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

# 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
```

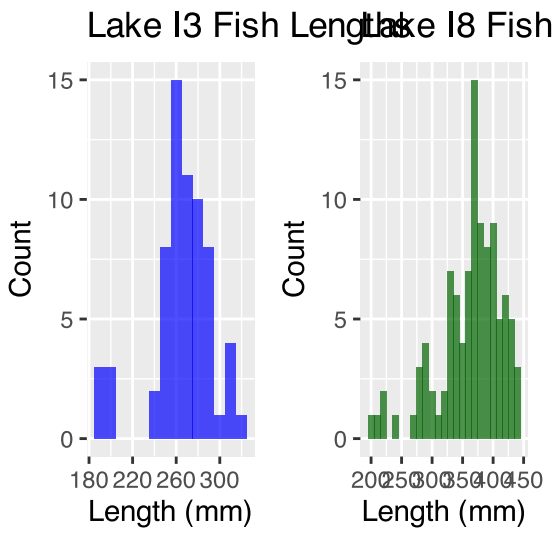
Part 1: Exploring the Data

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 I3
i3_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 I8
i8_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 patchwork
i3_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

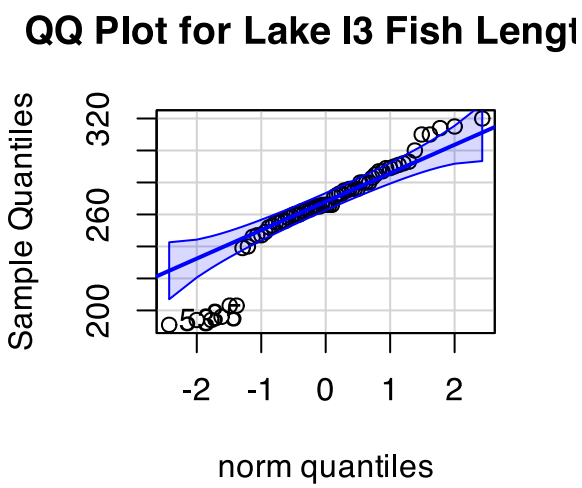
Before conducting a t-test, we need to check if our data meets the necessary assumptions:

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 fish
i3_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 distributed
qqPlot(i3_df$length_mm,
       main = "QQ Plot for Lake I3 Fish Lengths",
       ylab = "Sample Quantiles")
```

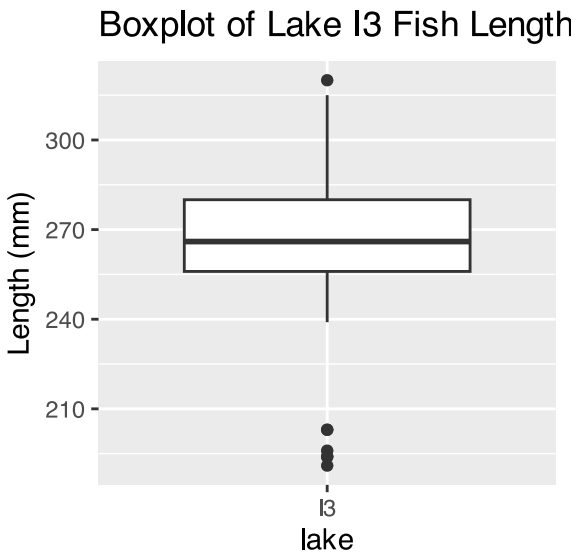



```
[1] 53 35
```

```
# 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 fish
i3_mean <- mean(i3_df$length_mm, na.rm = TRUE)
cat("Mean fish length in Lake I3:", round(i3_mean, 1), "mm\n")
```

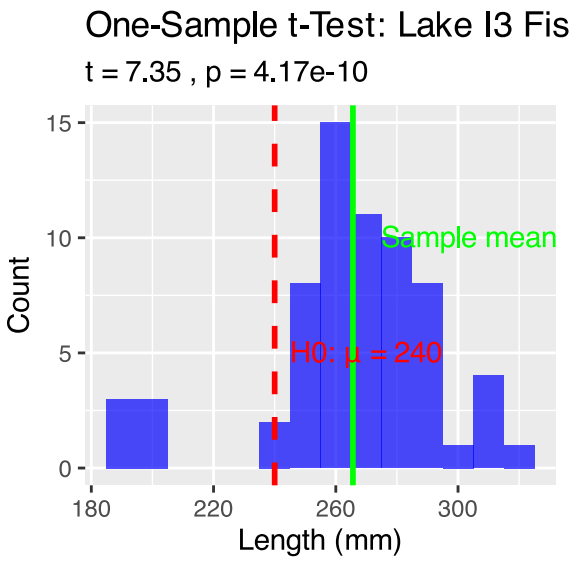
```
Mean fish length in Lake I3: 265.6 mm
```

```
# 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
```

```r
# 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:

$$95\% \text{ CI} = \bar{x} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$$

Where: - $\bar{x}$ is the sample mean - s is the sample standard deviation - n is the sample size - $t_{\alpha/2,n-1}$ 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 statistics
i3_stats <- grayling_summary %>% filter(lake == "I3")
i3_mean <- i3_stats$mean_length
i3_se <- i3_stats$se_length
i3_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 interval
i3_ci_lower <- i3_mean - t_critical * i3_se
i3_ci_upper <- i3_mean + t_critical * i3_se

# Display the confidence interval
cat("95% Confidence Interval for Lake I3 fish mean length:",
    round(i3_ci_lower, 1), "to", round(i3_ci_upper, 1), "mm\n")
```

