PLS 120: Applied Statistics in Agricultural Sciences

Final Project Workspace

Optional Alternative to Final Exam

Complete Statistical Analysis Guide

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

mnarimani@ucdavis.edu

November 2024

Contents

Important Links

Essential Project Resources

Course Website

All course materials available at:

Course Website Link

Final Project Binder Environment

Access your project workspace:

Final Project Binder Link

Welcome to Your Final Project

This comprehensive final project is an **optional alternative to the final exam**. Students may choose to complete either the traditional final exam or this hands-on project. The project brings together all statistical concepts learned throughout PLS 120, allowing you to conduct a complete statistical analysis using your own agricultural dataset while demonstrating mastery of R programming, statistical testing, and scientific communication.

Project Overview

Project Option

Choose Your Assessment Method:
Option 1: Traditional Final Exam (150 points)
Option 2: Final Project Workspace (150 points)

This project is completely optional! Students who prefer hands-on analysis over traditional testing may choose this comprehensive project as their final assessment.

Project Objectives

Demonstrate Mastery Of:

- **Data Management** - Import, clean, and organize agricultural datasets
- **Exploratory Analysis** - Summarize and visualize data patterns
- **Statistical Testing** - Apply appropriate tests for research questions
- **Interpretation** - Draw meaningful conclusions from results
- **Communication** - Present findings in professional format
- **R Programming** - Use functions, packages, and best practices

Pre-Installed Analysis Tools

Your Binder environment includes comprehensive R packages for professional statistical analysis:

Data Manipulation & Import:
`dplyr, tidyr, readr, readxl, data.table`

Statistical Analysis:
`car, agricolae, multcomp, emmeans, broom, lmtest, nortest`

Visualization:
`ggplot2, plotly, corrplot, pheatmap`

Report Generation:
`knitr, rmarkdown, DT`

Advanced Modeling:
`MASS, lawstat` (plus ability to install additional packages)

Getting Started

Step 1: Launch Your Environment

1. Click the **Final Project Binder Link** above
2. Wait 2-5 minutes for environment setup
3. Navigate to `Final_Project_Workspace.ipynb`
4. Begin your analysis following the structured workflow

Step 2: Upload Your Data

Data Upload Process:

1. Use the **Upload button** in left panel (folder icon area)
2. Select your data file (CSV, Excel, or other formats)
3. File appears in left panel when ready
4. Load data using appropriate R functions:
   - CSV: `data <- read.csv("filename.csv")`
   - Excel: `data <- read_excel("filename.xlsx")`

Supported Data Formats:
CSV, Excel (.xlsx, .xls), Tab-delimited, and other common formats

Project Workflow

Phase 1: Data Preparation & Exploration

Essential Steps:

1. **Load Required Libraries**
   - Load `ggplot2, dplyr, readr, knitr`
   - Add specialized packages as needed
2. **Import and Examine Data**
   - Use `str(data)` to check structure
   - Apply `head(data)` and `summary(data)`
   - Check for missing values with `is.na()`
3. **Data Cleaning**
   - Convert categorical variables to factors
   - Handle missing values appropriately
   - Check for and address outliers
4. **Exploratory Visualization**
   - Create histograms for continuous variables
   - Generate boxplots for group comparisons
   - Build scatter plots for relationships

Phase 2: Descriptive Statistics

Calculate Summary Statistics:

- **Central Tendency:** mean, median, mode
- **Variability:** variance, standard deviation, coefficient of variation
- **Distribution:** quantiles, skewness, normality tests
- **Group Comparisons:** statistics by treatment/category

Key R Functions:
`mean()`, `median()`, `var()`, `sd()`, `quantile()`, `summary()`
`group_by()` `%>%` `summarise()` for grouped statistics

Phase 3: Statistical Testing

Choose Appropriate Tests Based on Data:

- **Two-sample t-tests:** Compare means between groups
- **ANOVA:** Compare means across multiple groups
- **Chi-square tests:** Test associations in categorical data
- **Correlation analysis:** Examine relationships between variables
- **Regression analysis:** Model relationships and predictions

Hypothesis Testing Framework:

1. State null and alternative hypotheses
2. Check test assumptions
3. Perform appropriate statistical test
4. Interpret p-values and confidence intervals
5. Draw conclusions in agricultural context

Phase 4: Advanced Analysis

Sophisticated Statistical Methods:

- **Multiple Regression:** Model complex relationships
- **ANOVA with Post-hoc Tests:** Detailed group comparisons
- **Non-parametric Tests:** When assumptions aren't met
- **Effect Size Calculations:** Practical significance assessment
- **Power Analysis:** Sample size and study design evaluation

Model Validation:
Check residuals, assess assumptions, validate predictions

Phase 5: Results Visualization

Publication-Quality Graphics:

- **Treatment Comparisons:** Boxplots with significance indicators
- **Relationships:** Scatter plots with regression lines
- **Distributions:** Histograms and density plots
- **Correlations:** Correlation matrices and heatmaps
- **Model Results:** Coefficient plots and diagnostic plots

Visualization Best Practices:
Clear titles, axis labels, legends, and appropriate color schemes

Project Deliverables

Required Components

Your Final Project Must Include:

1. **Student Information** (Complete in raw text cells)
2. **Data Description** (Source, variables, sample size)
3. **Research Questions** (Clear, testable hypotheses)
4. **Statistical Analysis** (Appropriate tests with justification)
5. **Results Visualization** (Professional plots and tables)
6. **Interpretation** (Agricultural significance and implications)
7. **Limitations** (Study constraints and future directions)
8. **Conclusions** (Evidence-based recommendations)

Submission Requirements

Submit TWO Files to Canvas:

1. **HTML Report** (File → Save and Export Notebook As → HTML)
   - Complete formatted report with all outputs
   - Professional presentation of results
   - All plots and tables included
2. **Jupyter Notebook** (File → Download → .ipynb)
   - Complete code for reproducibility
   - Backup of your analysis
   - All cells executed with outputs

File Naming Convention:

`LastName_FirstName_PLS120_FinalProject.html`
`LastName_FirstName_PLS120_FinalProject.ipynb`

Evaluation Criteria

Grading Rubric

Technical Competency (40%):

- Appropriate statistical test selection
- Correct R code implementation
- Proper assumption checking
- Accurate calculations and outputs

Data Analysis Quality (30%):

- Thorough exploratory analysis
- Meaningful visualizations
- Comprehensive statistical testing
- Appropriate model validation

Interpretation & Communication (20%):

- Clear agricultural context
- Correct statistical interpretation
- Practical significance discussion
- Professional presentation

Scientific Rigor (10%):

- Proper hypothesis formation
- Acknowledgment of limitations
- Evidence-based conclusions
- Reproducible methodology

Agricultural Applications

Potential Research Areas:

- **Crop Production** - Yield comparisons, variety trials, treatment effects
- **Soil Science** - Nutrient analysis, pH effects, organic matter studies
- **Plant Breeding** - Genetic line comparisons, trait correlations
- **Pest Management** - Treatment efficacy, resistance studies
- **Environmental Impact** - Sustainability metrics, carbon footprint
- **Economic Analysis** - Cost-benefit analysis, profitability studies
- **Quality Assessment** - Product quality, post-harvest analysis
- **Precision Agriculture** - Spatial analysis, technology adoption

Statistical Methods Reference

Common Agricultural Statistics

Descriptive Statistics:

```
mean(data$variable)
sd(data$variable)
quantile(data$variable)
```

Group Comparisons:

```
t.test(group1, group2)
aov(response ~ treatment, data)
TukeyHSD(aov_model)
```

Relationships:

```
cor(x, y)
lm(y ~ x, data)
plot(x, y)
```

Visualization:

```
ggplot(data, aes(x, y)) + geom_point()
ggplot(data, aes(x, y)) + geom_boxplot()
```

Tips for Success

Best Practices:

- **Start Early** - Allow time for data exploration and analysis
- **Check Assumptions** - Verify test requirements before proceeding
- **Document Everything** - Add comments explaining your reasoning
- **Visualize First** - Always plot data before statistical testing
- **Interpret Practically** - Consider agricultural significance
- **Save Frequently** - Download work regularly to avoid loss
- **Test Code** - Run cells in order and check outputs
- **Seek Help Early** - Use office hours for guidance

Common Pitfalls to Avoid

Avoid These Mistakes:

- Using inappropriate statistical tests for data type
- Ignoring test assumptions (normality, independence)
- Over-interpreting non-significant results
- Confusing statistical and practical significance
- Poor data visualization choices
- Inadequate description of methods and data
- Missing or incorrect interpretation of results
- Failing to save work before closing Binder

Troubleshooting

Technical Issues

Common Problems and Solutions:

- **Binder won't load:** Refresh page, clear browser cache
- **Package not found:** Use `install.packages("package_name")`
- **Code error:** Check syntax, run cells in order
- **Data won't load:** Verify file name and format
- **Lost work:** Always download files before closing
- **Plot not showing:** Check ggplot syntax and data

Learning Objectives

By completing this final project, you will demonstrate ability to:

- Import and manage agricultural datasets in R
- Conduct comprehensive exploratory data analysis
- Select and apply appropriate statistical tests
- Create professional data visualizations
- Interpret statistical results in agricultural context
- Communicate findings effectively to stakeholders
- Use R programming for reproducible research
- Make evidence-based recommendations for agricultural practices

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

Additional Resources:
- Course website for all materials
- Canvas discussion forum for peer help
- R documentation: `?function_name`
- Online R tutorials and Stack Overflow

Last updated: November 2024 | PLS 120 - Applied Statistics in Agriculture | UC Davis
Final Project: Complete Statistical Analysis