08\_Class\_Activity

Bill Perry

# In class activity 8: Study Design and Power Analysis

## Introduction

This document demonstrates key concepts in experimental design using ecological examples, focusing on:

1. **Formulating research questions**
2. **Understanding different study designs**
3. **Recognizing proper replication vs. pseudoreplication**
4. **Designing appropriate controls**
5. **Conducting power analysis** (a priori and post hoc)
6. **Planning sampling strategies**

We’ll work with simulated pine needle data to practice these concepts.

Let’s start by exploring these concepts with hands-on examples!

# **Part 1:** Load Required Packages

# 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 3.2.1  
✔ 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()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(patchwork) # For combining plots  
library(pwr) # For power analysis  
  
# Set seed for reproducible results  
set.seed(42)

|  |
| --- |
| Package Overview |
| * **tidyverse**: Collection of packages for data science (includes ggplot2, dplyr, etc.) * **patchwork**: Easily combine multiple ggplot2 plots * **pwr**: Functions for power analysis and sample size calculations |

# **Part 2:** Formulating Research Questions

Before we design any study, we need clear research questions. Let’s practice with pine needle ecology.

|  |
| --- |
| Activity 1: Research Question Practice |
| Think about pine trees on campus. Write down 2-3 specific research questions about:   * - Pine needle characteristics (length, density, color) * - Environmental factors (wind, sunlight, soil) * - Tree health or growth   **Example questions:**   * - Does wind exposure affect pine needle length? * - Do pine needles on south-facing branches differ from north-facing branches? * - Does tree size influence needle density?   **Your questions:**   1. 1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2. 2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3. 3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

# **Part 3:** Understanding Study Design Types

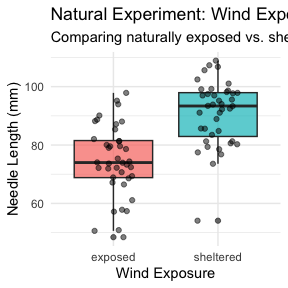
Let’s simulate data for different types of studies to understand their strengths and limitations.

## Natural Experiment: Wind Exposure Study

# Simulate pine needle data from naturally exposed and sheltered locations  
# This represents a "natural experiment" - we didn't manipulate wind exposure  
  
# Create data for exposed locations (shorter needles due to wind stress)  
exposed\_data <- data.frame(  
 location = rep(paste0("Exposed\_", 1:5), each = 8),  
 wind\_exposure = "exposed",  
 needle\_length\_mm = rnorm(40, mean = 75, sd = 10),  
 tree\_id = rep(1:5, each = 8)  
)  
  
# Create data for sheltered locations (longer needles, less wind stress)  
sheltered\_data <- data.frame(  
 location = rep(paste0("Sheltered\_", 1:5), each = 8),  
 wind\_exposure = "sheltered",   
 needle\_length\_mm = rnorm(40, mean = 90, sd = 12),  
 tree\_id = rep(6:10, each = 8)  
)  
  
# Combine the datasets  
natural\_exp\_data <- rbind(exposed\_data, sheltered\_data)  
  
# Look at the first few rows  
head(natural\_exp\_data)

location wind\_exposure needle\_length\_mm tree\_id  
1 Exposed\_1 exposed 88.70958 1  
2 Exposed\_1 exposed 69.35302 1  
3 Exposed\_1 exposed 78.63128 1  
4 Exposed\_1 exposed 81.32863 1  
5 Exposed\_1 exposed 79.04268 1  
6 Exposed\_1 exposed 73.93875 1

