06\_Class\_Activity

Bill Perry

# In class activity 6:

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

* H test for a single population
* 1- and 2-sided tests
* H test for two populations
* **Assumptions of parametric tests**

## Today’s focus:

* MORE R CODE!!!
* Parametric Statistical test assumptions and how to do them
* What to do if they fail?
  + transformations
  + NON PARAMETRIC TESTS - just as good or better

# This code will accompany and reinforce Lecture 6

# 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")

Rows: 1502 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (3): sampling\_site, species, lake  
dbl (2): length\_mm, mass\_g  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Examine the first few rows  
head(df)

# A tibble: 6 × 5  
 sampling\_site species length\_mm mass\_g lake   
 <chr> <chr> <dbl> <dbl> <chr>  
1 I8 lake trout 515 1400 I8   
2 I8 lake trout 468 1100 I8   
3 I8 lake trout 527 1550 I8   
4 I8 lake trout 525 1350 I8   
5 I8 lake trout 517 1300 I8   
6 I8 lake trout 607 2100 I8

# **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.  
  
# Create a summary table for all pine needles  
df\_summary <- df %>%  
 group\_by(lake) %>%   
 summarize(  
 mean\_length = mean(length\_mm, na.rm=TRUE),  
 sd\_length = sd(length\_mm, na.rm=TRUE),  
 n = sum(!is.na(length\_mm)),  
 se\_length = sd\_length / sqrt(n)  
 )  
  
print(df\_summary)

# A tibble: 6 × 5  
 lake mean\_length sd\_length n se\_length  
 <chr> <dbl> <dbl> <int> <dbl>  
1 I8 438. 90.3 87 9.69  
2 Island Lake 698. 121. 10 38.2   
3 N 01 405. 87.7 658 3.42  
4 NE 12 348. 127. 323 7.05  
5 NE 14 417. 110. 135 9.42  
6 Toolik 381. 98.4 241 6.34

# Now calculate summary statistics by wind exposure  
# YOUR CODE HERE

# **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.

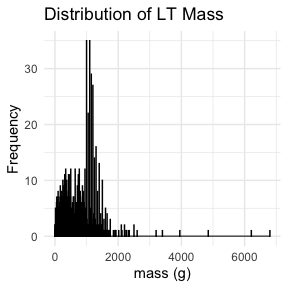
Effective data visualization helps us understand:

* 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  
ggplot(df, aes(x = mass\_g)) +  
 geom\_histogram(binwidth = 2, fill = "steelblue", color = "black") +  
 labs(title = "Distribution of LT Mass",  
 x = "mass (g)",  
 y = "Frequency") +  
 theme\_minimal()

Warning: Removed 17 rows containing non-finite outside the scale range  
(`stat\_bin()`).



# 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

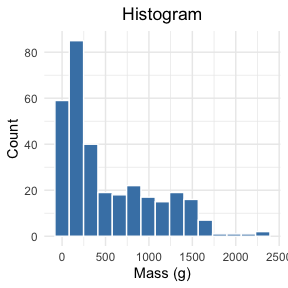
## We need to select only that data

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

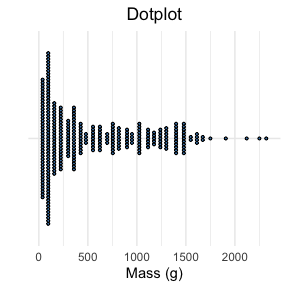
# A tibble: 6 × 5  
 sampling\_site species length\_mm mass\_g lake   
 <chr> <chr> <dbl> <dbl> <chr>  
1 NE 12 lake trout 206 83 NE 12  
2 NE 12 lake trout 191 48 NE 12  
3 NE 12 lake trout 228 98 NE 12  
4 NE 12 lake trout 208 65 NE 12  
5 NE 12 lake trout 228 90 NE 12  
6 NE 12 lake trout 215 86 NE 12

## Make a new histogram

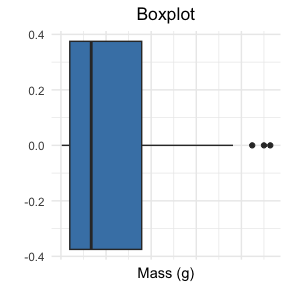
# 1. Histogram  
ne12\_histo\_plot <- ggplot(ne12\_data, aes(x = mass\_g)) +  
 geom\_histogram(bins = 15, fill = "steelblue", color = "white") +  
 labs(title = "Histogram", x = "Mass (g)", y = "Count") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))  
ne12\_histo\_plot



