45)

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a non-parametric statistical test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. It is the non-parametric equivalent of the paired t-test.

When to Use

- When you have two related samples.
- When the differences between pairs are not normally distributed.

Steps

1. **Calculate Differences**: For each pair of observations, calculate the difference.
2. **Rank Differences**: Ignore the sign and rank the absolute differences.
3. **Apply Signs to Ranks**: Attach the sign of the differences to their ranks.
4. **Calculate Test Statistic**: Sum the ranks of the positive and negative differences separately and use these to calculate the test statistic.
5. **Compare to Critical Value**: Compare the calculated test statistic to a critical value from the Wilcoxon signed-rank table to determine significance.

Example

Imagine we want to test whether a new diet plan affects the weights of a group of individuals. We measure their weights before and after the diet plan.

Subject	Before (kg)	After (kg)	Difference (After - Before)	Absolute Difference	Rank	Signed Rank
1	70	68	-2	2	1	-1
2	80	78	-2	2	1	-1
3	75	72	-3	3	2	-2
4	85	83	-2	2	1	-1
5	60	62	2	2	1	1

Sum of positive ranks: 1 Sum of negative ranks: 1 + 1 + 2 + 1 = 5

We would compare the smaller of these sums (1) against the critical value for N=5N=5 at a chosen significance level (e.g., 0.05) from the Wilcoxon signed-rank table.

Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon (MWW) test, also known simply as the Mann-Whitney U test or the Wilcoxon rank-sum test, is a non-parametric test used to compare two independent samples to determine whether they come from the same distribution.

When to Use

- When you have two independent samples.
- When the data is not normally distributed.

Steps

1. **Combine and Rank Data**: Combine the data from both samples and rank them together.
2. **Sum Ranks**: Calculate the sum of ranks for each sample.
3. **Calculate Test Statistic (U)**: Use the rank sums to calculate the U statistic.
4. **Compare to Critical Value**: Compare the U statistic to the critical value from the Mann-Whitney U table.

Example

Suppose we want to test if two teaching methods affect test scores differently. We have two groups of students, each taught with a different method.

Method A Scores	Rank	Method B Scores	Rank
85	1	92	2
87	3	88	4
90	6	85	1
93	7	78	5
95	9	80	8

Sum of ranks for Method A: 1 + 3 + 6 + 7 + 9 = 26

Sum of ranks for Method B: 2 + 4 + 1 + 5 + 8 = 20

The U statistic is calculated using the following formula:

$$U = R1 - \frac{n1(n1+1)}{2}$$

where $R1$ is the sum of ranks for the first sample, and $n1$ is the sample size for the first sample.

For Method A:

$$U1 = 26 - \frac{5(5+1)}{2} = 26 - 15 = 11$$

For Method B:

$$U2 = 20 - \frac{5(5+1)}{2} = 20 - 15 = 5$$

We then compare the smaller U value (5) against the critical value for $n1$=5$n1$=5 and $n2$=5$n2$=5 at a chosen significance level (e.g., 0.05) from the Mann-Whitney U table.

Conclusion

Both tests serve specific purposes based on the nature of the data and the relationship between the samples. The Wilcoxon signed-rank test is used for related samples, while the Mann-Whitney-Wilcoxon test is used for independent samples. These tests provide robust alternatives to their parametric counterparts (paired t-test and independent t-test) when the data does not meet the assumptions of normality.

42)

Nonparametric Methods in Time Series Analysis

Nonparametric methods in time series analysis are techniques that do not assume a specific parametric form for the underlying data distribution. These methods are flexible and can adapt to a wide range of data patterns, making them useful when the data does not meet the assumptions of parametric models, such as normality, linearity, or homoscedasticity.

Examples of Nonparametric Methods in Time Series Analysis

1. **Sign Test**: Used to determine if there is a difference between two paired samples, based on the sign of the differences.
2. **Wilcoxon Signed-Rank Test**: Used to compare two related samples to assess whether their population mean ranks differ.
3. **Mann-Whitney-Wilcoxon Test**: Used to compare two independent samples to determine whether they come from the same distribution.
4. **Kruskal-Wallis Test**: Used to determine whether there are statistically significant differences between the medians of three or more independent groups.
5. **Rank Correlation**: Measures the strength and direction of the association between two ranked variables.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a nonparametric statistical test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. It is the nonparametric equivalent of the paired t-test.

When to Use

- When you have two related samples.
- When the differences between pairs are not normally distributed.

Steps

1. **Calculate Differences**: For each pair of observations, calculate the difference.
2. **Rank Differences**: Ignore the sign and rank the absolute differences.
3. **Apply Signs to Ranks**: Attach the sign of the differences to their ranks.

4. **Calculate Test Statistic**: Sum the ranks of the positive and negative differences separately and use these to calculate the test statistic.
5. **Compare to Critical Value**: Compare the calculated test statistic to a critical value from the Wilcoxon signed-rank table to determine significance.

Detailed Example

Imagine we want to test whether a new diet plan affects the weights of a group of individuals. We measure their weights before and after the diet plan.

Subject	Before (kg)	After (kg)	Difference (After - Before)	Absolute Difference	Rank	Signed Rank
1	70	68	-2	2	1	-1
2	80	78	-2	2	1	-1
3	75	72	-3	3	2	-2
4	85	83	-2	2	1	-1
5	60	62	2	2	1	1

Sum of positive ranks: 1 Sum of negative ranks: 1 + 1 + 2 + 1 = 5

The test statistic TT is the smaller of these sums, so T=1T=1.

To determine if this result is statistically significant, we compare TT to the critical value from the Wilcoxon signed-rank table for N=5N=5 at a chosen significance level (e.g., 0.05). For small samples, exact critical values can be used, and for larger samples, the distribution approaches a normal distribution.

Conclusion

The Wilcoxon signed-rank test is a powerful nonparametric method for comparing two related samples. It provides a robust alternative to the paired t-test when the differences between pairs do not meet the assumptions of normality. This makes it particularly useful in time series analysis for detecting changes between paired observations without assuming a specific distribution for the differences.

44)

Time Series Data

Week	1	2	3	4	5	6
Value	20	18	14	16	11	13

Naive Method Forecasts

Using the naive method, the forecast for each week is the value from the previous week. Hence:

- Forecast for Week 2 = Value of Week 1 = 20
- Forecast for Week 3 = Value of Week 2 = 18
- Forecast for Week 4 = Value of Week 3 = 14
- Forecast for Week 5 = Value of Week 4 = 16
- Forecast for Week 6 = Value of Week 5 = 11

Forecast Values

Week	2	3	4	5	6
Actual Value	18	14	16	11	13
Forecast Value	20	18	14	16	11

Errors

Week	2	3	4	5	6
Actual Value	18	14	16	11	13
Forecast Value	20	18	14	16	11
Error (Actual - Forecast)	18 - 20 = -2	14 - 18 = -4	16 - 14 = 2	11 - 16 = -5	13 - 11 = 2

Week	2	3	4	5	6
Absolute Error	2	4	2	5	2
Squared Error	4	16	4	25	4
Absolute Percentage Error	218×100≈11.11%182×100≈11.11%	414×100≈28.57%144×100≈28.57%	216×100≈12.50%162×100≈12.50%	511×100≈45.45%115×100≈45.45%	213×100≈15.38%132×100≈15.38%

a. Mean Absolute Error (MAE)

a. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\text{Error}_i| = \frac{2+4+2+5+2}{5} = \frac{15}{5} = 3$$

b. Mean Squared Error (MSE)

b. Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{Error}_i)^2 = \frac{4+16+4+25+4}{5} = \frac{53}{5} = 10.6$$

c. Mean Absolute Percentage Error (MAPE)

c. Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{Error}_i}{\text{Actual Value}_i}\right| \times 100 = \frac{11.11+28.57+12.50+45.45+15.38}{5} = \frac{11}{}$$

d. Forecast for Week 7

Using the naive method, the forecast for Week 7 is the actual value from Week 6:

Forecast for Week 7=Value of Week 6=13Forecast for Week 7=Value of Week 6=13

Summary

- **Mean Absolute Error (MAE)**: 3
- **Mean Squared Error (MSE)**: 10.6
- **Mean Absolute Percentage Error (MAPE)**: 22.60%
- **Forecast for Week 7**: 13

This detailed calculation provides all the necessary forecast accuracy measures using the naive method for the given time series data.

42)

In time series analysis, understanding the underlying data patterns is crucial for accurate modeling and forecasting. Here are some common types of data patterns that can be identified in time series plots:

1. **Trend**:
    - **Definition**: A long-term increase or decrease in the data.
    - **Characteristics**: The trend can be linear or nonlinear.
    - **Example**: A steady rise in the average global temperature over several decades.
2. **Seasonality**:
    - **Definition**: Regular and predictable changes that recur at specific periods (e.g., daily, monthly, yearly).
    - **Characteristics**: Seasonal patterns are consistent and repeat over time.
    - **Example**: Retail sales increasing during the holiday season each year.
3. **Cycle**:
    - **Definition**: Long-term oscillations or fluctuations that are not of a fixed period and are influenced by economic, financial, or natural conditions.
    - **Characteristics**: Cyclic patterns have a variable length and amplitude.
    - **Example**: Business cycles with alternating periods of economic expansion and contraction.
4. **Irregular/Noise**:
    - **Definition**: Random variations or residuals that do not follow a pattern.
    - **Characteristics**: These variations are unpredictable and do not repeat over time.
    - **Example**: Daily stock price movements influenced by unforeseen events.
5. **Stationarity**:

- **Definition**: The statistical properties of the series (mean, variance) do not change over time.
- **Characteristics**: Stationary series have a constant mean and variance.
- **Example**: White noise series with no discernible pattern.