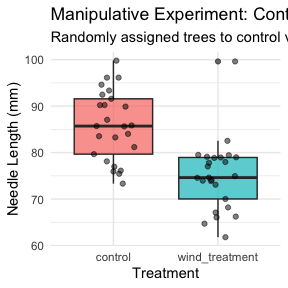
# Visualize the natural experiment data  
natural\_plot <- natural\_exp\_data %>%   
 ggplot(aes(x = wind\_exposure, y = needle\_length\_mm, fill = wind\_exposure)) +  
 geom\_boxplot(alpha = 0.7) +  
 geom\_jitter(width = 0.2, alpha = 0.5) +  
 labs(title = "Natural Experiment: Wind Exposure Effect",  
 subtitle = "Comparing naturally exposed vs. sheltered locations",  
 x = "Wind Exposure",  
 y = "Needle Length (mm)") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
natural\_plot



|  |
| --- |
| Natural Experiments: Pros and Cons |
| **Advantages:** - Realistic conditions - Large scale possible - Cost-effective  **Disadvantages:** - Cannot control confounding variables - Cannot determine causation direction - Many unmeasured factors might influence results  **Question:** What other factors besides wind might differ between “exposed” and “sheltered” locations? |

## Manipulative Experiment: Controlled Wind Study

# Simulate a controlled experiment where we manipulate wind exposure  
# All trees start similar, then we apply treatments  
  
# Create data for control group (normal conditions)  
control\_data <- data.frame(  
 treatment = "control",  
 needle\_length\_mm = rnorm(25, mean = 85, sd = 8),  
 tree\_id = 1:25  
)  
  
# Create data for wind treatment (artificial wind exposure)  
wind\_treatment\_data <- data.frame(  
 treatment = "wind\_treatment",  
 needle\_length\_mm = rnorm(25, mean = 78, sd = 8),  
 tree\_id = 26:50  
)  
  
# Combine the datasets  
manipulative\_data <- rbind(control\_data, wind\_treatment\_data)  
  
# Visualize the manipulative experiment  
manipulative\_plot <- manipulative\_data %>%   
 ggplot(aes(x = treatment, y = needle\_length\_mm, fill = treatment)) +  
 geom\_boxplot(alpha = 0.7) +  
 geom\_jitter(width = 0.2, alpha = 0.5) +  
 labs(title = "Manipulative Experiment: Controlled Wind Treatment",  
 subtitle = "Randomly assigned trees to control vs. wind treatment",  
 x = "Treatment",  
 y = "Needle Length (mm)") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
manipulative\_plot

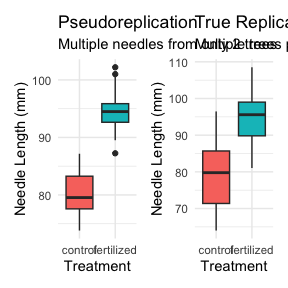


|  |
| --- |
| Manipulative Experiments: Key Features |
| **Advantages:** - Can establish causation - Control confounding variables - Random assignment eliminates bias  **Disadvantages:** - Often smaller scale - May not reflect natural conditions - Can be expensive and logistically challenging  **Key Question:** Which experiment gives stronger evidence for causation? |

# **Part 4:** Identifying Proper Replication

One of the most common mistakes in ecological studies is pseudoreplication. Let’s practice identifying true replication vs. pseudoreplication.

# Example 1: Pseudoreplication - multiple measurements from same trees  
pseudo\_data <- data.frame(  
 treatment = rep(c("fertilized", "control"), each = 20),  
 tree\_id = rep(c("Tree\_A", "Tree\_B"), each = 20), # Only 2 trees total!  
 needle\_length\_mm = c(  
 rnorm(20, mean = 95, sd = 5), # Tree A (fertilized)  
 rnorm(20, mean = 80, sd = 5) # Tree B (control)  
 ),  
 measurement = rep(1:20, times = 2)  
)  
  
# Example 2: True replication - multiple trees per treatment  
true\_rep\_data <- data.frame(  
 treatment = rep(c("fertilized", "control"), each = 20),  
 tree\_id = paste0("Tree\_", 1:40), # 40 different trees  
 needle\_length\_mm = c(  
 rnorm(20, mean = 95, sd = 8), # 20 fertilized trees  
 rnorm(20, mean = 80, sd = 8) # 20 control trees  
 )  
)  
  
