariance

1. Covariance is **not a bounded measure**. Hence it is difficult to tell what it means to say "large" or "small".
2. The value of the covariance is affected by the **units of measurement** of $X$ and $Y$ since the covariance measures the strength and direction in terms of the units of measurement of $X$ and $Y$.

# Covariance: Example

**Example 7**

The data below shows the number of years spent in college ($X$) and the subsequent yearly income ($Y$), in thousands of Ghana cedis for Six people. Can one conclude that the number of years spent in college affects yearly income?

| Person | number of years spent in college (X) | yearly income (Y) |
|---|---|---|
| 1 | 0 | 15 |
| 2 | 1 | 15 |
| 3 | 3 | 30 |
| 4 | 4 | 25 |
| 5 | 4 | 30 |
| 6 | 6 | 35 |

# Covariance: Solution to Example 7

- The table given can be extended as follows:

| Person | X | Y | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|---|
| 1 | 0 | 15 | 0 | 225 | 0 |
| 2 | 1 | 15 | 1 | 225 | 15 |
| 3 | 3 | 30 | 9 | 400 | 60 |
| 4 | 4 | 25 | 16 | 625 | 100 |
| 5 | 4 | 30 | 16 | 900 | 120 |
| 6 | 6 | 35 | 36 | 1225 | 210 |
| **Total** | 18 | 140 | 78 | 3600 | 505 |

# Solution to Example 7 Contd.

- Now if the data was from the population the **population covariance** between the two random variables $X$ and $Y$ would be:

$$Cov(X, Y) = \frac{(6)(505) - (18)(140)}{6(6)} = \frac{3030 - 2520}{36} = \frac{510}{36} = 14.1667 \tag{39}$$

- and if the data was from the sample and we were estimating the population, then the **sample covariance** between two random variables $X$ and $Y$ would be:

$$Cov(X, Y) = \frac{(6)(505) - (18)(140)}{6(5)} = \frac{3030 - 2520}{30} = \frac{510}{30} = 17 \tag{40}$$

# Solution to Example 7

**Interpreting the result**

- Since the covariance is larger than zero (0), it implies there exists a positive linear relationship between number of years spent in college an the yearly income.
- However, it will be difficult to tell the strength of this linear relationship since the covariance measure is not bounded.

# Measures of Linear Association: Correlation

- A new measure called **Correlation** overcomes the drawbacks of Covariance.
- That is to say **Correlation** is a measure which is free of units of measurements of the two variables and also a bounded measure.
- Correlation refers to the degree to which two quantifiable variables in which numbers are meaningful are linearly related.
- The main result of a correlation is called the **Correlation Coefficient** ($r$).

# Properties of Correlation

1. It ranges from $-1$ to $+1$. $(-1 \leqslant r \leqslant +1)$
2. The closer $r$ is to $+1$ or $-1$ , the more closely the two variables are related. As $r$ gets closer to 0 , it implies that the relationship between the two variables will be weaker.
3. If $r = 0$ , it means there is **no relationship** between the variables. If $r$ is **positive**, it means that as one variable gets larger, the other gets larger, and vice versa. If $r$ is **negative** it means that as one variable gets larger, the other gets smaller, and vice versa (called **inverse or indirect correlation)**.
4. When $r$ is squared, the resulting value is equal to the percentage of variation in one variable that is related to the variation in the other variable

# Guide to Interpreting the Strength of Correlation

- Whilst the sign indicates the direction of the correlation,the **absolute value of the correlation indicates the strength (magnitude) of the correlation**.
- The following guide can be used to interpret or comment on the strength of the correlation:
- $0.00 \leqslant r \leqslant 0.19$: **Very Weak**
- $0.20 \leqslant r \leqslant 0.39$: **Weak**
- $0.40 \leqslant r \leqslant 0.59$: **Moderate**
- $0.60 \leqslant r \leqslant 0.79$: **Strong**
- $0.80 \leqslant r \leqslant 0.99$: **Very Strong**
  **NB:** $r = 1$ implies a perfect relationship.

