

2

Statistical and Probabilistic Foundations for Bioassay

2.1: Learning Objectives

1. Review the distributional assumptions (normal, binomial) underlying bioassay analysis.
2. Derive least-squares and maximum-likelihood estimators for simple linear dose–response regressions.
3. Construct pooled variance estimates across preparations.
4. Understand the delta method as a tool for approximating the variance of nonlinear functions of estimators (e.g., ratios).
5. Implement basic regression and variance computations in R and Python.

2.2: Why Statistical Theory Matters More in Bioassay

Ordinary regression theory is necessary but not sufficient for bioassay. What is special is that the *parameter of ultimate interest* — the relative potency ρ — is a **ratio** of linear combinations of regression coefficients, not a coefficient itself. Ratios of normally distributed random variables do not have a normal distribution, and in extreme cases (when the denominator is not significantly different from zero) they do not even possess a finite variance. This chapter builds the tools — pooled variance estimation and the delta method — that prepare us for Fieller’s exact treatment of ratio-type estimators in Chapter 3.

2.3: The Simple Linear Dose–Response Model

For a single preparation, we assume observations Y_i at dosage metameter x_i ($i = 1, \dots, m$; possibly several replicate responses Y_{ij} , $j = 1, \dots, n_i$, at each x_i) follow

$$Y_{ij} = \alpha + \beta x_i + e_{ij}, \quad e_{ij} \stackrel{iid}{\sim} N(0, \sigma^2).$$

Ordinary least squares gives

$$\hat{\beta} = \frac{\sum_i n_i (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_i n_i (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x},$$

with residual variance estimator $s^2 = \frac{\sum_{ij} (y_{ij} - \hat{\alpha} - \hat{\beta} x_i)^2}{N - 2}$,
 $N = \sum n_i$.

Under normality, $\hat{\alpha}, \hat{\beta}$ are also the maximum-likelihood estimators, and

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_i n_i (x_i - \bar{x})^2}, \quad \text{Var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_i n_i (x_i - \bar{x})^2} \right]$$

2.4: Pooling Variance Estimates Across Preparations

When standard (S) and test (T) preparations are assayed with the same design, each with residual sum of squares SS_S, SS_T on f_S, f_T degrees of freedom, the standard practice (assuming a common error variance, itself a testable assumption) is to pool:

$$s^2 = \frac{SS_S + SS_T}{f_S + f_T}.$$

This pooled s^2 is used throughout the potency computation, including inside Fieller's theorem. The assumption of a common σ^2 across preparations should be checked (e.g., an F -test of SS_S/f_S vs SS_T/f_T , or a Bartlett's test when there are several groups — see Chapter 9).

2.5: The Delta Method

Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ be jointly asymptotically normal with mean (θ_1, θ_2) and covariance matrix Σ . For a smooth function $g(\theta_1, \theta_2)$, a first-order Taylor expansion around (θ_1, θ_2) gives the approximate variance

$$\text{Var}[g(\hat{\theta}_1, \hat{\theta}_2)] \approx \left(\frac{\partial g}{\partial \theta_1}\right)^2 \text{Var}(\hat{\theta}_1) + \left(\frac{\partial g}{\partial \theta_2}\right)^2 \text{Var}(\hat{\theta}_2) + 2 \frac{\partial g}{\partial \theta_1} \frac{\partial g}{\partial \theta_2} \text{Cov}(\hat{\theta}_1, \hat{\theta}_2).$$

For the ratio $R = A/B$ with A, B independent,

$$\text{Var}(R) \approx \frac{1}{B^2} \text{Var}(A) + \frac{A^2}{B^4} \text{Var}(B) = \frac{R^2}{1} \left[\frac{\text{Var}(A)}{A^2} + \frac{\text{Var}(B)}{B^2} \right].$$

This is the working formula used, for example, in Section 1.1's cat-tolerance illustration and in the direct-assay chapter.

