Advanced Data Visualization Lecture Notes (Lecture 6 - 2 Hours)

1. Introduction to Hypothesis Testing

Statistical hypothesis testing is a fundamental method used to determine whether there is enough evidence in a data sample to infer that a certain condition is true for the entire population. It is widely used in statistics for decision-making, especially when analyzing data visualizations.

- **Key Terms**:
  - **Null Hypothesis (H_0)**: A statement that there is no effect or no difference; it is the default or starting assumption.
  - **Alternative Hypothesis (H_1)**: A statement that contradicts the null hypothesis, suggesting an effect or difference.
  - **Significance Level (α)**: The probability threshold below which the null hypothesis is rejected (commonly set at 0.05).
  - **P-Value**: The probability that the observed data would occur under the null hypothesis.

2. Procedure for Hypothesis Testing

Step 1: Formulate Hypotheses

- **Null Hypothesis (H_0)**: This is the hypothesis you aim to test, generally stating no effect or no relationship between variables.
- **Alternative Hypothesis (H_1)**: The hypothesis stating there is an effect or a relationship.

Example:

In a study examining the effect of a new drug, the hypotheses might be:

- H_0: The drug has no effect on patients' recovery.
- H_1: The drug has a positive effect on patients' recovery.

Step 2: Choose a Significance Level (α)

- The significance level represents the risk of rejecting the null hypothesis when it is actually true (Type I error). A common choice is **$\alpha = 0.05$**.

Step 3: Select the Appropriate Statistical Test

- **T-Test**: Used to compare means between two groups (independent or paired).
- **Chi-Square Test**: Used for categorical data to test relationships between variables.

- **ANOVA**: Used to compare means across three or more groups.
- **Regression Analysis**: Used to understand the relationship between independent and dependent variables.
- **Correlation Test**: Used to test the strength and direction of the relationship between two continuous variables.

Step 4: Collect and Visualize Data

- Before performing a statistical test, data is often visualized to observe trends, outliers, and relationships. Visualizations help to make initial observations:
  - **Bar Charts**: To compare categorical data.
  - **Boxplots**: To visualize data spread and identify outliers.
  - **Histograms**: To analyze data distribution.
  - **Scatter Plots**: To see relationships between variables.

Step 5: Perform the Hypothesis Test

1. **Compute the Test Statistic**: Use the appropriate formula for the statistical test (e.g., t-statistic for a t-test, chi-square statistic, etc.).
2. **Calculate the P-Value**: This indicates the probability of observing the data given that the null hypothesis is true.
3. **Compare the P-Value to α**:
   - If **P-value ≤ α**, reject the null hypothesis.
   - If **P-value > α**, fail to reject the null hypothesis.

Step 6: Interpret Results

- **Reject H_0**: If the p-value is less than the significance level, you conclude that there is enough evidence to support the alternative hypothesis.
- **Fail to Reject H_0**: If the p-value is greater than the significance level, there is insufficient evidence to support the alternative hypothesis.

3. Visualizations and Observations

- **Bar Chart Observations**: Shows differences in group frequencies or proportions. For example, if visualizing drug effects across different groups, it may show how many patients recover in each group.
- **Boxplot Observations**: Can reveal whether there are significant differences in data spread (i.e., variance) between groups.
- **Scatter Plot Observations**: Can suggest positive, negative, or no correlation between two variables (e.g., dosage and recovery time).

Example:

- If using a scatter plot to observe the relationship between drug dosage and recovery time, you might:

    - **Null Hypothesis (H_0)**: There is no relationship between dosage and recovery time.
    - **Alternative Hypothesis (H_1)**: There is a relationship between dosage and recovery time.

If the plot shows a clear downward trend and a subsequent regression analysis yields a p-value less than 0.05, you would reject the null hypothesis and conclude that dosage does indeed affect recovery time.

4. Reporting the Results

- Present the visualization (e.g., a bar chart or scatter plot) along with the hypothesis test result.
- Explain whether you rejected or failed to reject the null hypothesis and what that implies about the data.

Example:

"The scatter plot shows a negative correlation between drug dosage and recovery time. The regression analysis yielded a p-value of 0.03, suggesting that we reject the null hypothesis ($\alpha = 0.05$). This indicates that there is a statistically significant relationship between dosage and recovery time."

5. Common Pitfalls

- **Misinterpreting P-Value**: A small p-value indicates strong evidence against H_0, not proof that H_1 is true.
- **Type I and Type II Errors**:
    - **Type I error** (false positive): Rejecting H_0 when it's true.
    - **Type II error** (false negative): Failing to reject H_0 when H_1 is true.

6. Conclusion

Hypothesis testing, when combined with visualizations, helps make informed decisions based on data. By following a structured process, we can test assumptions and observe trends effectively.

Key Takeaways

- Understand the hypothesis testing framework (H_0 vs. H_1).
- Use appropriate visualizations to observe patterns in data.
- Conduct statistical tests to confirm or reject hypotheses based on data.
- Interpret p-values in the context of the significance level and report conclusions accurately.