PCED

Module 3 -Statistical Analysis

Topics

1. Descriptive Statistics
2. Inferential Statistics
3. Key Concepts to Review
4. Preparation Tips

Descriptive Statistics

Measures of Central Tendency

Mean: The average of a dataset.
Median: The middle value when the data is ordered.
Mode: The most frequently occurring value(s).

Measures of Variability

Range: Difference between the maximum and minimum values.
Variance: The average of squared differences from the mean.
Standard Deviation (SD): The square root of the variance, indicating spread around the mean.

Inferential Statistics

Simple Regression Analysis

Understand the relationship between two variables:
- Independent variable (predictor).
- Dependent variable (response).

Linear regression equation: $y=mx+c$, where:
- y: Dependent variable.
- x: Independent variable.
- m: Slope (rate of change).
- c: Intercept (value of y when x=0).

Basics of correlation:

Correlation coefficient (r): Measures the strength and direction of the linear relationship between two variables (range: -1 to +1).

Key Concepts to Review

Population vs. Sample:

Population: Entire group being studied.
Sample: Subset of the population used for analysis.

Hypothesis Testing:

Null hypothesis (H0): Assumes no effect or relationship.
Alternative hypothesis (Ha): Assumes an effect or relationship exists.
P-value: Probability of observing the data given that H0 is true (commonly compared to a significance level, e.g., 0.05).

Confidence Intervals:

Range within which the true population parameter is likely to fall.

Preparation Tips

Practice calculating mean, median, mode, variance, and standard deviation from sample datasets.

Use example problems to apply simple regression concepts, such as predicting y given x.

Review scatterplots to visually interpret relationships between variables.

Understand when and how to use t-tests or z-tests (even if they’re not explicitly required, familiarity helps).

Examples

1. Find the mean, median, mode, and standard deviation for the dataset:
   4,8,6,5,3,7,9,124,8,6,5,3,7,9,12

2. For the dataset above, calculate the range and interpret what it tells you about variability.

More on : - https://www.khanacademy.org/math/statistics-probability

Descriptive Statistics Problem

Dataset:

$$4, 8, 6, 5, 3, 7, 9, 12$$

1. Mean

The formula for the mean is:

$$\text{Mean} = \frac{\text{Sum of all data points}}{\text{Number of data points}}$$

$$\text{Mean} = \frac{4 + 8 + 6 + 5 + 3 + 7 + 9 + 12}{8} = \frac{54}{8} = 6.75$$

2. Median

The median is the middle value of an ordered dataset.

- Arrange the data:

$$3, 4, 5, 6, 7, 8, 9, 12$$

- Since the dataset has an even number of values ($n = 8$), take the average of the two middle numbers:

$$\text{Median} = \frac{6 + 7}{2} = 6.5$$

2

3. Mode

The mode is the most frequently occurring value.

- Each number appears once, so there is **no mode**.

4. Range

The range is the difference between the maximum and minimum values:

$$\text{Range} = \text{Maximum} - \text{Minimum} = 12 - 3 = 9$$

5. Variance and Standard Deviation

Variance formula:

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

Standard Deviation:

$$\text{SD} = \sqrt{\text{Variance}}$$

- Calculate deviations:

$$x_i - \bar{x} = \{-2.75, 1.25, -0.75, -1.75, -3.75, 0.25, 2.25, 5.25\}$$

- Square deviations:

$$\{7.5625, 1.5625, 0.5625, 3.0625, 14.0625, 0.0625, 5.0625, 27.5625\}$$

- Sum of squared deviations:

$$59.1875$$

- Variance:

$$\text{Variance} = \frac{59.1875}{8} = 7.3984$$

- Standard Deviation:

$$\text{SD} = \sqrt{7.3984} \approx 2.72$$

Inferential Statistics

Simple Regression Practice Problem

Consider the following dataset:

x (Independent Variable)	y (Dependent Variable)
1	3
2	5
3	7
4	9

1. Find the **regression equation** $y = mx + c$:
   - Calculate the **slope** (m):

$$m = \frac{\text{Covariance}(x, y)}{\text{Variance}(x)}$$

   - Calculate the **intercept** (c).