6. **Non-Stationarity**:
    - **Definition**: The statistical properties of the series change over time.
    - **Characteristics**: Non-stationary series exhibit trends, seasonality, or varying variance.
    - **Example**: Time series of economic data with an upward trend and increasing variance over time.
7. **Level Shift**:
    - **Definition**: Sudden and permanent change in the mean level of the series.
    - **Characteristics**: The series abruptly changes to a new level and remains there.
    - **Example**: A sudden jump in sales due to a major product launch.
8. **Structural Break**:
    - **Definition**: Points in time where the data's statistical properties change.
    - **Characteristics**: These breaks indicate significant changes in the underlying process generating the data.
    - **Example**: A policy change affecting economic indicators.
9. **Volatility Clustering**:
    - **Definition**: Periods of high volatility followed by periods of low volatility.
    - **Characteristics**: Volatility tends to cluster, with large changes followed by large changes (of either sign) and small changes followed by small changes.
    - **Example**: Financial time series like stock returns showing periods of high and low market volatility.
10. **Outliers/Anomalies**:
    - **Definition**: Data points that deviate significantly from the rest of the observations.
    - **Characteristics**: Outliers can be due to measurement errors, sudden shocks, or rare events.
    - **Example**: A sudden spike in temperature due to an unusual weather event.

Visual Identification

These patterns can often be visually identified by plotting the time series data. A time series plot (a graph of the data points in time order) can reveal trends, seasonal effects, cyclic behaviors, and potential outliers or anomalies. Analysts and statisticians use these visual cues to choose appropriate models and methods for forecasting and analysis.

41)

Definition: Lag-one sample autocorrelation is a measure of the correlation between each value in a time series and the value that immediately precedes it. In other words, it quantifies the extent to which the value of a variable at time tt is related to its value at time $t−1t−1$.

Formula: The lag-one sample autocorrelation $r1r1$ is computed as:

$$r_1 = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- $xixi$ is the value of the time series at time ii.
- $x¯x¯$ is the mean of the time series.
- nn is the number of observations in the time series.

Steps to Calculate:

1. Calculate the mean of the time series.
2. Compute the numerator: the sum of the product of deviations of each value and its preceding value from the mean.
3. Compute the denominator: the sum of the squared deviations of each value from the mean.
4. Divide the numerator by the denominator to get the lag-one sample autocorrelation.

When to Use Lag-One Sample Autocorrelation

1. Detecting Serial Dependence:

- Lag-one sample autocorrelation is used to detect serial dependence in time series data. Serial dependence means that current values in a series depend on previous values.

2. Identifying Patterns:

- Identifying if there are patterns or trends in the time series. A significant positive autocorrelation indicates that high values tend to follow high values and low values tend to follow low values, suggesting a trend.
- A significant negative autocorrelation indicates that high values tend to follow low values and vice versa, suggesting a cyclical pattern.

3. Model Selection in Time Series Analysis:

- It helps in selecting appropriate models for time series forecasting. For example, if significant autocorrelation is detected, autoregressive models (AR models) might be appropriate.

4. Evaluating Model Residuals:

- When evaluating time series models, lag-one autocorrelation of residuals can help assess the adequacy of the model. Ideally, residuals should show no significant autocorrelation, indicating that the model has captured the serial dependencies in the data.

5. Financial Time Series Analysis:

- In financial markets, lag-one autocorrelation is used to analyze stock prices, returns, and other financial indicators to identify momentum or mean-reversion properties.

Given data series:

{23.32,32.33,32.88,28.98,33.16,26.33,29.88,32.69,18.98,21.23,26.66,29.89}{23.32,32.33,32.88,28.98,33.16,26.33,29.88,32.69,18.98,21.23,26.66,29.89}

Step 1: Calculate the mean of the data series

$x^{-}=1n\sum i=1nxix^{-}=n1i=1\sum nxi$

where nn is the number of data points.

$x^{-}=23.32+32.33+32.88+28.98+33.16+26.33+29.88+32.69+18.98+21.23+26.66+29.8912x^{-}=1223.32+32.33+32.88+28.98+33.16+26.33+29.88+32.69+18.98+21.23+26.66+29.89$

$x^{-}=336.3312=28.0275x^{-}=12336.33=28.0275$

Step 2: Compute the numerator of the autocorrelation formula

The numerator is the sum of the product of each pair of lagged values:

$$\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})$$

Let's compute this step-by-step:

$(23.32-28.0275)(32.33-28.0275)=(-4.7075)(4.3025)=-20.25520625(23.32-28.0275)(32.33-28.0275)=(-4.7075)(4.3025)=-20.25520625$

(32.33−28.0275)(32.88−28.0275)=(4.3025)(4.8525)=20.85755625(32.33−28.0275)(32.88−28.0275)=(4.3025)(4.8525)=20.85755625

(32.88−28.0275)(28.98−28.0275)=(4.8525)(0.9525)=4.62085625(32.88−28.0275)(28.98−28.0275)=(4.8525)(0.9525)=4.62085625

(28.98−28.0275)(33.16−28.0275)=(0.9525)(5.1325)=4.88555625(28.98−28.0275)(33.16−28.0275)=(0.9525)(5.1325)=4.88555625

(33.16−28.0275)(26.33−28.0275)=(5.1325)(−1.6975)=−8.70985625(33.16−28.0275)(26.33−28.0275)=(5.1325)(−1.6975)=−8.70985625

(26.33−28.0275)(29.88−28.0275)=(−1.6975)(1.8525)=−3.14520625(26.33−28.0275)(29.88−28.0275)=(−1.6975)(1.8525)=−3.14520625

(29.88−28.0275)(32.69−28.0275)=(1.8525)(4.6625)=8.63730625(29.88−28.0275)(32.69−28.0275)=(1.8525)(4.6625)=8.63730625

(32.69−28.0275)(18.98−28.0275)=(4.6625)(−9.0475)=−42.16795625(32.69−28.0275)(18.98−28.0275)=(4.6625)(−9.0475)=−42.16795625

(18.98−28.0275)(21.23−28.0275)=(−9.0475)(−6.7975)=61.47105625(18.98−28.0275)(21.23−28.0275)=(−9.0475)(−6.7975)=61.47105625

(21.23−28.0275)(26.66−28.0275)=(−6.7975)(−1.3675)=9.29830625(21.23−28.0275)(26.66−28.0275)=(−6.7975)(−1.3675)=9.29830625

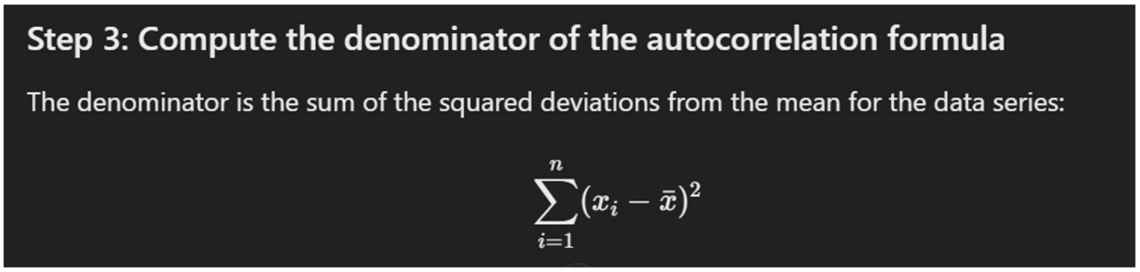
(26.66−28.0275)(29.89−28.0275)=(−1.3675)(1.8625)=−2.54601875(26.66−28.0275)(29.89−28.0275)=(−1.3675)(1.8625)=−2.54601875

Sum of the products:

−20.25520625+20.85755625+4.62085625+4.88555625−8.70985625−3.14520625+8.63730625−42.16795625+61.47105625+9.29830625−2.54601875=32.946365−20.25520625+20.85755625+4.62085625+4.88555625−8.70985625−3.14520625+8.63730625−42.16795625+61.47105625+9.29830625−2.54601875=32.946365

Step 3: Compute the denominator of the autocorrelation formula

The denominator is the sum of the squared deviations from the mean for the data series:



Let's compute this step-by-step:

(23.32−28.0275)2=(−4.7075)2=22.16120625(23.32−28.0275)2=(−4.7075)2=22.16120625

(32.33−28.0275)2=(4.3025)2=18.50280625(32.33−28.0275)2=(4.3025)2=18.50280625

(32.88−28.0275)2=(4.8525)2=23.53930625(32.88−28.0275)2=(4.8525)2=23.53930625

(28.98−28.0275)2=(0.9525)2=0.90725625(28.98−28.0275)2=(0.9525)2=0.90725625

(33.16−28.0275)2=(5.1325)2=26.34130625(33.16−28.0275)2=(5.1325)2=26.34130625

(26.33−28.0275)2=(−1.6975)2=2.88125625(26.33−28.0275)2=(−1.6975)2=2.88125625

(29.88−28.0275)2=(1.8525)2=3.43175625(29.88−28.0275)2=(1.8525)2=3.43175625

(32.69−28.0275)2=(4.6625)2=21.73930625(32.69−28.0275)2=(4.6625)2=21.73930625

(18.98−28.0275)2=(−9.0475)2=81.85625625(18.98−28.0275)2=(−9.0475)2=81.85625625

(21.23−28.0275)2=(−6.7975)2=46.19825625(21.23−28.0275)2=(−6.7975)2=46.19825625

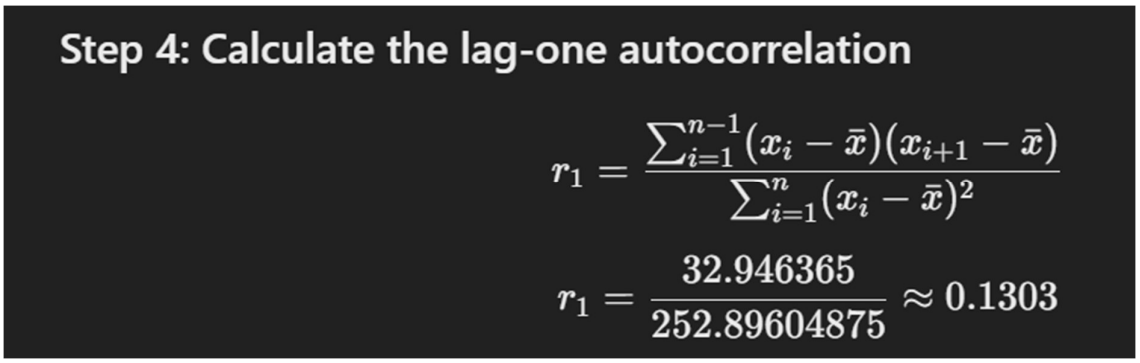
(26.66−28.0275)2=(−1.3675)2=1.86955625(26.66−28.0275)2=(−1.3675)2=1.86955625