# Create comparison plots  
pseudo\_plot <- pseudo\_data %>%  
 ggplot(aes(x = treatment, y = needle\_length\_mm, fill = treatment)) +  
 geom\_boxplot() +  
 labs(title = "Pseudoreplication",  
 subtitle = "Multiple needles from only 2 trees",  
 x = "Treatment", y = "Needle Length (mm)") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
true\_plot <- true\_rep\_data %>%  
 ggplot(aes(x = treatment, y = needle\_length\_mm, fill = treatment)) +  
 geom\_boxplot() +  
 labs(title = "True Replication",   
 subtitle = "Multiple trees per treatment",  
 x = "Treatment", y = "Needle Length (mm)") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
# Combine plots  
pseudo\_plot + true\_plot



|  |
| --- |
| Pseudoreplication Alert! |
| **Pseudoreplication occurs when:** - You treat subsamples as independent when they’re not - Multiple measurements from the same experimental unit - Replication at wrong scale for your hypothesis  **Common examples:** - Multiple leaves from one plant - Multiple water samples from one lake - Multiple plots within one treatment area  **Why it’s bad:** - Underestimates variability - Inflates sample size artificially - Increases Type I error (false positives) |

## Activity: Identify Replication Issues

|  |
| --- |
| Activity 2: Replication Practice |
| For each scenario, identify if there’s proper replication or pseudoreplication:  **Scenario A:** Testing fertilizer effects by using 1 large pot with fertilizer containing 10 pine seedlings, and 1 control pot with 10 seedlings. - **Your answer:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ - **Fix:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Scenario B:** Testing altitude effects by measuring needle length on 5 trees at 1000m elevation and 5 trees at 2000m elevation. - **Your answer:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ - **Fix:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Scenario C:** Testing soil pH by measuring 20 needles each from 10 trees in acidic soil and 10 trees in basic soil. - **Your answer:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ - **Fix:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

# **Part 5:** Power Analysis - Planning Your Study

Power analysis helps us determine how many samples we need to detect an effect if it really exists.

## A Priori Power Analysis (Before Data Collection)

# Scenario: We want to detect a difference in needle length between   
# fertilized and control trees  
  
# Based on pilot data, we expect:  
control\_mean <- 80 # mm  
fertilized\_mean <- 90 # mm   
pooled\_sd <- 12 # mm  
  
# Calculate effect size (Cohen's d)  
effect\_size <- abs(fertilized\_mean - control\_mean) / pooled\_sd  
cat("Effect size (Cohen's d):", round(effect\_size, 2), "\n")

Effect size (Cohen's d): 0.83

# Interpret effect size  
if(effect\_size < 0.2) {  
 interpretation <- "small"  
} else if(effect\_size < 0.5) {  
 interpretation <- "small-medium"   
} else if(effect\_size < 0.8) {  
 interpretation <- "medium-large"  
} else {  
 interpretation <- "large"  
}  
cat("This is a", interpretation, "effect size\n")

This is a large effect size

# Calculate required sample size for 80% power  
power\_result <- pwr.t.test(  
 d = effect\_size, # Effect size  
 sig.level = 0.05, # Alpha level (significance)  
 power = 0.8, # Desired power (80%)  
 type = "two.sample" # Two-sample t-test  
)  
  
print(power\_result)

Two-sample t test power calculation   
  
 n = 23.60467  
 d = 0.8333333  
 sig.level = 0.05  
 power = 0.8  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

cat("\nWe need", ceiling(power\_result$n), "trees per group for 80% power\n")

We need 24 trees per group for 80% power

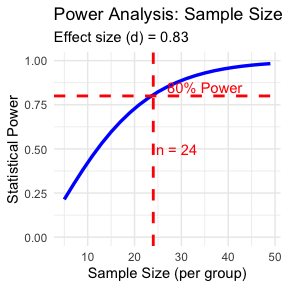
|  |
| --- |
| Understanding Effect Size (Cohen’s d) |
| * **d = 0.2**: Small effect (subtle difference) * **d = 0.5**: Medium effect (moderate difference) * **d = 0.8**: Large effect (substantial difference)   **Cohen’s d formula:** d = (Mean₁ - Mean₂) / Pooled Standard Deviation |

## Visualizing Power Curves

# Create a power curve showing relationship between sample size and power  
sample\_sizes <- seq(5, 50, by = 2)  
  
# Calculate power for each sample size  
power\_values <- sapply(sample\_sizes, function(n) {  
 power\_test <- pwr.t.test(n = n, d = effect\_size, sig.level = 0.05, type = "two.sample")  
 return(power\_test$power)  
})  
  