Dataset:

x	y
1	3
2	5
3	7
4	9

Step 1: Find the slope (m)

$$m = \frac{n\sum(xy) - \sum(x)\sum(y)}{n\sum(x^2) - (\sum(x))^2}$$

- $\sum(x) = 1 + 2 + 3 + 4 = 10$
- $\sum(y) = 3 + 5 + 7 + 9 = 24$
- $\sum(xy) = (1)(3) + (2)(5) + (3)(7) + (4)(9) = 70$
- $\sum(x^2) = 1^2 + 2^2 + 3^2 + 4^2 = 30$
- $n = 4$

- $n = 4$

$$m = \frac{4(70) - (10)(24)}{4(30) - (10)^2} = \frac{280 - 240}{120 - 100} = \frac{40}{20} = 2$$

Step 2: Find the intercept (c)

$$c = \frac{\sum(y) - m\sum(x)}{n}$$

$$c = \frac{24 - 2(10)}{4} = \frac{24 - 20}{4} = 1$$

Step 3: Regression Equation

$$y = 2x + 1$$

Step 4: Predict y when $x = 5$

$$y = 2(5) + 1 = 10 + 1 = 11$$

Regression

Regression: It predicts the continuous output variables based on the independent input variable.

Like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc

Inferential Statistics

statistics is a nuanced field that encompasses collecting, analyzing, interpreting, and presenting numerical data. It's invaluable for drawing broad conclusions from large populations where detailed measurements aren't feasible.

Inferential statistics involves drawing conclusions or making inferences about a population based on data collected from a sample of that population. Here's how it works:

Sampling: You start by collecting data from a subset of the population you're interested in studying. This subset is called a sample.

Analysis: After collecting data, you use various statistical techniques. This might include calculating measures like means, standard deviations, correlations, or regression coefficients.

Inferential Statistics

Inference: Once you've analyzed the sample data, you make inferences or generalizations about the population from which the sample was drawn. These inferences are based on the assumption that the sample is representative of the population.

Inferential statistics includes hypothesis testing, confidence intervals, and regression analysis, among other techniques. These methods help researchers determine whether their findings are statistically significant and whether they can generalize their results to the larger population.

Types of Inferential Statistics

1. Hypothesis Testing

Hypothesis testing is a fundamental technique in inferential statistics. It involves testing a hypothesis about a population parameter, such as a mean or proportion, using sample data. The process typically involves setting up null and alternative hypotheses and conducting a statistical test to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

Example: A researcher might hypothesize that the average income of people in a certain city is greater than $50,000 per year. They would collect a sample of incomes, conduct a hypothesis test, and determine whether the data provide enough evidence to support or reject this hypothesis.

Types of Inferential Statistics

T-Test

The T-test is used when the population standard deviation is unknown or the sample size is small (typically $n < 30$). It's based on the Student's t-distribution, which has thicker tails than the standard normal distribution.

There are two main types of t-tests: the independent samples t-test (for comparing means of two independent groups) and the paired samples t-test (for comparing means of two related groups).

The formula for the t-test statistic is similar to the Z-test, but it uses the sample standard deviation instead of the population standard deviation.

Example: A researcher wants to determine if there is a significant difference in exam scores between two groups of students. They collect exam scores from each group and use the t-test to compare the means.

Types of Inferential Statistics

Z-Test

The Z-test is a statistical test to determine whether the means of two populations differ when the population variance is known, and the sample size is large (typically n > 30). It's based on the standard normal distribution (Z-distribution).

The Z-test statistic follows a standard normal distribution under the null hypothesis.

Example: A researcher wants to determine if the mean height of a population is significantly different from 65 inches. They collect a large sample of heights with a known population standard deviation and use the Z-test to compare the sample mean to the population mean.

Types of Inferential Statistics

F-Test

The F-test is used to compare the variances of two populations or more than two populations. It's commonly used in the analysis of variance (ANOVA) to test for differences among means of multiple groups.

The F-test statistic follows an F-distribution, which is positively skewed and takes on only non-negative values.

In ANOVA, the F-test compares the variance between groups to the variance within groups. If the ratio of these variances is sufficiently large, it suggests that the groups' means are different.

Example: A researcher wants to determine if there are differences in the effectiveness of three teaching methods on student performance. They collect performance data from students taught using each method and use ANOVA, which utilizes the F-test, to compare the variances between and within the groups.

