05\_Class\_Activity

Bill Perry

# In-Class Activity 5: Probability and Statistical Inference

## What did we do last time?

In our previous activity, we:

* Created and interpreted frequency distributions (histograms)
* Compared data between groups using side-by-side histograms
* Explored how sample size affects our understanding of populations
* Created density plots and calculated probabilities

## Today’s focus:

Today we’ll focus on:

* t-distribution and when to use it
* Calculating and interpreting standard error
* Creating confidence intervals
* Conducting one-sample and two-sample t-tests
* Understanding statistical assumptions and their importance

# Setup

First, let’s load the packages and data we’ll be using:

# Load required packages
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 3.5.2 ✔ tibble 3.3.0
✔ lubridate 1.9.4 ✔ tidyr 1.3.1
✔ purrr 1.1.0
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag() masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(patchwork) # For combining plots
library(car) # For diagnostic tests (QQ plots)

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

 recode

The following object is masked from 'package:purrr':

 some

# Read in the data files
g\_df <- read\_csv("data/gray\_I3\_I8.csv")

Rows: 168 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (2): lake, species
dbl (3): site, 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.

p\_df <- read\_csv("data/pine\_needles.csv")

Rows: 48 Columns: 6
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (4): date, group, n\_s, wind
dbl (2): tree\_no, length\_mm

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.

# Look at the first few rows of each dataset
head(g\_df)

# A tibble: 6 × 5
 site lake species length\_mm mass\_g
 <dbl> <chr> <chr> <dbl> <dbl>
1 113 I3 arctic grayling 266 135
2 113 I3 arctic grayling 290 185
3 113 I3 arctic grayling 262 145
4 113 I3 arctic grayling 275 160
5 113 I3 arctic grayling 240 105
6 113 I3 arctic grayling 265 145

head(p\_df)

# A tibble: 6 × 6
 date group n\_s wind tree\_no length\_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: Exploring the Data

Before conducting statistical tests, it’s important to understand your data.

|  |
| --- |
|  Practice Exercise 1: Creating Histograms |
| Let’s create histograms of fish lengths from each lake to visualize their distributions.# Create a histogram for Lake I3i3\_hist <- g\_df %>%  filter(lake == "I3") %>% ggplot(aes(length\_mm)) +  geom\_histogram(binwidth = 10, fill = "blue", alpha = 0.7) + labs(title = "Lake I3 Fish Lengths", x = "Length (mm)", y = "Count")# Create a histogram for Lake I8i8\_hist <- g\_df %>%  filter(lake == "I8") %>% ggplot(aes(length\_mm)) +  geom\_histogram(binwidth = 10, fill = "darkgreen", alpha = 0.7) + labs(title = "Lake I8 Fish Lengths", x = "Length (mm)", y = "Count")# Display the histograms side by side using patchworki3\_hist + i8\_hist# CAN YOU THINK OF AN EASIER WAY? |

## Now, let’s calculate summary statistics for each lake:

# Calculate summary statistics for both lakes
grayling\_summary <- g\_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),
 .groups = "drop"
 )

# Display the summary statistics
grayling\_summary

# A tibble: 2 × 5
 lake mean\_length sd\_length n se\_length
 <chr> <dbl> <dbl> <int> <dbl>
1 I3 266. 28.3 66 3.48
2 I8 363. 52.3 102 5.18

# Part 3: Testing Assumptions

Before conducting a t-test, we need to check if our data meets the necessary assumptions:

1. **Normality**: The data should be approximately normally distributed
2. **Independence**: Observations should be independent
3. **No extreme outliers**: Outliers can heavily influence t-test results

Let’s check the normality assumption for Lake I3 fish lengths:

|  |
| --- |
|  Practice Exercise 2: Checking Normality |
| # Filter for Lake I3 fishi3\_df <- g\_df %>% filter(lake == "I3")# Create a QQ plot to check normality# QQ plots compare our data to a theoretical normal distribution# Points should roughly follow the line if data is normally distributedqqPlot(i3\_df$length\_mm,  main = "QQ Plot for Lake I3 Fish Lengths", ylab = "Sample Quantiles")[1] 53 35 |

# Also perform a formal test of normality using the Shapiro-Wilk test
# Null hypothesis: Data is normally distributed
# If p > 0.05, we don't reject the assumption of normality
shapiro\_test <- shapiro.test(i3\_df$length\_mm)
print(shapiro\_test)