# Measures of Linear Association: Correlation Contd.

- **Note: Correlation does not imply causation**. That is when working with correlation remember never to assume that a correlation means that **a change in one variable causes a change in the other**.
- The following three types of correlation coefficients that are usually measured in statistics will be considered:

1. The **Pearsons Product Moment Correlation Coefficient (PMCC)**
2. The **Spearmans Rank Correlation Coefficient (SRCC)**
3. The **Kendalls Rank Correlation Coefficient (KRCC)**

# Pearsons Product Moment Correlation Coefficient (PMCC)

- This is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.
- Given a set of observations from two random variables $X$ and $Y$ , the **Pearsons Product Moment Correlation Coefficient (PMCC)** is defined as:

$$R_{X,Y} = \frac{S_{XY}}{S_X S_Y} \tag{41}$$

- where $S_{XY}$ is the **covariance** between $X$ and $Y$, $S_X$ is the **standard deviation** of $X$ and $S_Y$ is the **standard deviation** of $Y$ .

# Pearsons Product Moment Correlation Coefficient (PMCC) Contd.

- **Computationally** the PMCC is given as:

$$R_{X,Y} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}} \tag{42}$$

# Assumptions of the Pearsons Product Moment Correlation Coefficient (PMCC)

- The calculation of Pearsons correlation coefficient requires the following data assumptions to hold:

1. The data should be $n$ pairs of data from a random sample and it should be at least on the interval level of measurement (interval or ratio level).
2. The two variable should be linearly related.
3. The data should be bi-variate normally distributed. In practice this assumption is checked by requiring both variables to be individually normally distributed (which is a by-product consequence of bi-variate normality).
4. The data should not have a skewed distribution.
5. There should be no outliers.

# PMCC: Example

**Example 8**

The data below shows the number of years spent in college ($X$) and the subsequent yearly income ($Y$), in thousands of Ghana cedis for Six people. Calculate the Pearson's Product Moment Correlation Coefficient between the two variables and comment on it.

| Person | number of years spent in college (X) | yearly income (Y) |
|---|---|---|
| 1 | 0 | 15 |
| 2 | 1 | 15 |
| 3 | 3 | 30 |
| 4 | 4 | 25 |
| 5 | 4 | 30 |
| 6 | 6 | 35 |

# Pearsons Product Moment Correlation Coefficient (PMCC): Solution to Example 8

- The table given can be extended as follows:

| Person | X | Y | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|---|
| 1 | 0 | 15 | 0 | 225 | 0 |
| 2 | 1 | 15 | 1 | 225 | 15 |
| 3 | 3 | 30 | 9 | 400 | 60 |
| 4 | 4 | 25 | 16 | 625 | 100 |
| 5 | 4 | 30 | 16 | 900 | 120 |
| 6 | 6 | 35 | 36 | 1225 | 210 |
| **Total** | 18 | 140 | 78 | 3600 | 505 |

# Solution to Example 8 Contd.

- The **Pearsons Product Moment Correlation Coefficient** between the two random variables $X$ and $Y$ would be:

$$R_{X,Y} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}} \tag{43}$$

$$R_{X,Y} = \frac{(6)(505) - (18)(140)}{\sqrt{[(6)(78) - (18)^2][(6)(3600) - (140)^2]}} \tag{44}$$

$$R_{X,Y} = \frac{3030 - 2520}{\sqrt{[468 - 324][21600 - 19600]}} = \frac{510}{\sqrt{[144][2000]}} \tag{45}$$

$$R_{X,Y} = \frac{510}{\sqrt{288000}} = \frac{510}{536.6563} = 0.9503 \tag{46}$$

# Solution to Example 8 Contd.

$$R_{X,Y} = \frac{510}{\sqrt{288000}} = \frac{510}{536.6563} = 0.9503 \tag{47}$$

## Comment

There is a **strong positive relationship** between number of years spent in college and the yearly income of a person. In other words, the more years you spent in college, the higher your yearly income.

# Spearman's Rank Correlation Coefficient (SPCC)

- If your data does not meet the assumptions required to use the Pearsons Product Moment Correlation Coefficient (PMCC), then use Spearman's Rank Correlation Coefficient.
- The Spearman's Rank Correlation ($R_s$), sometimes referred to as Spearmans rho, after the Greek letter rho ($\rho$), is a non-parametric measure of correlation which make use of the ranks of the paired data. The Spearman's Rank correlation measures the strength and direction of association between two **ranked variables** (i.e. the Spearman's correlation between two variables is equal to the Pearson correlation between the rank values of those two variables).
- While Pearsons correlation assesses linear relationships, Spearmans correlation assesses monotonic relationships (whether linear or not).