(29.89−28.0275)2=(1.8625)2=3.46825625(29.89−28.0275)2=(1.8625)2=3.46825625

Sum of the squared deviations:

22.16120625+18.50280625+23.53930625+0.90725625+26.34130625+2.88125625+3.43175625+21.73930625+81.85625625+46.19825625+1.86955625+3.46825625=252.8960487522.16120625+18.50280625+23.53930625+0.90725625+26.34130625+2.88125625+3.43175625+21.73930625+81.85625625+46.19825625+1.86955625+3.46825625=252.89604875

Step 4: Calculate the lag-one autocorrelation



$r1$=32.946365252.89604875≈0.1303$r1$=252.8960487532.946365≈0.1303

Conclusion

The lag-one sample autocorrelation of the given time series data is approximately 0.13030.1303.

40) Already done!

39)

Given Data

Temperature (x)	Customer (y)
98	15
87	12
90	10
85	10
95	16
75	7

Step 1: Calculate the Means

Calculate the means of x (Temperature) and y (Customer).

$$\bar{x} = \frac{98 + 87 + 90 + 85 + 95 + 75}{6} = \frac{530}{6} = 88.33$$

$$\bar{y} = \frac{15 + 12 + 10 + 10 + 16 + 7}{6} = \frac{70}{6} = 11.67$$

$$\text{Cov}(x, y) = \frac{125.6277}{6} = 20.93795$$

Step 3: Calculate the Correlation Coefficient

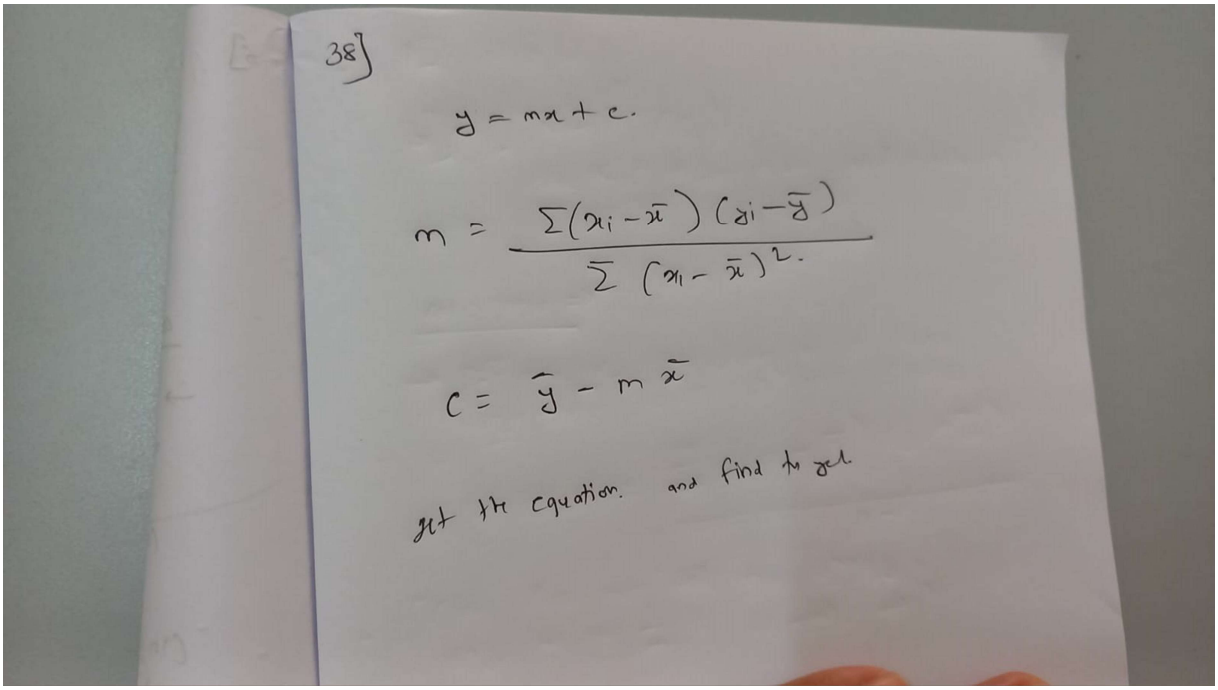
The formula for the correlation coefficient r is:

$$r = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

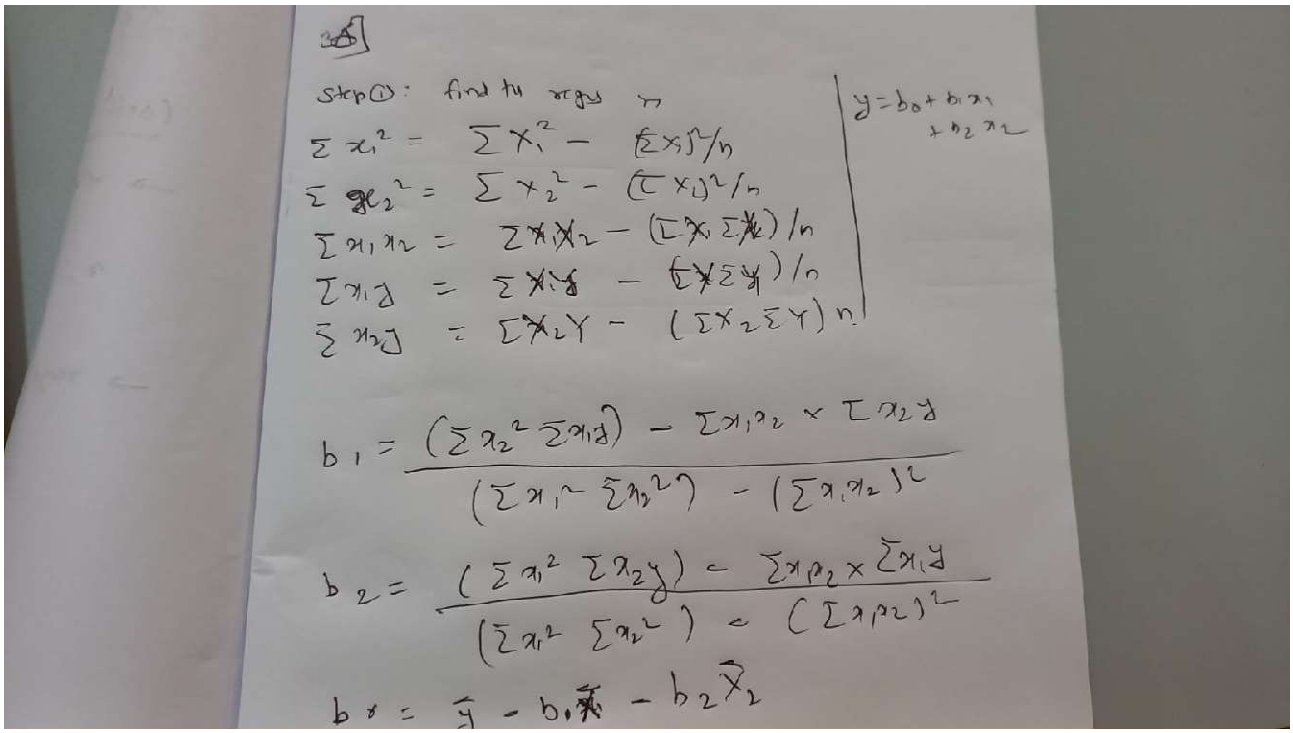
Where σ_x and σ_y are the standard deviations of x and y respectively.

