06\_Class\_Activity

Bill Perry

# In class activity 6:

## What did we do last time in activity 5?

* stuff here
* stuff here
* stuff here
* stuff here

## Today’s focus:

* stuff here
* stuff here
* stuff here
  + stuff here
  + stuff here

# This code will accompany and reinforce Lecture X

# Install packages if needed (uncomment if necessary)  
# install.packages("readr")  
# install.packages("tidyverse")  
# install.packages("car")  
# install.packages("here")  
  
# Load libraries  
library(car) # For diagnostic tests

Loading required package: carData

library(patchwork)  
library(tidyverse) # For data manipulation and visualization

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
✔ purrr 1.0.4

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
✖ dplyr::recode() masks car::recode()  
✖ purrr::some() masks car::some()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Load the pine needle data  
# Use here() function to specify the path  
# df <- read\_csv("data/lake\_trout.csv")  
  
# Examine the first few rows  
# head(df)

# **Part 1:** Exploratory Data Analysis

Before conducting hypothesis tests, we should always explore our data to understand its characteristics.

Let’s calculate summary statistics and create visualizations.

**Activity: Calculate basic summary statistics for pine needle length**

# YOUR TASK: Calculate summary statistics for pine needle length  
# Hint: Use summarize() function to calculate mean, sd, n, etc.

# **Part 1:** Visualizing the Data

**Activity: Create visualizations of pine needle length**

Create a histogram and a boxplot to visualize the distribution of pine needle length values.

Effective data visualization helps us understand:

* The central tendency
* The spread of the data
* Potential outliers
* Shape of distribution

# Your Task

# YOUR TASK: Create a histogram of lke trout mass  
# Hint: Use ggplot() and geom\_histogram()  
  
# Histogram of all pine needle lengths  
  
  
# how can you do this by wind to see both plots

Now to make the various plots we talk about only for lake NE 12

## We need to select only that data

# # Filter for Lake NE 12  
# ne12\_data <- df %>%   
# filter(lake == "NE 12") %>%  
# filter(!is.na(mass\_g)) # Remove any NA values  
#   
# head(ne12\_data)

## Use Patchwork to combine the plots

# # Combine all plots using patchwork  
# combined\_stats\_plot <- (ne12\_histo\_plot + ne12\_dot\_plot) / (ne12\_box\_plot + ne12\_qq\_plot) +  
# plot\_annotation(  
# title = "Lake NE 12 Trout Mass Distribution",  
# subtitle = paste("n =", nrow(ne12\_data), "fish samples"),  
# theme = theme(plot.title = element\_text(hjust = 0.5),  
# plot.subtitle = element\_text(hjust = 0.5))  
# )  
#   
# # Display the combined plot  
# combined\_stats\_plot

# **Part 1:** Single Sample T-Test

We want to test if the mean pine needle length on the windward side differs from 55mm.

**Activity: Define hypotheses and identify assumptions**

H₀: μ = 55 (The mean pine needle length on windward side is 55mm) H₁: μ ≠ 55 (The mean pine needle length on windward side is not 55mm)

## Assumptions for t-test:

1. Data is normally distributed
2. Observations are independent
3. No significant outliers

# **Part 1:** Testing Assumptions

Before conducting our t-test, we need to verify that our data meets the necessary assumptions.

**Activity: Test the normality assumption**

Methods to test normality:

* Visual methods:
  + QQ plots or histograms
  + Statistical tests: Shapiro
  + Wilk test

# Assumptions in R - qqplots

# Filter for just windward side needles

## Shapiro Wilk

# Shapiro-Wilk test  
# shapiro\_test <- shapiro.test(windward\_data$len\_mm)  
# print(shapiro\_test)

# Check for outliers using boxplot  
# YOUR CODE HERE

# **Part 1:** Conducting the Single Sample T-Test

Now that we’ve checked our assumptions, we can perform the single sample t-test.

**Activity: Conduct a single sample t-test to compare windward needle length to 55mm** **What is probability of getting sample at least as far from 55mm as our sample mean?**

This is our p-value, which helps us decide whether to reject the null hypothesis.

# Calculate summary statistics for windward needles  
# windward\_summary <- windward\_data %>%  
# summarize(  
# mean\_length = mean(len\_mm),  
# sd\_length = sd(len\_mm),  
# n = n(),  
# se\_length = sd\_length / sqrt(n)  
# )  
#   
# print(windward\_summary)

# Your Task

# YOUR TASK: Conduct a single sample t-test  
# t\_test\_result <- t.test(windward\_data$len\_mm, mu = 55 )  
# print(t\_test\_result)

# Calculate t-statistic manually   
# YOUR CODE HERE: t = (sample\_mean - hypothesized\_mean) / (sample\_sd / sqrt(n))  
  
# can you do this manually or manually with R?

# **Part 2:** Two Sample T-Test

stuff here

Question: **Is there a significant difference in needle length between the windward and leeward sides?**

# your task

# # YOUR TASK: Create a plot using stat\_summary to show means and standard errors  
# ggplot(pine\_data, aes(x = wind, y = len\_mm, color = wind)) +  
# stat\_summary(fun = mean, geom = "point") +  
# stat\_summary(fun.data = mean\_se, geom = "errorbar", width = 0.2) +  
# labs(title = "Mean Pine Needle Length by Wind Exposure",  
# x = "Wind Exposure",  
# y = "Mean Length (mm)") +  
# theme\_minimal()

# **Part 3:** Paired T-Test (Extended Activity)

asdfasdfasd

# your task

# # YOUR TASK: Create a plot using stat\_summary to show means and standard errors  
# ggplot(pine\_data, aes(x = wind, y = len\_mm, color = wind)) +  
# stat\_summary(fun = mean, geom = "point") +  
# stat\_summary(fun.data = mean\_se, geom = "errorbar", width = 0.2) +  
# labs(title = "Mean Pine Needle Length by Wind Exposure",  
# x = "Wind Exposure",  
# y = "Mean Length (mm)") +  
# theme\_minimal()

# **Final Activity:**

**Common assumptions for t-tests:**

1. Normality: Data comes from normally distributed populations
2. Equal variances (for two-sample tests)
3. Independence: Observations are independent
4. No outliers: Extreme values can influence results

What can we do if our data violates these assumptions?

Alternatives when assumptions are violated:

* Data transformation (log, square root, etc.)
* Non-parametric tests
* Bootstrapping approaches

# **Summary and Conclusions**

In this activity, we’ve:

1. Formulated hypotheses about pine needle length
2. Tested assumptions for parametric tests
3. Conducted one-sample and two-sample t-tests
4. Visualized data using appropriate methods
5. Learned how to interpret and report t-test results

**Key takeaways:**

* Always check assumptions before conducting tests
* Visualize your data to understand patterns
* Report results comprehensively
* Consider alternatives when assumptions are violated

# Reflection Questions

After completing the activities, discuss these questions with your group:

1. How does sample size affect our confidence in estimating the population mean?
2. Why is the t-distribution more appropriate than the normal distribution when working with small samples?
3. When comparing two populations, what can we learn from looking at confidence intervals versus performing a t-test?
4. How would you explain the concept of statistical significance to someone who has never taken a statistics course?