# Assumptions of the Spearman's Rank Correlation Coefficient

- The calculation of Spearman's Rank Correlation Coefficient requires the following data assumptions to hold:

1. The data should be at least on the ordinal level of measurement (i.e. ordinal, interval or ratio level).
2. The $n$ pairs of data should constitute a random sample
3. The underlying distributions for the two variables should be far from bi-variate normal distributions.
4. The underlying distribution is skewed.
5. The underlying distribution is influenced by outliers.

# Spearman's Rank Correlation Coefficient Contd.

- Consider two variables $X$ and $Y$ that meet the assumptions of using a Spearman's Rank Correlation Coefficient such that we have $n$ pairs $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$;
- Let $R(X_i)$ represent the rank of $X_i$, where each rank is an integer, from 1 through $n$ , indicating **relative magnitude**.
- The measurements may be **ranked from high to low** (e.g. $R(X_i) = 1$ is assigned to the largest of the $X_i$ , $R(X_i) = 2$ is assigned to the next largest of the $X_i$ , and so on, with $R(X_i) = n$ is assigned to the least of the $X_i$ ).
- **Alternatively**, the measurements may be **ranked from low to high** ( $R(X_i) = 1$ is assigned to the least t of the $X_i$ , $R(X_i) = 2$ is assigned to the next least of the $X_i$ , and so on, with $R(X_i) = n$ is assigned to the largest of the $X_i$ ).

# Spearman's Rank Correlation Coefficient Contd.

- Similarly, the same argument could be applied to the $Y_i$ such that **sequence of ranking (either high to low or low to high) is the same as for** $R(X_i)$.
- Then **Spearman's Rank Correlation Coefficient** is defined as:

$$R_s = \frac{12\sum_{i=1}^{n} R(X_i)R(Y_i)}{n^3 - n} - \frac{3(n+1)}{n-1} \tag{48}$$

- **Alternatively (most often)**, the Spearman's rank correlation coefficient,$R_s$ can be define in terms of the **difference,** $d_i$**, for each pair of ranks** ($d_i = R(X_i) - R(Y_i)$) as below:

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{49}$$

# Properties of Spearman's rank correlation coefficient, $R_s$

- An $R_s = 0$ (no correlation) indicates that the magnitudes of the ranks of one variable are independent of the magnitudes of the ranks of the second variable.
- A positive value of $R_s$ (positive correlation) indicates that the $R(X_i)$'s tend to increase as the $R(Y_i)$'s increase; a negative $R_s$ (negative correlation) indicates that the $R(X_i)$'s tend to decrease as the $R(Y_i)$'s increase.

# Properties of Spearman's rank correlation coefficient Contd.

- If the sequence of ranks were identical for the two variables, we would say that there was a perfect positive correlation, and $R_s = 1$ .
- If the magnitudes of the ranks for one variable vary inversely with the magnitudes of the ranks of the other variable, we have a perfect negative correlation, where $R_s = -1$.

# Spearman's rank Correlation Coefficient: Example

**Example 9:**

The data below represents the body length (in centimeters) and weight (in grams) of adult snakes donated by $(X)$ and $(Y)$, respectively. Calculate the Spearman's rank correlation coefficient between the two variables and comment on it.

| Snake | X = Length ( cm ) | Y = Weight ( g ) |
|---|---|---|
| 1 | 60 | 136 |
| 2 | 69 | 198 |
| 3 | 66 | 194 |
| 4 | 64 | 140 |
| 5 | 54 | 93 |
| 6 | 67 | 172 |
| 7 | 59 | 116 |
| 8 | 65 | 174 |
| 9 | 63 | 145 |

# Spearman's rank Correlation Coefficient: Solution to Example 9

- By ranking from high to low (assigning $R(X_i) = 1$ to the largest of the $X_i$ , $R(X_i) = 2$ to the next largest of the $X_i$ , and so on, and doing similar to the $R(Y_i)$ (assigning $R(Y_i) = 1$ to the largest of the $Y_i$ , $R(Y_i) = 2$ to the next largest of the $X_i$ ,and so on),the table given can be extended as follows:

# Spearman's rank Correlation Coefficient: Solution to Example 9 Contd.

| Snake | X | Y | $R(X_i)$ | $R(Y_i)$ | $d_i = R(X_i) - R(Y_i)$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| 1 | 60 | 136 | 7 | 7 | 0 | 0 |
| 2 | 69 | 198 | 1 | 1 | 0 | 0 |
| 3 | 66 | 194 | 3 | 2 | 1 | 1 |
| 4 | 64 | 140 | 5 | 6 | -1 | 1 |
| 5 | 54 | 93 | 9 | 9 | 0 | 0 |
| 6 | 67 | 172 | 2 | 4 | -2 | 4 |
| 7 | 59 | 116 | 8 | 8 | 0 | 0 |
| 8 | 65 | 174 | 4 | 3 | 1 | 1 |
| 9 | 63 | 145 | 6 | 5 | 1 | 1 |
| **Total** | | | | | $\sum d_i = 0$ | $d_i^2 = 8$ |

# Solution to Example 9 Contd.

- Then **Spearman's Rank Correlation Coefficient** bewteen the variables the Length ($X$) and Weight ($Y$) will be:

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{50}$$

$$R_s = 1 - \frac{(6)(8)}{9^3 - 9} \tag{51}$$

$$R_s = 1 - \frac{48}{720} \tag{52}$$

$$R_s = 1 - 0.0667 = 0.9333 \tag{53}$$

- **Comment:** There is a strong positive relationship between the length and weight of snakes.

# Spearman's Rank Correlation Coefficient with Ties in Ranks and how it is treated

- **Ties** occurs if two or more values of a particular variable are the same value.
- When that happens,the ranks of the tied values is set equal to the **mean of the ranks they would have received if there were no repetition in the ordered data set**.
- **NB:** if there are ties in both variables, the ties are dealt with separately or independently.
- **Example 10:** Given the variable $X$ with values 10, 8, 7, 8,6, 9, 6. Then 1st rank is assigned to 10 because it is the biggest value, then 2nd to 9.
- Now there is a repetition of 8 twice. Since both values are same, the same rank will be assigned which would be **average of the ranks that we would have assigned if there were no repetition**.

# Spearman's Rank Correlation Coefficient with Ties in Ranks and how it is treated Contd.

- Thus, both 8 will receive the average of 3 and 4, 3.5 (i.e. $\frac{(3+4)}{2} = 3.5$).
- Then 5th rank is given to 7.
- Again,there is a repetition of 6 twice. Hence, the same rank will be assigned which would be **average of the ranks that we would have assigned if there were no repetition**. Thus, both 6 will receive the average of 6 and 7, 6.5 (i.e. $\frac{(6+7)}{2} = 6.5$).

# Adjustment in the Formular of the Spearman's Rank Correlation Coefficient to account for Ties in Ranks.

- The formular for the Spearman's rank correlation coefficient, $R_s$ adjusted when there are ties as below:

$$R_s = 1 - \frac{6(\sum_{i=1}^{n} d_i^2 - \sum_{i=1}^{m}(t_x)_i - \sum_{i=1}^{r}(t_y)_i)}{\sqrt{[(n^3 - n) - 2\sum_{i=1}^{m}(t_x)_i][(n^3 - n) - 2\sum_{i=1}^{m}(t_y)_i]}} \quad (54)$$

- where

$$\sum_{i=1}^{m}(t_x)_i = \frac{\sum(t_i^3 - t_i)}{12} \tag{55}$$

- and $m$ is number of sets or groups of ties in $X$; $t_i$ is the number of observations tied at each group of ties in variable $X$ and the summation is over all groups of tied $Xs$.
- and

$$\sum_{i=1}^{r}(t_y)_i = \frac{\sum(t_i^3 - t_i)}{12} \tag{56}$$

- so that $r$ is number of sets or groups of ties in $Y$; $t_i$ is the number of observations tied at each group of ties in variable $Y$ and the summation is over all groups of tied $Ys$.
- If $\sum_{i=1}(t_x)_i$ and $\sum_{i=1}^{m}(t_y)_i$ are both zero, then equation (49) becomes equal to equation (54).

