Biostatistics — Case Study 1

Effect of Drug Dose × Time on Protein Expression

SupBiotech September 2025

Study Description

Balanced dataset with several groups and multiple timepoints; some values are missing. The dataset is provided as an Excel file with multiple sheets.

Objectives

1. Explore the dataset, handle missingness, and summarise cell counts.
2. Check distributional and variance assumptions and contrast the decision with CLT.
3. Apply several hypothesis-testing approaches: at least two two-level comparisons and one multi-level comparison.
4. Fit a regression model and compare conclusions to the test-based approaches.
5. Construct a tagged matrix of p-values to compare approaches, and provide a concise rationale.
6. Produce all visualisations at the end: boxplots, violin plots, and mean $\pm$ CI; include a model view and a comparison visual.

Tasks

Task 1: data preparation

Import the Excel data with multiple sheets and prepare it for analysis. Decide how to handle missing values. Summarise the structure and describe the outcome variable.

Task 2: Assumptions and CLT

Specify a preliminary formula with `??`. Apply a distributional check with `??` and a variance check with `??`. State whether CLT supports a parametric path and justify the choice.
(!! Remember to check the quantity of your groups)

Task 3: Multiple approaches and model

Run at least two two-level comparisons and one multi-level comparison (names not provided here). Fit a regression model (obligatory). If appropriate, perform a post-hoc analysis. Provide a brief comparison of approaches and the model.

Task 4: Comparison matrix

Extract comparable statistics (e.g., p-values) from each approach, build a tagged matrix, and include a short interpretation.

Working Rules and Deliverables

Group Policy

Groups of **3–4**. Tasks may be split or done jointly. If split, each member's contribution must be annotated; if joint, state this clearly.

Use of AI Tools

AI tools may be used only as a **study aid**. Any external output must be checked, understood, and adapted to this dataset.

Submission Format

Use the **Research Methodology Progress** template. Include **full code** and all figures. Annotate contributions if split.

Starter Code

Replace every `??` with the chosen paths, variables, tests, and functions. No comments inside code.

data import

```
?? <- read.("??/??.")
```

Descriptive statistics

```
n_all <- nrow(??)
n_na  <- sum(is.na(??$??))
tab   <- with(??, table(??, ??))
desc1 <- tapply(??$??, list(??$??), function(x) c(n=length(x),
   mean=mean(x,na.rm=TRUE), sd=sd(x,na.rm=TRUE)))
desc2 <- tapply(??$??, list(??$??, ??$??), function(x) c(n=length(x),
   mean=mean(x,na.rm=TRUE), sd=sd(x,na.rm=TRUE)))
n_all; n_na; tab; desc1; desc2
```

Assumptions and CLT

```
spec0    <- ??(?? ~ ?? * ??, ??a = ??)
check_1  <- ??(??)
check_2  <- ??(?? ~ ??, ??a = ??)
spec0; check_1; check_2
```

Two-level comparisons (at least two)

```
test_1 <- ??(?? ~ ??, ??a = subset(??, ?? == "??"))
test_2 <- ??(?? ~ ??, ??a = subset(??, ?? == "??"))
test_1; test_2
```

Multi-level comparison

```
test_multi <- ??(?? ~ ??, ??a = ??)
test_multi
```

Regression model (obligatory)

```
model <- ??(?? ~ ?? * ??, ??a = ??)
model_out <- ??(model)
model; model_out
```

Post-hoc (if applicable)

```
posthoc <- ??(??)
posthoc
```

P-value matrix and comparison visual

```
p_1 <- ??
p_2 <- ??
p_multi <- ??
p_model <- ??
p_mat <- matrix(c(p_two_1, p_two_2, p_multi, p_model), nrow=1)
colnames(p_mat) <- c("two_1","two_2","multi","model")
rownames(p_mat) <- "p"
p_mat
barplot(-log10(p_mat), beside=TRUE, ylab="-log10(p)")
image(t(-log10(p_mat)), axes=FALSE)
axis(1, at=seq(0,1,length.out=ncol(p_mat)), labels=colnames(p_mat))
axis(2, at=0.5, labels=rownames(p_mat))
```

Visualisations (end only)

Boxplots

```
boxplot(?? ~ ??, ??a = subset(??, ?? == "??"))
boxplot(?? ~ ??, ??a = subset(??, ?? == "??"))
boxplot(?? ~ ??, ??a = ??)
```

Violin plots

```
if (!"vioplot" %in% rownames(installed.packages()))
    install.packages("vioplot", quiet=TRUE)
library(vioplot)
x1 <- ??$??[??$??=="??"]
x2 <- ??$??[??$??=="??"]
vioplot::vioplot(x1, x2, names=c("??","??"))
```

Mean ± CI

```
m  <- tapply(??$??, list(??$??, ??$??), mean, na.rm=TRUE)
se <- tapply(??$??, list(??$??, ??$??), function(x)
    sd(x,na.rm=TRUE)/sqrt(sum(!is.na(x))))
lo <- m - 1.96*se
hi <- m + 1.96*se
x  <- seq_len(length(m))
plot(x, as.vector(m), ylim=range(c(lo,hi),na.rm=TRUE), xaxt="n", pch=19,
    xlab="??-??", ylab="mean  95% CI")
arrows(x, as.vector(lo), x, as.vector(hi), angle=90, code=3, length=0.05)
axis(1, at=x, labels=paste(rep(dimnames(m)[[1]],
    each=length(dimnames(m)[[2]])), rep(dimnames(m)[[2]],
    times=length(dimnames(m)[[1]])), sep="-"))
```

Model view

```
plot(fitted(model) ~ ??, ??a = ??)
```

Brief rationale

Provide a concise comparison of the approaches used and the regression model, indicating how conclusions align or differ and how CLT informed the decision regarding parametric choices.