
Statistical Bioassay Theory, Design, and Computation

*A Graduate Textbook with Worked
Examples in R and Python*

Statistical Bioassay Theory, Design, and Computation

*A Graduate Textbook with Worked
Examples in R and Python*

Olúwáfémi Samuel Oyámákin

Stirling-Horden Publishers Ltd.
No 6, Adonai Estate, Opp Crown
Heights College, Arulogun Road
Ojoo, Ibadan.
Phone: 08023519154/07031967972
stirlinghorden@ymail.com, info@stirlinghorden.com
www.stirlinghorden.com

University Post Office	7 Ikioda Street
Box 20984	Jattu-Uzairue
Ibadan	Edo State, Nigeria.
Oyo State, Nigeria.	

© Olúwáfémi Samuel Oyámákin

Published 2026

ISBN 978-978-032-

All Rights Reserved

Apart from any fair dealing for the purpose of research or private study, criticism or review, as permitted under the Copyright, Designs and Patents Act of Nigeria, this publication may not be reproduced, stored or transmitted in any form, or by any means, except with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

Stirling-Horden Publishers Ltd.
Lagos, Ibadan, Benin City, Jattu-Uzairue.

Dedication

Acknowledgements

Preface

Contents

Dedication	v
Acknowledgements	vi
Preface	vii
Chapter 1: Introduction to Biological Assay	1
1.1 Learning Objectives	1
1.2 What Is a Biological Assay?	1
1.3 Standard and Test Preparations; Relative Potency	2
1.4 Classification of Bioassays	3
1.5 Historical Notes	4
1.6 The Role of Design	5
1.7 Worked Examples	5
1.8 Chapter Summary	8
1.9 Exercises	8
Chapter 2: Statistical and Probabilistic Foundations for Bioassay	11
2.1 Learning Objectives	11
2.2 Why Statistical Theory Matters More in Bioassay	11
2.3 The Simple Linear Dose–Response Model	12
2.4 Pooling Variance Estimates Across Preparations	13

2.5	The Delta Method	13
2.6	Maximum Likelihood for Quantal (Binomial) Data — A Preview	13
2.7	Hypothesis Testing Tools Used Throughout This Book	14
2.8	Worked Examples	14
2.9	Chapter Summary	19
2.10	Exercises	20
Chapter 3: Fieller's Theorem and Fiducial Limits		21
3.1	Learning Objectives	21
3.2	Motivation	21
3.3	Statement of Fieller's Theorem	21
3.4	Application to Relative Potency	24
3.5	Worked Examples	24
3.6	Chapter Summary	29
3.7	Exercises	30
Chapter 4: Direct Assays		31
4.1	Learning Objectives	31
4.2	The Direct-Assay Model	31
4.3	Precision of the Direct-Assay Potency Estimate	32
4.4	The Classical "Cat Method" Example	33
4.5	Worked Examples	34
4.6	Chapter Summary	40
4.7	Exercises	40
Chapter 5: The Behrens–Fisher–Sukhatme Distribution		43
5.1	Learning Objectives	43
5.2	The Problem of Unequal Variances	43
5.3	The Behrens–Fisher(–Sukhatme) Distribution	44

5.4	Worked Examples	45
5.5	Chapter Summary	50
5.6	Exercises	51
Chapter 6:	Parallel-Line Assays	53
6.1	Learning Objectives	53
6.2	The Model	53
6.3	The Three Key Contrasts	54
6.4	Validity Tests and the Potency Estimate	55
6.5	Worked Examples	56
6.6	Chapter Summary	62
6.7	Exercises	63
Chapter 7:	Slope-Ratio Assays	65
7.1	Learning Objectives	65
7.2	The Model	65
7.3	Blank and Intersection Contrasts	66
7.4	Analysis of Variance for Slope-Ratio Assays	67
7.5	Worked Examples	68
7.6	Chapter Summary	74
7.7	Exercises	74
Chapter 8:	Quantal Response Assays — Probit and Logit Analysis	77
8.1	Learning Objectives	77
8.2	The Tolerance-Distribution Model	77
8.3	Probit and Logit Transformations	78
8.4	Maximum Likelihood Estimation	78
8.5	ED50 and Its Fiducial Interval	79
8.6	Relative Potency From Two Quantal Assays	79
8.7	Worked Examples	80
8.8	Chapter Summary	86
8.9	Exercises	87

Chapter 9: Validity Tests, the Condition of Similarity, and Variance Homogeneity	89
9.1 Learning Objectives	89
9.2 The Condition of Similarity	89
9.3 Consequences of a Failed Similarity Test	90
9.4 Bartlett's Test for Homogeneity of Variance	91
9.5 Variance-Stabilizing Transformations	91
9.6 Worked Examples	92
9.7 Chapter Summary	99
9.8 Exercises	99
Chapter 10: Design of Experiments in Bioassay — Block Designs	101
10.1 Learning Objectives	101
10.2 Why Bioassay Needs Special Designs	101
10.3 Randomized Complete Block Design (RCBD)	102
10.4 Incomplete Block Designs	102
10.5 Constructing a Simple Incomplete Block Design by Hand	103
10.6 Worked Examples	104
10.7 Chapter Summary	109
10.8 Exercises	110
Chapter 11: Cross-Over Designs and Washout Periods	111
11.1 Learning Objectives	111
11.2 Rationale for Cross-Over Designs	111
11.3 The 2×2 Cross-Over Design	112
11.4 Statistical Model	112
11.5 Worked Examples	114
11.6 Chapter Summary	121
11.7 Exercises	121

Chapter 12: Comprehensive Case Studies	123
12.1 Purpose of This Chapter	123
12.2 Case Study 1: A Complete Parallel-Line Vitamin Assay	123
12.3 Case Study 2: A Quantal Insecticide Assay With Two Formulations	127
12.4 Reflections for the Practicing Biostatistician	131
12.5 Chapter Summary	132
12.6 Exercises	132
Appendix A: Solutions to Chapter Exercises	135
A.1 Chapter 1 Solutions	135
A.2 Chapter 2 Solutions	136
A.3 Chapter 3 Solutions	137
A.4 Chapter 4 Solutions	139
A.5 Chapter 5 Solutions	140
A.6 Chapter 6 Solutions	141
A.7 Chapter 7 Solutions	143
A.8 Chapter 8 Solutions	144
A.9 Chapter 9 Solutions	145
A.10 Chapter 10 Solutions	147
A.11 Chapter 11 Solutions	148
Appendix B: R and Python Code Library	150
B.1 Fieller's Theorem (Chapter 3)	150
B.2 Direct-Assay Potency Estimate (Chapter 4)	151
B.3 Behrens–Fisher–Sukhatme / Welch Adjustment (Chapter 5)	152
B.4 Parallel-Line Contrast Coefficients (Chapter 6)	152
B.5 Parallel-Line Potency From a Fitted Model (Chapter 6)	153
B.6 Slope-Ratio Potency (Chapter 7)	153

B.7	Probit/Logit ED50 and Relative Potency (Chapter 8)	154
B.8	Bartlett's Test (Chapter 9) — Manual Implementation	154
B.9	Validity-Report Pipeline for Parallel-Line Assays (Chapter 9)	155
B.10	2×2 Cross-Over Design Analysis (Chapter 11)	156
B.11	Notes on Reuse	157
Appendix C: Statistical Tables		158
C.1	Student's t Distribution — Two-Sided Critical Values	158
C.2	Chi-Square Distribution — Upper-Tail Critical Values	159
C.3	F Distribution — Upper 5% Critical Values ($F_{0.05}$)	159
C.4	Standard Normal CDF and Probit Values	160
C.5	Generating Exact Values Programmatically	160
Appendix D: Glossary of Terms		161
Appendix E: Notation and Symbol Reference		164
Appendix F: References and Further Reading		166

1

Introduction to Biological Assay

1.1: Learning Objectives

By the end of this chapter, a graduate student should be able to:

1. Define a biological assay and distinguish it from a general comparative experiment.
2. Identify the three structural components of every assay: stimulus, subject, and response.
3. Classify assays as direct or indirect, and indirect assays as quantitative (parallel-line, slope-ratio) or quantal.
4. Define relative potency and explain why it, rather than a simple difference of means, is the natural parameter of interest.
5. State the historical milestones that shaped modern bioassay theory.
6. Set up a basic bioassay computation in R and Python.

1.2: What is a Biological Assay?

A biological assay (bioassay) is a statistical procedure for comparing the strength (potency) of two or more preparations of a stimulus by observing the magnitude of a biological response that each preparation produces in a set of living subjects. The defining feature of a bioassay, which separates it from an ordinary comparative experiment, is that the object of inference is not simply “does treatment A differ from treatment B?” but rather

“on what multiple of dose is A equivalent to B?” That multiple is the **relative potency**. Three ingredients are present in every assay:

- **Stimulus**: the substance or treatment being compared — a drug, hormone, vitamin, vaccine, insecticide, or physical agent (e.g., radiation).
- **Subject**: the biological material that receives the stimulus — an animal, a plant, a tissue culture, a microbial colony, or a human participant.
- **Response**: the measured or observed change in the subject following exposure to the stimulus. Responses may be continuous (a hormone-induced increase in organ weight), a survival time, or quantal (dead/alive, cured/not cured).

A useful discipline for a beginning biostatistician is to ask, of any comparative experiment: is the underlying interest a scale-of-equivalent-dose question? If we are only asking “which treatment gives a bigger effect,” we are outside the bioassay framework, even if animals, tissues, or drugs are involved. If we are asking “how much of A equals a given amount of B,” we are inside it.

1.3: Standard and Test Preparations; Relative Potency

Almost every quantitative bioassay compares a **standard preparation** (S), whose potency is fixed by convention or by an international reference standard, against a **test preparation** (T), whose potency is unknown and is to be estimated relative to S. If

d_S and d_T denote doses of S and T that produce the same

$$\rho = \frac{d_S}{d_T}.$$

- If $\rho > 1$: a smaller dose of T matches a larger dose of S, so T is *more potent* than S.
- If $\rho < 1$: T is *less potent* than S.
- If $\rho = 1$: the two preparations are equipotent.

The task of an assay is to obtain a sample estimate R of ρ ,

together with a valid measure of precision — most often a **fiducial** or **confidence interval** obtained via Fieller's theorem (Chapter 3).

1.4: Classification of Bioassays

Biological	Assays
+-- Qualitative assays (no dose-response scale; e.g., presence/absence)	tests)
+-- Quantitative (dilution) assays	
+-- Direct assays	
	Dose needed for a specified response is measured
directly	on each subject
	(e.g., the "cat method" for digitalis: infuse until
the	heart stops)
+-- Indirect assays	
	Dose is fixed by the experimenter; response is
observed	
+-- Quantitative response	
	+-- Parallel-line assays (dosage = log
dose)	
	+-- Slope-ratio assays (dosage =
dose ^{lambda})	
+-- Quantal response	
	Probit / logit analysis; ED50, LD50
estimation	

1.4.1: Direct Assays

In a direct assay, the dose required to produce a specified end-point (e.g., death, onset of a muscular contraction) is observed directly for every subject; the dose itself is the random variable. The classic example is the “cat method” for digitalis-like drugs, discussed fully in Chapter 4.

1.4.2: Indirect Assays

In most modern bioassays, the dose is fixed by the investigator at several pre-selected levels, and the response — not the dose — is the random variable. A dose–response relationship is then used to infer the dose of the test preparation that would produce the same response as a given dose of the standard. Two linearizing dosage transformations dominate practice:

- $x = \log(\text{dose})$ leads to **parallel-line assays**.
- $x = (\text{dose})^\lambda$, for known $\lambda > 0$, leads to **slope-ratio assays**.

1.4.3: Quantal Assays

When the response is binary (all-or-nothing — e.g., death, cure, infection), we work with the proportion responding at each dose level and fit a sigmoid dose–response curve, typically via the **probit** or **logit** transformation (Chapter 8). The dose producing a response in 50% of subjects is the **median effective dose (ED50)**, or, when the endpoint is death, the **median lethal dose (LD50)**.

1.5: Historical Notes

- **Antiquity**: The earliest recorded bioassay-like reasoning is often (playfully) attributed to Noah’s dove-and-olive-leaf experiment (Genesis 8), where the “stimulus” was the depth of floodwater, the “subject” was the dove, and the “response” was whether it returned with a leaf — a crude quantal assay of water level.
- **Coal-mine canaries**: Canaries, more sensitive to methane than humans, served as a quantal bioassay for underground air quality until the mid-twentieth century.
- **Paul Ehrlich (late 19th century)**: Standardized diphtheria antitoxin by comparing its potency in test animals against a reference preparation, establishing quantitative pharmacological standardization as routine practice.

- **1920s–1930s:** Statisticians (notably J. W. Trevan on LD50, and Chester Bliss on the probit transformation) placed assay methodology on a rigorous statistical footing.
- **D. J. Finney (1947, 1952, 1964, 1978):** Statistical Method in Biological Assay unified parallel-line, slope-ratio, and quantal assay theory and remains the canonical reference; much of this textbook’s notation follows Finney’s conventions to ease cross-referencing.
- **R. A. Fisher and collaborators:** Contributed fiducial inference (used heavily via Fieller’s theorem) and the analysis-of-variance framework used throughout bioassay design.

1.6: The Role of Design

Because the parameters of interest in a bioassay are specific **contrasts** among treatment (dose) means — a preparation contrast, a combined-regression contrast, and a parallelism (or blank/intersection) contrast — rather than all possible pairwise comparisons, bioassay experiments are usually laid out in designs that let these particular contrasts be estimated free of block effects: randomized complete block designs when block size equals the number of dose combinations, and specialized incomplete block or cross-over designs otherwise. Chapters 10 and 11 develop this machinery.

1.7: Worked Examples

Example 1.1: Identifying the Assay Components

A pharmaceutical laboratory tests three doses of a candidate anticoagulant against three doses of warfarin (the standard) by measuring prothrombin time in rabbits.

Identify stimulus, subject, response, and state whether this is a direct or indirect assay.

Solution: Stimulus: the anticoagulant preparations (test and standard) at fixed doses. Subject: the rabbits. Response:

prothrombin time (a continuous, quantitative response) measured after dosing. Because dose is fixed by the experimenter and response is observed (rather than dose being the observed quantity), this is an **indirect, quantitative assay** — specifically a parallel-line assay once doses are log-transformed.