# Create data frame for plotting  
power\_df <- data.frame(  
 sample\_size = sample\_sizes,  
 power = power\_values  
)  
  
# Create power curve plot  
power\_curve\_plot <- ggplot(power\_df, aes(x = sample\_size, y = power)) +  
 geom\_line(color = "blue", size = 1.2) +  
 geom\_hline(yintercept = 0.8, linetype = "dashed", color = "red", size = 1) +  
 geom\_vline(xintercept = ceiling(power\_result$n), linetype = "dashed", color = "red", size = 1) +  
 annotate("text", x = ceiling(power\_result$n) + 5, y = 0.5,   
 label = paste("n =", ceiling(power\_result$n)), color = "red") +  
 annotate("text", x = 35, y = 0.85, label = "80% Power", color = "red") +  
 ylim(0, 1) +  
 labs(title = "Power Analysis: Sample Size vs. Statistical Power",  
 subtitle = paste("Effect size (d) =", round(effect\_size, 2)),  
 x = "Sample Size (per group)",  
 y = "Statistical Power") +  
 theme\_minimal()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.

power\_curve\_plot



## Post Hoc Power Analysis (After Data Collection)

# Imagine we collected data with n = 15 per group but found no significant difference  
# Was our study adequately powered?  
  
observed\_n <- 15  
  
# Calculate the power we actually had  
actual\_power <- pwr.t.test(  
 n = observed\_n,  
 d = effect\_size,  
 sig.level = 0.05,  
 type = "two.sample"  
)  
  
print(actual\_power)

Two-sample t test power calculation   
  
 n = 15  
 d = 0.8333333  
 sig.level = 0.05  
 power = 0.5962064  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

cat("\nWith n =", observed\_n, "per group, we only had",   
 round(actual\_power$power \* 100, 1), "% power\n")

With n = 15 per group, we only had 59.6 % power

if(actual\_power$power < 0.8) {  
 cat("This study was underpowered! A non-significant result might be due to insufficient sample size.\n")  
} else {  
 cat("This study had adequate power. A non-significant result likely reflects no true effect.\n")  
}

This study was underpowered! A non-significant result might be due to insufficient sample size.

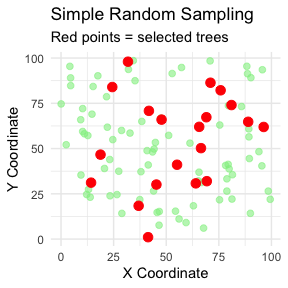
|  |
| --- |
| Activity 3: Power Analysis Practice |
| **Scenario:** You want to study the effect of drought stress on pine needle length. Based on literature, you expect: - Control trees: mean = 85mm, SD = 10mm - Drought-stressed trees: mean = 75mm, SD = 10mm  **Calculate the following:**  # Your turn! Fill in the values and run the code  # Step 1: Calculate effect size control\_mean <- 9 drought\_mean <- 99  pooled\_sd <- 999  effect\_size <- abs(control\_mean - drought\_mean) / pooled\_sd print(paste("Effect size:", round(effect\_size, 2)))  # Step 2: Calculate required sample size for 80% power power\_result <- pwr.t.test(  d = effect\_size,  sig.level = 0.05,  power = 0.8,  type = "two.sample" )  print(power\_result) print(paste("Required sample size:", ceiling(power\_result$n), "trees per group"))  # Step 3: What if you can only collect 12 trees per group? limited\_power <- pwr.t.test(  n = 12,  d = effect\_size,   sig.level = 0.05,  type = "two.sample" )  print(paste("Power with n=12:", round(limited\_power$power \* 100, 1), "%"))  **Questions:** 1. What is the effect size for this drought study? 2. How many trees do you need per group for 80% power? 3. If you can only sample 12 trees per group, what power will you have? |

# **Part 6:** Sampling Design Strategies

Different research questions require different sampling approaches. Let’s explore the main types.