After calculating every thing according to the formulas we wiil see that the values are storng releating in posttive manner

38,37,34)



36)

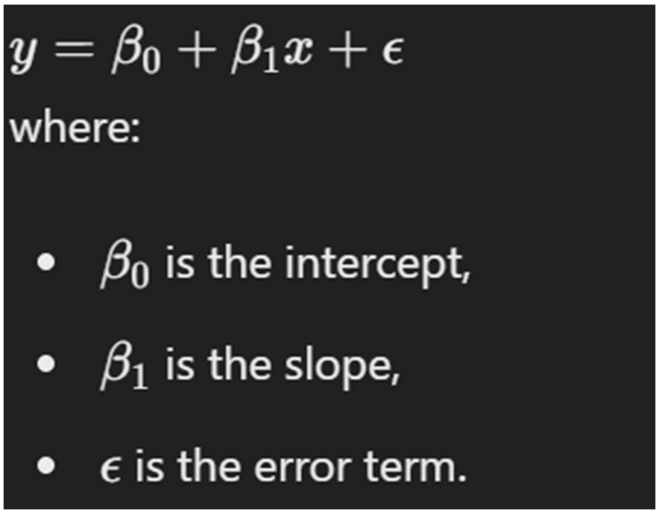


35)

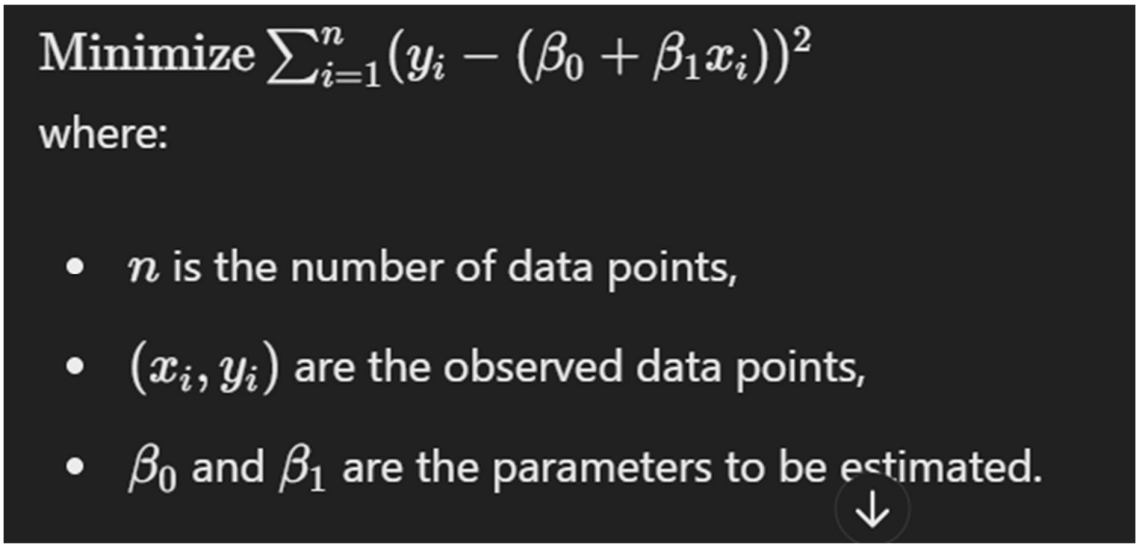
he least squares method is a mathematical technique used to find the best-fitting line or curve for a set of data points by minimizing the sum of the squares of the vertical deviations (residuals) from each data point to the line or curve. It's widely used in various fields, including statistics, economics, engineering, and machine learning.

Explanation of the Least Squares Method:

1. **Objective**: The primary goal of the least squares method is to find a line (in simple linear regression) or a hyperplane (in multiple linear regression) that minimizes the sum of squared differences between the observed and predicted values of the dependent variable.
2. **Mathematical Formulation**: For a simple linear regression model with one independent variable xx and one dependent variable yy, the equation of the line is typically represented as:



3. **Minimization Criterion**: The least squares method minimizes the sum of squared residuals (vertical distances between observed and predicted values):



4. **Estimating the Coefficients**: To find the best-fitting line, the method calculates the coefficients $\beta 0 \beta 0$ and $\beta 1 \beta 1$ that minimize the sum of squared residuals. This can be done using calculus or matrix algebra.

5. **Assessing Goodness of Fit**: Once the coefficients are estimated, the goodness of fit of the model is assessed using metrics such as R-squared, adjusted R-squared, and p-values.

Two Basic Categories of Least Squares Problems:

1. **Linear Least Squares**: In linear least squares problems, the relationship between the dependent variable and the independent variables is linear. This includes simple linear regression (one independent variable) and multiple linear regression (multiple independent variables).

2. **Nonlinear Least Squares**: In nonlinear least squares problems, the relationship between the dependent variable and the independent variables is nonlinear. This category includes models where the parameters appear nonlinearly, such as exponential, logarithmic, polynomial, or other nonlinear functions. Solving nonlinear least squares problems often requires iterative methods due to the nonlinearity of the model.

These categories cover a wide range of applications, and the least squares method is a powerful and versatile tool for estimating relationships between variables and making predictions based on data.

33)

In a simple linear regression model, where there is one independent variable and one dependent variable, several tests can be used to determine whether a linear association exists between them. These tests help assess the significance of the relationship and the goodness of fit of the model. Here are some common tests:

1. **F-test**: The F-test is used to assess the overall significance of the linear regression model. It tests whether at least one of the independent variables in the model has a non-zero coefficient. In the context of a simple linear regression model, this test evaluates whether the slope coefficient ($\beta 1 \beta 1$) is significantly different from zero. If the p-value of the F-test is below a chosen significance level (commonly 0.05), it suggests that there is a significant linear relationship between the independent and dependent variables.

2. **t-test for the Slope Coefficient**: The t-test is used to determine the significance of individual coefficients in the regression model. In the case of a simple linear regression model, the t-test assesses whether the slope coefficient ($\beta 1 \beta 1$) is significantly different from zero. A low p-value (typically less than 0.05) indicates that the independent variable has a significant effect on the dependent variable, implying a linear association.

3. **Coefficient of Determination ($R2R2$)**: $R2R2$ represents the proportion of the variance in the dependent variable that is explained by the independent variable(s). It ranges from 0 to 1, where 1 indicates a perfect fit. $R2R2$ can be interpreted as the percentage of variation in the dependent variable that is explained by the independent variable(s). A higher $R2R2$ value suggests a stronger linear relationship between the variables.

4. **Adjusted $R2R2$**: Adjusted $R2R2$ is a modified version of $R2R2$ that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary variables and provides a more accurate measure of the model's goodness of fit, particularly when comparing models with different numbers of predictors.

5. **Residual Analysis**: Residual analysis involves examining the residuals (the differences between observed and predicted values) to assess whether they are randomly distributed around zero. A plot of residuals versus predicted values should not show any discernible pattern or trend. Additionally, a histogram or Q-Q plot of residuals can be used to check for normality assumptions.

These tests collectively provide insights into the presence and strength of a linear association between the dependent and independent variables in a simple linear regression model, helping to determine the validity and reliability of the model's predictions.

32)

Simple linear regression is a statistical technique used to model the relationship between a single independent variable (predictor) and a single dependent variable (outcome). The goal of simple linear regression is to establish a linear relationship between the two variables and use this relationship to make predictions or infer the impact of changes in the independent variable on the dependent variable.

Components of a Simple Linear Regression Model:

1. **Dependent Variable (*yy*)**: The dependent variable, also known as the response variable, is the variable being predicted or explained by the independent variable. It is denoted by yy.
2. **Independent Variable (*xx*)**: The independent variable, also known as the predictor variable or regressor, is the variable used to predict or explain variation in the dependent variable. It is denoted by xx.
3. **Linear Relationship**: Simple linear regression assumes that the relationship between the independent and dependent variables is linear, meaning that a change in the independent variable is associated with a proportional change in the dependent variable. This relationship is represented by a straight line.
4. **Regression Equation**: The regression equation for a simple linear regression model is typically represented as: $y=\beta 0+\beta 1x+\epsilon y=\beta 0+\beta 1x+\epsilon$ where:
    - $\beta 0\beta 0$ is the intercept of the line (the value of yy when xx equals zero),
    - $\beta 1\beta 1$ is the slope of the line (the change in yy for a one-unit change in xx),
    - $\epsilon\epsilon$ is the error term, representing the difference between the observed and predicted values of yy.
5. **Assumptions**: Simple linear regression relies on several assumptions, including:

- Linearity: The relationship between xx and yy is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of the residuals (differences between observed and predicted values) is constant across all levels of the independent variable.
- Normality: The residuals are normally distributed.

6. **Estimation of Parameters**: The parameters $\beta 0 \theta 0$ and $\beta 1 \theta 1$ are estimated from the data using statistical methods such as the method of least squares. The goal is to find the line that minimizes the sum of squared differences between the observed and predicted values of yy.
7. **Interpretation of Results**: Once the regression coefficients are estimated, they can be interpreted to understand the relationship between the variables. The intercept ($\beta 0 \theta 0$) represents the predicted value of yy when xx equals zero, while the slope ($\beta 1 \theta 1$) represents the change in yy for a one-unit change in xx.

Simple linear regression provides a straightforward way to analyze the relationship between two variables and make predictions based on this relationship. It serves as the foundation for more complex regression techniques, such as multiple linear regression, which involve multiple independent variables.

31)

Linear regression and multiple regression are both statistical techniques used to model the relationship between variables, but they differ in terms of the number of independent variables they involve and the complexity of the relationships they can represent.

Linear Regression:

1. **Number of Independent Variables**: Linear regression involves only one independent variable (predictor). It models the relationship between this single independent variable and a dependent variable.
2. **Equation**: The equation for a simple linear regression model is: $y=\beta 0+\beta 1x+\epsilon y=в0+в1x+\epsilon$ where yy is the dependent variable, xx is the independent variable, $\beta 0в0$ is the intercept, $\beta 1в1$ is the slope, and $\epsilon\epsilon$ is the error term.
3. **Model Complexity**: Linear regression models represent linear relationships between variables. They assume that changes in the independent variable are associated with proportional changes in the dependent variable.
4. **Applications**: Linear regression is commonly used for predictive modeling and understanding the relationship between two variables, such as predicting house prices based on square footage or analyzing the impact of advertising spending on sales.