 Shapiro-Wilk normality test

data: i3\_df$length\_mm
W = 0.91051, p-value = 0.0001623

# Check for outliers using a boxplot
i3\_df %>%
 ggplot(aes(x = lake, y = length\_mm)) +
 geom\_boxplot() +
 labs(title = "Boxplot of Lake I3 Fish Lengths",
 y = "Length (mm)")



|  |
| --- |
|  Tip |
| How to interpret these results:* The QQ plot: Points should follow the straight line if data is normally distributed
* Shapiro-Wilk test: If p > 0.05, we don’t reject the assumption of normality
* Boxplot: Look for points beyond the whiskers as potential outliers
 |

# Part 4: One-Sample t-Test

A one-sample t-test compares a sample mean to a specific value.

Let’s test if the mean fish length in Lake I3 differs from 240mm:

|  |
| --- |
|  Practice Exercise 3: One-Sample t-Test |
| # Calculate the mean of I3 fishi3\_mean <- mean(i3\_df$length\_mm, na.rm = TRUE)cat("Mean fish length in Lake I3:", round(i3\_mean, 1), "mm\n")Mean fish length in Lake I3: 265.6 mm |

# Perform a one-sample t-test
# H0: μ = 240 (The mean fish length is 240mm)
# H1: μ ≠ 240 (The mean fish length is not 240mm)
t\_test\_result <- t.test(i3\_df$length\_mm, mu = 240)

# Display the test results
t\_test\_result

 One Sample t-test

data: i3\_df$length\_mm
t = 7.3497, df = 65, p-value = 4.17e-10
alternative hypothesis: true mean is not equal to 240
95 percent confidence interval:
 258.6481 272.5640
sample estimates:
mean of x
 265.6061

# Create a visualization of the test
i3\_df %>%
 ggplot(aes(x = length\_mm)) +
 geom\_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
 geom\_vline(xintercept = 240, color = "red", linetype = "dashed", linewidth = 1) +
 geom\_vline(xintercept = i3\_mean, color = "green", linewidth = 1) +
 annotate("text", x = 240, y = 5, label = "H0: μ = 240", color = "red", hjust = -0.1) +
 annotate("text", x = i3\_mean, y = 10, label = paste("Sample mean =", round(i3\_mean, 1)),
 color = "green", hjust = -0.1) +
 labs(title = "One-Sample t-Test: Lake I3 Fish Lengths",
 subtitle = paste("t =", round(t\_test\_result$statistic, 2),
 ", p =", format.pval(t\_test\_result$p.value, digits = 3)),
 x = "Length (mm)",
 y = "Count")



|  |
| --- |
|  Tip |
| Interpret the results:1. What was the null hypothesis? H0: μ = 240mm
2. What was the alternative hypothesis? H1: μ ≠ 240mm
3. What does the p-value tell us? (Is it < 0.05?)
4. Should we reject or fail to reject the null hypothesis?
5. What is the practical interpretation for biologists?
 |

# Part 5: Confidence Intervals

A confidence interval gives us a range of plausible values for the population mean.

For a 95% confidence interval using the t-distribution:

$$95\% CI=‾\pm t\_{α/2,n−1}×\frac{s}{\sqrt{n}}$$

Where: - $‾$ is the sample mean - $s$ is the sample standard deviation - $n$ is the sample size - $t\_{α/2,n−1}$ is the critical t-value with n-1 degrees of freedom

|  |
| --- |
|  Practice Exercise 4: Calculating Confidence Intervals |
| Let’s calculate the 95% confidence interval for Lake I3 fish lengths:# Extract sample statisticsi3\_stats <- grayling\_summary %>% filter(lake == "I3")i3\_mean <- i3\_stats$mean\_lengthi3\_se <- i3\_stats$se\_lengthi3\_n <- i3\_stats$n# Find the critical t-value for 95% confidence with n-1 degrees of freedom# qt(0.975, df) gives the t-value for a 95% confidence interval (two-tailed)t\_critical <- qt(0.975, df = i3\_n - 1)cat("Critical t-value for", i3\_n-1, "degrees of freedom:", round(t\_critical, 3), "\n")Critical t-value for 65 degrees of freedom: 1.997 # Calculate the confidence intervali3\_ci\_lower <- i3\_mean - t\_critical \* i3\_sei3\_ci\_upper <- i3\_mean + t\_critical \* i3\_se# Display the confidence intervalcat("95% Confidence Interval for Lake I3 fish mean length:",  round(i3\_ci\_lower, 1), "to", round(i3\_ci\_upper, 1), "mm\n")95% Confidence Interval for Lake I3 fish mean length: 258.6 to 272.6 mm# Compare this to a confidence interval using the normal approximation (z = 1.96)z\_ci\_lower <- i3\_mean - 1.96 \* i3\_sez\_ci\_upper <- i3\_mean + 1.96 \* i3\_secat("95% CI using normal approximation:",  round(z\_ci\_lower, 1), "to", round(z\_ci\_upper, 1), "mm\n")95% CI using normal approximation: 258.8 to 272.4 mm |

