Unit 1 Notes

2025-04-14

Table of contents

1 Unit 1: Exploring One-Variable Data

1.1 Variation in Categorical and Quantitative Variables

1.1.1 Definition

Definition 1.1. A **categorical variable** is a variable that can take on one of a limited, and usually fixed, number of possible values.

1.1.2 Example Problem

Example 1.1. Identify whether the following variable is categorical or quantitative: "Type of pet owned."

1.2 Representing Data Using Tables or Graphs

1.2.1 Properties

Properties 1.1. - Tables can be used to summarize categorical data. - Graphs such as bar charts and pie charts can visually represent categorical data.

1.2.2 Figure

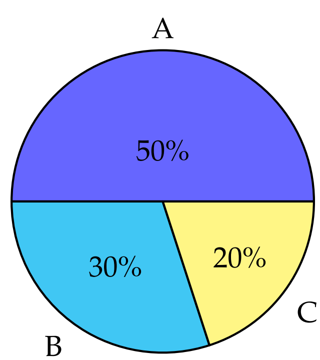

Figure 1: Example of a pie chart.

1.3 Calculating and Interpreting Statistics

1.3.1 Theorem

Theorem 1.1. *The mean of a set of data is calculated as the sum of all data points divided by the number of data points.*

1.4 Describing and Comparing Distributions of Data

1.4.1 Example Problem

Example 1.2. Compare the distributions of two datasets using histograms.

1.5 The Normal Distribution

1.5.1 Definition

Definition 1.2. The **normal distribution** is a continuous probability distribution characterized by a bell-shaped curve.

1.5.2 Figure

Figure 2: Normal distribution curve.

2 Unit 2: Exploring Two-Variable Data

2.1 Comparing Representations of 2 Categorical Variables

2.1.1 Example Problem

Example 2.1. Create a contingency table to compare two categorical variables.

2.2 Calculating Statistics for 2 Categorical Variables

2.2.1 Properties

Properties 2.1. - Chi-square tests can be used to determine if there is a significant association between two categorical variables.

2.3 Representing Bivariate Quantitative Data Using Scatter Plots

2.3.1 Figure

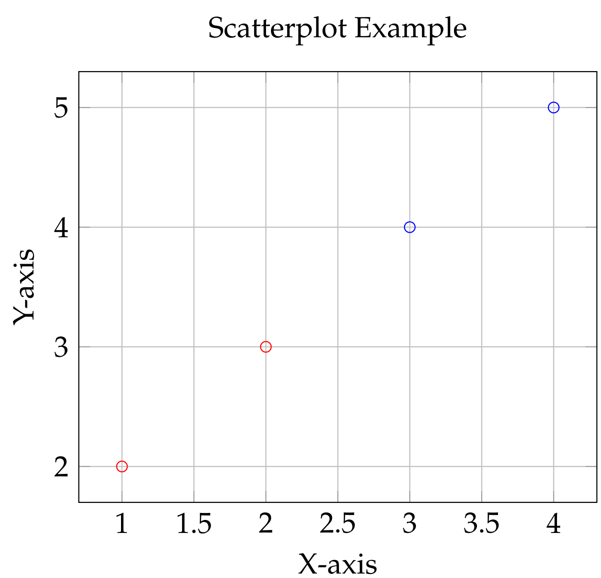

Figure 3: Example of a scatter plot.

2.4 Describing Associations in Bivariate Data and Interpreting Correlation

2.4.1 Theorem

Theorem 2.1. *The correlation coefficient r measures the strength and direction of a linear relationship between two variables.*

2.5 Linear Regression Models

2.5.1 Example Problem

Example 2.2. Fit a linear regression model to a dataset and interpret the slope and intercept.

2.6 Residuals and Residual Plots

2.6.1 Figure

Figure 4: Residual plot.

2.7 Departures from Linearity

2.7.1 Properties

Properties 2.2. - Departures from linearity can be identified using residual plots.

3 Unit 3: Collecting Data

3.1 Planning a Study

3.1.1 Definition

Definition 3.1. A **study** is a planned investigation designed to answer specific research questions.

3.2 Sampling Methods

3.2.1 Example Problem

Example 3.1. Describe the differences between simple random sampling and stratified sampling.

3.3 Sources of Bias in Sampling Methods

3.3.1 Properties

Properties 3.1. - Bias can occur due to non-random sampling methods.

3.4 Designing an Experiment

3.4.1 Theorem

Theorem 3.1. *Random assignment of subjects to treatment groups helps to control for confounding variables.*

3.5 Interpreting the Results of an Experiment

3.5.1 Example Problem

Example 3.2. Interpret the results of a randomized controlled trial.