Multiple Regression:

1. **Number of Independent Variables**: Multiple regression involves two or more independent variables. It models the relationship between these multiple independent variables and a dependent variable.
2. **Equation**: The equation for a multiple regression model is: $y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta kxk+\epsilon y=в0+в1x1+в2x2+\ldots+вkxk+\epsilon$ where yy is the dependent variable, $x1,x2,\ldots,xkx1,x2,\ldots,xk$ are the independent variables, $\beta 0в0$ is the intercept, $\beta 1,\beta 2,\ldots,\beta kв1,в2,\ldots,вk$ are the slopes for each independent variable, and $\epsilon\epsilon$ is the error term.

3. **Model Complexity**: Multiple regression models are more complex than simple linear regression models because they can capture the combined effects of multiple independent variables on the dependent variable. They allow for nonlinear relationships and interactions between variables.
4. **Applications**: Multiple regression is used when analyzing the relationship between a dependent variable and multiple independent variables, such as predicting a person's salary based on their education level, years of experience, and location.

In summary, linear regression involves modeling the relationship between one independent variable and a dependent variable, while multiple regression involves modeling the relationship between multiple independent variables and a dependent variable. Multiple regression allows for more complex relationships and provides a more comprehensive analysis of the factors influencing the dependent variable.

Aspect	**Linear Regression**	**Multiple Regression**
Number of Independent Variables	One independent variable (predictor).	Two or more independent variables.
Equation	$y=\beta 0+\beta 1x+\epsilon y=\beta 0+\beta 1x+\epsilon$	$y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta kxk+\epsilon y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta kxk+\epsilon$
Model Complexity	Represents linear relationships.	Allows for nonlinear relationships and interactions between variables.
Applications	Predictive modeling and understanding the relationship between two variables.	Analyzing the relationship between a dependent variable and multiple independent variables.

30)

The least squares method is a statistical technique used to estimate the parameters of a mathematical model by minimizing the sum of the squared differences between the observed and predicted values of a dependent variable. It's commonly used in regression analysis to find the best-fitting line or curve for a set of data points.

Steps in the Least Squares Method:

1. **Define the Model**: First, we need to define the mathematical model that describes the relationship between the independent and dependent variables. In the context of linear regression, the model equation is typically represented as: $y=\beta 0+\beta 1x+\epsilon y=\beta 0+\beta 1x+\epsilon$ Here, yy is the dependent variable, xx is the independent variable, $\beta 0\beta 0$ is the intercept, $\beta 1\beta 1$ is the slope, and $\epsilon\epsilon$ is the error term.

2. **Calculate the Residuals**: For each data point, calculate the difference between the observed value of the dependent variable and the value predicted by the model. These differences are called residuals.

3. **Square the Residuals**: Square each residual to ensure that negative and positive differences do not cancel each other out when summing them up.

4. **Sum the Squared Residuals**: Sum up all the squared residuals to obtain the total sum of squares (SSR). This represents the total deviation of the observed values from the predicted values.

5. **Minimize the Sum of Squared Residuals**: Find the values of the parameters (intercept and slope) that minimize the sum of squared residuals. This is typically done using calculus or optimization algorithms.

6. **Estimate the Parameters**: Once the sum of squared residuals is minimized, we obtain the estimated values of the parameters ($\beta 0\beta 0$ and $\beta 1\beta 1$) that define the best-fitting line or curve for the data.

Example:

Let's consider a simple example of fitting a linear regression model to a set of data points. Suppose we have the following data:

x	y
1	2
2	3
3	5
4	4
5	6

We want to find the equation of the line that best fits these data points using the least squares method.

1. **Define the Model**: The model equation for linear regression is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

2. **Calculate the Residuals**: For each data point, calculate the difference between the observed value of y and the value predicted by the model:

$$\text{Residual} = y_{\text{observed}} - (\beta_0 + \beta_1 x)$$

3. **Square the Residuals**: Square each residual:

$$\text{Squared Residual} = (\text{Residual})^2$$

Sum the Squared Residuals: Sum up all the squared residuals to obtain the total sum of squares (SSR).

4. **Minimize the Sum of Squared Residuals**: Find the values of $\beta 0\hat{\beta}0$ and $\beta 1\hat{\beta}1$ that minimize the sum of squared residuals.

5. **Estimate the Parameters**: Once the sum of squared residuals is minimized, we obtain the estimated values of $\beta 0\hat{\beta}0$ and $\beta 1\hat{\beta}1$, which define the equation of the best-fitting line for the data.

In this example, the least squares method would be used to find the values of $\beta 0\hat{\beta}0$ and $\beta 1\hat{\beta}1$ that minimize the sum of squared residuals, resulting in the equation of the best-fitting line for the given data points.

29)

A goodness of fit test is a statistical method used to assess how well an observed data set fits a particular theoretical distribution or model. It helps determine whether the observed data deviates significantly from the expected values under the assumed model, providing insights into the validity and reliability of the model.

Purpose of Goodness of Fit Test:

The primary purpose of a goodness of fit test is to evaluate whether the observed data matches the expected distribution or model. It helps answer questions such as:

- How well does the assumed model describe the observed data?
- Are there any systematic patterns or discrepancies between the observed and expected values?
- Are the deviations between the observed and expected values statistically significant?

Key Concepts:

1. **Null Hypothesis ($H0H0$)**: The null hypothesis states that there is no significant difference between the observed data and the expected distribution or model. It assumes that any discrepancies are due to random variation.
2. **Alternative Hypothesis ($H1H1$)**: The alternative hypothesis contradicts the null hypothesis and suggests that the observed data significantly deviates from the expected distribution or model.
3. **Test Statistic**: The test statistic is a numerical value calculated from the observed and expected data, representing the degree of discrepancy between them. Common test statistics include the chi-square statistic, Kolmogorov-Smirnov statistic, and Anderson-Darling statistic.
4. **Critical Value or P-value**: Based on the chosen significance level ($\alpha\alpha$), a critical value or p-value is calculated. If the test statistic exceeds the critical value or if the p-value is less than

$\alpha\alpha$, the null hypothesis is rejected, indicating a poor fit between the observed data and the expected distribution.

Common Goodness of Fit Tests:

1. **Chi-Square Test**: The chi-square goodness of fit test is used to assess whether observed categorical data follows a specified theoretical distribution. It compares the observed frequencies to the expected frequencies under the null hypothesis.
2. **Kolmogorov-Smirnov Test**: The Kolmogorov-Smirnov test is used to evaluate the goodness of fit between the observed data and a continuous theoretical distribution. It compares the cumulative distribution function of the observed data to that of the theoretical distribution.
3. **Anderson-Darling Test**: The Anderson-Darling test is a variation of the Kolmogorov-Smirnov test that provides more sensitivity to differences in the tails of the distribution. It is often used for testing the fit of normality or other specific distributions.

Interpretation:

- If the test statistic exceeds the critical value or if the p-value is less than the chosen significance level ($\alpha\alpha$), the null hypothesis is rejected. This indicates that the observed data significantly deviates from the expected distribution, suggesting a poor fit.
- If the test statistic does not exceed the critical value or if the p-value is greater than $\alpha\alpha$, the null hypothesis is not rejected. This suggests that there is insufficient evidence to conclude that the observed data deviates significantly from the expected distribution, indicating a good fit.

In summary, a goodness of fit test provides a formal and objective way to evaluate the adequacy of a model or distribution in describing the observed data, helping researchers make informed decisions about the validity of their analyses.

28)

A:

Lec 30.pdf

B:

Lec 31.pdf

27)

To test the hypothesis that the dice is fair, we need to check if the observed frequencies of the scores are consistent with the expected frequencies under the assumption of a fair dice.

For a fair octahedral dice, each of the 8 possible scores (1, 2, 3, 4, 5, 6, 7, 8) has an equal probability of 1/8 to occur. If the dice is rolled n times, the expected frequency for each score is n/8.

Given information:

- Total number of rolls = 8 + 10 + 11 + 7 + 12 + 14 + 10 + 7 = 79

Step 1: Calculate the expected frequency for each score under the assumption of a fair dice.

Expected frequency for each score = Total number of rolls / Number of possible scores

Expected frequency = 79 / 8 = 9.875

Step 2: Calculate the test statistic (chi-square statistic) to compare the observed and expected frequencies.

Chi-square statistic = Σ [(Observed frequency - Expected frequency)^2 / Expected frequency]

Score	Observed Frequency	Expected Frequency	(Observed - Expected)^2 / Expected
1	8	9.875	0.3489
2	10	9.875	0.0001
3	11	9.875	0.1444

4	7	9.875	0.8164
5	12	9.875	0.4489
6	14	9.875	1.6900
7	10	9.875	0.0001
8	7	9.875	0.8164

Chi-square statistic = 0.3489 + 0.0001 + 0.1444 + 0.8164 + 0.4489 + 1.6900 + 0.0001 + 0.8164 = 4.2652

Step 3: Compare the calculated chi-square statistic with the critical value from the chi-square distribution table for the desired level of significance (e.g., 0.05) and the appropriate degrees of freedom (df = number of scores - 1 = 7).

The critical value for chi-square distribution with df = 7 and $\alpha = 0.05$ is approximately 14.07.

Since the calculated chi-square statistic (4.2652) is less than the critical value (14.07), we fail to reject the null hypothesis that the dice is fair.

Therefore, based on the given data, there is no evidence to suggest that the dice is biased or unfair at the 5% significance level.

