Monte Carlo Analysis of Ant Nest Counts in Field and Forest Habitats

EA030 Reflection

October 8, 2025

Introduction

In ecological research, we often want to understand whether observed patterns are **statistically unusual** or could have occurred by chance. For example, if we count more ant nests in fields than forests, is this a real ecological pattern, or could it simply be random variation?

Why Not Just Use a t-test?

Traditional parametric tests like the t-test rely on several assumptions:[1]

- **Normality:** Data should be approximately normally distributed
- **Independence:** Observations should be independent of each other
- **Equal variance:** Groups should have similar variability
- **Adequate sample size:** Generally need at least 20-30 observations per group

With ecological data, especially from field studies, we often have:

- Small sample sizes (expensive/time-consuming to collect)
- Skewed distributions (count data often aren't bell-shaped)
- Unequal variances between groups
- Unknown distributions

Enter Monte Carlo Simulations

Monte Carlo simulations provide a flexible, *non-parametric* approach[2] that works by:

1. Taking your actual observed data
2. Randomly shuffling it many times (typically 1,000-10,000 times)
3. Creating a **null distribution**: what patterns would we expect if there were no real difference between habitats?
4. Comparing your observed pattern to this null distribution

Key concept: If habitat doesn't matter, then randomly reassigning "Field" and "Forest" labels to our nest counts should produce similar patterns to what we observed. If our observed pattern is very unusual compared to random shuffling, that suggests habitat *does* matter!

[1] Parametric tests make specific assumptions about the mathematical form of the data distribution, usually assuming it follows a normal (bell-shaped) curve.

[2] Non-parametric methods don't assume a specific distribution shape and work by comparing observed data to patterns generated by randomization.

Learning Goals

By the end of this activity, learners will be able to:

1. Understand the conceptual foundation of Monte Carlo simulations for hypothesis testing
2. Explain why Monte Carlo methods are useful for ecological data
3. Compute summary statistics (mean and standard deviation) for count data
4. Perform Monte Carlo randomizations to generate null distributions
5. Calculate and interpret p-values from simulation results
6. Visualize simulation results using histograms with observed values highlighted
7. Distinguish between statistical significance and ecological importance

Statistical Background: The Null Hypothesis

In this analysis, our **null hypothesis (H_0)** is:[3]

There is no difference in ant nest density between Field and Forest habitats.

Our **alternative hypothesis (H_A)** is:

Ant nest density differs between Field and Forest habitats.

The Logic: If the null hypothesis is true (habitat doesn't matter), then the labels "Field" and "Forest" are arbitrary. We could randomly reassign them to our 10 observations and get similar results. By doing this thousands of times, we create a distribution of what differences we'd expect *by chance alone*.

P-value interpretation:[4] The proportion of randomizations that produce a difference as large or larger than our observed difference. A small p-value (typically < 0.05) suggests our observed pattern is unusual under random chance.

Method & Analysis Workflow

Step 1: Input Data

First, let's look at our data:[5]

```
# ---------------------------
# Step 1: Input ant nest counts
# ---------------------------
```

[3] The null hypothesis represents the "nothing interesting is happening" scenario that we're testing against.

[4] A p-value is NOT the probability that the null hypothesis is true! It's the probability of seeing data this extreme IF the null hypothesis were true.

[5] These data represent counts from 10 quadrats (sampling plots): 6 in Forest habitat and 4 in Field habitat.

```
nest_counts <- data.frame(
  Habitat = c("Forest", "Forest", "Forest", "Forest", "Forest",
              "Forest", "Field", "Field", "Field", "Field"),
  Nests = c(9, 6, 4, 6, 7, 10, 12, 9, 12, 10)
)

# Display the data
print(nest_counts)
```

```
##    Habitat Nests
## 1   Forest     9
## 2   Forest     6
## 3   Forest     4
## 4   Forest     6
## 5   Forest     7
## 6   Forest    10
## 7    Field    12
## 8    Field     9
## 9    Field    12
## 10   Field    10
```

Quick exploration questions:

- How many quadrats were sampled in each habitat?
- What's the range of nest counts in each habitat?
- Do the Field counts look consistently higher than Forest counts?

Step 2: Compute Observed Statistics

Now we'll calculate summary statistics for each habitat. This gives us our **observed pattern** to compare against randomization.

Hint: We'll use the `dplyr` package for data wrangling. Think about: What column contains the groups? What column contains the values to summarize?

YOUR TASK: Fill in the blanks below to compute mean and standard deviation by habitat:

```
library(dplyr)

observed_stats <- nest_counts %>%
  group_by(______) %>%           # What column has habitat types?
  summarise(
    N = length(______),          # Count observations
    Mean = mean(______),         # Calculate mean
    SD = sd(______)              # Calculate standard deviation
  )
```

Challenge: Before running the code, predict: Which habitat do you expect to have a higher mean? Why?