Confidence Intervals

Confidence intervals provide a range of values within which a population parameter is likely to lie and a level of confidence associated with that range. They are often used to estimate the true value of a population parameter based on sample data. The width of the confidence interval depends on the sample size and the desired level of confidence.

Example: A pollster might use a confidence interval to estimate the proportion of voters who support a particular candidate. The confidence interval would give a range of values within which the true proportion of supporters is likely to lie, along with a confidence level such as 95%.

Regression Analysis

Regression analysis examines the relationship between one or more independent variables and a dependent variable. It can be used to predict the dependent variable's value based on the independent variables' values. Regression analysis also allows for testing hypotheses about the strength and direction of the relationships between variables.

Example: A researcher might use regression analysis to examine the relationship between hours of study and exam scores. They could then use the regression model to predict exam scores based on the hours studied.

Analysis of Variance (ANOVA)

ANOVA is a statistical technique that compares means across two or more groups. It tests whether there are statistically significant differences between the groups' means. ANOVA calculates both within-group variance (variation within each group) and between-group variance (variation between the group means) to determine whether any observed differences are likely due to chance or represent true differences between groups.

Example: A researcher might use ANOVA to compare the effectiveness of three different teaching methods on student performance. They would collect data on student performance in each group and use ANOVA to determine whether there are significant differences in performance between the groups.

Chi-Square Tests

Chi-square tests are used to determine whether there is a significant association between two categorical variables. They compare the observed frequency distribution of the data to the expected frequency distribution under the null hypothesis of independence between the variables.

Example: A researcher might use a chi-square test to examine whether there is a significant relationship between gender and voting preference. They would collect data on the gender and voting preferences of a sample of voters and use a chi-square test to determine whether gender and voting preference are independent.

Difference

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarizes and describes characteristics of a data set.	Makes inferences or predictions about populations based on sample data.
Focus	It focuses on describing the data (e.g., mean, median, mode).	It focuses on generalizing populations using sample data.
Population vs. Sample	Analyzes data from the entire population.	Analyzes data from a sample of the population.
Examples	Mean, median, mode, standard deviation, histograms.	Hypothesis testing, confidence intervals, regression analysis.
Application	Used to understand and visualize data.	Used to test hypotheses, make predictions, and draw conclusions.

Difference

Sample Size Requirement	Can analyze any size of the data set.	Often, it requires a sufficiently large sample size for accuracy.
Generalizability	Descriptive statistics do not make predictions beyond the data set.	Inferential statistics allow for predictions about the population.
Goal	To summarize and present data in a meaningful way.	To conclude or make predictions about populations.

Problem

A company claims that the average weight of its product is 50g. A sample of 30 products has a mean weight of 48g and a standard deviation of 5g. Test the company’s claim at a 95% confidence level.

Step 1: State the Hypotheses

1. **Null Hypothesis (H_0)**: The true mean weight is $\mu = 50$.
2. **Alternative Hypothesis (H_a)**: The true mean weight is $\mu \neq 50$ (two-tailed test).

Step 2: Identify the Test Statistic

Since the sample size is $n = 30$, which is reasonably large ($n \geq 30$), we can use a **z-test**. The test statistic formula is:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{x}$: Sample mean = 48g
- μ: Population mean = 50g
- s: Sample standard deviation = 5g
- n: Sample size = 30

Step 3: Calculate the Test Statistic

1. **Compute the standard error (SE):**

$$SE = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{30}} \approx \frac{5}{5.477} \approx 0.912$$

2. **Calculate the z-value:**

$$z = \frac{\bar{x} - \mu}{SE} = \frac{48 - 50}{0.912} = \frac{-2}{0.912} \approx -2.19$$

Step 4: Find the Critical Value

For a **95% confidence level** and a two-tailed test, the critical z-values are:

$$z = \pm 1.96$$

Step 5: Compare and Draw Conclusions

1. The calculated z-value is -2.19.
2. The critical values are -1.96 and $+1.96$.
3. Since $-2.19 < -1.96$, the z-value falls in the rejection region.

Results

Step 6: State the Result

Reject the null hypothesis (H_0). There is enough evidence to suggest that the true mean weight is significantly different from 50g.

Interpretation

The company's claim that the average weight is 50g is likely incorrect based on the sample data.