



Environmental Analysis Program

Pomona College
Biogeochemistry Laboratory

SOP-L806

Statistical Analysis Templates (R)

Standard Operating Procedure

SOP Number	SOP-L806
Revision	0.1
Author(s)	Marc Los Huertos
Date	2026-06-26
Organization	EA Program, Pomona College

FOR INTERNAL USE ONLY

This document is the property of the Environmental Analysis Program at Pomona College.
Unauthorized distribution is not permitted.

Table of contents

1 Purpose	2
2 Scope	2
3 Responsibilities	2
4 Definitions	2
5 Safety and Hazards	3
6 Equipment and Materials	3
6.1 Software	3
6.2 Core R Packages	3
7 Procedure	3
7.1 Analysis Workflow Overview	3
7.2 Quarto Document Header	4
7.3 Exploratory Data Analysis Template	4
7.4 Two-Sample Comparisons	5
7.5 One-Way ANOVA Template	6
7.6 Linear Regression Template	6
7.7 Mixed-Effects Model Template	7
7.8 Multivariate / Community Analysis Template	8
7.9 Figure Formatting Standards	8
7.10 Reporting Statistical Results	9
8 Quality Control and Acceptance Criteria	9
9 Waste Disposal	9
10 References	9
11 Training Record	10
12 Appendix A: Decision Guide for Choosing a Statistical Test	10
13 Appendix B: Common Data Transformations	10

1. Purpose

This SOP provides standardized R analysis templates for common statistical procedures used in EA Program research. It establishes reproducible workflows for exploratory data analysis, hypothesis testing, regression modeling, and reporting, ensuring that all analyses are transparent, well-documented, and consistent across projects (Wickham et al., 2023; Zuur et al., 2010).

2. Scope

- **Applies to:** All quantitative statistical analyses performed for EA Program research, coursework, and publications
- **Does not apply to:** Bioinformatics sequence analysis (see [SOP-L705](#)); GIS-specific spatial statistics (see [SOP-L804](#)); instrument calibration curve fitting (see method-specific SOPs)
- **Location:** RStudio Server (rstudio.pomona.edu)

3. Responsibilities

Role	Responsibility
Lab PI	Approve analytical approach; review statistical methods in manuscripts
Analyst / Student	Select appropriate methods; follow templates; document all decisions
Data Reviewer	Verify that assumptions are checked and reported
Statistical Advisor	Consult on complex designs (mixed models, multivariate methods)

4. Definitions

- **ANOVA (Analysis of Variance):** A statistical test that partitions total variance in a response variable into components attributable to one or more categorical factors; tests whether group means differ significantly; assumes normality and homogeneity of variance.
- **Confidence interval:** A range of values computed from sample data that, under repeated sampling, contains the true population parameter with a stated probability (e.g., 95%); wider intervals reflect greater uncertainty.
- **Effect size:** A standardized measure of the magnitude of a difference or relationship, independent of sample size; examples include Cohen's d (mean differences), r^2 (explained variance), and Hedge's g; should always accompany p-values.
- **Mixed model:** A statistical model containing both fixed effects (factors whose levels are deliberately chosen) and random effects (factors whose levels represent a random sample from a larger population); appropriate when data have hierarchical or repeated-measures structure.
- **p-value:** The probability of observing a test statistic as extreme as the one obtained (or more extreme) if the null hypothesis is true; does not measure the probability that the null hypothesis is true; must be interpreted with effect size.

- **Power:** The probability of correctly rejecting the null hypothesis when it is false; increases with sample size, effect size, and significance level; should be considered in study design to ensure adequate sample sizes.
- **Residual:** The difference between an observed value and the value predicted by a statistical model; examination of residuals is the primary diagnostic for checking model assumptions.

5. Safety and Hazards

Statistical analysis is a computational activity with no physical hazards.

6. Equipment and Materials

6.1. Software

Item	Access	Notes
R (version 4.x)	RStudio Server	Primary analysis platform
RStudio IDE	rstudio.pomona.edu	Integrated development environment
Quarto	Pre-installed	Report generation
Git / GitHub	github.com/marclos	Version control

6.2. Core R Packages

Package	Purpose
tidyverse	Data wrangling and visualization
broom	Tidy model output
car	Diagnostic tests (Levene, VIF)
lme4	Mixed-effects models
emmeans	Estimated marginal means and contrasts
vegan	Community ecology and ordination
MuMIn	Model selection and multi-model inference
performance	Model diagnostics and assumption checks
report	Automated statistical reporting
ggplot2	Publication-quality figures
knitr / kableExtra	Formatted tables

7. Procedure

7.1. Analysis Workflow Overview

Every statistical analysis should follow this sequence:

1. State the research question and hypotheses
2. Import and validate data (see [SOP-L803](#))
3. Conduct exploratory data analysis (EDA)
4. Check assumptions for the intended statistical test
5. Fit the model
6. Evaluate model diagnostics
7. Interpret results and generate figures/tables
8. Write up methods and results in a Quarto document
9. Commit the complete analysis to Git