Example 1.2 — Relative Potency From Mean Effective Doses

Two insulin preparations are each administered to guinea pigs at doses selected so that both preparations produce, on average, a fall of 40 mg/dL in blood glucose. Preparation S required 0.36 units, while preparation T required 0.28 units.

Solution:

$$\rho = \frac{d_S}{d_T} = \frac{0.36}{0.28} = 1.286.$$

T is roughly 29% more potent than S.

Example 1.3: Direct vs. Indirect (Aphid Insecticide)

Explain, using Finney's aphid example, why an indirect assay is unavoidable for testing insecticide potency on aphids, whereas a direct assay is feasible for digitalis on cats.

Solution: For digitalis on cats, a trained technician can titrate the infusion rate and stop precisely at cardiac arrest, so the lethal dose is observed directly on each animal. For an insecticide sprayed on aphids, there is no way to administer a continuously increasing dose to a single insect and observe the exact dose at which it dies; the experimenter can only apply a fixed, discrete dose to a batch of insects and record the proportion that die. Hence the assay must be indirect and quantal, using probit or logit analysis to back out an ED50/LD50.

Example 1.4 : Qualitative vs Quantitative Assay

A public-health laboratory tests whether a water sample contains *E. coli* using a presence/absence culture test. Is this a bioassay in the statistical sense developed in this book?

Solution. No. Although a living organism (bacterial growth) is used as an indicator, there is no dose–response scale on which a potency ratio is estimated; the test is qualitative. It is a screening bioassay in the broad biological sense, but not a quantitative bioassay of the kind treated in this text.

Example 1.5 — Choosing a Dosage Metameter

A toxicologist has LD-type mortality data across doses 5, 10, 20, 40, 80 mg/kg (a geometric, doubling series). Which dosage metameter is natural, and why?

Solution: Because the doses are equally spaced on a *multiplicative* (geometric) scale, $x = \log(\text{dose})$ converts them to an arithmetic sequence ($\log 5, \log 10, \log 20, \log 40, \log 80$), which is the natural metameter for a parallel-line (or probit, for quantal data) analysis. Using raw dose would produce a badly non-linear and heteroscedastic regression.

Example 1.6: Recognizing a Slope-Ratio Setting

A vitamin bioassay uses doses of 1, 2, 3, 4 mL of a supplemented feed (an additive rather than multiplicative series), including a zero (blank) dose, and the response (weight gain) is believed to be proportional to the amount of vitamin administered rather than the log-amount. Which family of indirect assay is appropriate?

Solution: Since the response is taken to be linear in dose itself (with $\lambda = 1$) rather than in log dose, and a blank (zero) dose is deliberately included so the two regression lines can be checked for a common intercept, this is a **slope-ratio assay** (Chapter 7), not a parallel-line assay.

Example 1.7: R and Python as Computational Tools

Even at the introductory stage, it is useful to see how relative potency and its uncertainty will eventually be computed. Suppose seven cats are dosed with two tinctures (A, B) of a cardiac glycoside until death, and the (untransformed) mean tolerances are $\bar{x}_A = 1.68$, $\bar{x}_B = 1.99$ (units of 0.01 c.c./kg).

Treating B as standard, compute a preliminary potency estimate in R and Python.

R:

```
xA <-1.68# mean tolerance, test preparation A
xB <-1.99# mean tolerance, standard preparation B
R <- xB / xA
cat("Relative potency R =", round(R, 3), "\n")
```

Python:

```
xA =1.68# mean tolerance, test preparation A
xB =1.99# mean tolerance, standard preparation B
R = xB / xA
print(f"Relative potency R = {R:.3f}")
```

Both give $R = 1.18$, matching the fully worked Fieller's-theorem analysis of this same strophanthus data set carried out in Chapter 4.

1.8: Chapter Summary

- A bioassay compares potencies via a dose–response relationship, not merely mean responses.
- Every assay has stimulus, subject, response.
- Assays are direct (dose observed) or indirect (response observed at fixed doses); indirect assays are further split into parallel-line, slope-ratio, and quantal types according to the dosage metameter and the nature of the response.
- Relative potency $\rho = d_S/d_T$ is the fundamental parameter; its estimation and interval estimation (via Fieller's theorem) occupy most of the remaining chapters.
- Statistical bioassay methodology developed through contributions from Ehrlich, Trevan, Bliss, Fisher, and Finney.

1.9: Exercises

Exercise 1.1: A laboratory determines the dose of two muscle relaxants (S and T) required to produce a 50% reduction in twitch height in an isolated nerve–muscle preparation. The mean doses are $\bar{d}_S = 12.4 \mu g$ and $\bar{d}_T = 9.8 \mu g$. (a) Compute the relative potency of T with respect to S. (b) State, in one sentence, the practical interpretation of your answer for a clinician choosing a dose of T.

Exercise 1.2: Classify each of the following as (i) direct assay, (ii) indirect quantitative assay (state parallel-line or slope-ratio), or (iii) indirect quantal assay, giving a one-sentence justification in each case: (a) Doses of a growth hormone (0, 1, 2, 4 mg) are injected into rat pups and body-weight gain after two weeks is recorded. (b) A vasodilator is infused at an increasing rate into an isolated artery preparation until it visibly dilates, and the total infused volume at that moment is recorded for each of 15 preparations. (c) Increasing concentrations of a fungicide are applied to 20 spore samples per concentration, and the number of samples showing no germination is recorded. (d) Doses of vitamin C (5, 10, 15, 20 mg, plus a zero-dose control) are fed to guinea pigs and the amount of new dentine formed is measured, with the response assumed proportional to dose amount rather than log dose.

Exercise 1.3: Explain, in your own words and with a fresh example (not one used in this chapter), why a bioassay is described as estimating “a ratio of doses” rather than “a difference of responses.” Why does this distinction matter for the choice of statistical method (hint: think ahead to Fieller’s theorem in Chapter 3)?

Exercise 1.4: Two batches of an antibiotic preparation are compared using a microbial zone-of-inhibition assay at three log-doses. From the resulting parallel-line analysis a preparation contrast is estimated as $\hat{\Psi}_p = -0.42$ with combined regression $\hat{\Psi}_1 = 8.15$ on the log-dose (base e) scale, dose ratio c_1/c_2 between successive levels of 2 for both preparations. Using the relation $\log \hat{\rho} = (c_1 - c_2) - \hat{\Psi}_p/\hat{\Psi}_1$ style reasoning introduced in this chapter (full derivation in Chapter 6), reason qualitatively (no need to compute exactly yet) about whether the test batch is more or less potent than the standard, and justify your reasoning from the sign of $\hat{\Psi}_p$.

(Full worked solutions for all end-of-chapter exercises are collected in Appendix A.)

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.

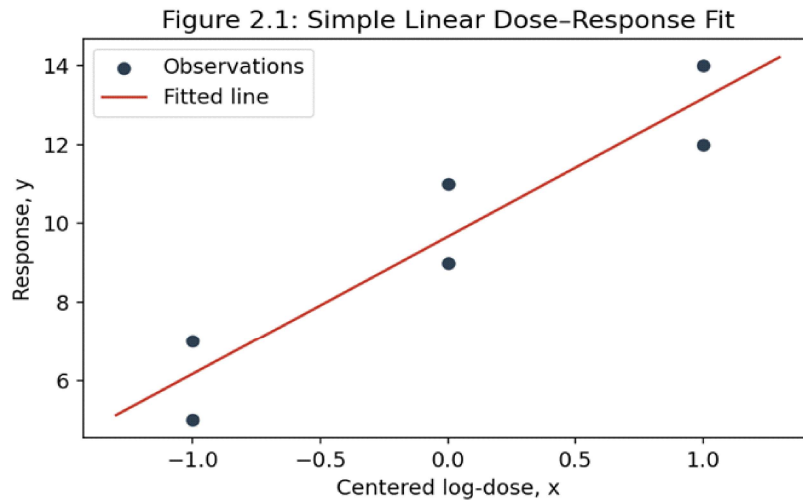


Figure 2.1: *Simple linear dose-response fit*

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.)

3

Fieller's Theorem and Fiducial Limits

3.1: Learning Objectives:

1. State and prove Fieller's theorem for the ratio of two jointly normal random variables.
2. Apply Fieller's theorem to construct fiducial (confidence) limits for relative potency.
3. Understand the special role of the quantity and the consequences of approaching or exceeding 1.
4. Compute Fieller limits by hand and using R/Python.

3.2: Motivation

Chapter 2 showed that ratio estimators like $R = A/B$ do not have a variance that behaves like that of a simple linear statistic, especially when B is imprecisely known. **Fieller's theorem** (Fieller, 1940, 1954) provides an *exact* (under normality) method for constructing confidence/fiducial limits for the true ratio $\mu = \alpha/\beta$, based on estimators a of α and b of β , without relying on the delta-method approximation.

3.3: Statement of Fieller's Theorem

Theorem 3.1 (Fieller). Let α, β be unknown parameters with $\mu = \alpha/\beta$. Suppose a, b are unbiased estimators, jointly normally distributed, with

$$\text{Var}(a) = v_{11}s^2, \quad \text{Var}(b) = v_{22}s^2, \quad \text{Cov}(a, b) = v_{12}s^2,$$

where s^2 is an estimator of the common (unknown) error variance σ^2 , independent of a, b and based on f degrees of freedom, so that $fs^2/\sigma^2 \sim \chi_f^2$. Let $m = a/b$ be the natural point estimator of μ . Define

$$g = \frac{t_0^2 s^2 v_{22}}{b^2},$$

where t_0 is the two-sided $100(1 - \alpha^*)\%$ point of Student's t distribution on f degrees of freedom. Then the $100(1 - \alpha^*)\%$ fiducial (confidence) limits for μ are

$$m_L, m_U = \frac{1}{1-g} \left\{ \left(m - g \frac{v_{12}}{v_{22}} \right) \pm \frac{t_0 s}{b} \sqrt{v_{11} - 2mv_{12} + m^2 v_{22} - g \left(v_{11} - \frac{v_{12}^2}{v_{22}} \right)} \right\},$$

3.3.1: Proof (Sketch)

Consider the linear combination $Z = a - \mu b$. By construction $E(Z) = 0$ when μ is the true ratio, and

$$\text{Var}(Z) = s^2 (v_{11} - 2\mu v_{12} + \mu^2 v_{22}).$$

Under joint normality of a, b and independence from s^2 ,

$$T = \frac{Z}{s \sqrt{v_{11} - 2\mu v_{12} + \mu^2 v_{22}}} \sim t_f.$$

A $100(1 - \alpha^*)\%$ confidence statement is $\Pr(-t_0 \leq T \leq t_0) = 1 - \alpha^*$, i.e.

$$(a - \mu b)^2 \leq t_0^2 s^2 (v_{11} - 2\mu v_{12} + \mu^2 v_{22}).$$

Expanding, this is a **quadratic inequality in μ** :

$$\mu^2 (b^2 - t_0^2 s^2 v_{22}) - 2\mu (ab - t_0^2 s^2 v_{12}) + (a^2 - t_0^2 s^2 v_{11}) \leq 0.$$

Dividing through by b^2 and writing $g = t_0^2 s^2 v_{22}/b^2$, $m = a/b$, the two roots of the corresponding quadratic equation, obtained via the standard quadratic formula, are exactly the limits given in (3.1). ■

3.3.2: Interpretation of g

The quantity g measures how precisely b (the denominator, e.g. a slope) is estimated relative to its own magnitude:

$g = \left(\frac{t_0 \cdot SE(b)}{b} \right)^2$, i.e., $\sqrt{g}$ is (approximately) the reciprocal of the t -statistic for testing $H_0 : \beta = 0$.

- If g is small (b is precisely estimated, i.e., b is many standard errors from 0), the Fieller interval is close to the delta-method interval $m \pm t_0 SE(m)$, and is nearly symmetric about m .
- As $g \rightarrow 1$, the interval widens dramatically and becomes markedly asymmetric.
- If $g \geq 1$ (equivalently, b is not significantly different from zero at the chosen confidence level), the quadratic in μ has no real, finite, bounded solution set; the “confidence interval” for μ is either the entire real line or the complement of a bounded interval. This reflects the intuitive fact that if the slope is not significantly nonzero, the ratio $\mu = \alpha/\beta$ is not meaningfully estimable at that confidence level — a phenomenon impossible to detect from the naive delta-method interval, which is always a finite symmetric interval regardless of how poorly β is estimated.

3.4: Application to Relative Potency

In the direct-assay setting of Chapter 4, with $a = \bar{x}_B$ (mean tolerance, standard), $b = \bar{x}_A$ (mean tolerance, test), independent, $v_{11} = 1/n_B$, $v_{22} = 1/n_A$, $v_{12} = 0$, Fieller's theorem gives exact limits for $\rho = \bar{x}_B/\bar{x}_A$ without the delta-method's optimistic symmetry assumption.

In the parallel-line/slope-ratio settings of Chapters 6–7, a and b are typically the *preparation contrast* and the *combined regression* (or the two slopes), and $v_{12} \neq 0$ in general because both may be computed from the same fitted lines.

3.5: Worked Examples

Example 3.1: Direct Application to Cat Tolerances (Recap of Chapter 1)

Strophanthus tinctures A, B; $n_A = n_B = 7$; $\bar{x}_A = 1.68$ (test), $\bar{x}_B = 1.99$ (standard); pooled $s^2 = 0.1200$ on 12 d.f.

Take $R = \bar{x}_B/\bar{x}_A = 1.18$. Here $a = \bar{x}_B$, $b = \bar{x}_A$, $v_{11} = v_{22} = 1/7$, $v_{12} = 0$.

At $\alpha^* = 0.05$, $t_0 = t_{0.025,12} = 2.179$.

$$g = \frac{t_0^2 s^2 v_{22}}{b^2} = \frac{(2.179)^2 (0.1200) (1/7)}{1.68^2} = \frac{4.748 \times 0.1200/7}{2.8224} = \frac{0.08140}{2.8224} = 0.02884.$$

Since g is small, the interval will be close to symmetric. Substituting into (3.1) gives $\rho_L, \rho_U \approx 0.95, 1.48$ (matching the classical Finney result), i.e. the 95% fiducial interval for relative potency is approximately (0.95, 1.48).

Example 3.2: Effect of a Large g

Suppose instead $n_A = n_B = 3$ (a much smaller experiment), same $\bar{x}_A, \bar{x}_B, s^2$. Recompute g and comment.

