06\_Class\_Activity

Bill Perry

# In class activity 6:

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

* Understanding standard normal distributions and z-scores
* Calculating and interpreting standard error
* Creating confidence intervals
* Working with the Student’s t-distribution

## Today’s focus:

* Review more r code and add
* 1 sample t tests
* 2 sample t tests

# Goes with Lecture 6 as this is the code we will do

# 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  
library(tidyverse) # For data manipulation and visualization

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

# A tibble: 6 × 6  
 date group n\_s wind tree\_no len\_mm  
 <chr> <chr> <chr> <chr> <dbl> <dbl>  
1 3/20/25 cephalopods n lee 1 20  
2 3/20/25 cephalopods n lee 1 21  
3 3/20/25 cephalopods n lee 1 23  
4 3/20/25 cephalopods n lee 1 25  
5 3/20/25 cephalopods n lee 1 21  
6 3/20/25 cephalopods n lee 1 16

# **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  
pine\_summary <- pine\_data %>%  
 summarize(  
 mean\_length = mean(len\_mm),  
 sd\_length = sd(len\_mm),  
 n = n(),  
 se\_length = sd\_length / sqrt(n)  
 )  
  
print(pine\_summary)

# A tibble: 1 × 4  
 mean\_length sd\_length n se\_length  
 <dbl> <dbl> <int> <dbl>  
1 17.7 3.53 48 0.509

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

# **Part 1:** Visualizing the Data

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

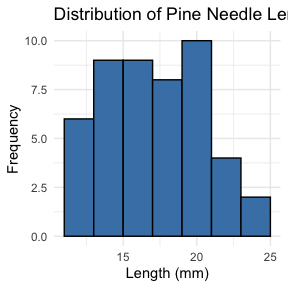
Create a histogram and a boxplot to visualize the distribution of pine needle length values.

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 pine needle length  
# Hint: Use ggplot() and geom\_histogram()  
  
# Histogram of all pine needle lengths  
ggplot(pine\_data, aes(x = len\_mm)) +  
 geom\_histogram(binwidth = 2, fill = "steelblue", color = "black") +  
 labs(title = "Distribution of Pine Needle Length",  
 x = "Length (mm)",  
 y = "Frequency") +  
 theme\_minimal()



# how can you do this by wind to see both plots

# Boxplot of pine needle length by wind exposure  
# YOUR CODE HERE

# **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.

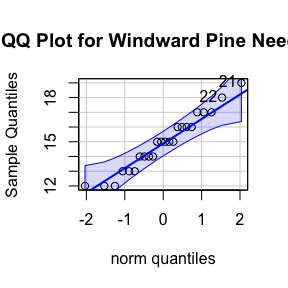
**Activity: Test the normality assumption**

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



[1] 21 22

## Shapiro Wilk

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

Shapiro-Wilk normality test  
  
data: windward\_data$len\_mm  
W = 0.96062, p-value = 0.451

# 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)

# A tibble: 1 × 4  
 mean\_length sd\_length n se\_length  
 <dbl> <dbl> <int> <dbl>  
1 14.9 1.91 24 0.390

# 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)

One Sample t-test  
  
data: windward\_data$len\_mm  
t = -102.85, df = 23, p-value < 2.2e-16  
alternative hypothesis: true mean is not equal to 55  
95 percent confidence interval:  
 14.11050 15.72284  
sample estimates:  
mean of x   
 14.91667

# 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)

# A tibble: 2 × 5  
 wind mean\_length sd\_length n se\_length  
 <chr> <dbl> <dbl> <int> <dbl>  
1 lee 20.4 2.45 24 0.500  
2 wind 14.9 1.91 24 0.390

# 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"])

# A tibble: 1 × 1  
 difference  
 <dbl>  
1 -5.5

# 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

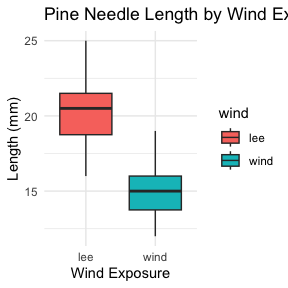
[1] -5.5

# **Part 2:** Visualizing Group Differences

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

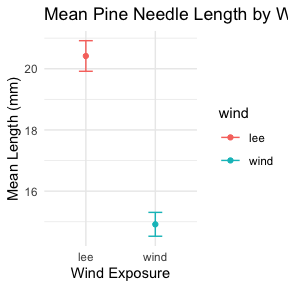
* Side-by-side boxplots
* Violin plots
* Error bar plots

# 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:

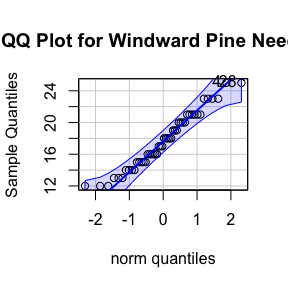
1. Normality within each group
2. Equal variances between groups (for standard t-test)
3. Independent observations

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")



[1] 4 28

# 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:")

[1] "Shapiro-Wilk test for leeward data:"

print(shapiro\_lee)

Shapiro-Wilk normality test  
  
data: lee\_data$len\_mm  
W = 0.95477, p-value = 0.3425

# 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)

# A tibble: 2 × 4  
 wind shapiro\_stat shapiro\_p\_value normal\_distribution  
 <chr> <dbl> <dbl> <chr>   
1 lee 0.955 0.343 Normal   
2 wind 0.961 0.451 Normal

# 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)

Levene's Test for Homogeneity of Variance (center = median)  
 Df F value Pr(>F)  
group 1 1.2004 0.2789  
 46

# 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:")

[1] "Standard two-sample t-test:"

print(t\_test\_result)

Two Sample t-test  
  
data: len\_mm by wind  
t = 8.6792, df = 46, p-value = 3.01e-11  
alternative hypothesis: true difference in means between group lee and group wind is not equal to 0  
95 percent confidence interval:  
 4.224437 6.775563  
sample estimates:  
 mean in group lee mean in group wind   
 20.41667 14.91667

# 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?