Let's view our results:

```
print(observed_stats)

## # A tibble: 2 x 4
##   Habitat     N  Mean    SD
##   <chr>   <int> <dbl> <dbl>
## 1 Field       4  10.8  1.5
## 2 Forest      6   7    2.19
```

Interpretation questions:

- What is the mean number of nests in Fields vs. Forests?
- Which habitat shows more variability (higher SD)?
- Does the difference seem large or small relative to the standard deviations?

Step 3: Calculate the Observed Difference

Our **test statistic**[6] will be the absolute difference in mean nest counts:

```
# Order data by habitat for easier viewing
nest_counts <- nest_counts[order(nest_counts$Habitat), ]

# Calculate observed difference
observed_diff <- abs(
  observed_stats$Mean[observed_stats$Habitat == "Forest"] -
  observed_stats$Mean[observed_stats$Habitat == "Field"]
)

cat("Observed difference in means:", observed_diff, "\n")

## Observed difference in means: 3.75
```

Think about it: Why do we use the absolute value? **Hint:** We care if there's a difference in *either* direction.

Step 4: Monte Carlo Randomization

Here's where the magic happens! We'll randomly shuffle habitat labels 1,000 times and calculate the difference each time.

The Algorithm:

1. Take the 10 nest counts (our actual data)
2. Randomly assign 6 to "Forest" and 4 to "Field" (matching our sample sizes)

[6] A test statistic is a single number that summarizes the pattern we're interested in. Here, it's the absolute difference between habitat means.

3. Calculate the difference in means
4. Repeat 1,000 times
5. This creates our **null distribution**

Hint: Think of it like shuffling a deck of cards 1,000 times. Each shuffle represents one way the data *could* have looked if habitat truly didn't matter.

YOUR TASK: Study this code and add comments explaining what each section does:

```
nsimul <- 1000
dif_sim <- numeric(nsimul)      # Create empty vector to store results

set.seed(123)  # Makes results reproducible

for(i in 1:nsimul){
  # What does sample() do here?
  habitat_sim <- sample(nest_counts$Habitat)

  # What data structure are we creating?
  nest_counts_sim <- data.frame(
    Habitat = habitat_sim,
    Nests = nest_counts$Nests
  )

  # What statistic are we calculating?
  dif_sim[i] <- abs(
    mean(nest_counts_sim$Nests[nest_counts_sim$Habitat == "Field"]) -
    mean(nest_counts_sim$Nests[nest_counts_sim$Habitat == "Forest"])
  )
}
```

Check your understanding:

- How many randomizations did we perform?
- What does `set.seed(123)` do? Why is it useful?
- What values does `dif_sim` contain after the loop?

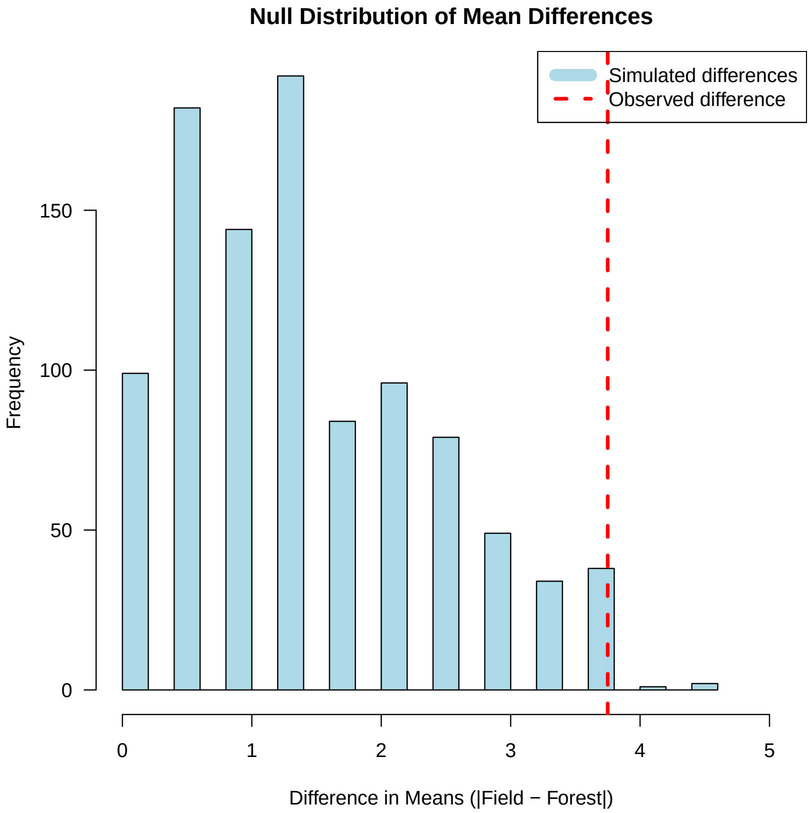
Step 5: Visualize the Null Distribution

Visualization helps us understand where our observed difference falls relative to chance expectations.

```r
par(las=1, mar=c(4, 4, 3, 1))
hist(dif_sim,
     main="Null Distribution of Mean Differences",
     xlab="Difference in Means (|Field - Forest|)",
     ylab="Frequency",
     col="lightblue",
     breaks=30,
     xlim=c(0, max(c(dif_sim, observed_diff)) + 0.5))

# Add observed difference as red line
abline(v=observed_diff, col="red", lwd=3, lty=2)

# Add legend
legend("topright",
       legend=c("Simulated differences", "Observed difference"),
       col=c("lightblue", "red"),
       lwd=c(10, 3),
       lty=c(1, 2))
```



Interpretation guide:

- The histogram shows what differences we'd expect by chance
- The red line shows our actual observed difference
- If the red line is far in the tail (far right), our observation is un-

usual

- If the red line is near the center, our observation is typical of random chance

Questions:

- Where does the observed difference fall relative to the null distribution?
- What does this suggest about whether habitat affects ant nest density?
- What shape is the null distribution? Why?

