Unit 2: Working with Two and Three Variables

Introduction

Understanding the relationship between multiple variables is crucial in data exploration and visualization. This chapter focuses on handling two and three variables effectively through various techniques, including scatterplots, transformations, and contingency tables.

1. Relationship between Two Variables

Definition

The relationship between two variables is a fundamental concept in statistics and data analysis. Understanding how one variable influences or is associated with another helps in predicting outcomes, drawing insights, and making informed decisions.

Types of Relationships

a. Positive Relationship

In a positive relationship, as one variable increases, the other variable also tends to increase.

Example: Height and weight. Generally, taller individuals weigh more.

Visualization

A scatterplot of height (X-axis) versus weight (Y-axis) would show an upward trend.

b. Negative Relationship

In a negative relationship, as one variable increases, the other variable tends to decrease.

Example: The amount of gas in a tank and the distance you can drive. As the gas decreases, the distance you can drive also decreases.

Visualization

A scatterplot would show a downward trend.

c. No Relationship

When two variables do not show any consistent pattern or association, they are said to have no relationship.

Example: The amount of ice cream sold and the height of a person. These two variables are likely unrelated.

Visualization

A scatterplot would show a random pattern with no discernible trend.

2. Correlation Coefficient

The **Pearson correlation coefficient (r)** quantifies the strength and direction of a linear relationship between two variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- n = number of data points
- Σxy = sum of the product of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

- **Values of r:**
  - 1: Perfect positive correlation
  - −1: Perfect negative correlation
  - 0: No correlation

Example

Let's calculate the correlation coefficient for the data on hours studied and exam scores:

Hours Studied (X)	Exam Score (Y)
1	50
2	55
3	70
4	80
5	90

Step 1: Calculate sums

- $n = 5$
- $\Sigma x = 1 + 2 + 3 + 4 + 5 = 15$
- $\Sigma y = 50 + 55 + 70 + 80 + 90 = 345$
- $\Sigma xy = (1 \cdot 50) + (2 \cdot 55) + (3 \cdot 70) + (4 \cdot 80) + (5 \cdot 90) = 50 + 110 + 210 + 320 + 450 = 1140$
- $\Sigma x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 1 + 4 + 9 + 16 + 25 = 55$
- $\Sigma y^2 = 50^2 + 55^2 + 70^2 + 80^2 + 90^2 = 2500 + 3025 + 4900 + 6400 + 8100 = 24925$

Step 2: Substitute into the formula

$$r = \frac{5(1140) - (15)(345)}{\sqrt{[5 \cdot 55 - 15^2][5 \cdot 24925 - 345^2]}}$$

Calculating the numerator:

$$5(1140) - (15)(345) = 5700 - 5175 = 525$$

Calculating the denominator:

- For x:

$$5 \cdot 55 - 15^2 = 275 - 225 = 50$$

- For y:

$$5 \cdot 24925 - 345^2 = 124625 - 119025 = 5600$$

Combine the Result:

$$\sqrt{50 \cdot 5600} = \sqrt{280000} \approx 529.15$$

Finally:

$$r \approx \frac{525}{529.15} \approx 0.993$$

Interpretation

- r=0.993 indicates a very strong positive correlation between hours studied and exam scores, suggesting that as study hours increase, exam scores also tend to increase.