26)

Lec 31.pdf

The level of significance and the rejection region are closely related concepts in hypothesis testing. Here's how they are connected:

1. **Level of Significance ($\alpha\alpha$)**:
    - The level of significance, often denoted by $\alpha\alpha$, is the probability of rejecting the null hypothesis ($H0H0$) when it is actually true. It represents the threshold for deciding whether the observed data is statistically significant.
    - Common values for $\alpha\alpha$ are 0.05, 0.01, and 0.10.
2. **Rejection Region**:
    - The rejection region (or critical region) is the set of all values of the test statistic that leads to the rejection of the null hypothesis.
    - This region is determined based on the level of significance $\alpha\alpha$. For a given $\alpha\alpha$, the rejection region is chosen such that the probability of the test statistic falling within this region is $\alpha\alpha$ if the null hypothesis is true.

Relationship:

- The level of significance $\alpha\alpha$ defines the size of the rejection region. For example, if α=0.05α=0.05, it means that there is a 5% chance of rejecting the null hypothesis when it is true. Consequently, the rejection region is determined so that the probability of the test statistic falling into this region is 0.05.

Types of Tests:

- **One-Tailed Test**: In a one-tailed test, the rejection region is located entirely in one tail of the distribution of the test statistic.
    - If α=0.05α=0.05, the rejection region is the extreme 5% in one tail of the distribution.
- **Two-Tailed Test**: In a two-tailed test, the rejection region is split between both tails of the distribution.
    - If α=0.05α=0.05, the rejection region is divided into two parts, with 2.5% in each tail, making up a total of 5%.

Example:

Suppose we are performing a z-test for a population mean:

- **Two-Tailed Test**:
    - If α=0.05α=0.05, the rejection region would be the values of the z-statistic that are less than −1.96−1.96 or greater than 1.961.96 because the area in the tails of the standard normal distribution corresponding to 0.05 (split as 0.025 in each tail) lies beyond these z-values.

- **One-Tailed Test**:
    - If α=0.05α=0.05, for a test where we are only concerned with values greater than a certain point, the rejection region would be the values of the z-statistic greater than 1.6451.645, which corresponds to the upper 5% of the standard normal distribution.

In summary, the level of significance $\alpha\alpha$ determines the size and placement of the rejection region in the distribution of the test statistic, thus directly influencing the decision rule for hypothesis testing.

25,23)

In the given hypothesis test, we have:

H0: ϕ0 = 5 (null hypothesis)

H1: ϕ1 > 5 (alternative hypothesis)

where ϕ represents the parameter being tested.

The test is a right-tailed test because the alternative hypothesis (H1: ϕ1 > 5) specifies that the parameter value is greater than the value specified in the null hypothesis (ϕ0 = 5).

In a right-tailed test, the critical region (the region of the test statistic where we reject the null hypothesis) lies in the right tail of the probability distribution. This means that we will reject the null hypothesis if the observed test statistic falls in the upper tail of the distribution, indicating that the parameter value is significantly larger than the value specified in the null hypothesis.

Right-tailed tests are used when the alternative hypothesis suggests that the parameter value is greater than the null value. In this case, we are interested in finding evidence that the parameter ϕ is greater than 5, which corresponds to the right tail of the distribution.

On the other hand, if the alternative hypothesis were H1: ϕ1 < 5, it would be a left-tailed test, where the critical region would lie in the left tail of the distribution. And if the alternative hypothesis were H1: ϕ1 ≠ 5, it would be a two-tailed test, where the critical region would lie in both tails of the distribution.

In summary, the given hypothesis test with H1: ϕ1 > 5 is a right-tailed test because we are interested in finding evidence that the parameter ϕ is greater than the value specified in the null hypothesis.

24)

To identify the dependent and independent attributes, we need to understand the relationship between the variables being studied.

In the statement "How does the amount of makeup one applies affect how clear their skin is?", we can identify the following attributes:

Independent Attribute (Variable):

- The amount of makeup one applies

Dependent Attribute (Variable):

- How clear their skin is

The independent attribute (variable) is the one that is varied or controlled in an experiment or study. In this case, it is "the amount of makeup one applies." This variable is considered independent because its values are determined by the researcher or the person applying the makeup.

The dependent attribute (variable) is the one that is observed or measured for changes in response to variations in the independent variable. In this statement, it is "how clear their skin is." The clarity of the skin is the outcome or effect that may depend on the amount of makeup applied.

The statement is essentially asking how the independent attribute (amount of makeup applied) influences or affects the dependent attribute (clarity of the skin).

In summary:

Independent Attribute (Variable): The amount of makeup one applies

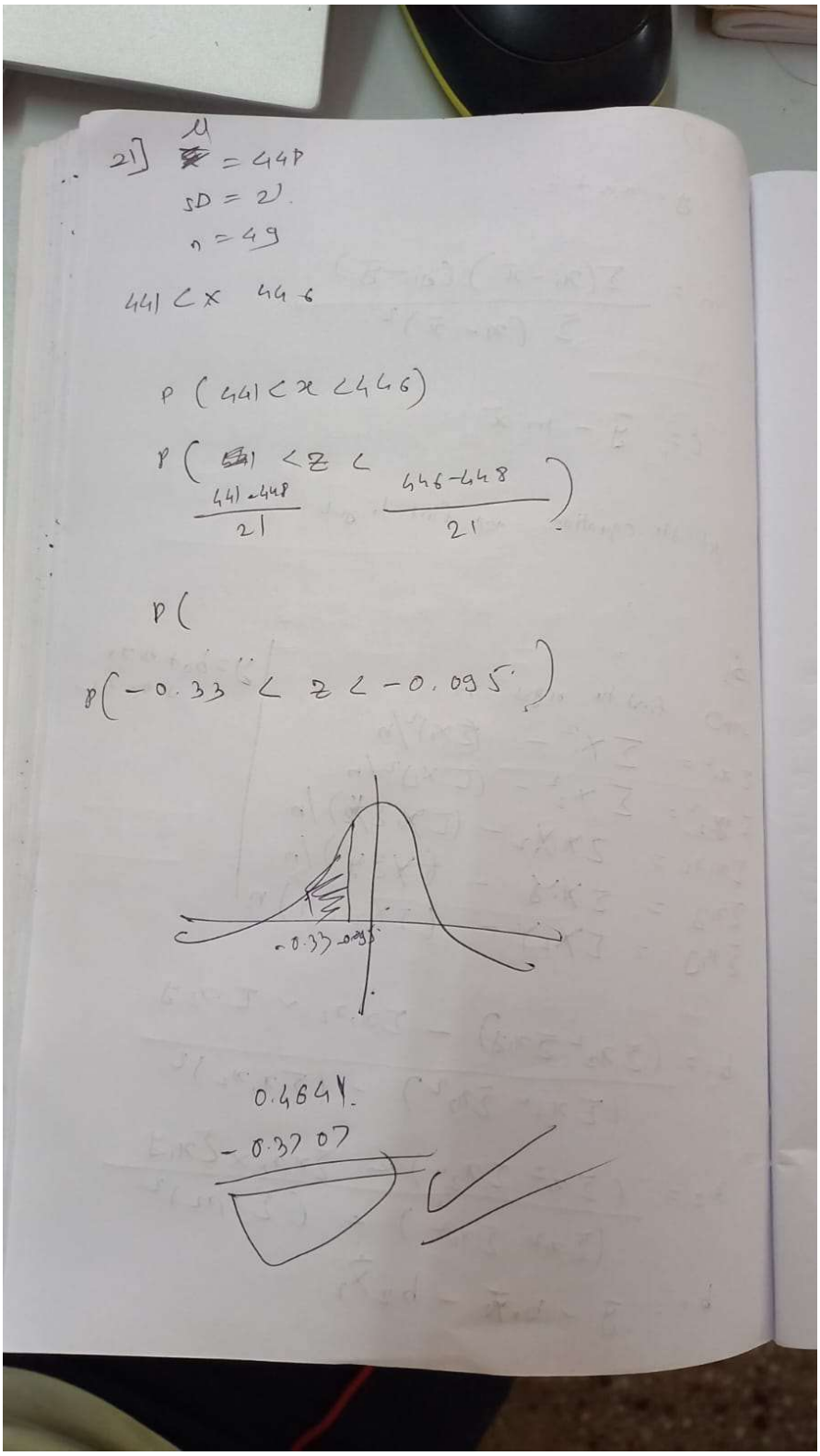
Dependent Attribute (Variable): How clear their skin is

The relationship being studied is whether the amount of makeup applied (independent variable) has an effect on the clarity of the skin (dependent variable).

22)

"D:\OneDrive - Vidyalankar Institute of Technology\Study_material\Semester VI\QA\Notes\QA Chapter 2 Datacollection and sampling methods.pptx"

21)



Note this is wrong as I have used the the sd of population pls refer the next example

20)

To solve this problem, we can use the central limit theorem and the standard normal distribution (Z-distribution). According to the central limit theorem, the distribution of sample means approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution.

Given:

- Population mean ($\mu\mu$): \$75.00
- Population standard deviation ($\sigma\sigma$): \$8.00
- Sample size (nn): 30
- Desired sample mean ($x^{-}x^{-}$): \$77.00

First, we need to calculate the standard error of the mean ($SESE$), which represents the standard deviation of the sample means:

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE = \frac{8.00}{\sqrt{30}}$$
$$SE \approx \frac{8.00}{5.48}$$
$$SE \approx 1.46$$

Next, we calculate the Z-score corresponding to the desired sample mean ($x^{-}x^{-}$) using the formula:

$$Z = \frac{\bar{x} - \mu}{SE}$$

$$Z = \frac{77.00 - 75.00}{1.46}$$
$$Z \approx \frac{2.00}{1.46}$$
$$Z \approx 1.37$$

Now, we can use the standard normal distribution table or a calculator to find the probability that a Z-score is less than or equal to 1.37. This represents the probability of observing a sample mean of \$77.00 or more.