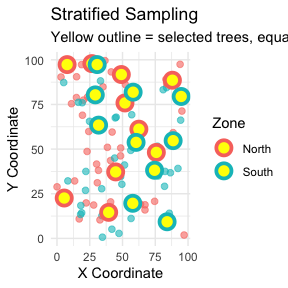
## Simple Random Sampling

# Simulate a campus with pine trees scattered randomly  
set.seed(123)  
campus\_trees <- data.frame(  
 tree\_id = 1:100,  
 x\_coordinate = runif(100, 0, 100), # Random x positions  
 y\_coordinate = runif(100, 0, 100), # Random y positions  
 needle\_length = rnorm(100, mean = 80, sd = 12)  
)  
  
# Simple random sampling: select 20 trees randomly  
random\_sample\_ids <- sample(1:100, size = 20, replace = FALSE)  
random\_sample <- campus\_trees[campus\_trees$tree\_id %in% random\_sample\_ids, ]  
  
# Visualize sampling design  
campus\_plot <- ggplot(campus\_trees, aes(x = x\_coordinate, y = y\_coordinate)) +  
 geom\_point(color = "lightgreen", size = 2, alpha = 0.6) +  
 geom\_point(data = random\_sample, color = "red", size = 3) +  
 labs(title = "Simple Random Sampling",  
 subtitle = "Red points = selected trees",  
 x = "X Coordinate", y = "Y Coordinate") +  
 theme\_minimal()  
  
campus\_plot



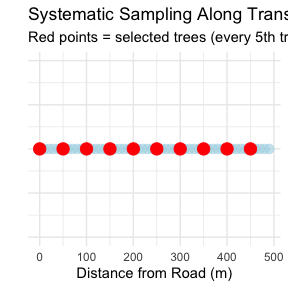
## Stratified Sampling

# Simulate campus with different zones (north vs south)  
set.seed(124)  
stratified\_trees <- data.frame(  
 tree\_id = 1:100,  
 x\_coordinate = runif(100, 0, 100),  
 y\_coordinate = runif(100, 0, 100),  
 zone = ifelse(runif(100) > 0.5, "North", "South"),  
 needle\_length = rnorm(100, mean = 80, sd = 12)  
)  
  
# Add zone effect to needle length  
stratified\_trees$needle\_length[stratified\_trees$zone == "South"] <-   
 stratified\_trees$needle\_length[stratified\_trees$zone == "South"] + 8  
  
# Stratified sampling: sample equally from each zone  
north\_trees <- stratified\_trees[stratified\_trees$zone == "North", ]  
south\_trees <- stratified\_trees[stratified\_trees$zone == "South", ]  
  
# Sample 10 from each zone  
north\_sample <- north\_trees[sample(nrow(north\_trees), 10), ]  
south\_sample <- south\_trees[sample(nrow(south\_trees), 10), ]  
stratified\_sample <- rbind(north\_sample, south\_sample)  
  
# Visualize stratified sampling  
stratified\_plot <- ggplot(stratified\_trees, aes(x = x\_coordinate, y = y\_coordinate, color = zone)) +  
 geom\_point(size = 2, alpha = 0.6) +  
 geom\_point(data = stratified\_sample, size = 4, shape = 21, fill = "yellow", stroke = 2) +  
 labs(title = "Stratified Sampling",  
 subtitle = "Yellow outline = selected trees, equal sampling from each zone",  
 x = "X Coordinate", y = "Y Coordinate", color = "Zone") +  
 theme\_minimal()  
  
stratified\_plot



## Systematic Sampling

# Systematic sampling along a transect  
set.seed(125)  
transect\_trees <- data.frame(  
 tree\_id = 1:50,  
 distance\_m = seq(0, 490, by = 10), # Trees every 10m along transect  
 needle\_length = rnorm(50, mean = 80, sd = 10)  
)  
  
# Add distance effect (trees farther from road have longer needles)  
transect\_trees$needle\_length <- transect\_trees$needle\_length +   
 (transect\_trees$distance\_m \* 0.02)  
  