3. Scatterplots

Creating a Scatterplot

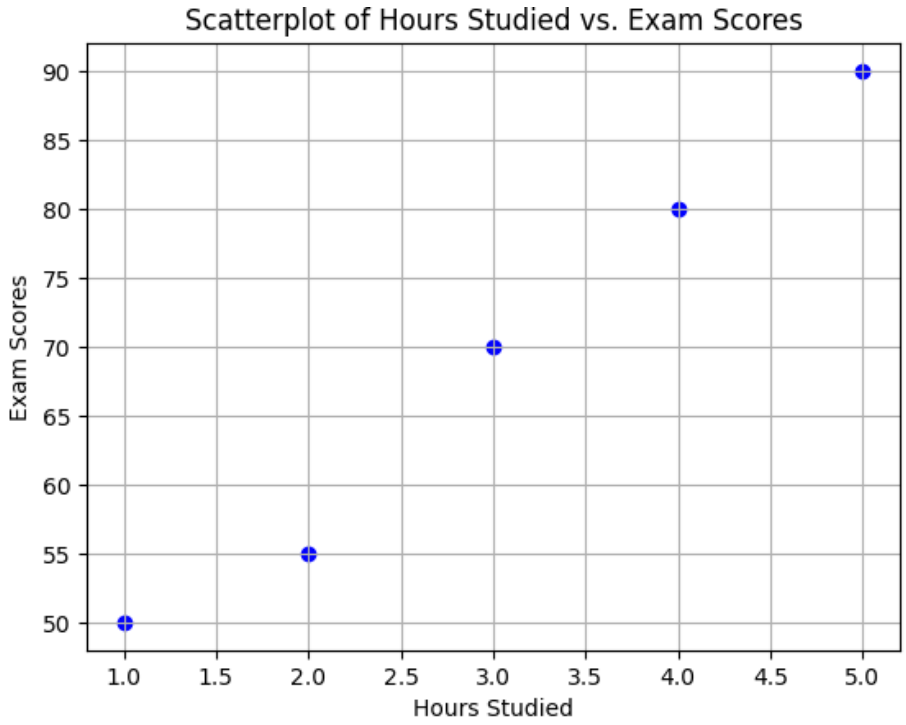
Scatterplots visually represent the relationship between two quantitative variables. Here's how to create one based on our example:

1. **Data Points:**
    - (1, 50)
    - (2, 55)
    - (3, 70)
    - (4, 80)
    - (5, 90)
2. **Axes:**
    - **X-axis:** Hours Studied
    - **Y-axis:** Exam Scores
3. **Plot the Points:**
    - Mark each pair of values on the graph.

Visual Representation:

As mentioned earlier, a simple scatterplot would show points rising from the lower left to the upper right, confirming the strong positive correlation.

Diagram: Scatterplot of Hours Studied vs. Exam Scores



4. Percentage Tables

Definition

Percentage tables summarize categorical data, showing the proportion of each category relative to the total.

Construction

1. Count the frequency of each category.
2. Calculate the percentage for each category.

Example

Suppose we have survey results on favorite fruits:

Fruit	Frequency	Percentage
Apples	40	40%
Bananas	30	30%
Cherries	30	30%

Calculating Percentages:

1. Total Frequency = $40 + 30 + 30 = 100$
2. Percentage Calculation:
   - Apples: $\frac{40}{100} \times 100 = 40\%$
   - Bananas: $\frac{30}{100} \times 100 = 30\%$
   - Cherries: $\frac{30}{100} \times 100 = 30\%$

5. Analyzing Contingency Tables

Definition

Contingency tables, also known as cross-tabulations or crosstabs, are powerful tools for analyzing the relationship between two or more categorical variables. They summarize the frequency distribution of the variables and allow for the examination of potential associations and interactions.

1. Structure of a Contingency Table

a. Definition

A contingency table displays the frequency counts of the occurrences of combinations of the values of two categorical variables. Each cell in the table represents the count of observations for a specific combination of categories.

b. Example

Consider a study examining the relationship between smoking status (smoker vs. non-smoker) and lung disease (presence vs. absence):

	Lung Disease Present	Lung Disease Absent	Total
Smoker	30	70	100
Non-Smoker	10	90	100
Total	40	160	200

2. Interpreting the Table

a. Marginal Totals

The totals at the end of each row and column (the margins) give insight into the distribution of each variable independently.

- **Row totals** indicate the number of smokers and non-smokers.
- **Column totals** indicate the number of individuals with and without lung disease.

b. Joint Frequency

The count in each cell represents the joint frequency of the combination of categories. For example, there are 30 smokers with lung disease.

c. Conditional Frequency

Conditional frequencies show the proportion of one variable given the other. For instance, the proportion of smokers with lung disease can be calculated as:

$$P(\text{Lung Disease} \mid \text{Smoker}) = \frac{\text{Number of Smokers with Disease}}{\text{Total Number of Smokers}} = \frac{30}{100} = 0.3$$

d. Relative Frequencies

Relative frequencies express counts as proportions of the total, providing insight into the distribution of categories. For example, the relative frequency of smokers with lung disease is:

$$P(\text{Smoker and Lung Disease}) = \frac{30}{200} = 0.15$$

3. Statistical Tests for Independence

a. Chi-Squared Test

The Chi-squared test is commonly used to determine if there is a significant association between two categorical variables.

