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
- understand α alpha and $\boldsymbol{\beta}$ beta errors
- do more
  - 1 sample t tests
  - 2 sample t tests

Goes with Lecture 6

```r
# Install packages if needed (uncomment if necessary)
# install.packages("readr")
# install.packages("tidyverse")
# install.packages("car")
# install.packages("here")

# Load libraries
library(patchwork)
library(car)          # For diagnostic tests
library(tidyverse)    # For data manipulation and visualization
```

```r
# Load the pine needle data
# Use here() function to specify the path
pine_data <- 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.
```

```r
# Examine the first few rows
head(pine_data)
```

```
# 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: 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 %>%
  group_by(wind) %>%
  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 / (n^0.5),
    t_critical = qt(0.975, df = n - 1),  # 95% CI uses 0.975 (two-tailed)
    ci_lower = mean_length - t_critical * se_length,
    ci_upper = mean_length + t_critical * se_length
  )

print(pine_summary)
```

```
# A tibble: 2 × 8
  wind  mean_length sd_length     n se_length t_critical ci_lower ci_upper
  <chr>       <dbl>     <dbl> <int>     <dbl>      <dbl>    <dbl>    <dbl>
1 lee          20.4      2.45    24     0.500       2.07     19.4     21.5
2 wind         14.9      1.91    24     0.390       2.07     14.1     15.7
```

```
# 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:

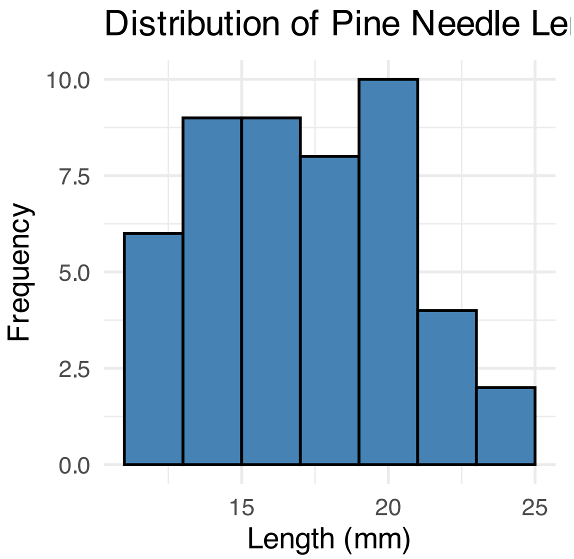
- The central tendency
- The spread of the data
- Potential outliers

- 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 = length_mm)) +
  geom_histogram(binwidth = 2, fill = "steelblue", color = "black") +
  labs(title = "Distribution of Pine Needle Length",
       x = "Length (mm)",
       y = "Frequency") +
  theme_minimal()
```



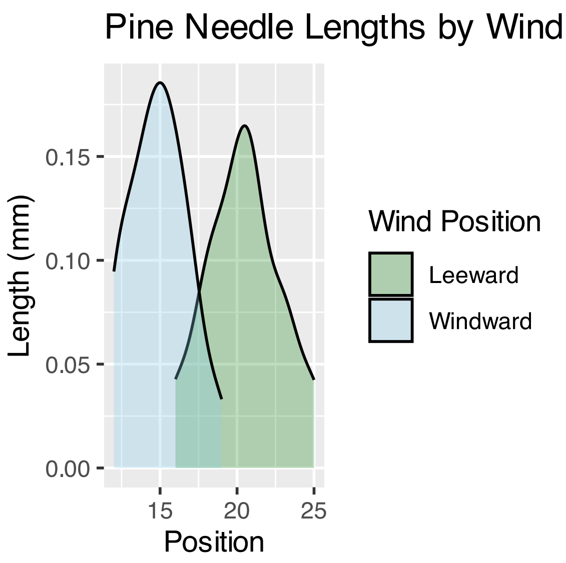
```
# how can you do this by wind to see both plots
```

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

Can you plot the density distributions for the two samples

```
# Histogram of all pine needle lengths
pine_data %>%
  ggplot(aes(x = length_mm, fill=wind)) +
  geom_density(color = "black", alpha=0.3, trim = TRUE) +
  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"))
```



what is the Effect size or difference in means?