# Systematic sampling: every 5th tree  
systematic\_sample <- transect\_trees[seq(1, 50, by = 5), ]  
  
# Visualize systematic sampling  
systematic\_plot <- ggplot(transect\_trees, aes(x = distance\_m, y = 1)) +  
 geom\_point(size = 3, alpha = 0.6, color = "lightblue") +  
 geom\_point(data = systematic\_sample, size = 4, color = "red") +  
 labs(title = "Systematic Sampling Along Transect",  
 subtitle = "Red points = selected trees (every 5th tree)",  
 x = "Distance from Road (m)", y = "") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_blank(), axis.ticks.y = element\_blank()) +  
 ylim(0.5, 1.5)  
  
systematic\_plot



|  |
| --- |
| Sampling Strategy Comparison |
| **Simple Random Sampling:** - Best for: General population estimates - Pros: Unbiased, simple analysis - Cons: May miss important subgroups  **Stratified Sampling:** - Best for: When you know there are distinct subgroups - Pros: Ensures representation of all strata - Cons: Requires prior knowledge of strata  **Systematic Sampling:** - Best for: Studying gradients or patterns - Pros: Good spatial coverage, easy to implement - Cons: Risk of bias if there’s hidden periodicity |

# **Part 7:** Putting It All Together - Design Your Own Study

|  |
| --- |
| Activity 4: Complete Study Design |
| **Research Question:** Does fertilizer application affect pine needle length?  **Design your study by answering these questions:**   1. **Study Type:** Will this be a natural experiment or manipulative experiment? Why?    * Your answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2. **Sample Size:** Using the following parameters, calculate required sample size:    * Expected control mean: 80mm    * Expected fertilized mean: 88mm    * Expected SD for both groups: 10mm    * Desired power: 80%   # Calculate effect size and sample size needed control\_mean <- 80 fertilized\_mean <- 88 pooled\_sd <- 10  effect\_size <- abs(fertilized\_mean - control\_mean) / pooled\_sd  power\_result <- pwr.t.test(  d = effect\_size,  sig.level = 0.05,  power = 0.8,  type = "two.sample" )  print(power\_result)   1. **Controls:** What controls will you include? Consider both positive and negative controls.    * Your answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2. **Randomization:** How will you randomize tree assignment to treatments?    * Your answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3. **Replication:** How will you ensure proper replication? What would be pseudoreplication?    * Proper replication: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_    * Pseudoreplication to avoid: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 4. **Independence:** What factors might violate independence? How will you address them?    * Your answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5. **Potential Confounds:** What other variables might affect needle length that you need to control for?    * Your answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

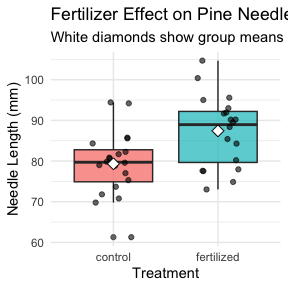
# **Part 8:** Analyzing Your Designed Study

Let’s simulate data from the study you designed and analyze it:

# Simulate data based on your study design  
set.seed(200)  
  
# Use the sample size you calculated (or use 20 if you didn't calculate)  
n\_per\_group <- 20 # Replace with your calculated sample size  
  
# Create the experimental data  
study\_data <- data.frame(  
 tree\_id = 1:(2 \* n\_per\_group),  
 treatment = rep(c("control", "fertilized"), each = n\_per\_group),  
 needle\_length\_mm = c(  
 rnorm(n\_per\_group, mean = 80, sd = 10), # Control group  
 rnorm(n\_per\_group, mean = 88, sd = 10) # Fertilized group  
 )  
)  
  
# Calculate summary statistics  
summary\_stats <- study\_data %>%  
 group\_by(treatment) %>%  
 summarise(  
 n = n(),  
 mean\_length = mean(needle\_length\_mm),  
 sd\_length = sd(needle\_length\_mm),  
 se\_length = sd\_length / sqrt(n)  
 )  
  
print(summary\_stats)

# A tibble: 2 × 5  
 treatment n mean\_length sd\_length se\_length  
 <chr> <int> <dbl> <dbl> <dbl>  
1 control 20 79.3 7.85 1.76  
2 fertilized 20 87.4 8.57 1.92