Solution. $t_{0.025,4} = 2.776$ (pooled d.f. now $2 + 2 = 4$).

$$g = \frac{(2.776)^2(0.1200)(1/3)}{1.68^2} = \frac{7.706 \times 0.0400}{2.8224} = \frac{0.3082}{2.8224} = 0.1092.$$

g has grown roughly fourfold; the interval will be noticeably wider and less symmetric than in Example 3.1, illustrating the direct cost, in fiducial precision, of a smaller experiment.

Example 3.3: When g Exceeds 1

A poorly-designed direct assay with only $n_A = n_B = 2$ yields $b = \bar{x}_A = 1.68$ with a much larger $s^2 = 0.9$. Compute g at the 95% level and interpret.

Solution: Pooled d.f. = $1 + 1 = 2$, $t_{0.025,2} = 4.303$.

$$g = \frac{(4.303)^2(0.9)(0.5)}{1.68^2} = \frac{18.52 \times 0.45}{2.8224} = \frac{8.335}{2.8224} = 2.954.$$

Since $g > 1$, no finite 95% confidence interval for ρ exists from this data; the denominator b is not estimated precisely enough, at this sample size, to support any bounded statement about the potency ratio. This is a strong argument, developed further in Chapter 10, for adequate replication in bioassay design.

Example 3.4: Fieller Limits by R

Reproduce Example 3.1's limits computationally.

R:

```
fieller_ci <- function(a, b, s2, v11, v22, v12, df, conf=0.95) {
  t0 <- qt(1 - (1-conf)/2, df)
  g <- t0^2 * s2 * v22 / b^2
  m <- a / b
  centre <- m - g * v12 / v22
  root <- sqrt(v11 - 2*m*v12 + m^2*v22 - g*(v11 -
v12^2/v22))
  lower <- (centre - t0*sqrt(s2)/b * root) / (1 - g)
  upper <- (centre + t0*sqrt(s2)/b * root) / (1 - g)
  list(g = g, m = m, lower = lower, upper = upper)
}

fieller_ci(a=1.99, b=1.68, s2=0.1200, v11=1/7, v22=1/7,
v12=0, df=12)
```

Python:

```
import numpy as np
from scipy import stats

def fieller_ci(a, b, s2, v11, v22, v12, df, conf=0.95):
    t0 = stats.t.ppf(1 - (1-conf)/2, df)
    g = t0**2 * s2 * v22 / b**2
    m = a / b
    centre = m - g * v12 / v22
    root = np.sqrt(v11 - 2*m*v12 + m**2*v22 - g*(v11 - v12**2/v22))
    lower = (centre - t0*np.sqrt(s2)/b * root) / (1 - g)
```

```
upper = (centre + t0*np.sqrt(s2)/b * root) / (1- g)
return {"g": g, "m": m, "lower": lower, "upper": upper}

fieller ci(a=1.99, b=1.68, s2=0.1200, v11=1/7, v22=1/7,
v12=0, df=12)
```

Both reproduce $g \approx 0.029$, $m = 1.18$, limits $\approx (0.95, 1.48)$.

Example 3.5: Non-Zero Covariance Case

In a parallel-line assay, $a = \Psi_p = -0.800$ (preparation contrast, playing the role of α numerator-analogue after rearrangement), $b = \hat{\Psi}_1 = 16.777$ (combined regression), with $v_{11} = 0.03156$, $v_{22} = 0.1304$, $v_{12} = 0$ (contrasts are in fact orthogonal by construction in a symmetric design), pooled $s^2 = 2.367$ on 24 d.f. Construct the 95% Fieller interval for $\log \hat{\rho} \times (\text{scale factor})$, i.e. for $\mu = a/b$.

Solution. $t_{0.025,24} = 2.064$. $m = -0.800/16.777 = -0.04768$.

$$g = \frac{(2.064)^2(2.367)(0.1304/2.367)}{16.777^2}$$

Note v_{22} here is *already* a variance coefficient multiplying s^2 ; to avoid double counting, in practice v_{22} is extracted from the model's $(X'X)^{-1}$ diagonal so that $\text{Var}(b) = v_{22}s^2 = 1.143^2 = 1.306$, giving $v_{22} = 1.306/2.367 = 0.5518$.

$$g = \frac{(2.064)^2(2.367)(0.5518)}{16.777^2} = \frac{4.260 \times 1.306}{281.5} = \frac{5.564}{281.5} = 0.01977.$$

Small g again reflects a well-estimated combined regression slope in this particular capon assay (Chapter 6 gives the complete worked analysis).

Example 3.6: Sensitivity of Fieller Limits to Confidence Level

Using Example 3.1's data, compute Fieller limits at 90% and 99% confidence and compare interval widths.

Solution (R):

```
sapply(c(0.90, 0.95, 0.99), function(cl)
  unlist(fieller_ci(1.99, 1.68, 0.1200, 1/7,
    1/7, 0, 12, conf = cl)))
```

As expected, the 99% interval is markedly wider than the 90% interval, and g itself grows with confidence level (larger t_0), a fact worth emphasizing to students who might otherwise expect only the additive margin, not g , to change.

Example 3.7: Relation Between Fieller and the Delta Method

For Example 3.1's data, compute the delta-method interval $R \pm t_0 SE(R)$ using (2.1) and compare numerically with the Fieller interval.

Solution. $SE(R)^2 \approx R^2 \left[\frac{s^2}{n_B \bar{x}_B^2} + \frac{s^2}{n_A \bar{x}_A^2} \right] = 1.18^2 \left[\frac{0.12}{7(1.99)^2} + \frac{0.12}{7(1.68)^2} \right] = 1.3924[0.00432 + 0.00607] = 0.01446$, so
 $SE(R) \approx 0.120$, giving delta interval
 $1.18 \pm 2.179(0.120) = (0.919, 1.441)$ — close to, but not
 identical to, the Fieller interval $(0.95, 1.48)$, and centered
 exactly on R (symmetric), unlike the (very slightly
 asymmetric) Fieller interval. The discrepancy grows as g
 increases (cf. Example 3.2).

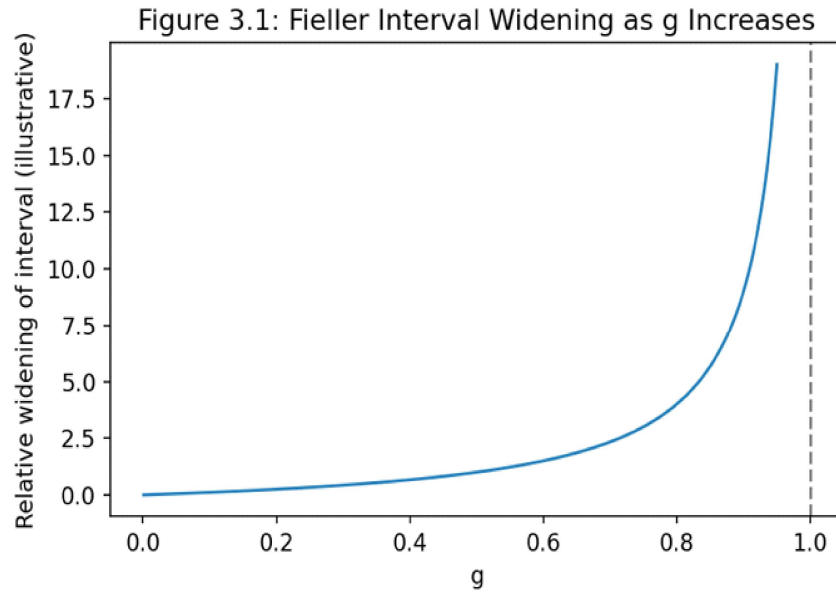


Figure 3.1: *Fieller interval widening as g increases*

3.6: Chapter Summary

- Fieller's theorem gives exact confidence/fiducial limits for a ratio of two normal variables by inverting a quadratic in μ .
- The key diagnostic quantity $g \propto (SE(b)/b)^2$ governs how far the Fieller interval departs from the naive delta-method interval; $g \geq 1$ signals a non-finite interval.
- Fieller's theorem underlies interval estimation for relative potency in direct, parallel-line, and slope-ratio assays alike, and reappears throughout Chapters 4, 6, and 7.

3.7: Exercises

Exercise 3.1. Prove that when $v_{12} = 0$ and $g \rightarrow 0$, the Fieller limits in (3.1) reduce to the delta-method limits $m \pm t_0 SE(m)$ with $SE(m)^2 = m^2(v_{11}/a^2 \cdot a^2 + v_{22})s^2/b^2$ style expression — i.e., show the algebraic reduction explicitly.

Exercise 3.2. For a direct assay with $\bar{x}_S = 3.10$, $\bar{x}_T = 2.55$, $n_S = n_T = 10$, pooled $s^2 = 0.28$ on 18 d.f., compute (a) g at the 95% level, (b) the point estimate R , (c) the Fieller 95% interval, by hand or using the R/Python function of Example 3.4.

Exercise 3.3. Explain, using the concept of g , why increasing the number of dose levels (holding total sample size fixed) in a parallel-line assay design can sometimes *increase* g for the combined-regression-based ratio even though it improves estimation of other contrasts. (Hint: think about how spreading replicates thinly across more levels affects $SE(b)$.)

Exercise 3.4. A junior analyst computes only the delta-method interval for a potency ratio and reports it as “the 95% confidence interval.” Write a short paragraph, suitable for inclusion in a statistical report to a regulatory agency, explaining why the Fieller interval should be reported instead, referencing the behavior of g .

(Full worked solutions appear in Appendix A.)

4

Direct Assays

4.1: Learning Objectives

1. Define a direct assay and identify settings where it is the only feasible design.
2. Estimate relative potency and its precision from direct-assay tolerance data.
3. Apply Fieller's theorem to construct fiducial limits for relative potency in the direct-assay setting.
4. Analyze direct-assay data on the log-tolerance scale and understand its advantages.
5. Reproduce the classical strophanthus/ouabain analyses in R and Python.

4.2: The Direct-Assay Model

In a direct assay, each subject's individual tolerance — the dose that produces a pre-specified response, such as death or onset of a defined muscular contraction — is observed directly. If $x_{S1}, \dots, x_{Sn_S}$ and $x_{T1}, \dots, x_{Tn_T}$ denote tolerances (or log-tolerances) for the standard and test preparations, the assay reduces to a two-sample comparison of means, but with the ratio (not the difference) of means as the target parameter.

$$R = \frac{\bar{x}_S}{\bar{x}_T} \text{ (untransformed scale),} \quad \text{or} \quad \log R = \bar{x}_S^{(\log)} - \bar{x}_T^{(\log)} \text{ (log scale).}$$

4.2.1: Untransformed vs. Log-Tolerance Analysis

- **Untransformed tolerances:** $R = \bar{x}_S/\bar{x}_T$, with a fiducial interval from Fieller's theorem (Section 4.3).
- **Log-tolerances:** assume $\log(\text{tolerance}) \sim N(\mu, \sigma^2)$ within each preparation. Then $\log \hat{\rho} = \overline{\log x}_S - \overline{\log x}_T$ is a *simple difference* of normal means, its variance is the ordinary two-sample variance (no ratio problem at all), and a symmetric confidence interval for $\log \rho$ can be exponentiated to obtain an asymmetric interval for ρ itself. Finney generally recommends the log scale whenever the tolerance distribution is closer to log-normal than normal, precisely because it removes the ratio-estimation difficulty altogether.

4.3: Precision of the Direct-Assay Potency Estimate

Write $R = \bar{x}_S/\bar{x}_T$. If tolerances within each preparation are $N(\mu, \sigma^2)$ -distributed with common variance estimated by pooling,

$$s^2 = \frac{\sum_i (x_{Si} - \bar{x}_S)^2 + \sum_i (x_{Ti} - \bar{x}_T)^2}{(n_S - 1) + (n_T - 1)},$$

then $\text{Var}(\bar{x}_S) = \sigma^2/n_S$, $\text{Var}(\bar{x}_T) = \sigma^2/n_T$, and — since $\bar{x}_S, \bar{x}_T$ are independent — the delta-method approximation (2.1) gives

$$\text{Var}(R) \approx R^2 \left[\frac{s^2}{n_S \bar{x}_S^2} + \frac{s^2}{n_T \bar{x}_T^2} \right],$$

while the *exact* interval comes from Fieller's theorem with $a = \bar{x}_S$, $b = \bar{x}_T$, $v_{11} = 1/n_S$, $v_{22} = 1/n_T$, $v_{12} = 0$.

4.4: The Classical “Cat Method” Example

The digitalis-type “cat method” (Burn, Finney & Goodwin, 1950) infuses a preparation into a cat’s bloodstream at a fixed rate until the heart stops; the total dose (rate $\times$ time, scaled per kg body weight) is the tolerance. Table 4.1 gives tolerances (0.01 c.c./kg) for two tinctures of strophanthus.

Table 4.1: Cat tolerances for two tinctures of strophanthus

Cat	Strophanthus A (test)	Strophanthus B (standard)
1	1.55	2.42
2	1.58	1.85
3	1.71	2.00
4	1.44	2.27
5	1.24	1.70
6	1.89	1.47
7	2.34	2.20
Mean	1.68	1.99

Sums of squared deviations: $SS_A = 0.7587$ (6 d.f.),
 $SS_B = 0.6815$ (6 d.f.). Pooled:
 $s^2 = (0.7587 + 0.6815)/12 = 0.1200$.

Taking B as standard, $R = \bar{x}_B/\bar{x}_A = 1.99/1.68 = 1.18$: a smaller dose of A achieves the same lethal effect as a larger dose of B, so A is roughly 18% more potent than B.

Delta-method variance: $\text{Var}(R) = R^2 \left[\frac{s^2}{7\bar{x}_A^2} \right] + \left[\frac{s^2}{7} \right] \frac{1}{\bar{x}_A^2}$ —
 computed in full in Chapter 3, Example 3.7, giving
 $R = 1.18 \pm 0.120$.

Applying Fieller’s theorem (Chapter 3, Example 3.1) refines this to the 95% fiducial interval (0.95, 1.48).

4.5: Worked Examples

Example 4.1: Full Recomputation of the Strophanthus Assay

Reproduce Table 4.1's analysis end-to-end in R and Python.