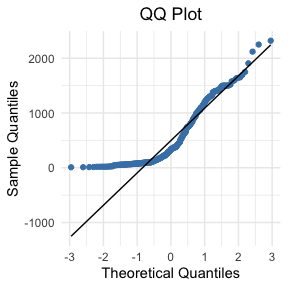
# 2. Dotplot  
ne12\_dot\_plot <- ggplot(ne12\_data, aes(x = mass\_g, y = "")) +  
 geom\_dotplot(binwidth = 60, stackdir = "center", fill = "steelblue", dotsize = 0.5) +  
 labs(title = "Dotplot", x = "Mass (g)", y = "") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.y = element\_blank(),  
 axis.ticks.y = element\_blank())  
ne12\_dot\_plot



# 3. Boxplot  
ne12\_box\_plot <- ggplot(ne12\_data, aes(y = mass\_g)) +  
 geom\_boxplot(fill = "steelblue") +  
 labs(title = "Boxplot", y = "Mass (g)", x = "") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5),  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank()) +  
 coord\_flip()  
ne12\_box\_plot

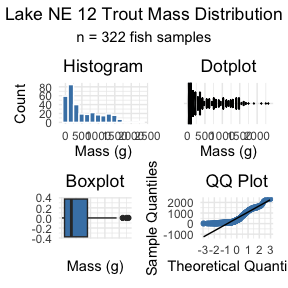


ne12\_qq\_plot <- ggplot(ne12\_data, aes(sample = mass\_g)) +  
 stat\_qq(color = "steelblue") +  
 stat\_qq\_line() +  
 labs(title = "QQ Plot", x = "Theoretical Quantiles", y = "Sample Quantiles") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))  
ne12\_qq\_plot



## 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 1:** Interpreting and Reporting Results

**Activity: Interpret the t-test results**

* What does the p-value tell us?
* Should we reject or fail to reject the null hypothesis?

**How to report this result in a scientific paper:**

“A two-tailed, one-sample t-test at α=0.05 showed that the mean pine needle length on the windward side (… mm, SD = …) [was/was not] significantly different from the expected 55 mm, t(…) = …, p = …”

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

Now, let’s compare pine needle lengths between windward and leeward sides of trees.

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

This requires a two-sample t-test.

Two-sample t-test compares means from two independent groups.

where:

* x̄₁ and x̄₂: These represent the sample means of the two groups you’re comparing
* s²ₚ: This is the pooled variance, calculated as: s²ₚ = [(n₁ - 1)s₁² + (n₂ - 1)s₂²] / (n₁ + n₂ - 2), where s₁² and s₂² are the sample variances of the two groups.
* **n₁ and n₂:** These are the sample sizes of the two groups.
* **√(1/n₁ + 1/n₂):** This represents the pooled standard error.

# **Part 2:** Exploratory Data Analysis by Group

**Activity: Calculate summary statistics grouped by wind exposure** Before conducting the test, we need to understand the data for each group.

# YOUR TASK: Calculate summary statistics by wind exposure  
# Hint: Use group\_by() and summarize()  
  
# group\_summary <- pine\_data %>%  
# group\_by(wind) %>%  
# summarize(  
# mean\_length = mean(len\_mm),  
# sd\_length = sd(len\_mm),  
# n = n(),  
# se\_length = sd\_length / sqrt(n)  
# )  
#   
# print(group\_summary)

# Alternative 1

# Calculate the difference in means  
# YOUR CODE HERE  
  
# # Assuming your dataframe is called df  
# group\_summary %>%  
# summarize(difference = mean\_length[wind == "wind"] - mean\_length[wind == "lee"])

# Alternative 2

# # Or alternatively using filter and pull:  
# lee\_mean <- group\_summary %>% filter(wind == "lee") %>% pull(mean\_length)  
# wind\_mean <- group\_summary %>% filter(wind == "wind") %>% pull(mean\_length)  
# difference <- wind\_mean - lee\_mean  
# difference