Null Hypothesis

The null hypothesis (H_0) states that there is no association between the variables (they are independent).

Steps to Conduct the Chi-Squared Test:

1. **Calculate Expected Frequencies**: The expected frequency for each cell can be calculated using:

$$E_{ij} = \frac{(\text{Row Total}_i) \times (\text{Column Total}_j)}{\text{Grand Total}}$$

2. **Calculate Chi-Squared Statistic**: The Chi-squared statistic is calculated as:

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

   Where O_{ij} is the observed frequency and E_{ij} is the expected frequency.

3. **Determine Degrees of Freedom**: Degrees of freedom (df) for a contingency table is given by:

$$df = (r - 1)(c - 1)$$

Where r is the number of rows and c is the number of columns.

4. **Compare to Critical Value**: Compare the Chi-squared statistic to the critical value from the Chi-squared distribution table at a given significance level (α=0.05). If $\chi 2$ is greater than the critical value, rejectH_0.

4. Visualizing Contingency Tables

a. Mosaic Plots

Mosaic plots visually represent the proportions of categories in a contingency table. Each rectangle represents a category, with the area proportional to the frequency.

b. Stacked Bar Charts

Stacked bar charts can show the distribution of one categorical variable across the levels of another variable, facilitating easy comparison.

c. Heatmaps

Heatmaps provide a color-coded representation of the frequencies in the contingency table, highlighting areas of higher and lower frequencies.

5. Applications of Contingency Tables

a. Market Research

Analyzing consumer preferences based on demographics (e.g., age and purchasing habits).

b. Healthcare

Studying the association between lifestyle factors (e.g., diet and exercise) and health outcomes.

c. Social Sciences

Examining the relationship between education level and employment status.

d. Environmental Studies

Investigating the relationship between exposure to pollutants and health effects.

6. Handling Several Batches

In data analysis, it's common to have multiple groups or batches of data. Each batch may represent different conditions or categories.

Example:

Suppose we are analyzing the effects of different fertilizers on plant growth. We have three batches:

- **Batch A**: Fertilizer Type 1
- **Batch B**: Fertilizer Type 2
- **Batch C**: No Fertilizer

Techniques:

- **Descriptive Statistics**: Calculate mean, median, and standard deviation for each batch.
- **Visualization**: Use box plots to compare distributions across batches.

Formula for Mean:

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

7. Scatterplot and Resistant Lines

Scatterplot

A scatterplot is a graphical representation of two quantitative variables. Each point represents an observation, with one variable on the x-axis and the other on the y-axis.

Example:

Consider the relationship between hours studied (X) and exam scores (Y).

Creating a Scatterplot:

- Plot points (X, Y) for each observation.
- Look for patterns, clusters, or trends.

Resistant Lines

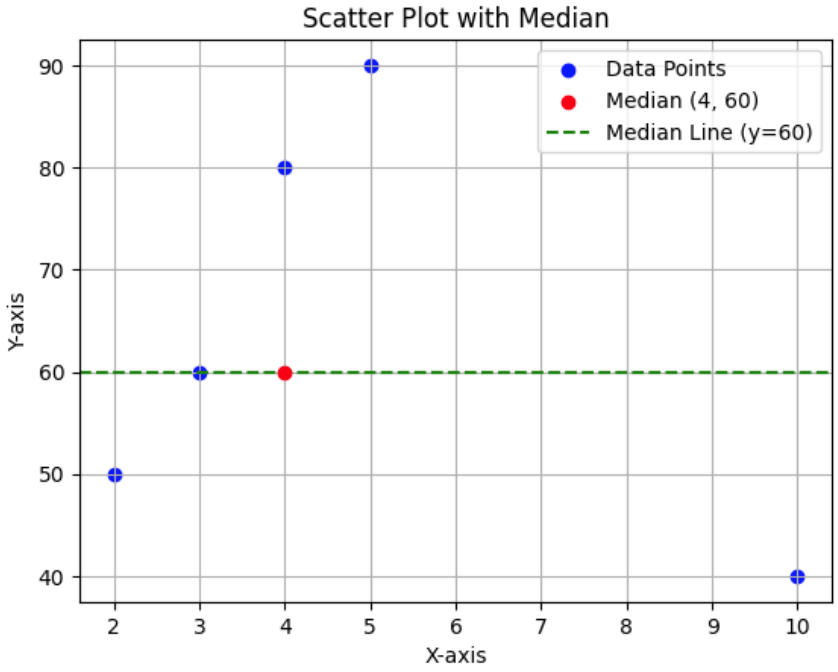
A resistant line (like the median line) minimizes the influence of outliers. This can be more informative than traditional least-squares regression.

