PLS 120: Applied Statistics in Agricultural Sciences

Sampling and Estimation

Week 5 Tutorial Guide

Mohammadreza Narimani
Department of Biological and Agricultural Engineering
University of California, Davis

mnarimani@ucdavis.edu

October 2025

Contents

Important Links

> **Essential Course Resources**
>
> ## Course Website
>
> **All course materials available at:**
>
> Course Website Link
>
> ## Interactive Binder Environment
>
> **Access Week 5 lab materials:**
>
> Week 5 Binder Link

Welcome to Week 5: Sampling and Estimation

This week, we explore **sampling distributions and statistical estimation** - critical skills for designing agricultural studies and interpreting research results. You'll learn to calculate sample sizes, construct confidence intervals, and understand estimation precision!

Reproducible Research with set.seed()

Understanding Reproducible Research

The set.seed() function is essential for creating reproducible research, especially when involving random number generation and sampling.

3.1.1 Why Reproducibility Matters

> **Benefits of Reproducible Research:**
> - Allows others to verify your results
> - Enables debugging of statistical analyses
> - Ensures consistent results across runs
> - Required for scientific publication
>
> **Usage:** `set.seed(value)`
> **Example:** `set.seed(123)`
> The specific number doesn't matter - use memorable values like 123, 42, or 2025

Z-Score Standardization

Understanding Z-Scores

Z-scores help us **standardize** data by converting values to standard deviations from the mean, making it easier to compare different datasets and calculate probabilities.

4.1.1 Z-Score Formula

Z-Score Formula:

$$z = \frac{x - \mu}{\sigma}$$

Where:
x = individual value
μ = population mean
σ = population standard deviation

R Implementation:
```
z_scores <- (data - mean(data)) / sd(data)
z_scores <- scale(data) # Alternative method
```

Normal Distribution Functions

4.2.1 Essential Normal Functions

pnorm() - Cumulative Probability:
```
pnorm(q, mean = 0, sd = 1)
```
Returns: $P(X \leq q)$ for normal distribution

qnorm() - Quantiles:
```
qnorm(p, mean = 0, sd = 1)
```
Returns: Value x such that $P(X \leq x) = p$

rnorm() - Random Generation:
```
rnorm(n, mean = 0, sd = 1)
```
Generates n random values from normal distribution

Confidence Intervals

Understanding Confidence Intervals

Confidence intervals provide an estimated range that is likely to include the true value of an unknown population parameter with a certain level of confidence.

5.1.1   Confidence Interval Formula

Confidence Interval for Means:

$$CI = \overline{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

Where:
$\overline{x}$ = sample mean
z = critical z-value
s = sample standard deviation
n = sample size

Common Z-Values:
90% confidence: z = 1.645
95% confidence: z = 1.96
99% confidence: z = 2.576

5.1.2   Margin of Error

Margin of Error Formula:

$$ME = z_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

R Implementation:
```
alpha <- 0.05
z_score <- qnorm(1 - alpha / 2)
margin_error <- z_score * (sample_sd / sqrt(sample_size))
lower_bound <- sample_mean - margin_error
upper_bound <- sample_mean + margin_error
```

Sample Size Calculations

Determining Required Sample Size

When designing experiments, calculating appropriate sample size is critical to minimize error and ensure sufficient statistical power.

6.1.1 Sample Size Formula for Proportions

Sample Size Formula:

$$n = \frac{z^2 \times p \times (1 - p)}{d^2}$$

Where:
n = required sample size
z = z-score for desired confidence level
p = expected proportion
d = desired margin of error

R Implementation:
```
z_score <- qnorm(1 - alpha / 2)
sample_size <- z_score^2 * p * (1 - p) / d^2
ceiling(sample_size) # Round up to whole number
```

Factors Affecting Sample Size

6.2.1 Trade-offs in Sample Size Planning

Key Relationships:
- Higher confidence level → Larger sample size needed
- Lower margin of error → Larger sample size needed
- Higher expected proportion (up to 0.5) → Larger sample size needed
- Cost vs. Precision: Balance practical constraints with statistical requirements

Example Scenarios:
90% confidence, 5% margin: n ≈ 271
95% confidence, 5% margin: n ≈ 385
99% confidence, 1% margin: n ≈ 16,641

Assignment 5 Overview

Assignment Structure (20 points total)

1. **Question 1: Basic Sample Size Calculation (6 points)**

   - Calculate sample size for 95% confidence and 5% margin of error
   - Explain margin of error concept
   - Interpret confidence intervals correctly

2. **Question 2: Effect of Prevalence Rate (5 points)**

   - Calculate sample sizes for different prevalence rates (0.5X, 2X, 3X)

   - Analyze how prevalence affects sample size requirements

3. **Question 3: Effect of Margin of Error (5 points)**

   - Calculate sample sizes for 2.5% and 7.5% margins of error
   - Understand precision trade-offs in sample size planning

4. **Question 4: Effect of Confidence Level (4 points)**

   - Compare sample sizes for 90% and 99% confidence levels
   - Analyze how confidence level affects sample size

Agricultural Applications

Real-World Applications:

- **Seed Germination Studies** - Estimate germination rates with confidence intervals for planting density planning

- **Crop Yield Estimation** - Sample fields to predict total harvest with known precision

- **Pest Occurrence Surveys** - Calculate sample sizes needed to detect pest outbreaks reliably

- **Quality Control Testing** - Design sampling plans to ensure product quality standards

- **Weather Pattern Analysis** - Use historical data to create prediction intervals for rainfall

- **Agricultural Insurance** - Quantify risks using sampling and probability distributions

Key Concepts Summary

Estimation Fundamentals

Confidence Interval Interpretation:
- A 95% confidence interval means: if we repeated our study 100 times, about 95 of the intervals would contain the true population parameter
- Your specific interval either contains μ or it doesn't - the confidence is about the method
- Wider intervals = more confidence, narrower intervals = more precision

Sample Size Principles:
- Larger samples create more precise estimates
- There's always a trade-off between precision and cost
- Plan sample size before data collection, not after

Statistical Inference

Population vs Sample:
Population = entire group of interest
Sample = subset used to make inferences about population

Point Estimate vs Interval Estimate:
Point estimate = single value (e.g., sample mean)
Interval estimate = range of plausible values (confidence interval)

Sampling Distribution:
Distribution of sample statistics across many samples from the same population

Getting Started

1. Launch Week 5 Binder environment

2. Navigate to `class_activity` folder

3. Open `Week5_Sampling_Estimation.ipynb`

4. Work through interactive exercises

5. Complete Assignment 5 in `assignment` folder

Learning Objectives

By the end of this week, you will be able to:

- Calculate required sample sizes for different precision levels

- Construct confidence intervals for means and proportions

- Understand the relationship between confidence level and margin of error

- Use normal distribution functions for statistical inference

- Apply z-score standardization to compare different datasets

- Design studies with appropriate statistical power

- Make informed decisions about resource allocation in research

Tips for Success

Best Practices:

- Use `ceiling()` to round up sample sizes to whole numbers

- Remember key z-values: 1.96 for 95%, 1.645 for 90%, 2.576 for 99%

- Understand trade-offs: higher confidence = larger sample size

- Verify calculations by checking formulas and parameter values

- Think practically about cost and feasibility of sample sizes

- Always interpret confidence intervals in context

Need Help?

Mohammadreza Narimani
Email: mnarimani@ucdavis.edu
Department of Biological and Agricultural Engineering, UC Davis
Office Hours: Thursdays 10 AM - 12 PM (Zoom)
Zoom Link: Join Office Hours