# Example 11: Spearman's Rank Correlation Coefficient with Ties in Ranks

**Example 11:** Calculate Spearman's rank correlation coefficient for the following data:

| No. | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | 10 | 6 |
| 2 | 15 | 25 |
| 3 | 14 | 12 |
| 4 | 25 | 18 |
| 5 | 14 | 25 |
| 6 | 14 | 40 |
| 7 | 20 | 10 |
| 8 | 22 | 7 |

# Spearman's rank Correlation Coefficient with ties: Solution to Example 11

- In this example, there are ties in both variable $X$ and $Y$. By ranking from high to low and accounting for the presence of ties,the table given can be extended as follows:

# Spearman's rank Correlation Coefficient with ties: Solution to Example 11 Contd.

| No. | $X_i$ | $Y_i$ | $R(X_i)$ | $R(Y_i)$ | $d_i = R(X_i) - R(Y_i)$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| 1 | 10 | 6 | 8 | 8 | 0 | 0 |
| 2 | 15 | 25 | 4 | 2.5 | 1.5 | 2.25 |
| 3 | 14 | 12 | 6 | 5 | 1 | 1 |
| 4 | 25 | 18 | 1 | 4 | -3 | 9 |
| 5 | 14 | 25 | 6 | 2.5 | 3.5 | 12.25 |
| 6 | 14 | 40 | 6 | 1 | 5 | 25 |
| 7 | 20 | 10 | 3 | 6 | -3 | 9 |
| 8 | 22 | 7 | 2 | 7 | -4 | 16 |
| **Total** | | | | | $\sum d_i = 0$ | $d_i^2 = 83.50$ |

# Solution to Example 11 Contd.

- And since there are ties in both variables, then Spearman's Rank Correlation Coefficient between the variables $(X)$ and $(Y)$ will be:

$$R_s = 1 - \frac{6(\sum_{i=1}^{n} d_i^2 - \sum_{i=1}^{m}(t_x)_i - \sum_{i=1}^{r}(t_y)_i)}{\sqrt{[(n^3 - n) - 2\sum_{i=1}^{m}(t_x)_i][(n^3 - n) - 2\sum_{i=1}^{m}(t_y)_i]}} \quad (57)$$

- Now

$$\sum_{i=1}^{m}(t_x)_i = \frac{\sum(t_i^3 - t_i)}{12} = \frac{\sum 3^3 - 3)}{12} = 2 \quad (58)$$

- since $t_i = 3$, because there is only one set of ties in variable $X$ (i.e 14 occurs thrice.

# Solution to Example 11 Contd.

and

$$\sum_{i=1}^{r}(t_y)_i = \frac{\sum(t_i^3 - t_i)}{12} = \frac{\sum(2^3 - 2)}{12} \tag{59}$$

- since $t_i = 2$, because there is one set of ties in variable $Y$ (i.e 25 occurs twice.

$$R_s = 1 - \frac{6(83.5 - 2 - 0.5)}{\sqrt{[(8^3 - 8) - 2(2)][(8^3 - 8) - 2(0.5)]}} \tag{60}$$

$$R_s = 1 - \frac{6(81)}{\sqrt{[504 - 4][504 - 1]}} \tag{61}$$

$$R_s = 1 - \frac{486}{\sqrt{[500][503]}} = 1 - \frac{486}{\sqrt{251500}} \tag{62}$$

# Solution to Example 11 Contd.

$$R_s = 1 - \frac{486}{501.4978} = 1 - 0.9690 = 0.031 \tag{63}$$

- **Comment:** There is a weak positive relationship between the variables $X$ and $Y$.

# Kendall's Rank Correlation (KRCC)

- This is also a non-parametric statistic that measures the strength of dependence between two variables. It is commonly referred to as **Kendalls Tau coefficient**, after the Greek letter tau ($\tau$).
- It is a measure of the similarity of the orderings of the data when ranked by each of the quantities.
- **Note** however, that the Kendall's rank correlation coefficient can be computed from actual observations without first ranking them (i.e. without converting the observations to ranks).
- It is based on the tendency of the two variables to move in the same or opposite directions.

# Kendall's Rank Correlation Contd.

- The Kendalls tau coefficient can be defined as:

$$r_k = \tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{64}$$

$$r_k = \tau = \frac{2(C - D)}{n(n-1)} \tag{65}$$

- where $C$ is the **number of concordant pairs**, $D$ is the **number of discordant pairs** and $n$ is the number of paired observations.