4 Unit 4: Probability, Random Variables, and Probability Distributions

4.1 Using Simulation to Estimate Probabilities

4.1.1 Example Problem

Example 4.1. Use a simulation to estimate the probability of rolling a sum of 7 with two dice.

4.2 Calculating the Probability of a Random Event

4.2.1 Theorem

Theorem 4.1. *The probability of an event A is given by $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$.*

4.3 Random Variables and Probability Distributions

4.3.1 Definition

Definition 4.1. A **random variable** is a variable whose possible values are outcomes of a random phenomenon.

4.4 The Binomial Distribution

4.4.1 Properties

Properties 4.1. - The binomial distribution describes the number of successes in a fixed number of independent Bernoulli trials.

4.5 The Geometric Distribution

4.5.1 Example Problem

Example 4.2. Calculate the probability of the first success occurring on the third trial in a geometric distribution.

5 Unit 5: Sampling Distributions

5.1 Variation in Statistics for Samples Collected from the Same Population

5.1.1 Theorem

Theorem 5.1. *The sampling distribution of a statistic is the distribution of the statistic over many random samples from the same population.*

5.2 The Central Limit Theorem

5.2.1 Definition

Definition 5.1. The **Central Limit Theorem** states that the sampling distribution of the sample mean will be approximately normal if the sample size is sufficiently large.

5.3 Biased and Unbiased Point Estimates

5.3.1 Properties

Properties 5.1. - An unbiased estimator is one whose expected value is equal to the population parameter being estimated.

5.4 Sampling Distributions for Sample Proportions

5.4.1 Example Problem

Example 5.1. Calculate the standard error of the sample proportion.

5.5 Sampling Distributions for Sample Means

5.5.1 Figure

Figure 5: Sampling distribution of the sample mean.

6 Unit 6: Inference for Categorical Data: Proportions

6.1 Constructing and Interpreting a Confidence Interval for a Population Proportion

6.1.1 Example Problem

Example 6.1. Construct a 95

6.2 Setting Up and Carrying Out a Test for a Population Proportion

6.2.1 Theorem

Theorem 6.1. *The test statistic for a population proportion is given by* $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

6.3 Interpreting a p-Value and Justifying a Claim About a Population Proportion

6.3.1 Properties

Properties 6.1. - A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis.

6.4 Type I and Type II Errors in Significance Testing

6.4.1 Definition

Definition 6.1. A **Type I error** occurs when the null hypothesis is rejected when it is actually true.

6.5 Confidence Intervals and Tests for the Difference of 2 Proportions

6.5.1 Example Problem

Example 6.2. Construct a confidence interval for the difference between two population proportions.

7 Unit 7: Inference for Quantitative Data: Means

7.1 Constructing and Interpreting a Confidence Interval for a Population Mean

7.1.1 Example Problem

Example 7.1. Construct a 95

7.2 Setting Up and Carrying Out a Test for a Population Mean

7.2.1 Theorem

Theorem 7.1. *The test statistic for a population mean is given by* $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

7.3 Interpreting a p-Value and Justifying a Claim About a Population Mean

7.3.1 Properties

Properties 7.1. - A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis.

7.4 Confidence Intervals and Tests for the Difference of 2 Population Means

7.4.1 Example Problem

Example 7.2. Construct a confidence interval for the difference between two population means.

8 Unit 8: Inference for Categorical Data: Chi-Square

8.1 The Chi-Square Test for Goodness of Fit

8.1.1 Example Problem

Example 8.1. Perform a chi-square test for goodness-of-fit to determine if a categorical variable follows a specified distribution.

8.2 The Chi-Square Test for Homogeneity

8.2.1 Theorem

Theorem 8.1. *The chi-square test for homogeneity tests whether the proportions of a categorical variable are the same across different levels of another categorical variable.*

8.3 The Chi-Square Test for Independence

8.3.1 Properties

Properties 8.1. - The chi-square test for independence tests whether two categorical variables are independent.

8.4 Selecting an Appropriate Inference Procedure for Categorical Data

8.4.1 Example Problem

Example 8.2. Determine the appropriate inference procedure for a given set of categorical data.

9 Unit 9: Inference for Quantitative Data: Slopes

9.1 Confidence Intervals for the Slope of a Regression Model

9.1.1 Example Problem

Example 9.1. Construct a confidence interval for the slope of a regression model.

9.2 Setting Up and Carrying Out a Test for the Slope of a Regression Model

9.2.1 Theorem

Theorem 9.1. *The test statistic for the slope of a regression model is given by* $t = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}}$.

9.3 Selecting an Appropriate Inference Procedure

9.3.1 Properties

Properties 9.1. - The appropriate inference procedure depends on the nature of the data and the research question.