# Create visualization  
study\_plot <- study\_data %>%  
 ggplot(aes(x = treatment, y = needle\_length\_mm, fill = treatment)) +  
 geom\_boxplot(alpha = 0.7) +  
 geom\_jitter(width = 0.2, alpha = 0.6) +  
 stat\_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white") +  
 labs(title = "Fertilizer Effect on Pine Needle Length",  
 subtitle = "White diamonds show group means",  
 x = "Treatment",  
 y = "Needle Length (mm)") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
study\_plot



# Conduct statistical test  
t\_test\_result <- t.test(needle\_length\_mm ~ treatment, data = study\_data)  
print(t\_test\_result)

Welch Two Sample t-test  
  
data: needle\_length\_mm by treatment  
t = -3.1164, df = 37.715, p-value = 0.003493  
alternative hypothesis: true difference in means between group control and group fertilized is not equal to 0  
95 percent confidence interval:  
 -13.364541 -2.837295  
sample estimates:  
 mean in group control mean in group fertilized   
 79.33243 87.43335

# Interpret results  
if(t\_test\_result$p.value < 0.05) {  
 cat("\nResult: Significant difference found!\n")  
 cat("Fertilizer significantly affects needle length (p =",   
 round(t\_test\_result$p.value, 4), ")\n")  
} else {  
 cat("\nResult: No significant difference found.\n")  
 cat("No evidence that fertilizer affects needle length (p =",   
 round(t\_test\_result$p.value, 4), ")\n")  
}

Result: Significant difference found!  
Fertilizer significantly affects needle length (p = 0.0035 )

# Calculate actual effect size observed  
observed\_effect\_size <- abs(diff(t\_test\_result$estimate)) /   
 sqrt(((n\_per\_group-1) \* var(study\_data$needle\_length\_mm[study\_data$treatment == "control"]) +   
 (n\_per\_group-1) \* var(study\_data$needle\_length\_mm[study\_data$treatment == "fertilized"])) /   
 (2\*n\_per\_group - 2))  
  
cat("Observed effect size (Cohen's d):", round(observed\_effect\_size, 2), "\n")

Observed effect size (Cohen's d): 0.99

# **Summary and Key Takeaways**

|  |
| --- |
| What We Learned Today |
| 1. **Study Design Matters:** Statistics cannot fix a poorly designed study 2. **Replication:** Must be at the appropriate scale for your research question 3. **Controls:** Essential for ruling out alternative explanations 4. **Power Analysis:** Plan your sample size before collecting data 5. **Sampling Strategy:** Choose the approach that best fits your research question 6. **Integration:** Good analysis flows naturally from good design   **Remember:** - Design before you collect data - Consider practical and logistical constraints - Be transparent about limitations - Correlation ≠ causation (especially in natural experiments) |

|  |
| --- |
| Common Pitfalls to Avoid |
| 1. **Pseudoreplication:** Taking multiple measurements from the same experimental unit 2. **Inadequate Power:** Collecting too few samples to detect meaningful effects 3. **Poor Controls:** Not controlling for important confounding variables 4. **Non-random Sampling:** Introducing bias through convenience sampling 5. **HARKing:** Hypothesizing After Results are Known   **The Golden Rule:** Plan your analysis when you plan your experiment! |

# **Practice Problems**

|  |
| --- |
| Challenge Problems |
| Try these on your own to test your understanding:  **Problem 1:** A researcher wants to test if tree diameter affects needle length. They measure diameter and needle length on 50 randomly selected pine trees. - What type of study is this? - What statistical test would be appropriate? - What are potential confounding variables?  **Problem 2:** An experiment tests three fertilizer types (A, B, control) with 5 trees per treatment. From each tree, they collect 20 needles and measure length. - Is this proper replication? - What’s the actual sample size for statistical purposes? - How would you redesign this study?  **Problem 3:** Calculate sample size needed to detect a 15mm difference in needle length between treatments, with SD = 20mm and 90% power.  # Your code here for Problem 3 effect\_size <- \_\_\_ power\_result <- pwr.t.test(  d = \_\_\_,  sig.level = \_\_\_,  power = \_\_\_,  type = "two.sample" ) |