Caveat: the delta method is only a large-sample approximation; it under-states uncertainty when B (the denominator) is not precisely estimated relative to its mean — precisely the situation that motivates Fieller's *exact* fiducial/confidence interval in Chapter 3.

2.6: Maximum Likelihood for Quantal (Binomial) Data — A Preview

When the response at dose x_i is a binomial count r_i out of n_i subjects responding, and $p_i = \Phi(\alpha + \beta x_i)$ (probit model) or $p_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$ (logit model), the log-likelihood is

$$\ell(\alpha, \beta) = \sum_i [r_i \log p_i + (n_i - r_i) \log(1 - p_i)].$$

There is no closed form for $(\hat{\alpha}, \hat{\beta})$; iteratively re-weighted least squares (equivalently, Fisher scoring / Newton–Raphson) is used, which we return to in full generality in Chapter 8.

2.7: Hypothesis Testing Tools Used Throughout This Book

- ***t*-tests** for individual contrasts (e.g., is the parallelism contrast zero?).
- ***F*-tests** via analysis of variance for composite hypotheses (linearity of regression, parallelism, blank/intersection validity).
- **Bartlett's test** for homogeneity of variance across groups (Chapter 9).
- **Fieller's theorem** for exact confidence/fiducial limits of a ratio (Chapter 3).
- **Behrens–Fisher–Sukhatme methods** when two variances cannot be assumed equal (Chapter 5).

2.8: Worked Examples

Example 2.1: Fitting a Simple Dose–Response Line by Hand

Doses (log scale) $x = -1, 0, 1$ with two replicate responses each: at $x = -1$: 5, 7; at $x = 0$: 9, 11; at $x = 1$: 14, 12.

Solution. $\bar{x} = 0$, $n_i = 2$ for all i . $\bar{y}_{-1} = 6$, $\bar{y}_0 = 10$, $\bar{y}_1 = 13$;
grand mean $\bar{y} = 9.667$.

$$\sum n_i(x_i - \bar{x})(\bar{y}_i - \bar{y}) = 2(-1)(6 - 9.667) + 2(0)(\cdot) + 2(1)(13 - 9.667) = 7.333 + 6.667 = 14.0$$

$$\sum n_i(x_i - \bar{x})^2 = 2(1) + 0 + 2(1) = 4. \quad \hat{\beta} = 14/4 = 3.5, \\ \hat{\alpha} = 9.667. \text{ Fitted line: } \hat{y} = 9.667 + 3.5x.$$

Example 2.2: Residual Variance and R/Python Fit

Using the data of Example 2.1, verify the fit computationally and obtain s^2 .

R:

```
x <-c(-1,-1,0,0,1,1)
y <-c(5,7,9,11,14,12)
fit <-lm(y ~ x)
summary(fit)
s2 <-summary(fit)$sigma^2
cat("Residual variance s^2 =", round(s2,3), "\n")
```

Python:

```
import numpy as np
import statsmodels.api as sm

x = np.array([-1,-1,0,0,1,1])
y = np.array([5,7,9,11,14,12])
X = sm.add_constant(x)
fit = sm.OLS(y, X).fit()
print(fit.summary())
resid_var = fit.mse_resid
print(f'Residual variance s^2 = {resid_var:.3f}')
```

Both confirm $\hat{\alpha} \approx 9.667$, $\hat{\beta} = 3.5$, with $s^2 \approx 2.33$ (4 residual d.f.).

Example 2.3: Pooling Variances From Two Preparations

Standard preparation: $SS_S = 12.4$ on 8 d.f. Test preparation: $SS_T = 9.8$ on 8 d.f. Obtain the pooled estimate and test whether pooling is justified.

Solution. $s^2 = (12.4 + 9.8)/(8 + 8) = 22.2/16 = 1.3875$.
 $F = 12.4/9.8 = 1.265$ on (8,8) d.f. Critical value

$F_{0.05,8,8} \approx 3.44$; since $1.265 < 3.44$, the variances are not significantly different, and pooling is justified.

