Utilizing R to visualize and analyze data

Marieke K Jones, PhD (Marieke@virginia.edu)
Research Data Specialist
Claude Moore Health Sciences Library
University of Virginia

Set-up

- Option A: go to https://github.com/mariekekjones/BIMS-bootcamp
  - Fork repo to your account
  - Clone repo
    - In Rstudio
      - New project → Version control → git → copy repo

- Option B: go to https://data.hsl.virginia.edu/workshop-materials
  - Download materials
  - Create local Bootcamp directory somewhere
  - In RStudio
    - New project → Existing Directory → [browse to Bootcamp]

Set-up

- Open R-Viz-Skeleton.R

- Ensure you have the tidyverse package installed and loaded:

```
install.packages("tidyverse")

library(tidyverse)
```

Agenda

- Set up
- dplyr review
- ggplot2

- Descriptive statistics
- T-tests
- ANOVA
- Linear Models
- Discrete variable stats
  - Chi square
  - Logistic regression

Assumptions of T-tests

- Random Sampling

- Independent Samples (violated in paired t-test)
  - Need to assess (think)

- Normality
  - Need to assess (plot or test)

- Equal variance
  - Need to assess (think, plot, test)

T-Tests

- Test the difference in 2 group means

- Independent Samples, unequal variance

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{(SE_1^2 + SE_2^2)}}$$

- Independent Samples, Equal variance

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- Paired Samples

$$\frac{\overline{d}}{\frac{s_d}{\sqrt{n_d}}}$$

T-tests, ANOVA, and linear models

- **T-test** = difference in 2 groups
- **ANOVA** = difference in 3+ groups
- **Linear Model** = effect of predictor variable on response

- T-tests are specific case of ANOVA and ANOVA is specific case of Linear Model

- **ANOVA & T**: Does mean response differ between levels of categorical predictor?
- **Linear Model**: Does response differ based on (a categorical) predictor?

Linear Regression Models

- Single predictor X

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Multiple predictors X_1 and X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Y = response variable
- X = predictor
- β_0 = y-intercept
- β_i = slope. What effect does one unit change in X do to Y?
- ϵ = residual error. Given X, slope, and y-intercept, model cannot perfectly predict Y. These are <u>assumed to be normally distributed</u> with a mean of 0 and a standard deviation σ

Assumptions of Linear Regression

- Random sampling
- Residuals are normally distributed
- Residuals show constant variance across levels of X

If you are starting to love this stuff

University of Virginia Health System | Claude Moore Health Sciences Library

R Ladies

useRs meetup

Cville Data Science Meetup

University of Virginia | School of Medicine Research Computing