Welcome to

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 10: One- and Two-Sample Hypothesis Tests

instats

Framework for Hypothesis Tests

- Hypothesis testing: make inferences about population based on a sample
- Null hypothesis (H_0): assumed true, no effect or difference
- Alternative hypothesis (H_A): claim against H_0 , need evidence to reject H_0

Comparing Hypothesis Tests to Criminal Trials

- **Null Hypothesis**: Similar to presumption of innocence ("innocent until proven guilty")
- **Alternative Hypothesis**: Similar to the prosecutor's claim ("guilty")
- Decision:
  - **Reject H_0**: Sufficient evidence to reject innocence (guilty verdict)
  - **Fail to reject H_0** : Not enough evidence to reject innocence (not guilty, but not necessarily innocent)
- **Type I Error**: Wrongly convicting an innocent person (rejecting H_0 when it is true)
- **Type II Error**: Letting a guilty person go free (failing to reject H_0 when H_A is true)
- Significance level (α) = "beyond a reasonable doubt"

instats

One-Sample Hypothesis Test

- Example: Test claim about population mean, μ
- Null hypothesis: $\mu = 3.6$ grams (average almond weight)
- Alternative hypothesis: $\mu < 3.6$ grams
- Goal: determine likelihood of observing a sample mean as extreme as observed

Types of Hypothesis Tests

- **Two-sided test**: evidence against H_0 comes from both directions (greater or less)
- **One-sided test**: evidence against H_0 comes from one direction only
- Examples:
    - left-sided test (H_A: $\mu < 3.6$)
    - right-sided test (H_A: $\mu > 3.6$)
    - two-sided test (H_A: $\mu \neq 3.6$)

instats

Steps in Hypothesis Testing

1. Define H_0 and H_A (null and alternative hypotheses)
2. Choose significance level α (e.g., 0.05)
3. Calculate test statistic (e.g., *t* statistic for one-sample tests)
4. Compare *p*-value calculated from test statistic to α
5. Make decision and interpret results in the context of the problem

Theory-Based Hypothesis Test Example

- Calculate sample mean and standard deviation
- Test statistic formula: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
- Calculate *p*-value for significance

Simulation-Based Hypothesis Testing

- Use permutation tests to simulate null hypothesis scenario
- Shuffle data to simulate "no effect" world
- Calculate test statistic for each shuffle to form null distribution

P-Value and Statistical Significance/Discernibility

- ***P*-value**: probability of observing a test statistic as extreme as the observed one, assuming H_0 is true
- If *p*-value < α, reject H_0 (evidence against H_0 is strong)
- Example: *p*-value = 0.03, α = 0.05 → reject H_0

Connection Between Hypothesis Testing and Confidence Intervals

- If the null value is outside the confidence interval, reject H_0
- Example: 95% CI does not contain 3.6 → reject H_0

Demo & Exercises

Q & A

STOP

Welcome to

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 11: Inference for Regression

instats

Introduction to Statistical Inference for Regression

- Regression helps us explore relationships between variables
- Inference allows us to make conclusions about these relationships
- Types of regression:
  - **Simple Linear Regression**: One predictor/regressor, one response
  - **Multiple Linear Regression**: More than one predictor/regressor for the response

Simple Linear Regression (SLR)

- Focuses on the relationship between one explanatory variable and one response variable
- Goal: Estimate how much the response variable changes with one unit increase in the predictor

Inference for SLR Coefficients

- Confidence intervals: Estimate the range of the true slope for one predictor
- Hypothesis testing: Check if the predictor has a significant effect on the response
  - Null hypothesis: No effect (slope = 0)
  - Alternative hypothesis: There is an effect (slope ≠ 0)

Multiple Linear Regression

- Multiple predictors are used to explain the response variable
- Each predictor explains a unique part of the variability in the response
- Coefficients are adjusted for the presence of other predictors
- Identifies the contribution of each predictor while controlling for others
- Inference involves understanding the significance and confidence intervals for each predictor's coefficient

instats

Inference for MLR Coefficients

- Each predictor has its own **coefficient**, representing its unique impact
- Confidence intervals: Estimate the range of the true coefficients for each predictor
- Hypothesis testing: Check if each predictor has a significant effect on the response, controlling for other predictors

Key Differences in Inference

- In SLR: We are assessing the relationship between one predictor and the response
- In MLR: We assess each predictor's effect **while controlling for the others**

Model Fit and Assumptions

- For both SLR and MLR:
  - The relationships should be linear
  - Residuals should appear random and normally distributed
  - LINE acrostic
- In MLR: Check for **multicollinearity** (predictors should not be highly correlated)

Hypothesis Tests for Partial Slopes

- Formulating hypothesis tests
    - Null and alternative hypotheses
    - Conducting t-tests for regression coefficients
    - p-values and statistical significance
    - Practical implications and decision-making

Interpreting Results in MLR

- Each coefficient in MLR shows the effect of one predictor **while holding the others constant**
- MLR can handle interaction terms and transformations
- Bootstrap methods can estimate confidence intervals when assumptions are violated
- Permutation tests can check the significance of predictors in MLR

Demo & Exercises

Q & A

STOP

Welcome to

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 12: Storytelling with Data

instats

Wrap-Up

- Journey through ModernDive (Second Edition)
- Explored key steps in the data science pipeline
- Applied tools and techniques for data wrangling, visualization, and modeling using computation and visualization as the drivers

instats

Review: Data Science with `tidyverse`

- Learned to visualize data using `ggplot2`
- Mastered data wrangling with `dplyr`
- Understood the concept of "tidy" data with `tidyr`
- Imported and cleaned real-world data

Review: Statistical Modeling with `moderndive`

- Fitted simple linear regression models
- Expanded to multiple regression models with more than one predictor
- Interpreted coefficients and effects of predictors on the response

Review: Statistical Inference with `infer`

- Understood the role of sampling variability and sample size
- Constructed confidence intervals using bootstrapping and theory-based approaches
- Conducted hypothesis testing using theory-based and permutation methods

Final Thoughts: Thinking with Data

- Data-driven mindset
    - Asking the right questions
    - Approaching problems methodically
- Analytical thinking
    - Breaking down complex issues
    - Identifying patterns and trends
- Critical thinking
    - Evaluating sources
    - Checking validity of conclusions

instats

Looking Forward: Telling Your Story with Data

- Use your new skills to tell data-driven stories
- Stay curious and continue to explore different data science workflows
- Keep refining your techniques and learning new tools

Short Walkthrough + Final Q & A

STOP