Example 2.4 : Delta-Method Variance of a Ratio

Let $A = 16.777$ (combined regression estimate) with $\text{Var}(A) = 1.143^2 = 1.306$, and B a preparation-log-ratio numerator $= 0.800$ with $\text{Var}(B) = 0.562^2 = 0.316$ (values from the testosterone-propionate capon assay of Chapter 6). Treat $R^* = B/A$ (the exponent of $\hat{\rho}$) and approximate its variance using (2.1), assuming independence.

Solution.

$$\text{Var}(R^*) \approx \left(\frac{R^*}{A} \right)^2 \cdot 0 \dots$$

Using (2.1) directly: $R^* = 0.800/16.777 = 0.04768$.

$$\text{Var}(R^*) \approx (R^*)^2 \left[\frac{\text{Var}(B)}{B^2} + \frac{\text{Var}(A)}{A^2} \right] = (0.04768)^2 \left[\frac{0.316}{0.64} + \frac{1.306}{281.5} \right] = 0.00227;$$

So $\text{SE}(R^*) \approx 0.0337$ — an approximate delta-method value; Chapter 3 shows how Fieller's theorem refines this into an exact interval for $\hat{\rho} = \exp(R^* \cdot \log(\text{dose ratio}))$.

Example 2.5: Checking Normality Assumptions With R/Python

For a set of 20 direct-assay tolerances, check approximate normality of log-tolerances using a Shapiro–Wilk test.

R:

```
tol <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34,2.42,1.85,2.00,
2.27,1.70,1.47,2.20,1.60,1.95,2.05,1.77,1.66,2.10)
shapiro.test(log(tol))
```

Python:

```

from scipy import stats
import numpy as np

tol =
np.array([1.55,1.58,1.71,1.44,1.24,1.89,2.34,2.42,1.85,2.00,
2.27,1.70,1.47,2.20,1.60,1.95,2.05,1.77,1.66,2.10])
w, p = stats.shapiro(np.log(tol))
print(f'Shapiro-Wilk W = {w:.3f}, p = {p:.3f}')

```

A p -value comfortably above 0.05 supports proceeding with the standard normal-errors assay model on the log scale.

Example 2.6: Covariance Between Regression Estimators

Show that $\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\bar{x} \text{Var}(\hat{\beta})$ in simple linear regression, and explain its relevance to potency-interval calculations.

Solution: Writing $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and noting $\text{Cov}(\bar{y}, \hat{\beta}) = 0$ (a standard OLS identity because $\bar{y}$ and $\hat{\beta}$ are computed from orthogonal contrasts of the data under the classical model), we get $\text{Cov}(\hat{\alpha}, \hat{\beta}) = \text{Cov}(\bar{y} - \hat{\beta}\bar{x}, \hat{\beta}) = -\bar{x} \text{Var}(\hat{\beta})$. This nonzero covariance is precisely why the general form of Fieller's theorem (Chapter 3) must carry a covariance term v_{12} rather than assuming independence, whenever the two quantities forming the ratio come from the same fitted line (as happens, e.g., for an ED50 computed from a single probit line).

Example 2.7: Simulating a Bioassay Data Set for Teaching

Generate a synthetic parallel-line data set in R and Python to illustrate the concepts before moving to real data in later chapters.

R:

```
set.seed(757)
logdose <- rep(log(c(20,40,80)), each =5)
alpha_S <-2; beta <-8
response_S <- alpha_S + beta*logdose
+rnorm(15, sd =1.5)
alpha_T <-2.3
response_T <- alpha_T + beta*logdose
+rnorm(15, sd =1.5)
sim_data <- data.frame(logdose
=rep(logdose,2),
response =c(response_S, response_T),
prep =rep(c("S","T"), each=15))
head(sim_data)
```

Python:

```
import numpy as np
import pandas as pd

rng = np.random.default_rng(757)
logdose = np.repeat(np.log([20,40,80]), 5)
alpha_S, beta =2, 8
resp_S = alpha_S + beta*logdose +
rng.normal(scale=1.5, size=15)
alpha_T =2.3
resp_T = alpha_T + beta*logdose +
rng.normal(scale=1.5, size=15)
sim_data = pd.DataFrame({
"logdose": np.tile(logdose, 2),
"response": np.concatenate([resp_S, resp_T]),
"prep": np.repeat(["S","T"], 15)
})
sim_data.head()
```

Such simulated data sets are useful for demonstrating that recovered slopes/intercepts scatter around the true values used to generate them, reinforcing the sampling-distribution ideas developed above.