From the standard normal distribution table, we find that the probability corresponding to Z = 1.37 is approximately 0.9147.

Therefore, the probability that the sample average expenditure per customer for this sample will be $77.00 or more is approximately 0.9147, or 91.47%.

19) QA

18)

Multiplication Law and Bayes' Theorem are two important concepts in probability theory. Here's a description of each, along with examples:

1. Multiplication Law:

The Multiplication Law, also known as the Product Rule, is used to calculate the probability of the intersection (joint occurrence) of two or more events. It states that the probability of two events occurring together is equal to the probability of one event multiplied by the conditional probability of the other event, given that the first event has occurred.

The formula for the Multiplication Law is:

$P(A \cap B) = P(A) \times P(B|A)$

Where:

- $P(A \cap B)$ is the probability of both events A and B occurring together
- $P(A)$ is the probability of event A occurring
- $P(B|A)$ is the conditional probability of event B occurring, given that event A has occurred

Example: Suppose you have a bag containing 3 red balls and 2 blue balls. Consider the following events:

A: Selecting a red ball

B: Selecting a blue ball

To find the probability of selecting a red ball and then a blue ball (without replacement):

$P(A) = 3/5$ (probability of selecting a red ball)

$P(B|A) = 2/4$ (conditional probability of selecting a blue ball, given that a red ball has been drawn and not replaced)

Using the Multiplication Law:

$P(A \cap B) = P(A) \times P(B|A)$

$P(A \cap B) = (3/5) \times (2/4) = 0.3$ or 30%

2. Bayes' Theorem:

Bayes' Theorem is a fundamental concept in probability theory that relates the conditional probabilities of two events. It provides a way to update the probability of an event based on new evidence or information.

The formula for Bayes' Theorem is:

$P(A|B) = (P(B|A) \times P(A)) / P(B)$

Where:

- $P(A|B)$ is the conditional probability of event A occurring, given that event B has occurred
- $P(B|A)$ is the conditional probability of event B occurring, given that event A has occurred
- $P(A)$ is the prior probability of event A occurring
- $P(B)$ is the probability of event B occurring

Example: Suppose a medical test for a certain disease has a 95% accuracy rate for people who have the disease (true positive rate) and a 90% accuracy rate for people who don't have the disease (true negative rate). If the prevalence of the disease in the population is 2%, what is the probability that a person has the disease given a positive test result?

Let:

A: Person has the disease

B: Positive test result

Given:

$P(B|A) = 0.95$ (true positive rate)

$P(A) = 0.02$ (prevalence of the disease)

$P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A) = (0.95 \times 0.02) + (0.1 \times 0.98) = 0.116$ (probability of a positive test result)

Using Bayes' Theorem:

$P(A|B) = (P(B|A) \times P(A)) / P(B)$

$P(A|B) = (0.95 \times 0.02) / 0.116 \approx 0.1638$ or 16.38%

Therefore, given a positive test result, the probability that a person has the disease is approximately 16.38%.

Bayes' Theorem is widely used in various fields, such as medical diagnosis, machine learning, and decision theory, to update probabilities based on new evidence or data.

17)

Lect 16.pdf

Lect 17.pdf

16)

Lect 17.pdf

refer 21

15)

Based on the updated data in the image, here are the correct solutions:

a) What is the probability that a machinist randomly selected from the polled group mildly supports the package? Total number of machinists polled = 9 + 11 + 2 + 4 + 4 = 30 Probability of a machinist mildly supporting the package = 11 / 30 ≈ 0.3667 or 36.67%

b) What is the probability that an inspector randomly selected from the polled group is undecided? Total number of inspectors polled = 10 + 3 + 2 + 8 + 7 = 30 Probability of an inspector being undecided = 2 / 30 ≈ 0.0667 or 6.67%

c) What is the probability that a worker (machinist or inspector) randomly selected from the polled group strongly or mildly supports the package? Total number of workers polled = 30 + 30 = 60 Number of workers who strongly or mildly support the package = 9 + 11 + 10 + 3 = 33 Probability of a worker strongly or mildly supporting the package = 33 / 60 ≈ 0.55 or 55%

d) What types of probability estimates are these? These are empirical probabilities or relative frequencies calculated based on the data collected from the poll. They provide an estimate of the true probabilities in the population, assuming that the sample (the polled group) is representative.

14)

To solve this problem, we can use Bayes' theorem, which states:

$$P(\text{Box } i|\text{ Red ball}) = \frac{P(\text{Red ball} \mid \text{Box } i) \times P(\text{Box } i)}{P(\text{Red ball})}$$

We need to find the probability that the second box was chosen given that a red ball was drawn ($P(\text{Box } 2 \mid \text{Red ball})$).

Given:

- $P(\text{Red ball} \mid \text{Box } 1) = \frac{3}{5}$
- $P(\text{Red ball} \mid \text{Box } 2) = \frac{4}{9}$
- $P(\text{Red ball} \mid \text{Box } 3) = \frac{2}{6} = \frac{1}{3}$
- $P(\text{Box } 1) = P(\text{Box } 2) = P(\text{Box } 3) = \frac{1}{3}$

First, let's calculate the probability of drawing a red ball from any box ($P(\text{Red ball})$):
$P(\text{Red ball}) = P(\text{Red ball} \mid \text{Box } 1) \times I\downarrow\text{ox } 1) + P(\text{Red ball} \mid \text{Box } 2) \times P(\text{Box } 2) +$
$P(\text{Red ball} \mid \text{Box } 3) \times P(\text{Box } 3)$

$= \frac{3}{5} \times \frac{1}{3} + \frac{4}{9} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3}$
$= \frac{1}{5} + \frac{4}{27} + \frac{1}{9}$
$= \frac{27}{135} + \frac{20}{135} + \frac{15}{135}$
$= \frac{62}{135}$

Now, let's use Bayes' theorem to find $P(\text{Box } 2 \mid \text{Red ball})$:
$P(\text{Box } 2 \mid \text{Red ball}) = \frac{P(\text{Red ball} \mid \text{Box } 2) \times P(\text{Box } 2)}{P(\text{Red ball})}$
$= \frac{\frac{4}{9} \times \frac{1}{3}}{\frac{62}{135}}$
$= \frac{4}{9} \times \frac{1}{3} \times \frac{135}{62}$
$= \frac{4 \times 135}{9 \times 3 \times 62}$
$= \frac{540}{558}$
≈ 0.9662

Therefore, the probability that the second box was chosen given that a red ball was drawn is approximately 0.9662, or 96.62%.

13)Basic

12)

Lect 16.pdf

11)

Scales of measurement, also known as levels of measurement or types of data, classify variables into different categories based on the nature of the information they represent. There are four primary scales of measurement: nominal, ordinal, interval, and ratio. These scales differ in terms of the properties they possess and the types of statistical analyses that can be performed on data measured at each level.

1. Nominal Scale:

- Nominal data represent categories or labels without any inherent order or ranking.
- Examples include gender (male, female), eye color (blue, brown, green), and marital status (married, single, divorced).
- Statistical operations such as counts and percentages are often used with nominal data. Frequencies and mode are common measures of central tendency.

2. Ordinal Scale:

- Ordinal data represent categories with a meaningful order or ranking.
- Examples include Likert scales (e.g., strongly disagree, disagree, neutral, agree, strongly agree), educational levels (e.g., high school, bachelor's degree, master's degree), and income categories (e.g., low, medium, high).
- In addition to measures used for nominal data, median and percentile ranks can be calculated for ordinal data.

3. Interval Scale:

- Interval data represent measurements where the differences between values are meaningful and consistent, but there is no true zero point.
- Examples include temperature measured in Celsius or Fahrenheit, calendar dates, and IQ scores.
- Interval data allow for addition and subtraction operations, but ratios of values are not meaningful. Statistical operations such as mean and standard deviation can be used.

4. Ratio Scale:

- Ratio data represent measurements where both the differences between values and the ratios of values are meaningful, and there is a true zero point.
- Examples include height, weight, age, income (in dollars), and number of items.
- All arithmetic operations (addition, subtraction, multiplication, division) are meaningful for ratio data. Statistical operations such as mean, median, mode, range, standard deviation, and coefficient of variation are applicable.

Comparison between Categorical and Quantitative Data:

- **Nature**: Categorical data represent groups or categories, while quantitative data represent numerical measurements.