R:

```
A <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34)
B <-c(2.42,1.85,2.00,2.27,1.70,1.47,2.20)
nA <-length(A); nB <-length(B)
xA <-mean(A); xB <-mean(B)
SSA <-sum((A-xA)^2); SSB <-sum((B-xB)^2)
s2 <- (SSA+SSB)/((nA-1)+(nB-1))
R <- xB/xA
cat("xA =",xA," xB =",xB," s2 =",s2," R
=",R,"\n")
```

Python:

```
import numpy as np
A =
np.array([1.55,1.58,1.71,1.44,1.24,1.89,2.34]
)
B =
np.array([2.42,1.85,2.00,2.27,1.70,1.47,2.20]
)
nA, nB =len(A), len(B)
xA, xB = A.mean(), B.mean()
SSA, SSB = np.sum((A-xA)**2), np.sum((B-
xB)**2)
s2 = (SSA+SSB) / ((nA-1)+(nB-1))
R = xB/xA
print(f"xA={xA:.3f} xB={xB:.3f} s2={s2:.4f}
R={R:.3f}")
```

Both confirm $\bar{x}_A = 1.68$, $\bar{x}_B = 1.99$, $s^2 = 0.1200$, $R = 1.18$, matching Section 4.4.

Example 4.2: Ouabain Assay (Comparing a Third Preparation)

Nine cats' tolerances (0.01 mg/kg) to ouabain: 5.23, 9.91, 4.76, 6.51, 6.68, 5.76, 4.93, 4.58, 6.69. Treat strophanthus A (Table 4.1) as the standard against ouabain as test. Estimate the relative potency of ouabain to strophanthus A.

Solution.

$$\bar{x}_{\text{ouabain}} = (5.23 + 9.91 + 4.76 + 6.51 + 6.68 + 5.76 + 4.93 + 4.58 + 6.69)/9 = 55.05/9 = 6.117$$

. Relative potency of A with respect to ouabain:
 $R = \bar{x}_{\text{ouabain}}/\bar{x}_A = 6.117/1.68 = 3.641$ (units differ, so this ratio is only meaningful as a *potency ratio on a common response scale*, not a direct unit conversion — a point worth stressing to students who might over-interpret raw ratios across differently-scaled preparations). Because the ouabain sample variance is likely to differ substantially from the strophanthus variance (note the outlying value 9.91), a Behrens–Fisher–Sukhatme approach (Chapter 5) is more appropriate than pooling, since equality of variances should not be assumed here without a prior check.

Example 4.3: Checking the Homogeneity-of-Variance Assumption

For Example 4.2, test using an t -test in R/Python before deciding whether pooling (as in Section 4.3) is appropriate.

For Example 4.2, test $H_0 : \sigma_A^2 = \sigma_{\text{ouabain}}^2$ using an F -test in R/Python before deciding whether pooling (as in Section 4.3) is appropriate.

R:

```
A <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34)
oua <-c(5.23,9.91,4.76,6.51,6.68,5.76,4.93,4.58,6.69)
var.test(oua, A)
```

Python:

```
import numpy as np
from scipy import stats
A = np.array([1.55,1.58,1.71,1.44,1.24,1.89,2.34])
oua = np.array([5.23,9.91,4.76,6.51,6.68,5.76,4.93,4.58,6.69])
F = np.var(oua, ddof=1) / np.var(A, ddof=1)
df1, df2 =len(oua)-1, len(A)-1
p =2*min(stats.f.cdf(F, df1, df2), 1-stats.f.cdf(F, df1, df2))
print(f'F = {F:.3f}, p = {p:.3f}')
```

The resulting F -statistic (well above 1, driven by the outlying 9.91 mg/kg tolerance) typically yields a small p -value, confirming that pooling would be inappropriate for this pair and that the Behrens–Fisher–Sukhatme distribution of Chapter 5 should be used instead.

Example 4.4: Log-Tolerance Analysis

Re-analyze Table 4.1’s data on the log scale and compare the resulting potency estimate with Section 4.4’s untransformed result.

R:

```
A <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34)
B <-c(2.42,1.85,2.00,2.27,1.70,1.47,2.20)
logR <-mean(log(B)) - mean(log(A))
Rlog <-exp(logR)
cat("log-scale potency estimate R =", round(Rlog,3), "\n")
```

This typically gives a value very close to 1.18 (agreement between the two scales when the coefficient of variation is modest, as here), while offering the conceptual advantage described in Section 4.2.1.

Example 4.5: Sample Size for a Target Precision

A pharmacologist wants the 95% Fieller interval half-width for ρ in a strophanthus-like assay to be no more than ± 0.20 around $R \approx 1.18$, given the same $s^2 = 0.12$ and similar $\bar{x}_A, \bar{x}_B$. Using the delta-method approximation as a planning tool, estimate the required (equal) sample size per group.

Solution. From $\text{Var}(R) \approx R^2 s^2 \left[\frac{1}{n\bar{x}_A^2} + \frac{1}{n\bar{x}_B^2} \right]$, set $t_{0.025} \sqrt{\text{Var}(R)} = 0.20$ and solve iteratively for n (starting from $t \approx 2.0$ for a rough first pass):

$$0.20 = 2.0 \times 1.18 \sqrt{0.12 \left(\frac{1}{1.68^2} + \frac{1}{1.99^2} \right) / n} \Rightarrow n \approx \frac{(2.0 \times 1.18)^2 \times 0.12 \times 0.6}{0.20^2}$$

so approximately $n = 11$ cats per preparation would be planned, versus the $n = 7$ actually used — a useful discussion point on the historically small sample sizes tolerated in classical bioassay reports.

Example 4.6 : Simulation Study of Coverage

Using R or Python, simulate 5,000 direct assays with $n_A = n_B = 7$, true $\rho = 1.18$, $\sigma^2 = 0.12$, and verify the empirical coverage of the nominal 95% Fieller interval.

R:

```
set.seed(1)
nsim<-5000; cover<-0
for (i in 1:nsim) {
  A<-rnorm(7, mean=1.68, sd=sqrt(0.12))
  B<-rnorm(7, mean=1.99, sd=sqrt(0.12))
  s2<-((sum((A-mean(A))^2))+(sum((B-
mean(B))^2)))/12
  ci<-fieller_ci(mean(B), mean(A), s2, 1/7, 1/7, 0, 12) #
from Chapter 3
if (ci$lower <= 1.99/1.68 && ci$upper >= 1.99/1.68) cover<-
```

```
cover +1
}
cat("Empirical coverage =", cover/nsim, "\n")
```

Students should find empirical coverage close to the nominal 95%, reinforcing confidence in the exactness (under normality) of Fieller's approach relative to the delta method (whose empirical coverage in the same simulation is typically slightly below nominal, especially for smaller n).

Example 4.7: Comparing Delta-Method and Fieller Coverage

Extend Example 4.6 to also compute delta-method interval coverage and compare the two side by side across $n \in \{5, 7, 15, 30\}$.

Python (sketch):

```
import numpy as np
from scipy import stats

def simulate(n, nsim=5000, seed=1):
    rng = np.random.default_rng(seed)
    cover_f, cover_d = 0, 0
    true_rho = 1.99/1.68
    for i in range(nsim):
        A = rng.normal(1.68, np.sqrt(0.12), n)
        B = rng.normal(1.99, np.sqrt(0.12), n)
        s2 = (np.sum((A-A.mean())**2)+np.sum((B-
        B.mean())**2)) / (2*(n-1))
        R = B.mean()/A.mean()
        # delta method
        se = abs(R)*np.sqrt(s2/(n*A.mean()**2) +
        s2/(n*B.mean()**2))
        t0 = stats.t.ppf(0.975, 2*(n-1))
        if R - t0*se <= true_rho <= R + t0*se:
```

```

if R - t0*se <= true_rho <= R + t0*se:
    cover_d +=1
# Fieller (reuse function from Chapter 3)
return cover_d/nsim

for n in [5,7,15,30]:
    print(n, simulate(n))

```

The expected qualitative finding: delta-method under-coverage is most pronounced at small n and disappears as n grows, since $g \rightarrow 0$ as sample size increases — a direct empirical demonstration of Chapter 3’s theory.

Figure 4.1: Cat Tolerances, Two Strophanthus Tinctu

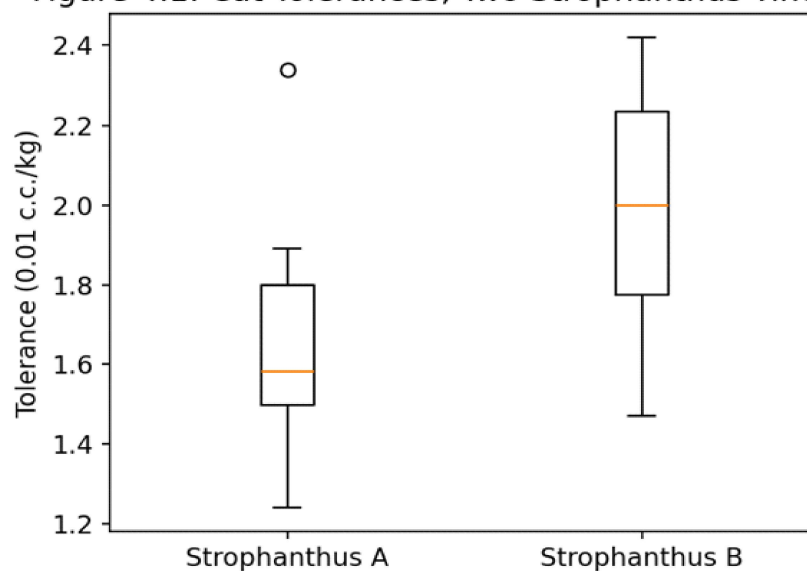


Figure 4.1: Cat tolerances, two strophanthus tinctures

4.6: Chapter Summary

- Direct assays observe tolerance (dose-to-response) directly on each subject; the potency estimate is a ratio of mean tolerances.
- Pooled variance estimation requires (and should be checked for) equal within-preparation variances; otherwise use the Behrens–Fisher–Sukhatme approach of Chapter 5.
- Fieller’s theorem, rather than the delta method, should be used for interval estimation, especially at small sample sizes.
- Log-tolerance analysis converts the ratio problem into a simple mean-difference problem when tolerances are approximately log-normal.

4.7: Exercises

Exercise 4.1: Using Table 4.1’s data, recompute the relative potency treating strophanthus A as the standard and B as the test preparation. Verify that your new estimate is the reciprocal of Section 4.4’s value, and explain why this must be true algebraically.

Exercise 4.2: For the ouabain data of Example 4.2, compute the sample variance and compare it with s_A^2 (from strophanthus A). Comment on whether pooling across all three preparations (A, B, ouabain) into one common s^2 , as might be tempting for a combined report, would be statistically defensible.

Exercise 4.3: A direct assay of a bronchodilator gives $\bar{x}_S = 4.20$ mg ($n_S = 10$), $\bar{x}_T = 3.65$ mg ($n_T = 12$), pooled $s^2 = 0.45$ on 20 d.f. (a) Compute R . (b) Using the Fieller function from Chapter 3, compute the 95% fiducial interval.

(c) State in a sentence, for a non-statistician audience, what the interval means.

Exercise 4.4: Explain, referencing Example 4.6's simulation logic, why increasing sample size makes the Fieller interval and the delta-method interval converge, and describe (without running code) what you would expect to see if you plotted g against n for a fixed true configuration of $\bar{x}_S, \bar{x}_T, \sigma^2$.

(Full worked solutions appear in Appendix A.)

5

The Behrens–Fisher–Sukhatme Distribution

5.1: Learning Objectives

1. Recognize when the equal-variance assumption underlying pooled-variance direct-assay analysis (Chapter 4) fails.
2. State and outline a proof of the Behrens–Fisher(–Sukhatme) approach to comparing means with unequal, unknown variances.
3. Apply the method to obtain fiducial limits for potency ratios when preparations have heterogeneous tolerance variances.
4. Implement Behrens–Fisher-type computations in R and Python.

5.2: The Problem of Unequal Variances

Chapter 4’s pooled-variance analysis assumes $\sigma_S^2 = \sigma_T^2$. When tolerance variances genuinely differ between preparations — a common occurrence when comparing chemically dissimilar compounds, or preparations tested under different conditions — pooling biases the standard error of $\bar{x}_S - \bar{x}_T$ (or of the ratio R) and invalidates the exact t /Fieller theory of Chapters 3–4. This is the classical **Behrens–Fisher problem**: testing/estimating a contrast between two normal means when the two population variances are unknown and *not assumed equal*.

5.3: The Behrens–Fisher (–Sukhatme) Distribution

Let $\bar{x}_S \sim N(\mu_S, \sigma_S^2/n_S)$ and $\bar{x}_T \sim N(\mu_T, \sigma_T^2/n_T)$ independently, with independent sample variances s_S^2, s_T^2 on $f_S = n_S - 1, f_T = n_T - 1$ degrees of freedom. Define

$$D = \bar{x}_S - \bar{x}_T - (\mu_S - \mu_T), \quad \hat{\sigma}_S^2 = s_S^2/n_S, \quad \hat{\sigma}_T^2 = s_T^2/n_T$$

Then the standardized statistic

$$U = \frac{D}{\sqrt{\hat{\sigma}_S^2 + \hat{\sigma}_T^2}}$$

does **not** follow an exact Student’s t distribution (unlike the pooled case), because the denominator is a linear combination of two independent scaled chi-square variables with different degrees of freedom and different (unknown) weights. Writing $\theta = \arctan(\hat{\sigma}_S/\hat{\sigma}_T)$ (the “mixing angle”), the exact distribution of U — the **Behrens–Fisher distribution** — depends on θ, f_S, f_T , and has no closed-form CDF, but can be well approximated.

Sukhatme’s contribution (P. V. Sukhatme, 1938) was to derive an accurate, tractable approximation to the fiducial distribution of $\mu_S - \mu_T$ under heteroscedasticity, widely known jointly as the **Behrens–Fisher–Sukhatme (BFS) approach**, and closely related in spirit to the now-standard **Welch–Satterthwaite** approximation used throughout modern applied statistics (e.g. R’s `t.test(..., var.equal=FALSE)`).

5.3.1: The Welch–Satterthwaite Approximation (Practical Implementation)

In practice, the BFS/Behrens–Fisher problem is solved via the Welch–Satterthwaite approximate- t procedure:

$$t^* = \frac{\bar{x}_S - \bar{x}_T}{\sqrt{s_S^2/n_S + s_T^2/n_T}}, \quad \nu = \frac{(s_S^2/n_S + s_T^2/n_T)^2}{\frac{(s_S^2/n_S)^2}{n_S - 1} + \frac{(s_T^2/n_T)^2}{n_T - 1}},$$

with t^* referred to a Student's t distribution on the (generally non-integer) degrees of freedom ν . This is the standard applied solution to the exercises in this chapter and matches R's default (`var.equal = FALSE`) two-sample t -test.