Practice Exercise: Calculate Effect size

We could also look at the difference in means… some cool code here

```
# Assuming your dataframe is called df
pine_summary %>%
  summarize(difference =  mean_length[wind == "lee"] -mean_length[wind == "wind"])
```

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

Part 1: 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.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$: These represent the sample means of the two groups you're comparing
- s^2_p: This is the pooled variance, calculated as: $s^2_p = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2)$, where s_1^2 and s_2^2 are the sample variances of the two groups.
- **n_1 and n_2:** These are the sample sizes of the two groups.
- **$\sqrt{(1/n_1 + 1/n_2)}$:** This represents the pooled standard error.

Part 1: 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:

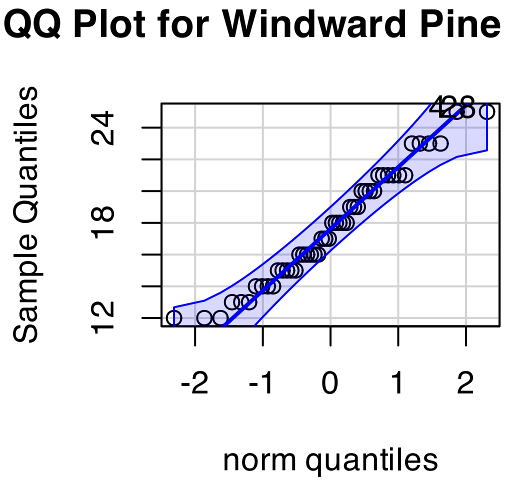
1. Normality within each group
2. Equal variances between groups (for standard t-test)
3. Independent observations

If assumptions are violated:

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

Your task

```r
# YOUR TASK: Test normality of windward pine needle lengths
# QQ Plot
qqPlot(pine_data$length_mm,
       main = "QQ Plot for Windward Pine Needles",
       ylab = "Sample Quantiles")
```



```
[1]  4 28
```

```r
# Testing normality for each group
# Leeward group
lee_data <- pine_data %>% filter(wind == "lee")
shapiro_lee <- shapiro.test(lee_data$length_mm)
print("Shapiro-Wilk test for leeward data:")
```

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

```r
print(shapiro_lee)
```

```

    Shapiro-Wilk normality test

data:  lee_data$length_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(length_mm)$statistic,
    shapiro_p_value = shapiro.test(length_mm)$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 Levene's Test

```
# Test for equal variances
# YOUR TASK: Conduct Levene's test for equality of variances
levene_test <- leveneTest(length_mm ~ wind, data = pine_data)
```

```
Warning in leveneTest.default(y = y, group = group, ...): group coerced to
factor.
```

```
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.

H_0: $\mu_1 = \mu_2$ (The mean needle lengths are equal)