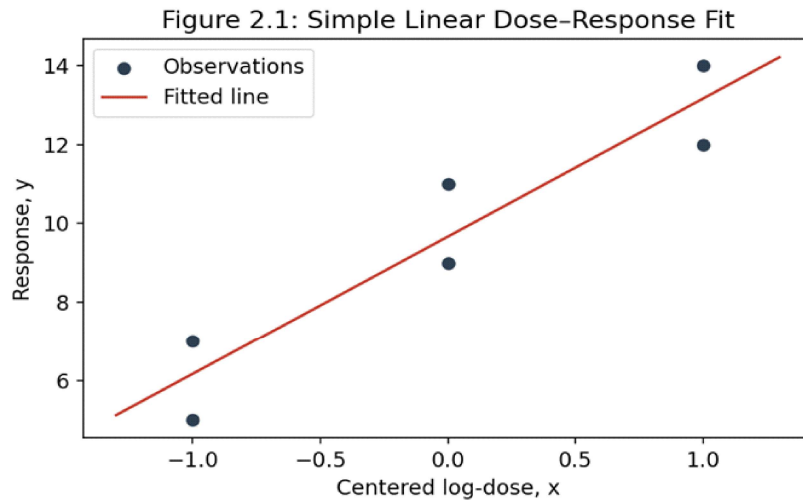


Figure 2.1: *Simple linear dose-response fit*

2.9: Chapter Summary

- Bioassay regression relies on the classical normal-errors linear model, with pooled variance estimates across preparations when variances are homogeneous.
- The delta method gives an approximate variance for ratio-type statistics but is unreliable when the denominator is imprecisely estimated — motivating Fieller’s exact approach (Chapter 3).
- Quantal-response likelihoods (probit/logit) require iterative maximum likelihood, foreshadowed here and developed fully in Chapter 8.
- R and Python lm/statsmodels tools reproduce classical hand calculations and extend naturally to simulation-based teaching.

2.10: Exercises

Exercise 2.1: Using the delta method formula (2.1), derive the variance of $R = \bar{x}_S/\bar{x}_T$ for two *independent* sample means with variances $\text{Var}(\bar{x}_S) = \sigma^2/n_S$, $\text{Var}(\bar{x}_T) = \sigma^2/n_T$, and express the result in terms of R , σ^2 , $\bar{x}_T$, n_S , n_T .

Exercise 2.2: Fit, by hand and then in R or Python, the regression of response on log-dose for the data: $x = \log(10), \log(20), \log(40)$, with two replicates each: $y = 4, 6$ ($x = \log 10$); $9, 11$ ($x = \log 20$); $15, 17$ ($x = \log 40$).

Report $\hat{\alpha}, \hat{\beta}, s^2$.

Exercise 2.3: Two independent residual sums of squares are $SS_1 = 18.6$ (10 d.f.) and $SS_2 = 31.2$ (10 d.f.). (a) Test whether pooling is justified at the 5% level. (b) If justified, compute the pooled variance estimate.

Exercise 2.4: Explain in your own words why the delta-method variance of a ratio $R = A/B$ becomes unreliable as B approaches zero, and connect this explanation to why a *confidence interval* for ρ (Chapter 3) can sometimes be one-sided or even the whole real line — a phenomenon that cannot occur with the simple symmetric interval $R \pm t \cdot SE(R)$ implied by the delta method.

(Full worked solutions appear in Appendix A.)