Robust Regression Formula: The resistant line can be determined using median values instead of mean.

1. Calculate the median of X and Y.
2. Determine the slopes and intercepts based on the median points.

Example Calculation:

1. **Data**: (2, 50), (3, 60), (4, 80), (5, 90), (10, 40)
2. **Medians**: Median X = 4, Median Y = 60.
3. Draw a line through the point (4, 60).



8. Transformations

Introduction

Data transformations are mathematical operations applied to datasets to modify their distributions, stabilize variance, and improve the relationship between variables. Transformations can help meet the assumptions of statistical tests, enhance interpretability, and facilitate better modeling of relationships.

Why Transform Data?

1. **Stabilizing Variance**: Many statistical methods assume homoscedasticity (constant variance). Transformations can help achieve this.
2. **Normalizing Data**: Many statistical techniques require data to be normally distributed. Transformations can reduce skewness and make data more Gaussian.
3. **Improving Relationships**: Transformations can help linearize relationships between variables, making it easier to model them.

Common Transformations:

Log Transformation: Log transformations are particularly useful when data are positively skewed. This transformation can help reduce the impact of large values (outliers) and make the data more normally distributed.

Formula: $Y' = \log(Y)$

When to Use:

- Data that follows a multiplicative model.
- When dealing with growth rates or financial data (like income).

Example

Consider a dataset of incomes where values range widely. An example income dataset might look like this:

Income ($)
2000
3000
5000
7000
100000

The income distribution is highly skewed due to the presence of a few high-income outliers.

Transformation Steps:

1. Apply the log transformation to each income value.
2. The transformed dataset will look more like a normal distribution.

Square Root Transformation: Square root transformations are useful for count data, particularly when the data has a Poisson distribution. This transformation can help stabilize variance.

Formula: $Y' = \sqrt{Y}$

When to Use:

- Count data where the variance increases with the mean (e.g., number of events in a time period).

Example

Imagine a dataset representing the number of calls received at a call center per hour:

Calls
1
2
3
4
25

Using the square root transformation helps reduce the variance by diminishing the effect of larger counts.

Transformation Steps:

1. Apply the square root transformation to each count.
2. The transformed counts will stabilize the variance.

Example:

Consider the data for income (Y) and expenditure (X):

- Original data is highly skewed; applying a log transformation can linearize the relationship.

9. Introducing a Third Variable

Understanding the Role of a Third Variable

When examining relationships between two variables, introducing a third variable can uncover additional insights. This third variable may interact with or confound the relationship, providing a more comprehensive understanding.

Example: Age as a Third Variable

Let’s consider the relationship between hours studied (X) and exam scores (Y). However, we also want to analyze how age (Z) influences this relationship.

Steps to Analyze:

1. **Identify the Variables**:
    - **Independent Variable (X)**: Hours studied.
    - **Dependent Variable (Y)**: Exam scores.
    - **Third Variable (Z)**: Age of students.
2. **Data Collection**:
    - Collect data on students' hours studied, their exam scores, and their ages.
3. **Visualizing the Relationships**:
    - Use a 3D scatterplot to visualize the relationship between hours studied and exam scores while incorporating age.

Visualization Techniques:

- **3D Scatterplot**:
    - The x-axis represents hours studied, the y-axis represents exam scores, and the z-axis represents age.
- **Color Coding**:
    - Use different colors to represent age groups (e.g., 18-25, 26-35).