# Kendall's Rank Correlation: Terminology and Computation

- Let $(X_i, Y_i)$ and $(X_j, Y_j)$ be a pair of observations.
- If $(X_j - X_i)$ and $(Y_j - Y_i)$ have the **same sign**, we say that the pair is **CONCORDANT**.
- Alternatively, if $(X_i - X_j)$ and $(Y_i - Y_j)$ have the same sign, we say that the pair is **CONCORDANT**.
- That is two pairs of observations are said to to form a concordant pair if the $X_i$'s and the $Y_i$'s change in the same direction
- For example the pair of observation $(2, 1)$ and $(3, 5)$ are **concordant** since

$$(X_j - X_i) = (3 - 2) = 1; (Y_j - Y_i) = (5 - 1) = 4 \tag{66}$$

have the same sign (i.e. both are **positive**). OR

$$(X_i - X_j) = (2 - 3) = -1; (Y_i - Y_j) = (1 - 5) = -4 \tag{67}$$

have the same sign (i.e. both are **negative**).

# Kendall's Rank Correlation: Terminology and Computation Contd.

- Let $(X_i, Y_i)$ and $(X_j, Y_j)$ be a pair of observations.
- If $(X_j - X_i)$ and $(Y_j - Y_i)$ have the **opposite signs**, we say that the pair is **DISCORDANT**.
- Alternatively, if $(X_i - X_j)$ and $(Y_i - Y_j)$ have the opposite signs, we say that the pair is **DISCORDANT**.
- That is two pairs of observations are said to to form a discordant pair if the $X_i$'s and the $Y_i$'s change in the opposite direction
- For example the pair of observation $(2, 3)$ and $(4, 1)$ are **discordant** since

$$(X_j - X_i) = (4 - 2) = 2; (Y_j - Y_i) = (1 - 3) = -3 \tag{68}$$

have the opposite signs (i.e. one is **positive** and the other is **negative**).

OR

$$(X_i - X_j) = (2 - 4) = -2; (Y_i - Y_j) = (3 - 1) = -4 \tag{69}$$

have the opposite signs (i.e. one is **negative** and the other is **positive**).

# Example 12: Kendall's Rank Correlation Coefficient

**Example 12:** The data below shows the ranking of 6 judges for two variables "Eloquence (X)" and "Intelligence(Y)" for a beauty pageant. Calculate the Kendall's rank correlation coefficient between the judges ranking.

| Judge | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| X | 4 | 6 | 2 | 3 | 1 | 5 |
| Y | 3 | 6 | 1 | 4 | 2 | 5 |

# Kendall's Rank Correlation Coefficient: Solution to Example 12

- From the table $n = 6$ . Let us denote a concordant pair as "+ "and a discordant pair as "-". Then the pairs of concordant and discordant pairs can be summarized in the table below

| **Pair** | Sign | **Pair** | Sign | **Pair** | Sign | **Pair** | Sign | **Pair** | Sign |
|---|---|---|---|---|---|---|---|---|---|
| AB | + | BC | + | CD | + | DE | + | EF | + |
| AC | + | BD | + | CE | - | DF | | | |
| AD | - | BE | + | CF | + | | | | |
| AE | + | BF | + | | | | | | |
| AF | + | | | | | | | | |

- From the table, the number of "+" signs is 13 . Hence the number of **concordant pairs** $C = 13$ . Also the number of "-" signs is 2 , hence the number of **discordant pairs** is $D = 2$.

- **Alternatively**, we can rearrange one of the variables in an ascending order or descending order. Then calculate the $S = C - D$ for each value and sum them.
- In this example, the variable $X_i$ is arranged in an ascending order as shown in the table below:

| Judge | E | C | D | A | F | B |
|---|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 | 6 |
| Y | 2 | 1 | 4 | 3 | 5 | 6 |

- The $S = C - D$ for the table above is
$S = C - D = (4-1)+(4-0)+(2-1)+(2-0)+(1-0) = 13-2 = 11$

# Kendall's Rank Correlation Coefficient: Solution to Example 12 Contd.

- Hence, the Kendalls tau coefficient is

$$r_k = \tau = \frac{2(C-D)}{n(n-1)} \tag{70}$$

$$r_k = \frac{2(13-2)}{6(6-1)} = \frac{2(11)}{6(5)} \tag{71}$$

$$r_k = \frac{11}{15} = 0.733 \tag{72}$$

- **Comment:** There is a strong positive relationship between the judges ranking.