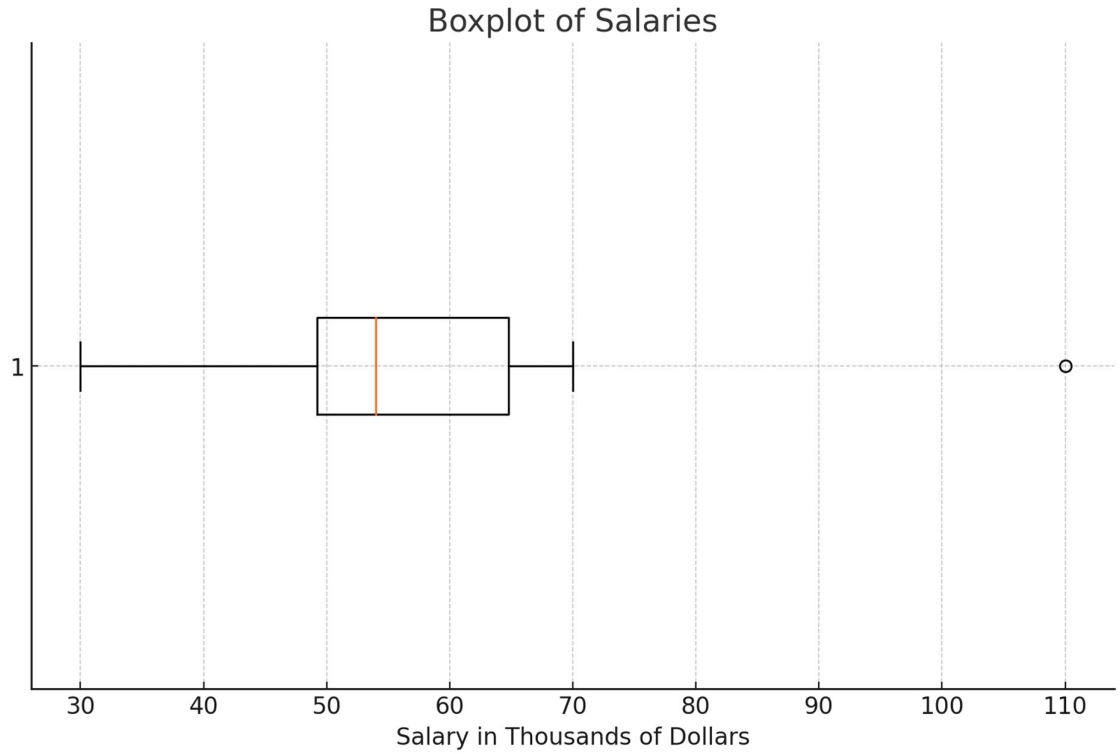
- **Representation**: Categorical data are typically represented by labels or names, while quantitative data are represented by numerical values.
- **Measurement Scales**: Categorical data are often measured at the nominal or ordinal scale, while quantitative data are measured at the interval or ratio scale.
- **Analysis**: Different statistical techniques are used to analyze categorical and quantitative data. For categorical data, measures of frequency, proportions, and association (e.g., chi-square test) are commonly used. For quantitative data, measures of central tendency, dispersion, and correlation (e.g., Pearson correlation coefficient) are used, along with parametric and nonparametric tests depending on the data distribution and research questions.

In summary, scales of measurement classify variables into different categories based on the nature of the information they represent, while categorical and quantitative data differ in terms of their representation, measurement scales, and the types of statistical analyses used to analyze them.

Module 1_ data & Statistics.pdf

DS part2.pdf

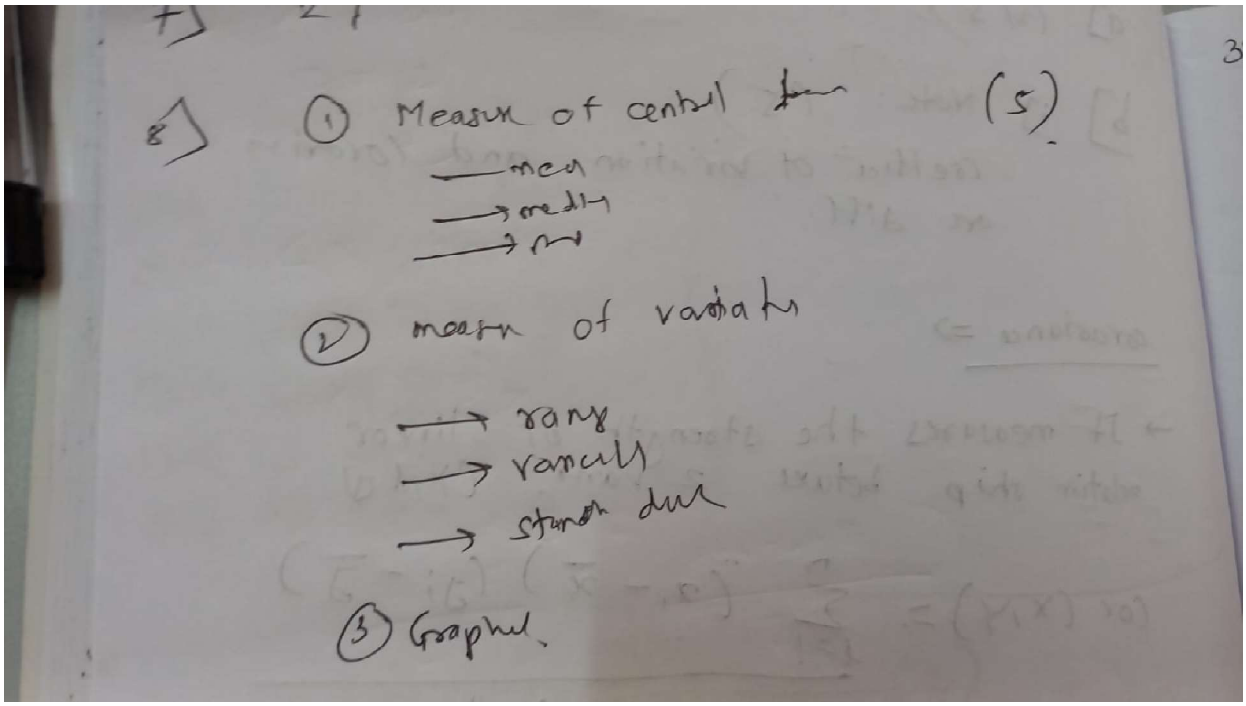
10) plot it!



10) Basic

9) Measures of variations.pdf

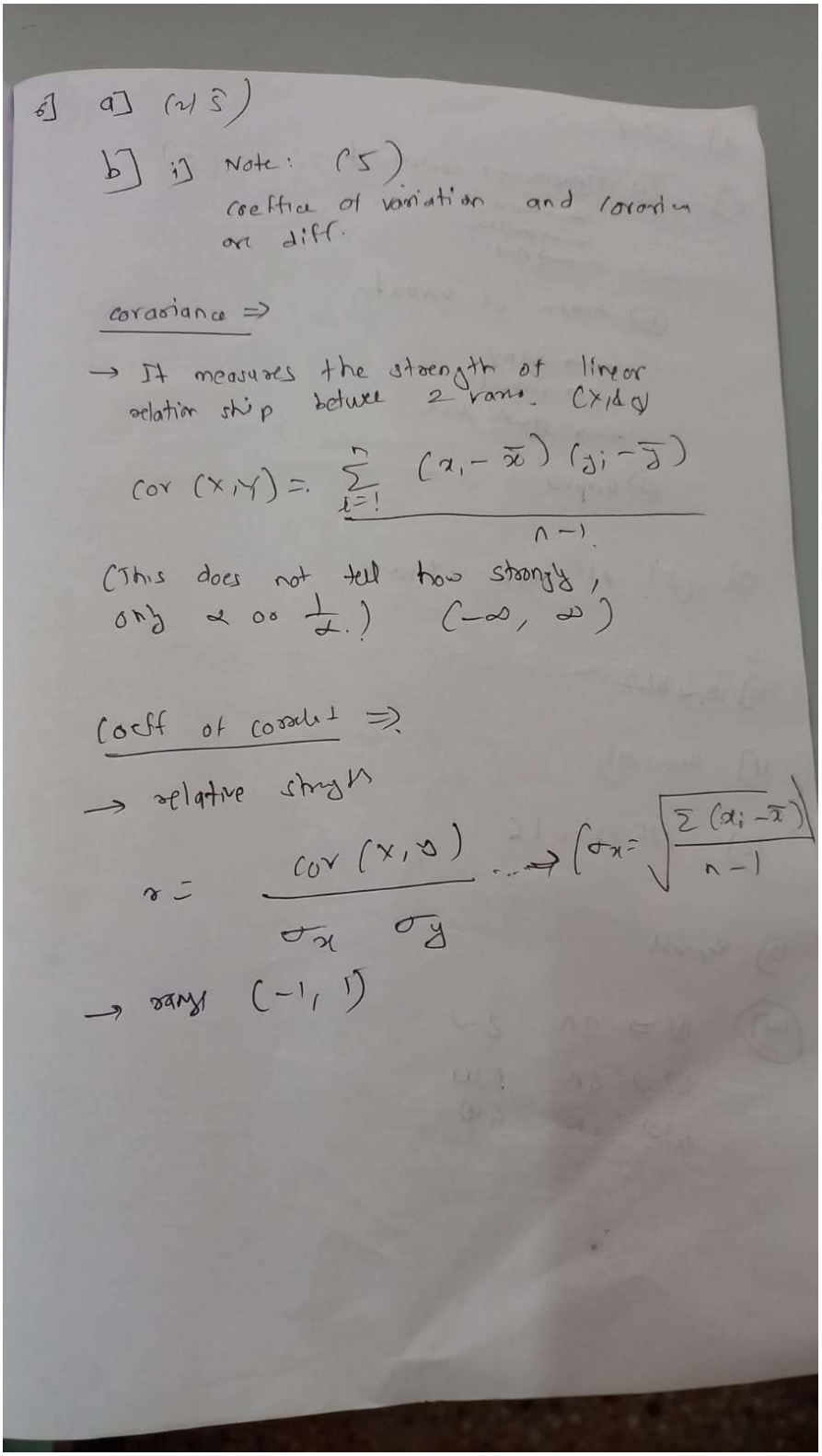
8)



7)

Basic

6)


5] a] (2)
b] i] Note: (5)
coeffic of variation and covarian are diff.
covariance ⇒
→ It measures the strength of linear relation ship betwee 2 vars. (X, & Y)
cov (X,Y) = $\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
(This does not tell how strongly, only ∝ as $\frac{1}{x}$.) (−∞, ∞)
Coeff of correlat ⇒
→ relative strgth
$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$ ⟶ ($\sigma_x = \sqrt{\frac{\Sigma (x_i - \bar{x})}{n-1}}$
→ range (−1, 1)

1-5) Basics