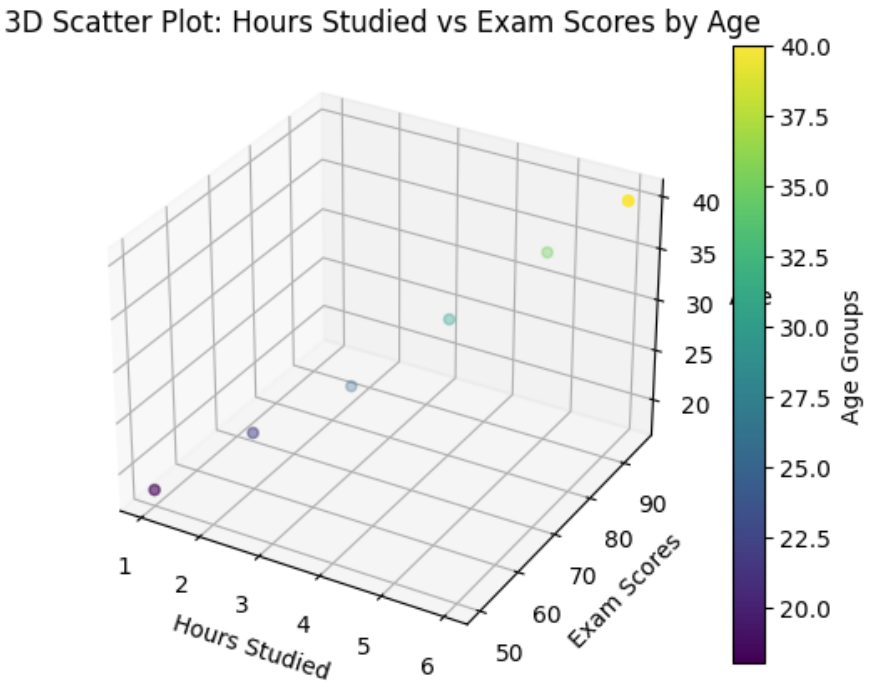
Example of a 3D Scatterplot:

Imagine a 3D scatterplot where:

- Data points are plotted based on hours studied and exam scores.
- Points for different age groups are colored distinctly.

Benefits of Introducing a Third Variable:

- **Understanding Interaction**: Age might change the effect of hours studied on exam scores. For instance, younger students may benefit more from studying than older students.
- **Controlling for Confounding**: By examining the third variable, researchers can control for confounding effects that might mislead the analysis.
- **Enhanced Interpretability**: Adding dimensions helps capture the complexity of real-world relationships, allowing for more nuanced conclusions.



10. Three Variable Contingency Tables and Beyond

Introduction

Three-variable contingency tables extend the concept of two-variable contingency tables to include an additional categorical variable, allowing for more complex analyses of relationships among variables. This enables researchers to examine interactions and associations in multi-dimensional categorical data, providing deeper insights into the relationships among variables.

1. Structure of Three Variable Contingency Tables

a. Definition

A three-variable contingency table organizes data to display the frequency counts of combinations of three categorical variables. Each cell in the table represents a count for a specific combination of categories from all three variables.

b. Example

Consider a study examining the relationship between:

- **Variable A**: Smoking status (Smoker vs. Non-Smoker)
- **Variable B**: Lung disease status (Disease Present vs. Disease Absent)
- **Variable C**: Age group (Young, Middle-aged, Old)

A possible three-variable contingency table might look like this:

Smoking Status	Age Group	Disease Present	Disease Absent	Total
Smoker	Young	10	20	30
Smoker	Middle-aged	15	15	30
Smoker	Old	20	10	30
Non-Smoker	Young	5	25	30
Non-Smoker	Middle-aged	10	20	30
Non-Smoker	Old	10	20	30
Total		70	105	175

Log-Linear Models

Log-linear models are often used for three or more categorical variables. These models can examine the relationships and interactions among variables while accounting for the structure of the contingency table.

Model Structure

A log-linear model can be represented as:

$$\log(E_{ijk}) = \mu + \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{ik} + \lambda_{jk} + \lambda_{ijk}$$

Where E_{ijk} is the expected frequency, μ is the overall mean, and λ terms represent the effects and interactions.