# Kendall's Rank Correlation Coefficient with Ties

- A pair $(X_i, Y_i)$ and $(X_j, Y_j)$ is said to be tied if $X_i = X_i$ or $Y_i = Y_j$. A tied pair is neither concordant nor discordant.
- When tied pairs arise in the data, an adjustment is made to the Kendall tau coefficient. The adjusted coefficient is called **Kendall tau-b coefficient** and is defined as:

# Kendall's Rank Correlation Coefficient with Ties Contd.

$$r_k = \tau_\beta = \frac{(C - D)}{\sqrt{(N - T)}\sqrt{(N - U)}} \tag{73}$$

where

$$N = \frac{n(n-1)}{2} \tag{74}$$

$$T = \frac{\sum_{i=1}^{m} t_i(t_i - 1)}{2} \tag{75}$$

$$U = \frac{\sum_{i=1}^{r} t_i(t_i - 1)}{2} \tag{76}$$

# Kendall's Rank Correlation Coefficient with Ties Contd.

- so that $m$ is number of sets or groups of ties in $X$; $t_i$ is the number of observations tied at each group of ties in variable $X$ and the summation is over all groups of tied $Xs$.
- and $r$ is number of sets or groups of ties in $Y$; $t_i$ is the number of observations tied at each group of ties in variable $Y$ and the summation is over all groups of tied $Ys$.
- and $C$ and $D$ is calculated in the usual way.

# Example 13: Kendall's Rank Correlation Coefficient with ties

**Example 13:** Calculate the Kendalls tau correlation coefficient for the data below.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 68 | 70 | 71 | 71 | 72 | 77 | 77 | 86 | 87 | 88 | 91 | 91 |
| $Y_i$ | 64 | 65 | 77 | 80 | 72 | 65 | 76 | 88 | 72 | 81 | 90 | 96 |

# Kendall's Rank Correlation Coefficient with ties: Solution to Example 13

- From the table $n = 12$ and variable $X$ is already arranged in an ascending order:

| $X_i$ | 68 | 70 | 71 | 71 | 72 | 77 | 77 | 86 | 87 | 88 | 91 | 91 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y_i$ | 64 | 65 | 77 | 80 | 72 | 65 | 76 | 88 | 72 | 81 | 90 | 96 |

- Then the $S = C - D$ for each value and is calculated as $S = C - D = (11 - 0) + (9 - 0) + (4 - 4) + (4 - 3) + (5 - 1) + (5 - 0) + (4 - 1) + (2 - 2) + (3 - 0) + (2 - 0) = 49 - 11 = 38.$

# Kendall's Rank Correlation Coefficient with ties: Solution to Example 13 Contd.

- Now from the data, 71 occurs twice, 77 occurs twice and 91 occurs for variable $X$.
- Thus $m = 3; t_1 = 2;\ t_2 = 2;\ t_3 = 2$. Hence,

$$T = \frac{\sum_{i=1}^{m} t_i(t_i - 1)}{2} \tag{77}$$

$$T = \frac{2(2-1) + 2(2-1) + 2(2-1)}{2} = \frac{6}{2} = 3 \tag{78}$$

- Similarly from the data, 65 occurs twice, and 72 occurs twice for variable $Y$.
- Thus $r = 2; t_1 = 2;\ t_2 = 2$. Hence,

$$U = \frac{\sum_{i=1}^{r} t_i(t_i - 1)}{2} \tag{79}$$

$$U = \frac{2(2-1) + 2(2-1)}{2} = \frac{4}{2} = 2 \tag{80}$$

# Kendall's Rank Correlation Coefficient with ties: Solution to Example 13 Contd.

And

$$N = \frac{n(n-1)}{2} = \frac{12(12-1)}{2} = 66 \tag{81}$$

- Hence, the Kendalls tau-b coefficient is:

$$r_k = \tau_\beta = \frac{(C-D)}{\sqrt{(N-T)}\sqrt{(N-U)}} \tag{82}$$

$$r_k = \tau_\beta = \frac{38}{\sqrt{(66-3)}\sqrt{(66-2)}} \tag{83}$$

$$r_k = \tau_\beta = \frac{38}{\sqrt{(63)}\sqrt{(64)}} = \frac{38}{\sqrt{(63)(64)}} \tag{84}$$

$$r_k = \tau_\beta = \frac{38}{\sqrt{(4032}} = \frac{38}{63.4980} = 0.5984 \tag{85}$$

- **Comment:** There is a moderate positive relationship between variables $X$ and $Y$.