# Visualize the confidence interval
ggplot() +
 geom\_errorbar(aes(x = "Lake I3",
 ymin = i3\_ci\_lower,
 ymax = i3\_ci\_upper),
 width = 0.2) +
 geom\_point(aes(x = "Lake I3", y = i3\_mean), size = 3) +
 labs(title = "Mean Fish Length with 95% Confidence Interval",
 subtitle = "Lake I3",
 x = NULL,
 y = "Length (mm)")



|  |
| --- |
|  Tip |
| Interpretation:* We are 95% confident that the true population mean fish length in Lake I3 falls within this interval
* Note the small difference between using the t-distribution vs. normal approximation
 |

# Part 6: Two-Sample t-Test

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

Let’s compare pine needle lengths between windward and leeward sides:

# Summarize pine needle data by wind exposure
pine\_summary <- p\_df %>%
 group\_by(wind) %>%
 summarize(
 mean\_length = mean(length\_mm),
 sd\_length = sd(length\_mm),
 n = n(),
 se\_length = sd\_length / sqrt(n)
 )

# Display the summary statistics
print(pine\_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

## Look a the plot of pine needles

# Create a boxplot to visualize the data
p\_df %>%
 ggplot(aes(x = wind, y = length\_mm, fill = wind)) +
 geom\_boxplot() +
 labs(title = "Pine Needle Lengths by Wind Exposure",
 x = "Position",
 y = "Length (mm)",
 fill = "Wind Position") +
 scale\_fill\_manual(values = c("lee" = "forestgreen", "wind" = "skyblue"),
 labels = c("lee" = "Leeward", "wind" = "Windward"))



Before conducting the t-test, we should check the assumptions:

|  |
| --- |
|  Practice Exercise 5: Check Assumptions for Two-Sample t-Test |
| # Separate data by groupswindward\_data <- p\_df %>% filter(wind == "wind")leeward\_data <- p\_df %>% filter(wind == "lee")# 1. Check for normality in each group using QQ plotsqqPlot(windward\_data$length\_mm, main = "QQ Plot: Windward Needles")[1] 21 22qqPlot(leeward\_data$length\_mm, main = "QQ Plot: Leeward Needles")[1] 4 16# 2. Check for equal variances using Levene's test# H0: Variances are equal# H1: Variances are not equallevene\_result <- leveneTest(length\_mm ~ wind, data = p\_df)Warning in leveneTest.default(y = y, group = group, ...): group coerced tofactor.print("Levene's Test for Homogeneity of Variance:")[1] "Levene's Test for Homogeneity of Variance:"print(levene\_result)Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F)group 1 1.2004 0.2789 46  |

|  |
| --- |
|  Tip |
| Interpreting the assumption checks:* QQ plots: Do points approximately follow the line for both groups?
* Levene’s test: If p > 0.05, we don’t reject the assumption of equal variances
 |