# **Part 2:** Visualizing Group Differences

**Activity: Create visualizations to compare the groups** Effective visualizations for group comparisons:

* Side-by-side boxplots
* Violin plots
* Error bar plots

# # YOUR TASK: Create boxplots to compare groups  
# ggplot(pine\_data, aes(x = wind, y = len\_mm, fill = wind)) +  
# geom\_boxplot() +  
# labs(title = "Pine Needle Length by Wind Exposure",  
# x = "Wind Exposure",  
# y = "Length (mm)") +  
# theme\_minimal()

# 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 2:** Testing Assumptions for Two-Sample T-Test

**Activity: Test assumptions for two-sample t-test**

For a two-sample t-test, we need to check:

1. Normality within each group
2. Equal variances between groups (for standard t-test)
3. Independent observations

If assumptions are violated:

* Welch’s t-test (unequal variances)
* Non-parametric alternatives (Mann-Whitney U test)

# your task

# # YOUR TASK: Test normality of windward pine needle lengths  
# # QQ Plot  
# qqPlot(pine\_data$len\_mm,   
# main = "QQ Plot for Windward Pine Needles",  
# ylab = "Sample Quantiles")

# # Testing normality for each group  
# # Leeward group  
# lee\_data <- pine\_data %>% filter(wind == "lee")  
# shapiro\_lee <- shapiro.test(lee\_data$len\_mm)  
# print("Shapiro-Wilk test for leeward data:")  
# print(shapiro\_lee)

# windward group

# Windward group  
# YOUR CODE HERE for windward group normality test

# Remember you can always do it in one go

# # there are always two ways  
# # Test for normality using Shapiro-Wilk test for each wind group  
# # All in one pipeline using tidyverse approach  
# normality\_results <- pine\_data %>%  
# group\_by(wind) %>%  
# summarize(  
# shapiro\_stat = shapiro.test(len\_mm)$statistic,  
# shapiro\_p\_value = shapiro.test(len\_mm)$p.value,  
# normal\_distribution = if\_else(shapiro\_p\_value > 0.05, "Normal", "Non-normal")  
# )  
#   
# # Print the results  
# print(normality\_results)

# Conduct a Levenes Test

# # Test for equal variances  
# # YOUR TASK: Conduct Levene's test for equality of variances  
# levene\_test <- leveneTest(len\_mm ~ wind, data = pine\_data)  
# print(levene\_test)

# Visual check for normality with QQ plots  
# YOUR CODE HERE

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

**Activity: Conduct a two-sample t-test**

Now we can compare the mean pine needle lengths between windward and leeward sides.

H₀: μ₁ = μ₂ (The mean needle lengths are equal) H₁: μ₁ ≠ μ₂ (The mean needle lengths are different)

Deciding between:

* Standard t-test (equal variances)
* Welch’s t-test (unequal variances)

# Based on our Levene’s test result.

# # YOUR TASK: Conduct a two-sample t-test  
# # Use var.equal=TRUE for standard t-test or var.equal=FALSE for Welch's t-test  
#   
# # Standard t-test (if variances are equal)  
# t\_test\_result <- t.test(len\_mm ~ wind, data = pine\_data, var.equal = TRUE)  
# print("Standard two-sample t-test:")  
# print(t\_test\_result)  
#   
# # Welch's t-test (if variances are unequal)  
# # YOUR CODE HERE  
#   
# # Calculate t-statistic manually (optional)  
# # YOUR CODE HERE: t = (mean1 - mean2) / sqrt((s1^2/n1) + (s2^2/n2))

# **Part 2:** Interpreting and Reporting Two-Sample T-Test Results

**Activity: Interpret the results of the two-sample t-test**

What can we conclude about the needle lengths on windward vs. leeward sides?

**How to report this result in a scientific paper:**

“A two-tailed, two-sample t-test at α=0.05 showed [a significant/no significant] difference in needle length between windward (M = …, SD = …) and leeward (M = …, SD = …) sides of pine trees, t(…) = …, p = ….”

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

If we collected data in pairs (same tree, different sides), we would use a paired t-test. **How would the analysis differ?**

1. We’d calculate the difference for each pair
2. Test if the mean difference equals zero
3. The paired approach often has more statistical power

Paired t-test formula:

where:

* is the mean difference
* is the standard deviation of differences
* is the number of pairs

# **Final Activity:** Assumptions of Parametric Tests

**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?