Step 6: Calculate the P-value

The p-value tells us: "What proportion of random shuffles produced a difference as large or larger than what we observed?"

Hint: Think about it: If we ran 1000 simulations and 30 of them had differences larger than our observed difference, what proportion is that?

Hint: What vector contains all the simulated differences? What value are we comparing them to? How many simulations did we run?

Challenge: Before calculating the p-value, try this exploratory step: Type `dif_sim >= observed_diff` into your R console. What does this produce? It returns a vector of TRUE/FALSE values! Then try `sum(dif_sim >= observed_diff)`. What happens when you sum logical values?[7]

YOUR TASK: Fill in the blanks to calculate the p-value:

```
# Count simulations >= observed difference
p_value <- sum(______ >= ______) / ______

cat("P-value:", p_value, "\n")
cat("Interpretation: ", p_value * 100, "% of random shuffles produced\n")
cat("a difference as large or larger than observed.\n")
```

Hint: What vector contains all the simulated differences? What value are we comparing them to? How many simulations did we run?

```
## P-value: 0.041
## Interpretation:  4.1 % of random shuffles produced
## a difference as large or larger than observed.
```

Statistical interpretation framework:[8]

- $p < 0.001$: Very strong evidence against null hypothesis

[7] In R, TRUE is treated as 1 and FALSE as 0 when summed. So `sum(dif_sim >= observed_diff)` counts how many simulations produced differences greater than or equal to the observed difference. Try viewing the first 20 values: `head(dif_sim >= observed_diff, 20)`

[8] These are conventional thresholds, but the exact cutoff is somewhat arbitrary. Consider effect size and ecological importance too!

- $p < 0.01$: Strong evidence against null hypothesis
- $p < 0.05$: Moderate evidence against null hypothesis
- $p < 0.10$: Weak evidence against null hypothesis
- $p \geq 0.10$: Insufficient evidence to reject null hypothesis

YOUR TASK: Based on your p-value, what do you conclude about ant nest density in these habitats?

Step 7: Additional Explorations

Challenge: Try these extensions to deepen your understanding:

A. Increase simulation number:

```
# Try nsimul <- 10000
# Does your p-value change? By how much?
```

B. Examine the most extreme randomizations:

```
# What's the largest difference we saw in 1000 randomizations?
max_random_diff <- max(dif_sim)
cat("Maximum random difference:", max_random_diff, "\n")

## Maximum random difference: 4.583333

cat("Observed difference:", observed_diff, "\n")

## Observed difference: 3.75
```

C. Calculate confidence intervals:

```
# 95% of random differences fall between what values?
ci_lower <- quantile(dif_sim, 0.025)
ci_upper <- quantile(dif_sim, 0.975)
cat("95% of random differences fall between", ci_lower, "and", ci_upper, "\n")

## 95% of random differences fall between 0 and 3.75
```

Biological Interpretation

Statistical significance is just the first step. Now we need to think ecologically:

- **Why might ant nests differ between habitats?**
  - Resource availability (food sources)
  - Microclimate differences (temperature, moisture)
  - Soil characteristics (easier to excavate?)
  - Predation pressure

  - Competition with other species

- **Is the difference biologically meaningful?**
  - A statistically significant difference might be too small to matter ecologically
  - Consider the natural variation in nest counts
  - Think about what size difference would affect ecosystem function
- **What are the limitations?**
  - Small sample size (only 10 quadrats total)
  - Potential confounding factors (were quadrats randomly placed?)
  - Temporal variation (what season was this?)
  - Species identity (different ant species might have different patterns)

Reflection Questions

1. Explain in your own words: What does a Monte Carlo simulation do, and why might it be preferred over traditional parametric tests (like t-tests) for ecological datasets with small sample sizes?[9]
2. What is the null hypothesis in this study? Based on your p-value and the visualization of the null distribution, what do you conclude about ant nest density in Field versus Forest habitats? Be sure to address both statistical significance and potential biological/ecological importance.[10]

Additional Resources

- **R Documentation:** `?sample`, `?hist`, `?dplyr`
- **Statistical concepts:** Research "permutation tests" and "resampling methods"
- **Ecological context:** Read about ant ecology and habitat preferences
- **Advanced:** Explore the `coin` package in R for permutation tests

[9] Hint: Think about the assumptions each method requires and how randomization creates a null distribution.

[10] Consider: Is a statistically significant result always ecologically meaningful?