## Now let’s conduct the two-sample t-test:

|  |
| --- |
|  Practice Exercise 6: Two-Sample t-Test |
| # Perform a two-sample t-test# H0: μ1 = μ2 (The mean needle lengths are equal)# H1: μ1 ≠ μ2 (The mean needle lengths are different)# var.equal=TRUE uses the standard t-test (pooled variance)# var.equal=FALSE uses Welch's t-test (for unequal variances)t\_test\_result <- t.test(length\_mm ~ wind, data = p\_df, var.equal = TRUE)# Display the test resultsprint(t\_test\_result) Two Sample t-testdata: length\_mm by windt = 8.6792, df = 46, p-value = 3.01e-11alternative hypothesis: true difference in means between group lee and group wind is not equal to 095 percent confidence interval: 4.224437 6.775563sample estimates: mean in group lee mean in group wind  20.41667 14.91667 # Calculate the mean differencemean\_diff <- pine\_summary$mean\_length[pine\_summary$wind == "lee"] -  pine\_summary$mean\_length[pine\_summary$wind == "wind"]cat("Mean difference (lee - wind):", round(mean\_diff, 2), "mm\n")Mean difference (lee - wind): 5.5 mm# Visualize the results with a mean and error bar plotggplot(pine\_summary, aes(x = wind, y = mean\_length, fill = wind)) + geom\_bar(stat = "identity", alpha = 0.7) + geom\_errorbar(aes(ymin = mean\_length - se\_length,  ymax = mean\_length + se\_length), width = 0.2) + scale\_fill\_manual(values = c("lee" = "forestgreen", "wind" = "skyblue"), labels = c("lee" = "Leeward", "wind" = "Windward")) + labs(title = "Pine Needle Lengths by Wind Exposure", subtitle = paste("t =", round(t\_test\_result$statistic, 2),  ", p =", format.pval(t\_test\_result$p.value, digits = 3)), x = "Position", y = "Mean Length (mm)", fill = "Wind Position")Interpret the results:1. What was the null hypothesis?
2. What was the alternative hypothesis?
3. What does the p-value tell us?
4. Should we reject or fail to reject the null hypothesis?
5. What is the practical interpretation for botanists?
 |

# Part 7: Comparing Fish Lengths Between Lakes

Let’s apply what we’ve learned to compare fish lengths between Lakes I3 and I8:

|  |
| --- |
|  Practice Exercise 7: Comparing Lakes |
| # Perform a two-sample t-test comparing I3 and I8# First check assumptions (variances)levene\_lakes <- leveneTest(length\_mm ~ lake, data = g\_df)Warning in leveneTest.default(y = y, group = group, ...): group coerced tofactor.print("Levene's Test for Lakes:")[1] "Levene's Test for Lakes:"print(levene\_lakes)Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F) group 1 13.705 0.0002907 \*\*\* 166 ---Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1# Perform the t-test with appropriate variance settinglakes\_t\_test <- t.test(length\_mm ~ lake, data = g\_df,  var.equal = (levene\_lakes$`Pr(>F)`[1] > 0.05))# Display the resultsprint(lakes\_t\_test) Welch Two Sample t-testdata: length\_mm by laket = -15.532, df = 161.63, p-value < 2.2e-16alternative hypothesis: true difference in means between group I3 and group I8 is not equal to 095 percent confidence interval: -109.32342 -84.66053sample estimates:mean in group I3 mean in group I8  265.6061 362.5980 # Create a visualizationggplot(g\_df, aes(x = lake, y = length\_mm, fill = lake)) + geom\_boxplot(alpha = 0.7) + labs(title = "Comparison of Fish Lengths Between Lakes", subtitle = paste("t =", round(lakes\_t\_test$statistic, 2),  ", p =", format.pval(lakes\_t\_test$p.value, digits = 3)), x = "Lake", y = "Length (mm)")Write your interpretation of the results:1. Is there a significant difference in fish lengths between lakes?
2. Which lake has longer fish on average?
3. How would you report this result in a scientific paper?
 |

# Part 8: Communicating Statistical Results

In scientific writing, it’s important to report statistical results clearly and consistently.

Here’s a standard format for reporting t-test results:

For a one-sample t-test: “A one-sample t-test showed that the mean fish length in Lake I3 (M = [mean], SD = [sd]) was [significantly/not significantly] different from 240 mm, t([df]) = [t-value], p = [p-value].”

For a two-sample t-test: “A two-sample t-test revealed that pine needle lengths on the leeward side (M = [mean1], SD = [sd1]) were [significantly/not significantly] [longer/shorter] than on the windward side (M = [mean2], SD = [sd2]), t([df]) = [t-value], p = [p-value].”

|  |
| --- |
|  Practice Exercise 8: Writing Statistical Results |
| Write properly formatted statements reporting the results of: 1. The one-sample t-test comparing Lake I3 fish to 240mm 2. The two-sample t-test comparing pine needle lengths 3. The two-sample t-test comparing fish lengths between lakesRemember to include: - Means and standard deviations for each group - The t-value with degrees of freedom - The p-value and whether the result is significant |

# Reflection Questions

1. How does the t-distribution differ from the normal distribution, and why does this matter for small samples?
2. What assumptions must be met to use a t-test, and what alternatives exist if these assumptions are violated?
3. What is the difference between statistical significance and practical importance?
4. How would the confidence interval change if we used a 99% confidence level instead of 95%?
5. How would you explain the concept of a p-value to someone with no statistical background?