7.2. Quarto Document Header

Begin every analysis with a standard Quarto header:

```
---  
title: "Analysis: [Project Name]"  
author: "[Your Name]"  
date: today  
format:  
  html:  
    toc: true  
    code-fold: true  
    embed-resources: true  
bibliography: references.bib  
---
```

7.3. Exploratory Data Analysis Template

```
library(tidyverse)  
  
# --- 1. Load and inspect data ---  
dat <- read_csv("data/processed/project_data.csv")  
glimpse(dat)  
  
# --- 2. Summary statistics ---  
dat %>%  
  group_by(treatment) %>%  
  summarize(  
    n = n(),  
    mean = mean(response, na.rm = TRUE),  
    sd = sd(response, na.rm = TRUE),  
    median = median(response, na.rm = TRUE),  
    min = min(response, na.rm = TRUE),  
    max = max(response, na.rm = TRUE)  
  )
```

```
# --- 3. Distributional plots ---
ggplot(dat, aes(x = response)) +
  geom_histogram(bins = 20, fill = "steelblue", color = "white") +
  facet_wrap(~ treatment) +
  labs(x = "Response Variable (units)", y = "Count") +
  theme_minimal()

ggplot(dat, aes(x = treatment, y = response)) +
  geom_boxplot(fill = "steelblue", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.4) +
  labs(x = "Treatment", y = "Response Variable (units)") +
  theme_minimal()

# --- 4. Missing data assessment ---
dat %>%
  summarize(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "variable",
               values_to = "n_missing") %>%
  filter(n_missing > 0)
```

7.4. Two-Sample Comparisons

```
# --- Check normality within groups ---
dat %>%
  group_by(treatment) %>%
  summarize(
    shapiro_p = shapiro.test(response)$p.value
  )

# --- Check equal variances ---
car::leveneTest(response ~ treatment, data = dat)

# --- Parametric: two-sample t-test ---
t_result <- t.test(response ~ treatment, data = dat, var.equal = FALSE)
t_result

# --- Nonparametric alternative: Wilcoxon rank-sum ---
wilcox.test(response ~ treatment, data = dat)

# --- Effect size ---
library(effectsiz)
cohens_d(response ~ treatment, data = dat)
```

7.5. One-Way ANOVA Template

```
# --- Fit ANOVA ---
aov_model <- aov(response ~ treatment, data = dat)

# --- Check assumptions ---
# Normality of residuals
shapiro.test(residuals(aov_model))
qqnorm(residuals(aov_model)); qqline(residuals(aov_model))

# Homogeneity of variance
car::leveneTest(response ~ treatment, data = dat)

# Residual vs. fitted plot
plot(aov_model, which = 1)

# --- ANOVA table ---
summary(aov_model)

# --- Post-hoc pairwise comparisons ---
library(emmeans)
emm <- emmeans(aov_model, ~ treatment)
pairs(emm, adjust = "tukey")

# --- Nonparametric alternative: Kruskal-Wallis ---
kruskal.test(response ~ treatment, data = dat)
```

Tip

Always check assumptions before interpreting p-values. If residuals are non-normal or variances are unequal, consider a transformation (log, square root) or switch to a nonparametric test. Document the reasoning in your analysis script (Zuur et al., 2010).

7.6. Linear Regression Template

```
# --- Fit linear model ---
lm_model <- lm(response ~ predictor1 + predictor2, data = dat)

# --- Model summary ---
summary(lm_model)

# --- Tidy output ---
library(broom)
tidy(lm_model, conf.int = TRUE)
```

```
glance(lm_model)

# --- Diagnostics ---
par(mfrow = c(2, 2))
plot(lm_model)
par(mfrow = c(1, 1))

# Check multicollinearity
car::vif(lm_model)

# --- Model performance ---
library(performance)
check_model(lm_model)

# --- Visualization ---
ggplot(dat, aes(x = predictor1, y = response)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "steelblue") +
  labs(x = "Predictor (units)", y = "Response (units)",
       caption = sprintf("R-squared = %.3f, p = %.4f",
                          summary(lm_model)$r.squared,
                          glance(lm_model)$p.value)) +
  theme_minimal()
```

7.7. Mixed-Effects Model Template

For data with nested or repeated-measures structure:

```
library(lme4)
library(lmerTest) # For p-values via Satterthwaite

# --- Fit mixed model ---
# Fixed effect: treatment
# Random effect: site (multiple samples per site)
mixed_model <- lmer(response ~ treatment + (1 | site_id), data = dat)

# --- Summary with p-values ---
summary(mixed_model)

# --- Marginal and conditional R-squared ---
MuMIn::r.squaredGLMM(mixed_model)

# --- Post-hoc comparisons ---
library(emmeans)
emm <- emmeans(mixed_model, ~ treatment)
pairs(emm, adjust = "tukey")
```

```
# --- Residual diagnostics ---  
plot(mixed_model)  
qqnorm(resid(mixed_model)); qqline(resid(mixed_model))
```