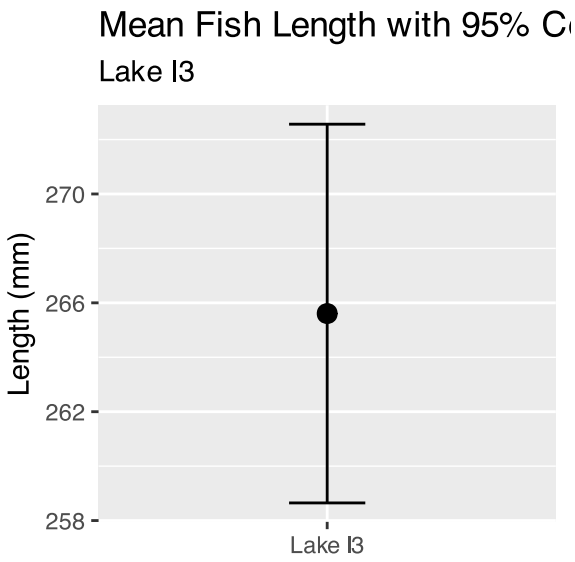
```
95% Confidence Interval for Lake I3 fish mean length: 258.6 to 272.6 mm
```

```
# Compare this to a confidence interval using the normal approximation (z = 1.96)
z_ci_lower <- i3_mean - 1.96 * i3_se
z_ci_upper <- i3_mean + 1.96 * i3_se

cat("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:

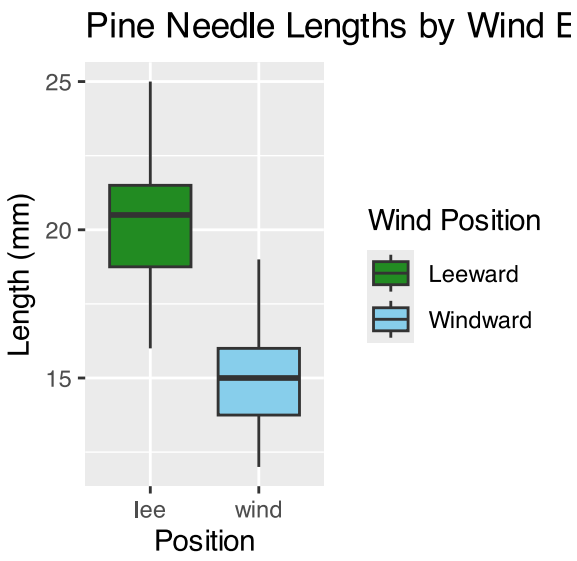
```
# Summarize pine needle data by wind exposure
pine_summary <- p_df %>%
  group_by(wind) %>%
  summarize(
    mean_length = mean(length_mm),
    sd_length = sd(length_mm),
    n = n(),
    se_length = sd_length / sqrt(n)
  )

# 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

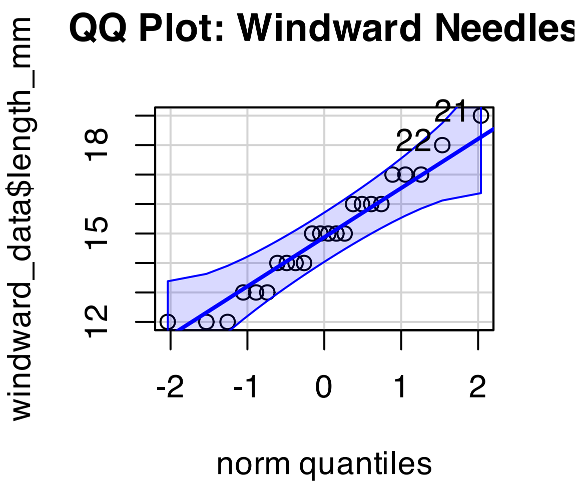
```
# Create a boxplot to visualize the data
p_df %>%
  ggplot(aes(x = wind, y = length_mm, fill = wind)) +
  geom_boxplot() +
  labs(title = "Pine Needle Lengths by Wind Exposure",
       x = "Position",
       y = "Length (mm)",
       fill = "Wind Position") +
  scale_fill_manual(values = c("lee" = "forestgreen", "wind" = "skyblue"),
                    labels = c("lee" = "Leeward", "wind" = "Windward"))
```