H_1: $\mu_1 \neq \mu_2$ (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(length_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:  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 t-statistic manually (optional - uggg - maybe )
# YOUR CODE HERE: t = (mean1 - mean2) / sqrt((s1^2/n1) + (s2^2/n2))
```

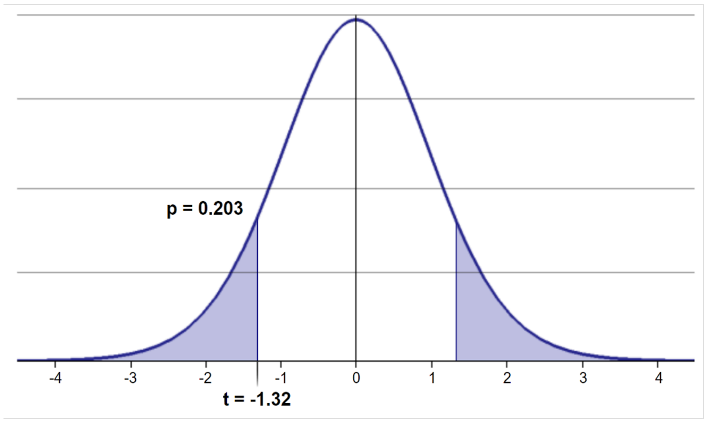
Part 2: Interpreting and Reporting Two-Sample T-Test Results

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 = ….”



What is Power

Statistical power represents the probability of detecting a true effect (rejecting the null hypothesis when it is false). In this case, with a power of 97%, there’s a 97% chance of detecting a true difference of 30 units between the means of the two groups if such a difference actually exists.

A power analysis like this is typically done for one of these purposes:

1. Before data collection to determine required sample size
2. After a study to evaluate if the sample size was adequate
3. To determine the minimum detectable effect size with the given sample

With 97% power, this test has excellent ability to detect the specified effect size. Generally, **80% power is considered acceptable**, so 97% indicates a very well-powered study for detecting a difference of 30mm between the groups.

```r
lee_df <- pine_data %>% filter(wind == "lee")
wind_df <- pine_data %>% filter(wind == "wind")
# Calculate power for detecting a 1 mm difference
wind_diff = 1

lee_n <- nrow(lee_df)
wind_n <- nrow(wind_df)

wind_sd_pooled <- sqrt((var(lee_df$length_mm) * (lee_n-1) +
                  var(wind_df$length_mm) * (wind_n-1)) /
                  (lee_n + wind_n - 2))

# Calculate power
wind_effect_size <- wind_diff / wind_sd_pooled  # Cohen's d
wind_df <- lee_n + wind_n - 2
wind_alpha <- 0.05
wind_power <- power.t.test(n = min(lee_n,wind_n),
                      delta = wind_effect_size,
                      sd = 1,  # Using standardized effect size
                      sig.level = 0.5,
                      type = "two.sample",
                      alternative = "two.sided")

# Display results
wind_power
```

```

     Two-sample t test power calculation

              n = 24
          delta = 0.4555423
             sd = 1
      sig.level = 0.5
          power = 0.8158402
    alternative = two.sided

NOTE: n is number in *each* group
```

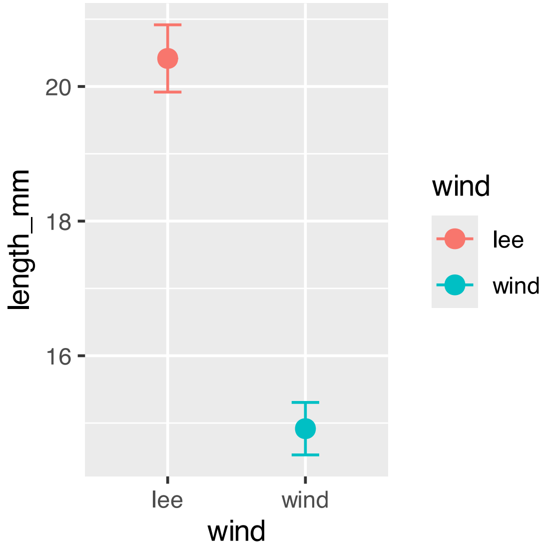
Now to make a final plot

Typically we will make a plot that has the mean and standard error on it to represent the data

your Task is to make this plot

```r
pine_mean_se <- pine_data %>%
  ggplot(aes(wind, length_mm, color = wind))+
  stat_summary(fun = "mean", na.rm=TRUE, geom="point", size = 3)+
  stat_summary(fun.data = "mean_se", width = 0.2, geom = "errorbar")

pine_mean_se
```



Summary and Conclusions

In this activity, we've:

1. Formulated hypotheses about pine needle length
2. Tested assumptions for parametric tests
3. Conducted a 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?
5. What do we do if assumptions FAIL‼!