7.8. Multivariate / Community Analysis Template

```
library(vegan)  
  
# --- Prepare community matrix ---  
# Rows = sites, columns = species/variables  
comm_matrix <- dat %>%  
  select(site_id, species, abundance) %>%  
  pivot_wider(names_from = species, values_from = abundance,  
              values_fill = 0) %>%  
  column_to_rownames("site_id")  
  
# --- Diversity indices ---  
diversity_h <- diversity(comm_matrix, index = "shannon")  
richness <- specnumber(comm_matrix)  
  
# --- Ordination (NMDS) ---  
nmfs <- metaMDS(comm_matrix, distance = "bray", k = 2,  
                trymax = 100)  
stressplot(nmfs)  
  
# Plot  
ordiplot(nmfs, type = "n")  
points(nmfs, col = as.factor(env_data$treatment), pch = 16)  
legend("topright", legend = levels(as.factor(env_data$treatment)),  
      col = 1:nlevels(as.factor(env_data$treatment)), pch = 16)  
  
# --- PERMANOVA ---  
adonis2(comm_matrix ~ treatment, data = env_data,  
        method = "bray", permutations = 999)
```

7.9. Figure Formatting Standards

All figures for reports and publications must follow these conventions:

1. Use `theme_minimal()` or a consistent custom theme
2. Label axes with variable names and units
3. Use colorblind-friendly palettes (e.g., `viridis`, `RColorBrewer` “Set2”)
4. Include figure captions using the `#| fig-cap:` chunk option
5. Save publication figures at 300 DPI as both PDF and PNG:

```
ggsave("analysis/output/fig_treatment_response.pdf",
       width = 6, height = 4, units = "in")
ggsave("analysis/output/fig_treatment_response.png",
       width = 6, height = 4, units = "in", dpi = 300)
```

7.10. Reporting Statistical Results

When writing results in a Quarto document, follow APA conventions:

- Report test statistics, degrees of freedom, and p-values: “Treatment had a significant effect on soil pH ($F(2, 27) = 4.53$, $p = 0.020$)”
- Report effect sizes alongside p-values
- Include confidence intervals for parameter estimates
- Round to a consistent number of decimal places (typically 2 for test statistics, 3 for p-values)
- Use `broom::tidy()` and `knitr::kable()` for formatted tables

```
library(broom)
library(knitr)

tidy(aov_model) %>%
  kable(digits = 3, caption = "ANOVA results for soil pH by treatment")
```

8. Quality Control and Acceptance Criteria

- Every analysis must include explicit assumption checks (normality, homoscedasticity, independence) documented in the script
- All models must include diagnostic plots
- Results must be reproducible: running the script from start to finish on the same data must produce identical output
- Analyses must be committed to Git with clear commit messages
- The PI or a designated reviewer must approve the analytical approach before results are included in publications

9. Waste Disposal

Not applicable (computational procedure).

10. References

Wickham, H., cCetinkaya-Rundel, M., & Golemund, G. (2023). *R for data science* (2nd ed.). O'Reilly Media. <https://r4ds.hadley.nz/>

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14. <https://doi.org/10.1111/j.2041-210X.2009.00001.x>

11. Training Record

Training is documented in the EA Program training log. Before conducting statistical analyses independently, personnel must:

1. Read this SOP in full
2. Complete the RStudio and GitHub orientation ([SOP-L801](#))
3. Complete the introductory R statistics module (see [RTricks](#))
4. Demonstrate a complete analysis workflow (EDA through reporting) on a practice dataset under supervision
5. Be signed off by the PI

12. Appendix A: Decision Guide for Choosing a Statistical Test

Research Question	Data Type	Groups	Recommended Test
Compare 2 group means	Continuous, normal	2	Two-sample t-test
Compare 2 group medians	Non-normal	2	Wilcoxon rank-sum
Compare 3+ group means	Continuous, normal	3+	One-way ANOVA
Compare 3+ group medians	Non-normal	3+	Kruskal-Wallis
Relationship between 2 variables	Both continuous	N/A	Linear regression / correlation
Multiple predictors	Continuous response	N/A	Multiple regression
Nested / repeated measures	Continuous	Any	Mixed-effects model
Community composition	Abundance matrix	Any	NMDS + PERMANOVA
Categorical association	Counts	Any	Chi-squared / Fisher's exact

13. Appendix B: Common Data Transformations

Transformation	R Code	When to Use
Log (natural)	<code>log(x)</code>	Right-skewed data; multiplicative relationships
Log (base 10)	<code>log10(x)</code>	Concentration data spanning orders of magnitude
Square root	<code>sqrt(x)</code>	Count data; moderate skew
Arcsine square root	<code>asin(sqrt(x))</code>	Proportions (0 to 1)

Transformation	R Code	When to Use
Box-Cox	<code>MASS::boxcox(model)</code>	Let the data determine the best transformation

 Caution

Always report whether data were transformed and provide results on both the transformed and original scales when possible. Transformations change the hypothesis being tested (e.g., log-transformed ANOVA compares geometric means, not arithmetic means).