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

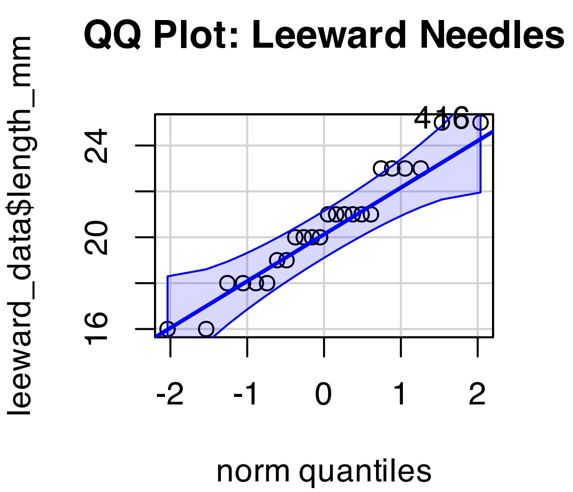
Practice Exercise 5: Check Assumptions for Two-Sample t-Test

```
# Separate data by groups
windward_data <- p_df %>% filter(wind == "wind")
leeward_data <- p_df %>% filter(wind == "lee")

# 1. Check for normality in each group using QQ plots

qqPlot(windward_data$length_mm, main = "QQ Plot: Windward Needles")
```




```
[1] 21 22
```

```
qqPlot(leeward_data$length_mm, main = "QQ Plot: Leeward Needles")
```




```
[1]  4 16
```

```
# 2. Check for equal variances using Levene's test
... group coerced to
factor Df F value Pr(>F)
# H1: Variances are not equal
```

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

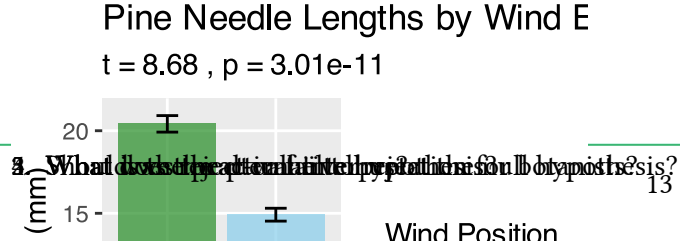
```
# Perform a two-sample t-test
# H0: μ1 = μ2 (The mean needle lengths are equal)
# H1: μ1 ≠ μ2 (The mean needle lengths are different)

# 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 results
print(t_test_result)
```

```
    Two Sample t-test

data:  length_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
```

```
# Calculate the mean difference
mean_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 plot
ggplot(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")
```



3. What does the p-value tell you about the hypothesis?

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 to
factor.
```

```
print("Levene's Test for Lakes:")
```

```
[1] "Levene's Test for Lakes:"
```

```
print(levene_lakes)
```

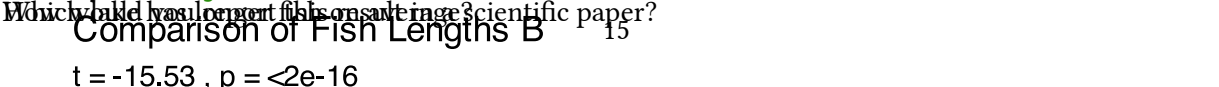
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  13.705 0.0002907 ***
      166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Perform the t-test with appropriate variance setting
lakes_t_test <- t.test(length_mm ~ lake, data = g_df,
                       var.equal = (levene_lakes$`Pr(>F)`[1] > 0.05))

# Display the results
print(lakes_t_test)
```

```

    Welch Two Sample t-test

data:  length_mm by lake
t = -15.532, df = 161.63, p-value < 2.2e-16
alternative hypothesis: true difference in means between group I3 and group I8 is not equal
to 0
95 percent confidence interval:
 -109.32342  -84.66053
sample estimates:
mean in group I3 mean in group I8
        265.6061         362.5980
```

```
# Create a visualization
ggplot(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)")
```

3. Which lake has longer fish on average?



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 lakes

Remember 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?