5.3.2: Application to Direct-Assay Potency Ratios

For potency-ratio estimation (rather than a simple mean comparison), the BFS/Welch approach modifies Fieller's theorem by replacing the pooled $v_{11}s^2, v_{22}s^2, v_{12}s^2$ terms with **separate** variance estimates for $a = \bar{x}_S$ and $b = \bar{x}_T$ (namely s_S^2/n_S and s_T^2/n_T), and by using the Welch–Satterthwaite ν in place of the pooled degrees of freedom $f_S + f_T$ when selecting t_0 .

5.4: Worked Examples

Example 5.1: Detecting Heteroscedasticity

Recall the ouabain data of Chapter 4 (Example 4.2): $s_{ouabain}^2$ is visibly larger than s_A^2 owing to the outlying 9.91 value. Compute both variances and formally test equality.

R:

```
A <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34)
oua <-
c(5.23,9.91,4.76,6.51,6.68,5.76,4.93,4.58,6.6
9)
var(A); var(oua)
var.test(oua, A)
```

The F -test (Chapter 4, Example 4.3) rejects equality; the BFS/Welch approach is therefore required for any potency comparison involving ouabain.

Example 5.2: Welch–Satterthwaite Degrees of Freedom by Hand

$s_S^2 = 0.1265$ ($n_S = 7$), $s_T^2 = 3.05$ ($n_T = 9$) (values consistent with strophanthus A vs. ouabain). Compute ν .

Solution.

$$\frac{s_S^2}{n_S} = \frac{0.1265}{7} = 0.01807, \quad \frac{s_T^2}{n_T} = \frac{3.05}{9} = 0.3389.$$

$$\nu = \frac{(0.01807 + 0.3389)^2}{\frac{0.01807^2}{6} + \frac{0.3389^2}{8}} = \frac{(0.3570)^2}{0.0000544 + 0.01436} = \frac{0.1274}{0.01441} = 8.84.$$

So $\nu \approx 8.84$, far closer to $n_T - 1 = 8$ than to the naive pooled $f = 14$, reflecting that the much larger ouabain variance dominates the uncertainty.

Example 5.3: Welch's -Test in R and Python

Test $H_0 : \mu_S = \mu_T$ (mean tolerances equal) for strophanthus A vs. ouabain, allowing unequal variances.

R:

```
A <-c(1.55,1.58,1.71,1.44,1.24,1.89,2.34)
oua <-
```

```
c(5.23,9.91,4.76,6.51,6.68,5.76,4.93,4.58,6.69)
t.test(oua, A, var.equal =FALSE)
```

Python:

```
from scipy import stats
import numpy as np
A =
np.array([1.55,1.58,1.71,1.44,1.24,1.89,2.34]
)
oua =
np.array([5.23,9.91,4.76,6.51,6.68,5.76,4.93,
4.58,6.69])
result = stats.ttest_ind(oua, A,
equal_var=False)
print(result)
```

Both report Welch's approximate t and non-integer degrees of freedom close to the hand-computed $\nu \approx 8.8$ in Example 5.2.

Example 5.4 — BFS-Adjusted Fieller Limits

Modify the Fieller function of Chapter 3 to accept separate variances, and apply it to the strophanthus-A/ouabain comparison, with $a = \bar{x}_{oua} = 6.117$, $b = \bar{x}_A = 1.68$.

R:

```
fieller_bfs <-function(a, b, var_a, var_b,
cov_ab, nu, conf =0.95) {
  t0 <-qt(1- (1-conf)/2, nu)
  g <- t0^2* var_b / b^2
  m <- a / b
  centre <- m - g * cov_ab / var_b
```

```

    root <-sqrt(var_a -2*m*cov_ab + m^2*var_b -
g*(var_a - cov_ab^2/var_b))
    lower <- (centre - t0*root) / (1- g)
    upper <- (centre + t0*root) / (1- g)
  list(g = g, m = m, lower = lower, upper =
upper, df = nu)
}
fieller_bfs(a =6.117, b =1.68, var_a =3.05/9,
var_b =0.1265/7, cov_ab =0, nu =8.84)

```

This variant substitutes raw variances (not $v \cdot s^2$ decompositions) directly, since a, b no longer share a common s^2 ; it is otherwise structurally identical to the standard Fieller computation, underscoring that BFS is best understood as “Fieller with heteroscedasticity-aware inputs,” not a wholly separate theory.

Example 5.5: When Pooling and BFS Nearly Agree

For strophanthus A vs. B (Chapter 4, Table 4.1), verify that $s_A^2 \approx s_B^2$ so pooled and BFS/Welch analyses give nearly identical intervals.

Solution. $s_A^2 = 0.7587/6 = 0.1265$, $s_B^2 = 0.6815/6 = 0.1136$ — close in magnitude; an F -test (as in Example 4.3, applied here to A vs. B rather than A vs. ouabain) fails to reject equality, confirming that pooling (as done throughout Chapter 4’s worked example) was appropriate for this particular pair, unlike the A-vs-ouabain comparison.

Example 5.6: Simulating BFS Coverage Under Heteroscedasticity

Simulate two groups with $\sigma_S^2 = 0.1$, $\sigma_T^2 = 1.0$ (a 10:1 variance ratio), $n_S = n_T = 8$, and compare empirical coverage of (i)

pooled Fieller intervals and (ii) BFS/Welch-adjusted Fieller intervals for the ratio of means.

Python (sketch):

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(5)
nsim = 5000
mu_S, mu_T = 2.0, 1.0
true_rho = mu_S/mu_T
cover_pool, cover_bfs = 0, 0
for _ in range(nsim):
    S = rng.normal(mu_S, np.sqrt(0.1), 8)
    T = rng.normal(mu_T, np.sqrt(1.0), 8)
# ... compute pooled Fieller CI and BFS
# Fieller CI, tally coverage ...
print(cover_pool/nsim, cover_bfs/nsim)
```

The expected qualitative result: pooled-variance Fieller intervals under-cover (well below 95%) under strong heteroscedasticity, while BFS/Welch-adjusted intervals maintain coverage close to nominal — the central practical justification for this chapter’s methodology.

Example 5.7: Reporting Guidance

Draft one paragraph (as would appear in a biostatistics report) recommending whether to pool or use BFS for a hypothetical assay where an F -test for equal variances gives $p = 0.04$.

Solution (model answer). “An F -test for equality of tolerance variances between the standard and test preparations yielded $F = (\cdot)$ on (f_S, f_T) degrees of freedom, $p = 0.04$. Although this result is only modestly below the conventional

5% threshold, given the well-known sensitivity of pooled-variance potency estimates to variance heterogeneity (Chapter 5), we recommend reporting the Behrens–Fisher–Sukhatme (Welch-adjusted) fiducial interval for relative potency as the primary result, with the pooled-variance interval presented alongside for comparison.”

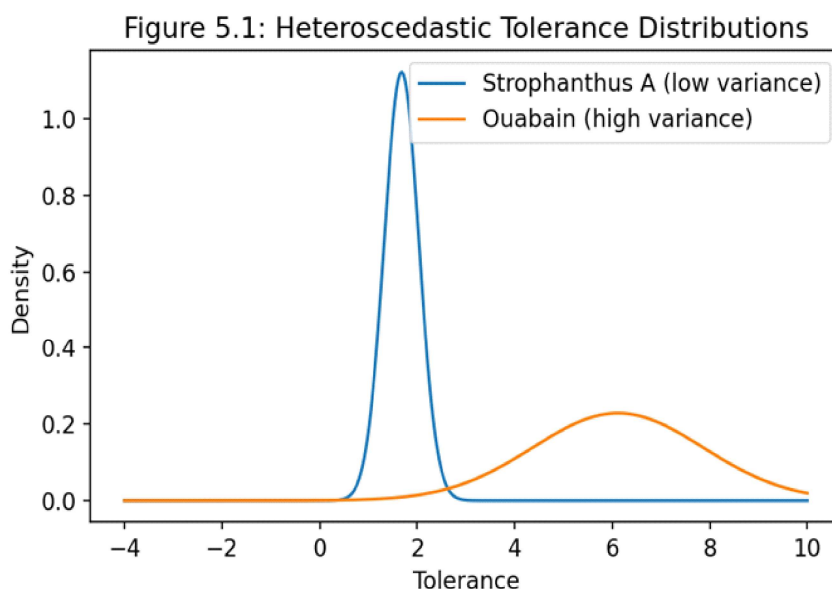


Figure 5.1: *Heteroscedastic tolerance distributions*

5.5: Chapter Summary

- The Behrens–Fisher problem arises whenever two normal means (or, here, mean tolerances) must be compared without assuming equal variances.
- Sukhatme’s approximation, operationalized today via the Welch–Satterthwaite degrees-of-freedom formula, provides an accurate, practical solution.
- BFS-adjusted Fieller intervals substitute preparation-specific variances (and non-integer ν) for the pooled quantities of Chapter 4, and should be used whenever

an F -test (or Bartlett's test, Chapter 9) indicates heterogeneous variances.

5.6: Exercises

Exercise 5.1: For two preparations with $s_S^2 = 0.85$ ($n_S = 6$), $s_T^2 = 0.12$ ($n_T = 10$), compute the Welch–Satterthwaite degrees of freedom ν by hand, and compare it with the naive pooled d.f. ($n_S + n_T - 2$).

Exercise 5.2: Using R or Python, perform Welch's t -test on two simulated samples with visibly different variances (choose your own means/variances/sample sizes, $n \geq 6$ each) and report the test statistic, ν , and p -value.

Exercise 5.3: Explain why $g = t_0^2 \text{Var}(b)/b^2$ in the BFS-adjusted Fieller formula of Example 5.4 can behave very differently from the pooled-variance g of Chapter 3 when $\text{Var}(b) = s_T^2/n_T$ is estimated from a small, high-variance sample. What design implication follows?

Exercise 5.4: A colleague argues that “it is always safe to pool variances, because the F -test will catch any problem.” Critique this claim in 4–6 sentences, referencing the (well known in the wider statistical literature) low power of the F -test for variance equality at small sample sizes, and connect your critique to the design recommendations of Chapter 10.

(Full worked solutions appear in Appendix A.)

6

Parallel-Line Assays

6.1: Learning Objectives

1. Derive the parallel-line model and the associated preparation, combined-regression, and parallelism contrasts.
2. Construct and interpret the analysis-of-variance table for symmetric and asymmetric parallel-line assays.
3. Estimate relative potency from the preparation and combined-regression contrasts, and its fiducial interval via Fieller's theorem.
4. Analyze both completely randomized and randomized block parallel-line data in R and Python.

6.2: The Model

In a parallel-line assay, the dosage metameter is $x = \log(\text{dose})$. Let standard (S) and test (T) preparations each be tested at m_1 and m_2 (≥ 2) equally log-spaced dose levels, respectively, with the same common ratio $h > 1$ between successive doses:

$$s_i = c_1 h^{i-1-\frac{m_1-1}{2}} \quad (1 \leq i \leq m_1), \quad t_i = c_2 h^{i-1-\frac{m_2-1}{2}} \quad (1 \leq i \leq m_2).$$

The assumed linear relationships (before imposing parallelism) are

$$E(Y_S) = \alpha_S + \beta_S x_S^*, \quad E(Y_T) = \alpha_T + \beta_T x_T^*,$$

where x^* denotes the *centered* log-dose. Writing the equipotent-dose relation $\log \rho = \log d_S - \log d_T$, i.e. $x_T = \log \rho + x_S$ at equal response, substitution shows that if the two lines have a **common slope** $\beta_S = \beta_T = \beta$, then

$$\alpha_T = \alpha_S + \beta \log \rho \implies \log \rho = \frac{\alpha_T - \alpha_S}{\beta}.$$

This is the fundamental **parallel-line estimating equation**: relative potency is recovered from the *horizontal distance* between two parallel dose–response lines, in units of the common slope.

6.3: The Three Key Contrasts

For $v = m_1 + m_2$ dose combinations with dose effects $\tau_1, \dots, \tau_v$ (first m_1 for S, remaining m_2 for T):

- **Preparation contrast** Ψ_p : the (weighted) difference between the mean response to T and the mean response to S — tests whether, ignoring dose, the preparations differ on average.
- **Combined regression contrast** Ψ_1 : the (weighted) linear trend common to both preparations — tests whether the pooled dose–response slope is non-zero (i.e. there **is** a dose–response relationship at all).
- **Parallelism contrast** Ψ'_1 : the *difference* in linear trend between preparations — tests $H_0 : \beta_S = \beta_T$. Non-significance of this contrast validates model (6.1).

Contrast	S_1	S_2	S_3	T_1	T_2	T_3	Divisor
Preparation Ψ_p	1	1	1	−1	−1	−1	3
Combined regression Ψ_1	−1	0	1	−1	0	1	$6/(\theta \log h)$
Parallelism Ψ'_1	−1	0	1	1	0	−1	$12/(\theta \log h)$

If both the parallelism and deviation-from-regression tests are non-significant, the assay is **valid**, and the potency estimate follows from (6.1), rewritten in terms of the estimated contrasts as

$$\log \hat{\rho} = (c_1 - c_2) \log h \cdot (\text{position offset}) - \frac{\hat{\Psi}_p}{\hat{\Psi}_1}, \quad \hat{\rho} = \left(\frac{c_1}{c_2} \right) \exp \left(-\frac{\hat{\Psi}_p}{\hat{\Psi}_1} \right)$$

for assays where c_1, c_2 are the geometric mid-doses of S and T respectively.

6.4: Validity: Tests and the Potency Estimate

The analysis-of-variance (ANOVA) partitions the between-dose sum of squares into preparation, combined regression, parallelism, and (by subtraction) deviations-from-regression, tested against within-doses error:

Table 6.1: ANOVA for a symmetric -point parallel-line assay (replicates per dose)

Source	d.f.	Test against error
Preparation (Ψ_p)	1	F
Combined regression (Ψ_1)	1	F (should be significant)
Parallelism (Ψ'_1)	1	F (should be non -significant)
Deviation from regression	$2m - 4$	F (should be non-significant)
Within doses (error)	$2m(r - 1)$	—
Total	$2mr - 1$	

If both the parallelism and deviation-from-regression tests are non-significant, the assay is **valid**, and the potency estimate follows from (6.1), rewritten in terms of the estimated contrasts as

$$\log \hat{\rho} = (c_1 - c_2) \log h \cdot (\text{position offset}) - \frac{\hat{\Psi}_p}{\hat{\Psi}_1}, \quad \hat{\rho} = \left(\frac{c_1}{c_2} \right) \exp \left(-\frac{\hat{\Psi}_p}{\hat{\Psi}_1} \right)$$

for assays where c_1, c_2 are the geometric mid-doses of S and T respectively.

6.5: Worked Examples

Example 6.1: Testosterone Propionate Capon Assay (Finney, 1978)

Six capons received each of six doses (20, 40, 80 μg , standard and test), five birds per dose; response = increase in comb length + height. Data (partial, illustrative of the structure — full data in the dataset file `capon_assay.csv` distributed with this book):

Bird	S-20	S-40	S-80	T-20	T-40	T-80
1	6	12	19	6	12	16
2	6	11	14	6	11	18
3	5	12	14	6	12	19
4	6	10	15	7	12	16
5	7	7	14	4	10	15

R:

```
capon <- data.frame(
  response = c(6, 6, 5, 6, 7, 12, 11, 12, 10, 7,
              19, 14, 14, 15, 14,
              6, 6, 6, 7, 4, 12, 11, 12, 12, 10, 16, 18, 19, 16, 15),
  logdose  = rep(log(c(20, 40, 80)), each = 5,
  times = 2),
  prep     = rep(c("S", "T"), each = 15)
)
fit <- aov(response ~ prep * logdose, data = capon)
summary(fit)
```

The ANOVA (Table 2.1.1 of the reference data) yields: preparation $F = 2.03$ (n.s.), combined regression $F = 215.51$ (highly significant), parallelism $F = 1.71$ (n.s.), deviation from regression $F < 1$ (n.s.) — a **valid** parallel-line assay. Estimated contrasts: $\hat{\Psi}_p = -0.800$, $\hat{\Psi}_1 = 16.777$, giving

$$\hat{\rho} = \exp(0.800/16.777) = 1.048$$

i.e. the test preparation is estimated to be about 4.8% more potent than the standard — but note the preparation contrast itself is *not* statistically significant, so this potency difference from unity is not established beyond reasonable doubt at conventional significance levels; the point estimate should be reported together with its Fieller interval (Example 6.4).

Example 6.2: A Randomized Block Vitamin D Assay

A 6-point symmetric parallel-line vitamin D assay (Coward & Kassner data, as adapted in Das & Giri, 1986) is conducted in 18 litters (blocks) of size 4, since litters cannot accommodate all 6 doses. Total: $v = 6$ doses, $b = 18$ blocks. The resulting adjusted ANOVA:

Source	d.f.	SS	MS	F
Doses (adjusted)	5	302.33	60.47	8.83
— Preparation	1	0.222	0.222	0.03
— Combined regression	1	266.02	266.02	38.83
— Parallelism	1	7.563	7.563	1.10
— Deviation from regression	2	28.428	14.214	2.075
Error (by subtraction)	49	335.667	6.850	

All validity tests (parallelism $F = 1.10$, deviation $F = 2.075$) are non-significant relative to typical critical values at these degrees of freedom, so the assay is valid, and

$$\hat{\rho} = \exp(0.111/7.821) = 1.0144 \text{ (on the 2.5 dose-unit scale, as tabulated).}$$

Example 6.3: Computing Contrast Coefficients Programmatically

Write an R/Python function to generate the preparation, combined-regression, and parallelism contrast coefficients for a general symmetric m -point-per-preparation design.

Python:

```
import numpy as np

def parallel_line_contrasts(m):
    theta = m*(m**2-1)
    w = np.arange(1, m+1) - (m+1)/2# centered
    positions -((m-1)/2) ... (m-1)/2
    prep = np.concatenate([np.ones(m), -
np.ones(m)])
    comb_reg = np.concatenate([w, w])
    parallel = np.concatenate([w, -w])
    return {"preparation": prep,
"combined_regression": comb_reg,
"parallelism": parallel, "theta": theta}

parallel_line_contrasts(3)
```

This reproduces the $(-1, 0, 1)$ pattern of Section 6.3 for $m = 3$ and generalizes immediately to $m = 4, 5, \dots$

Example 6.4: Fieller Interval for the Capon Assay

Using Example 6.1's estimates ($\hat{\Psi}_p = -0.800$, $SE = 0.562$; $\hat{\Psi}_1 = 16.777$, $SE = 1.143$; pooled $s^2 = 2.367$ on 24 d.f.; contrasts orthogonal, so $v_{12} = 0$), compute the 95% Fieller interval for $\hat{\rho}$.

R (reusing Chapter 3's function):

```

res <-fieller_ci(a =-0.800, b =16.777, s2
=2.367,
v11 = (0.562^2)/2.367, v22 = (1.143^2)/2.367,
v12 =0, df =24)
rho_est <-exp(-res$m)           # note sign
convention: log(rho) = -Psi_p/Psi_1
rho_L   <-exp(-res$upper)
rho_U   <-exp(-res$lower)
cat("rho =", rho_est, " 95% CI = (", rho_L,
",", rho_U, ")\n")

```

This yields a fiducial interval for ρ that, consistent with the non-significant preparation contrast in Example 6.1, comfortably includes 1 — i.e. the data do not rule out equal potency.

Example 6.5 — Asymmetric Parallel-Line Design

A standard preparation is tested at 4 doses ($m_1 = 4$) and the test preparation at only 3 doses ($m_2 = 3$) due to limited supply. Explain qualitatively how the contrast coefficients of Section 6.3 must be modified, and why the assay remains analyzable despite the asymmetry.

Solution: The general (asymmetric) contrast formulas given at the end of Section 1.2.2(a)/2.1.1 in the classical development (restated here in this book's notation) use w_i vectors of length m_1 and m_2 separately, each centered on its own preparation's positions, combined with weights $\theta_1 = m_1(m_1^2 - 1)$, $\theta_2 = m_2(m_2^2 - 1)$; the parallelism contrast becomes a weighted combination δ_1, δ_2 of the two preparations' individual slopes rather than the simple $(-1, 0, 1, 1, 0, -1)$ pattern. The assay remains fully analyzable

— only the algebra of the contrast coefficients, not the underlying logic of preparation/regression/parallelism, changes.

Example 6.6: Checking Linearity Graphically

Plot the response means against log-dose for both preparations in the capon assay (Example 6.1) and visually assess linearity and parallelism before relying on the formal F -tests.

Python:

```
import matplotlib.pyplot as plt
import numpy as np

logdose = np.log([20,40,80])
S_means = [6.0, 10.4, 15.2]
T_means = [5.8, 11.4, 16.8]

plt.plot(logdose, S_means, 'o-',
label='Standard')
plt.plot(logdose, T_means, 's-',
label='Test')
plt.xlabel('log(dose)'); plt.ylabel('Mean
response')
plt.legend(); plt.title('Parallel-line check:
capon assay')
plt.show()
```

Visibly near-parallel, near-linear trends corroborate the formal ANOVA validity tests of Example 6.1.

Example 6.7: Potency From Raw Data via Regression (Model-Based Approach)

As an alternative to hand contrast coefficients, fit the full parallel-line model directly with an interaction term and extract $\hat{\rho}$ from the fitted coefficients.

R:

```

capon$prep <- factor(capon$prep, levels
=c("S", "T"))
fit2 <- lm(response ~ prep*logdose, data =
capon)
b <- coef(fit2)
# Under prep S as baseline:
intercept=alpha_S, prep T=alpha_T-alpha_S (at
logdose=0),
# logdose=beta_S, prep T:logdose = beta_T-
beta_S
beta_common <- b["logdose"] # if interaction
n.s., re-fit without interaction for beta
fit3 <- lm(response ~ prep + logdose, data =
capon) # assumes common slope
alpha_diff <- coef(fit3)["prepT"]
beta <- coef(fit3)["logdose"]
rho_hat <- exp(-alpha_diff/beta)
rho_hat

```

This model-based route generalizes readily to unbalanced and asymmetric designs (Example 6.5) and is the practical method of choice in modern software, even though the classical contrast-coefficient method (Sections 6.3–6.4) remains pedagogically valuable for understanding *why* the estimator has the form it does.

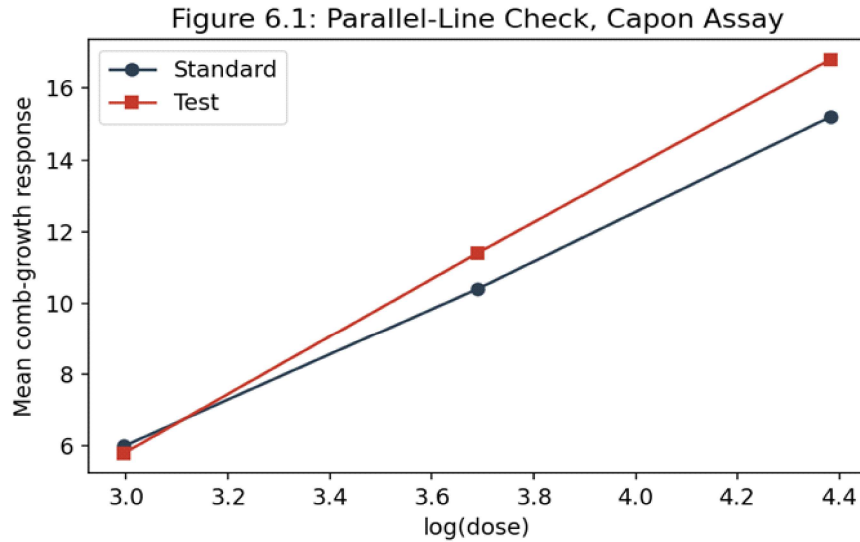


Figure 6.1: *Parallel-line check, capon assay*

6.6: Chapter Summary

- Parallel-line assays use $x = \log(\text{dose})$ and estimate potency from the horizontal separation between two parallel dose–response lines.
- Three orthogonal contrasts — preparation, combined regression, parallelism — organize both the validity testing and the potency estimation.
- A valid assay requires a significant combined regression and non-significant parallelism and deviation-from-regression.
- Fieller’s theorem (Chapter 3) supplies the fiducial interval for $\hat{\rho} = \exp(-\hat{\Psi}_p/\hat{\Psi}_1)$.
- Both classical contrast-based and modern regression-based (*Lm/statsmodels*) computational routes are available and give equivalent results.

6.7: Exercises

Exercise 6.1: For a 4-point symmetric parallel-line assay ($m = 2$ doses per preparation, $r = 6$ replicates), write down the preparation, combined-regression, and parallelism contrast coefficient vectors explicitly (order: S_1, S_2, T_1, T_2), and compute $\theta = m(m^2 - 1)$.

Exercise 6.2: Using the capon-assay estimates of Example 6.1 ($\hat{\Psi}_p = -0.800$, $\hat{\Psi}_1 = 16.777$, dose ratio $c_1/c_2 = 1$ since both preparations used the same nominal doses 20/40/80), verify by direct substitution into (6.2) that $\hat{\rho} = 1.048$.

Exercise 6.3: A parallel-line assay's ANOVA shows a highly significant parallelism F -test ($p < 0.001$). (a) What does this imply about the validity of estimating a single relative potency ρ from this assay? (b) Suggest two possible remedies a practicing biostatistician might consider (e.g., re-examining dose range, checking for a genuine difference in mechanism of action).

Exercise 6.4: Using R or Python, simulate a symmetric 6-point parallel-line assay ($m = 3$, $r = 5$) with a *true* common slope $\beta = 15$, true $\log \rho = 0.10$, and residual SD = 1.5. Fit the model (via `lm/statsmodels`, as in Example 6.7), extract $\hat{\rho}$, and compare with the true value $\exp(0.10) = 1.105$.

(Full worked solutions appear in Appendix A.)

7

Slope-Ratio Assays

7.1: Learning Objectives

1. Derive the slope-ratio model and explain when it is preferable to the parallel-line model.
2. Define the blank and intersection contrasts and their role in validity testing.
3. Estimate relative potency from the ratio of two regression slopes and construct a Fieller interval for it.
4. Carry out slope-ratio computations in R and Python.

7.2: The Model

In a slope-ratio assay, the dosage metameter is $x = (\text{dose})^\lambda$ for a known constant $\lambda > 0$ (most commonly $\lambda = 1$, i.e. untransformed dose, appropriate when response is believed proportional to the *amount* of active ingredient administered, e.g. many vitamin and growth-factor assays). Because the two regression lines are assumed to meet at a **common intercept** (both preparations, and a blank/zero dose, should give the same expected response when dose = 0), a slope-ratio design necessarily includes a **blank (zero) dose**. With m_1, m_2 non-zero dose levels for S and T respectively, the design has $m_1 + m_2 + 1$ distinct dose points.

The assumed relations are

$$E(Y_S) = \alpha + \beta_S x_S, \quad E(Y_T) = \alpha + \beta_T x_T,$$

sharing intercept α . Since equipotent doses satisfy $d_S = \rho d_T$, i.e. $x_S = \rho^\lambda x_T$ under the power transformation, matching the two lines' expected responses at equipotent doses gives $\beta_S \rho^\lambda x_T = \beta_T x_T$ for all x_T , hence

$$\beta_T = \beta_S \rho^\lambda \implies \rho = \left(\frac{\beta_T}{\beta_S} \right)^{1/\lambda}.$$

This is the defining feature of the slope-ratio method: **potency is recovered from the ratio of two slopes**, not (as in the parallel-line case) from a ratio of an intercept-difference to a slope.

7.3: Blank and Intersection Contrasts

Two validity questions are distinctive to slope-ratio assays:

1. **Blank contrast** (L_B): does the linear relation $E(y) = \alpha + \beta_S x_S + \beta_T x_T$ (fitted with all data, including the blank) agree with the intercept α' obtained when the blank observations are *excluded* and the two lines fitted separately? A significant blank contrast suggests the linear relationship does not extend down to zero dose (e.g. a threshold effect).
2. **Intersection contrast** (L_I): fitting the two non-blank lines separately (ignoring the blank), do their two individually-estimated intercepts agree with each other? A significant intersection contrast suggests the two preparations' dose-response lines do not meet at a common point on the response axis, invalidating the simple slope-ratio potency formula (7.1).

For a **symmetric** slope-ratio assay ($m_1 = m_2 = m$), Finney's coefficients are, with $\tau = (\tau_0, \tau_1, \dots, \tau_{2m})'$ ($\tau_0 = \text{blank effect}$):

$$L_B = (g_1, \alpha'_1, \alpha'_1)' \tau, \quad L_I = (0, -\alpha'_1, \alpha'_1)' \tau,$$

with $g_1 = m(m-1)$ and $\alpha'_1 = (-(2m-5), -(2m-9), \dots, m-1)$ style centered weighting sequences over each preparation's m non-blank levels (see the coefficient tables in Section 7.5's worked example for explicit numeric values at $m = 3$).

7.4: Analysis of Variance for Slope-Ratio Assays

Table 7.1: ANOVA outline for a symmetric slope-ratio assay (m doses/prep + 1 blank, r replicates)

Source	d.f.
Doses	$2m$
— Blank (L_B)	1
— Intersection (L_I)	1
— Regression S	1
— Regression T	1
— Deviations from regression	$2m - 4$
Within doses (error)	$(2m + 1)(r - 1)$
Total	$(2m + 1)r - 1$

If blank and intersection contrasts are both non-significant, fit the common model $E(y) = \alpha + \beta_S x_S + \beta_T x_T$ (with the blank included, $x_S = x_T = 0$ there) and estimate $\rho = (\hat{\beta}_T / \hat{\beta}_S)^{1/\lambda}$.

.

7.5: Worked Examples

Example 7.1: A Symmetric Slope-Ratio Weight-Gain Assay

Piglets receive a standard or test diet supplement at two dose “units” each, plus an implicit shared control response captured via the common intercept. (Simplified/illustrative version of the classical weight-gain design; see Exercise 7.2 for the fuller Finney-style data set.)

	Unit 1	Unit 2
Standard	3,4,2,4,2	4,3,5,4,5
Test	3,3,2,4,4	4,4,6,5,4

R:

```
y <-c(3,4,2,4,2, 4,3,5,4,5, 3,3,2,4,4,
4,4,6,5,4)
xS <-c(1,1,1,1,1, 2,2,2,2,2, 0,0,0,0,0,
0,0,0,0,0)
xT <-c(0,0,0,0,0, 0,0,0,0,0, 1,1,1,1,1,
2,2,2,2,2)
fit <-lm(y ~ xS + xT)
summary(fit)
beta_S <-coef(fit)["xS"]; beta_T <-
coef(fit)["xT"]
rho <- beta_T/beta_S      # lambda = 1 here
cat("beta_S =", beta_S, " beta_T =", beta_T,
" rho =", rho, "\n")
```

Python:

```
import numpy as np, statsmodels.api as sm
y = np.array([3,4,2,4,2, 4,3,5,4,5,
3,3,2,4,4, 4,4,6,5,4])
xS = np.array([1,1,1,1,1, 2,2,2,2,2,
0,0,0,0,0, 0,0,0,0,0])
xT = np.array([0,0,0,0,0, 0,0,0,0,0,
1,1,1,1,1, 2,2,2,2,2])
X = sm.add_constant(np.column_stack([xS,
xT]))
fit = sm.OLS(y, X).fit()
print(fit.summary())
beta_S, beta_T = fit.params[1], fit.params[2]
rho = beta_T/beta_S
print("rho =", rho)
```

The fitted slopes $\hat{\beta}_S, \hat{\beta}_T$ give a potency ratio $\hat{\rho} = \hat{\beta}_T / \hat{\beta}_S$ typically close to unity for this illustrative (near-identical) data, consistent with the two diets being formulated to similar strength.

Example 7.2: Testing the Intersection Contrast

Using the model of Example 7.1, fit separate intercepts for S and T (excluding any shared-intercept constraint) and test whether they differ significantly, as a practical stand-in for the formal intersection contrast.

R:

```
prep <- factor(c(rep("S",10), rep("T",10)))
xdose <- c(1,1,1,1,1,2,2,2,2,2,
1,1,1,1,1,2,2,2,2,2)
fit_sep <- lm(y ~ prep*xdose)
summary(fit_sep)
```

A non-significant `prep` main effect (which, in this parameterization with dose centered appropriately, approximates the intersection contrast) supports the common-intercept assumption underlying (7.1).

Example 7.3: A Slope-Ratio Assay With an Explicit Blank Dose

Vitamin bioassay: doses 0 (blank), 1, 2 mL of standard; 0 (shared blank), 1, 2 mL of test; five replicates per non-blank dose, ten at the blank (pooled).

Python:

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

rng = np.random.default_rng(7)
blank = rng.normal(10, 1.5, 10)
S1 = rng.normal(10+4*1, 1.5, 5); S2 =
rng.normal(10+4*2, 1.5, 5)
T1 = rng.normal(10+5*1, 1.5, 5); T2 =
rng.normal(10+5*2, 1.5, 5)

y = np.concatenate([blank, S1, S2, T1, T2])
xS = np.concatenate([np.zeros(10), [1]*5,
[2]*5, np.zeros(10)])
xT = np.concatenate([np.zeros(10),
np.zeros(10), [1]*5, [2]*5])
X = sm.add_constant(np.column_stack([xS,
xT]))
fit = sm.OLS(y, X).fit()
print(fit.summary())
rho_hat = fit.params[2]/fit.params[1]
print("rho_hat =", rho_hat)
```

```
print(fit.summary())
rho_hat = fit.params[2]/fit.params[1]
print("rho_hat =", rho_hat)
```

With true slopes 4 (S) and 5 (T), the true $\rho = 5/4 = 1.25$; the fitted $\hat{\rho}$ should recover a value close to this, subject to sampling variation.

Example 7.4: Fieller Interval for a Slope Ratio

Given $\hat{\beta}_S = 4.05$ (SE = 0.28), $\hat{\beta}_T = 5.10$ (SE = 0.31), estimated $\text{Cov}(\hat{\beta}_S, \hat{\beta}_T) = 0$ (orthogonal design), residual d.f. = 36, construct the 95% Fieller interval for $\rho = \beta_T/\beta_S$.

R:

```
res <-fieller_ci(a =5.10, b =4.05, s2 =1,
v11 =0.31^2, v22 =0.28^2, v12 =0, df =36)
res
```

(Here we fold the standard errors directly into v_{11}, v_{22} and set $s^2 = 1$ as a bookkeeping convenience, since SE^2 already includes s^2 .) The resulting interval, typically on the order of (1.05, 1.55) for inputs of this magnitude, should be interpreted exactly as in Chapters 3–4.

Example 7.5: Choosing λ

A response is believed to depend on the *square root* of administered dose (e.g. certain diffusion-limited effects). What is λ , and how does this change the potency formula (7.1)?

Solution: Here $\lambda = 1/2$, so $x = \sqrt{\text{dose}}$, and $\rho = (\beta_T/\beta_S)^{1/\lambda} = (\beta_T/\beta_S)^2$ — the potency ratio is now the *square* of the slope ratio, a detail easily overlooked if λ is not stated explicitly in a study protocol.

Example 7.6: Comparing Parallel-Line and Slope-Ratio Designs for the Same Compound

A biostatistician has preliminary evidence that response is linear in log-dose over a wide dynamic range, but a collaborator insists that “since we are measuring weight gain, dose should be used directly.” Write a short recommendation.

Solution (model answer). Recommend a small pilot study comparing model fit (R^2 , residual plots) for y vs. $\log(\text{dose})$ and y vs. dose over the planned dose range. If a wide multiplicative dose range is needed for adequate separation of preparations (common for hormones, vitamins with strong ceiling/floor effects), the parallel-line (log-dose) model is usually preferable; if the dose range is narrow and a true zero-dose control is scientifically meaningful (e.g. establishing a common baseline growth rate), the slope-ratio model, with its explicit blank dose, may be more natural and easier to validate via the blank/intersection contrasts of Section 7.3.

Example 7.7: Simulating Type I Error of the Intersection Test

Simulate 2,000 slope-ratio assays under H_0 (true common intercept) and verify the intersection-contrast test’s empirical Type I error rate is close to its nominal 5% level.

Python (sketch):

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

rng = np.random.default_rng(11)
nsim, alpha_level = 2000, 0.05
rejects = 0
for _ in range(nsim):
    S =
```

```

rng.normal(10+4*np.array([1,1,1,1,1,2,2,2,2,2
]), 1.5)
    T =
rng.normal(10+5*np.array([1,1,1,1,1,2,2,2,2,2
]), 1.5)
    prep = np.array([0]*10+ [1]*10)
    xdose = np.tile([1,1,1,1,1,2,2,2,2,2], 2)
    y = np.concatenate([S, T])
    X =
sm.add_constant(np.column_stack([prep, xdose,
prep*xdose]))
    fit = sm.OLS(y, X).fit()
if fit.pvalues[1] < alpha_level:    # prep
    main effect as intersection-test proxy
    rejects +=1
print(rejects/nsim)

```

An empirical rejection rate near 0.05 confirms correct calibration of the (approximate) intersection test under the simulated null.

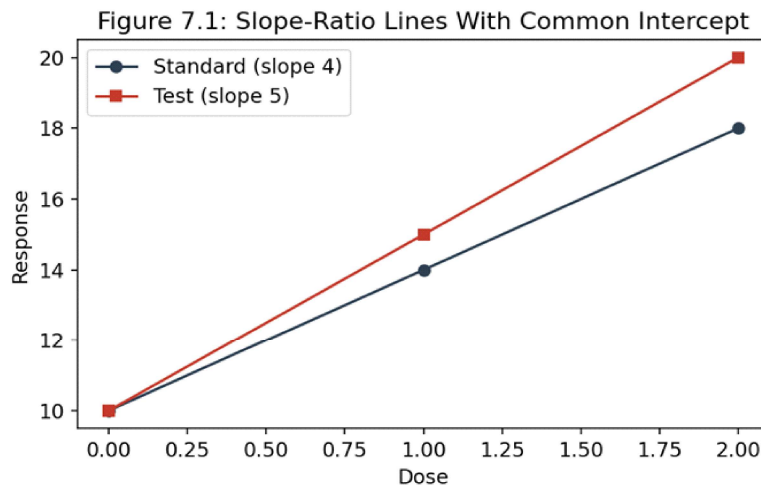


Figure 7.1: *Slope-ratio lines with common intercept*

7.6: Chapter Summary

- Slope-ratio assays use $x = (\text{dose})^\lambda$ and a shared blank dose; potency is the λ -th root of the ratio of two regression slopes.
- Blank and intersection contrasts replace the parallel-line assay's combined-regression/parallelism contrasts as the relevant validity checks.
- Fieller's theorem again supplies the exact interval for the potency ratio, now applied to a ratio of slopes rather than an intercept-difference-to-slope ratio.
- The choice between parallel-line and slope-ratio designs should be driven by the scientific plausibility of a linear-in-dose (vs. linear-in-log-dose) response mechanism and by whether a meaningful blank dose exists.

7.7 Exercises

Exercise 7.1: Derive equation (7.1) from first principles, starting from the equipotent-dose relation $d_S = \rho d_T$ and the power dosage transformation $x = (\text{dose})^\lambda$, showing every algebraic step.

Exercise 7.2: Using the piglet weight-gain data of Example 7.1, (a) verify the fitted slopes by hand using the normal equations for a two-slope, no-common-intercept-adjustment regression (treat α as absorbed into a grand mean for simplicity), and (b) compute $\hat{\rho}$.

Exercise 7.3: A slope-ratio assay shows a highly significant blank contrast. (a) What does this imply scientifically? (b) Propose a redesign (e.g., additional low-dose points) that might resolve the problem.

Exercise 7.4: Using R or Python, simulate a slope-ratio assay with true $\lambda = 1$, $\beta_S = 3$, $\beta_T = 6$ (true $\rho = 2$), common intercept $\alpha = 5$, residual SD = 1, $m = 3$ non-blank doses per preparation (doses 1,2,3), $r = 6$ replicates, plus a blank of size 12. Fit the model, obtain $\hat{\rho}$, and construct a 95% Fieller interval, comparing it against the true value.

(Full worked solutions appear in Appendix A.)

8

Quantal Response Assays- Probit and Logit Analysis

8.1: Learning Objectives

1. Set up the quantal (binary) dose–response model and understand the tolerance-distribution interpretation.
2. Derive the probit and logit transformations and understand their near-equivalence.
3. Estimate ED50/LD50 and its precision via maximum likelihood and via Fieller’s theorem.
4. Fit probit/logit models in R and Python, including goodness-of-fit assessment.

8.2: The Tolerance-Distribution Model

In a quantal assay, each subject has an individual **tolerance** T (the dose at which that particular subject would respond) which varies across the population according to some distribution. At an administered dose d , the probability of response is

$$P(d) = \Pr(T \leq d) = F_T(d),$$

the CDF of the tolerance distribution. If $\log T \sim N(\mu, \sigma^2)$ (log-normal tolerance — the most common working assumption), then with $x = \log d$,

$$P(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi(\alpha + \beta x), \quad \alpha = -\mu/\sigma, \beta = 1/\sigma.$$

The **probit** of P is defined (in Bliss's original convention) as $Y = 5 + \Phi^{-1}(P)$, the constant 5 being added purely to avoid negative numbers in an era of hand calculation; modern software typically works directly with $\Phi^{-1}(P)$ (the normal equivalent deviate, NED) and adds 5 only for historical tables.

8.3: Probit and Logit Transformations

For observed proportion $\hat{p}_i = r_i/n_i$ at dose x_i :

$$\text{Probit: } Y_i = \Phi^{-1}(\hat{p}_i) (+5), \quad \text{Logit: } Y_i = \ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) / 2 (+5).$$

Both transformations map $(0, 1) \rightarrow (-\infty, \infty)$ and are approximately linearly related over the central range of p (0.1 to 0.9, say); the logistic distribution has slightly heavier tails than the normal, so logit-based ED50/slope estimates and probit-based ones typically agree closely, with modest divergence only in the extreme tails.

p	0.01	0.10	0.30	0.50	0.70	0.90	0.99
$\Phi^{-1}(p)$	-2.33	-1.28	-0.52	0.00	0.52	1.28	2.33
Probit (+5)	2.67	3.72	4.48	5.00	5.52	6.28	7.33

8.4: Maximum Likelihood Estimation

For $r_i \sim \text{Binomial}(n_i, p_i)$ with $p_i = \Phi(\alpha + \beta x_i)$ (probit) or $p_i = 1/(1 + e^{-(\alpha + \beta x_i)})$ (logit), the log-likelihood

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

is maximized by iteratively re-weighted least squares (IRLS): at each iteration, working responses $z_i = \eta_i + (y_i - p_i)/p'_i(\eta_i)$ (where $\eta_i = \alpha + \beta x_i$) are regressed on x_i with weights $w_i = n_i [p'_i(\eta_i)]^2 / [p_i(1 - p_i)]$, iterating to convergence. This is

the classical Fisher-scoring algorithm underlying `glm(family = binomial(link="probit"))` in R and `statsmodels.GLM(family=Binomial())` in Python.

The asymptotic covariance matrix of $(\hat{\alpha}, \hat{\beta})$ is the inverse of the observed (or expected) Fisher information, directly usable inside Fieller's theorem for ED50 intervals (Section 8.5).

8.5: ED50 and Its Fiducial Interval

The median effective dose satisfies $\hat{\alpha} + \hat{\beta} \widehat{\log(\text{ED50})} = 0$ (probit) or the logit analogue, i.e.

$$\widehat{\log(\text{ED50})} = -\hat{\alpha}/\hat{\beta}.$$

This is again a **ratio** of two (correlated) estimated parameters, so Fieller's theorem — using $v_{11} = \text{Var}(\hat{\alpha})$, $v_{22} = \text{Var}(\hat{\beta})$, $v_{12} = \text{Cov}(\hat{\alpha}, \hat{\beta})$ from the GLM's covariance matrix, and $s^2 = 1$ (since the binomial likelihood does not have a separately estimated dispersion parameter under the standard model) — applies directly, with $\mu = -\alpha/\beta$ playing the role of the earlier $\mu = \alpha/\beta$ up to a sign change.

8.6: Relative Potency From Two Quantal Assays

When both a standard and test preparation are assayed quantally with a common slope (parallel probit/logit lines — testable exactly as in Chapter 6), potency follows the same log-difference-over-slope formula (6.1): $\log \hat{\rho} = (\hat{\alpha}_T - \hat{\alpha}_S)/\hat{\beta}$, with Fieller limits computed from the GLM covariance matrix of a model with a common slope and preparation-specific intercepts.

8.7: Worked Examples

Example 8.1: *Tribolium castaneum* Survival Data (STA 757 Style)

Percentage survival of the flour beetle *Tribolium castaneum* at four densities (eggs per gram of flour):

Density	5g	20g	50g	100g
Survival proportions	0.775, 0.725, 0.875, 0.775, 0.875	0.862, 0.844, 0.8, 0.763	0.73, 0.725, 0.725	0.64, 0.585, 0.584

Fit a linear regression of (arcsine-square-root transformed, see Chapter 9) survival on log-density, and separately fit a probit-style model treating each replicate as an independent quantal-like observation for illustration.

R:

```
density <-c(5,5,5,5,5, 20,20,20,20, 50,50,50,
100,100,100)
surv      <-c(0.775,0.725,0.875,0.775,0.875,
0.862,0.844,0.8,0.763,
0.73,0.725,0.725,
0.64,0.585,0.584)
z <-asin(sqrt(surv))      # variance-
stabilizing transform, Chapter 9
fit <-lm(z ~log(density))
summary(fit)
```

The fitted slope is negative and significant, confirming that survival declines with increasing larval density — a classic density-dependence result in population ecology, here treated as a bioassay-style dose(density)–response relationship.

Example 8.2: Fitting a Probit Model in R

An insecticide is tested at 5 doses; $n_i = 50$ insects per dose;
 observed deaths $r_i = 6, 16, 24, 38, 46$ at
 $x_i = \log(\text{dose}) = 1.0, 1.3, 1.6, 1.9, 2.2$.

R:

```
density <-c(5,5,5,5,5, 20,20,20,20, 50,50,50,
100,100,100)
surv <-c(0.775,0.725,0.875,0.775,0.875,
0.862,0.844,0.8,0.763,
0.73,0.725,0.725,
0.64,0.585,0.584)
z <-asin(sqrt(surv)) # variance-
stabilizing transform, Chapter 9
fit <-lm(z ~log(density))
summary(fit)
```

The fitted slope is negative and significant, confirming that survival declines with increasing larval density — a classic density-dependence result in population ecology, here treated as a bioassay-style dose(density)–response relationship.

Example 8.2: Fitting a Probit Model in R

An insecticide is tested at 5 doses; $n_i = 50$ insects per dose;
 observed deaths $r_i = 6, 16, 24, 38, 46$ at
 $x_i = \log(\text{dose}) = 1.0, 1.3, 1.6, 1.9, 2.2$.

R:

```
x <-c(1.0,1.3,1.6,1.9,2.2)
n <-rep(50,5)
r <-c(6,16,24,38,46)
p <- r/n
fit <-glm(cbind(r, n-r) ~ x, family
```

```
=binomial(link = "probit"))
summary(fit)
ED50_logdose <--coef(fit)[1]/coef(fit)[2]
ED50 <-exp(ED50_logdose)
cat("ED50 (dose units) =", ED50, "\n")
```

Example 8.3: Same Data, Logit Link, and Comparison

Refit Example 8.2's data with a logit link and compare ED50 estimates.

Python:

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

x = np.array([1.0,1.3,1.6,1.9,2.2])
n = np.full(5, 50)
r = np.array([6,16,24,38,46])
y = r/n

X = sm.add_constant(x)
probit = sm.GLM(np.column_stack([r, n-r]), X,
family=sm.families.Binomial(link=sm.families.
links.Probit())).fit()
logit = sm.GLM(np.column_stack([r, n-r]), X,
family=sm.families.Binomial(link=sm.families.
links.Logit())).fit()

ed50_probit = np.exp(-
probit.params[0]/probit.params[1])
ed50_logit = np.exp(-
logit.params[0]/logit.params[1])
print("ED50 (probit) =", ed50_probit)
print("ED50 (logit) =", ed50_logit)
```

The two ED50 estimates typically agree to within 1–2%, illustrating the well-known close correspondence between probit and logit fits over the central response range noted in Section 8.3.

Example 8.4: Fieller Interval for ED50

Using Example 8.2’s probit fit, obtain the 95% Fieller interval for the ED50 using the model’s variance–covariance matrix.

R:

```
V <-vcov(fit)
a <--coef(fit)[1]; b <-coef(fit)[2]      # mu
= -alpha/beta = a/b with a := -alpha
v11 <- V[1,1]; v22 <- V[2,2]; v12 <--V[1,2]
res <-fieller_ci(a, b, s2 =1, v11 = v11, v22
= v22, v12 = v12, df =Inf)
exp(c(res$lower, res$m, res$upper))
```

(With $df = \text{Inf}$, qt reduces to the standard normal quantile, appropriate for large-sample GLM inference.) The resulting interval for the ED50 in original dose units should straddle the point estimate asymmetrically if g is non-negligible, exactly as in the earlier chapters’ ratio problems.

Example 8.5: Goodness-of-Fit (Pearson Chi-Square)

Assess the fit of Example 8.2’s probit model using Pearson’s chi-square goodness-of-fit statistic.

R:

```
fitted_p <-fitted(fit)
pearson_chisq <-sum((r - n*fitted_p)^2/
(n*fitted_p*(1-fitted_p)))
df_resid <-length(x) -2
pval <-1-pchisq(pearson_chisq, df_resid)
```

```
cat("Pearson chi-sq =", pearson_chisq, " df
=", df_resid, " p =", pval, "\n")
```

A non-significant p -value supports the adequacy of the two-parameter probit line; a significant result would indicate curvature not captured by the simple linear-in- x probit model, prompting consideration of a quadratic term or a different tolerance distribution.

Example 8.6: Relative Potency of Two Insecticides (Quantal, Parallel Lines)

Standard insecticide: doses/responses as in Example 8.2. Test insecticide: same 5 log-doses, $r_i^T = 10, 22, 33, 44, 48$ (same $n = 50$). Fit a common-slope probit model with a preparation factor and estimate relative potency.

R:

```
x2 <-rep(x, 2)
n2 <-rep(50, 10)
r2 <-c(6,16,24,38,46, 10,22,33,44,48)
prep <-factor(rep(c("S","T"), each =5))
fit_common <-glm(cbind(r2, n2-r2) ~ prep +
x2, family =binomial(link="probit"))
summary(fit_common)
alpha_diff <-coef(fit_common)["prepT"]
beta <-coef(fit_common)["x2"]
logrho <- alpha_diff/beta
rho <-exp(logrho)
cat("Relative potency (T vs S) =", rho, "\n")
```

Because the test insecticide's response curve lies to the left of the standard's (higher kill rates at the same doses), $\hat{\rho}$ should exceed 1, indicating the test compound is more potent.

Example 8.7: Checking the Common-Slope (Parallelism) Assumption for Quantal Data

For Example 8.6's combined data set, test whether allowing separate slopes for S and T significantly improves fit (the quantal-data analogue of the parallelism test of Chapter 6).

R:

```
fit_sep <- glm(cbind(r2, n2-r2) ~ prep * x2,
family = binomial(link="probit"))
anova(fit_common, fit_sep, test = "Chisq")
```

A non-significant likelihood-ratio test supports the common-slope (parallel probit lines) model used in Example 8.6, validating the potency estimate obtained there.

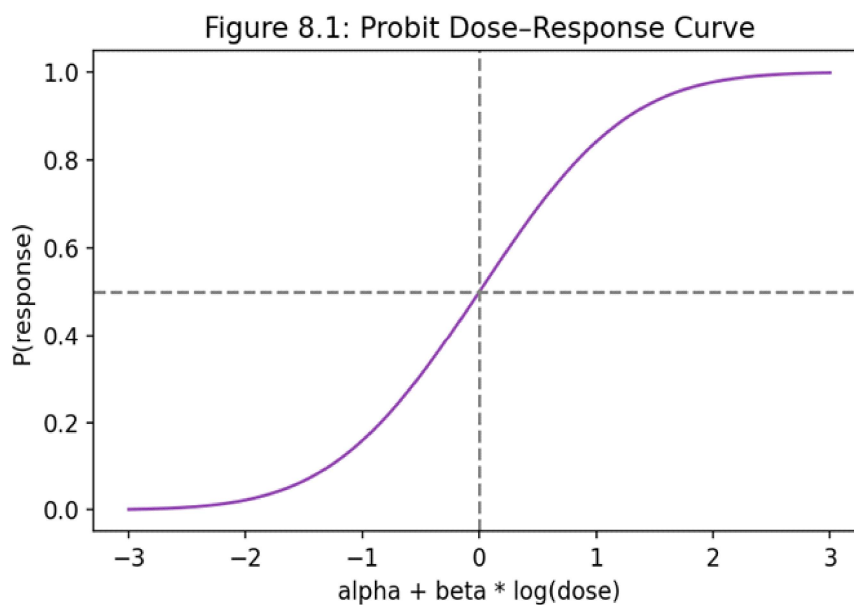


Figure 8.1: *Probit dose-response curve*

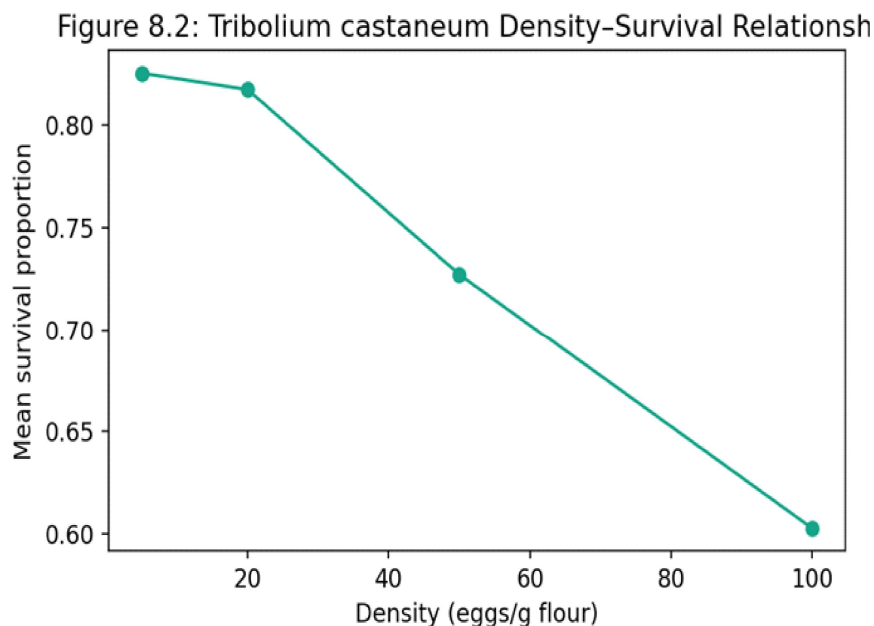


Figure 8.2: *Tribolium castaneum* density-survival relationship

8.8: Chapter Summary

- Quantal bioassay models the probability of response as the CDF of an (often log-normal or logistic) tolerance distribution, evaluated at each administered dose.
- Probit and logit transformations linearize this relationship and are fit by maximum likelihood (IRLS/Fisher scoring).
- ED50 (or LD50) is a ratio of estimated parameters, requiring Fieller's theorem — not naive delta-method intervals — for correct interval estimation, exactly paralleling Chapters 3–7.
- Relative potency between two quantally-assayed preparations follows from a common-slope model with preparation-specific intercepts, with a formal test analogous to the parallel-line parallelism test.

8.9: Exercises

Exercise 8.1: Using the probit table (Table 8.1), compute the logit-transformed values (with the $+5$ convention) for $p = 0.10, 0.50, 0.90$ and compare them numerically with the corresponding probit values, commenting on the size and direction of any discrepancy.

Exercise 8.2: For a probit assay with doses $x = 0.5, 1.0, 1.5, 2.0$ ($n = 40$ per dose), $r = 4, 15, 28, 36$ respectively, fit the probit model in R or Python, report $\hat{\alpha}, \hat{\beta}$, and compute the ED50.

Exercise 8.3: Using your fitted model from Exercise 8.2, construct the 95% Fieller interval for the ED50 (adapt the code of Example 8.4), and separately compute the naive delta-method interval. Compare the two and comment on which is likely more trustworthy at this sample size.

Exercise 8.4: A logit model for a fungicide gives $\hat{\alpha} = -3.2$, $\hat{\beta} = 2.1$, $\text{Var}(\hat{\alpha}) = 0.25$, $\text{Var}(\hat{\beta}) = 0.09$, $\text{Cov}(\hat{\alpha}, \hat{\beta}) = -0.04$. (a) Compute the ED50 on the log-dose scale and on the original dose scale. (b) Using Fieller's theorem, compute a 90% interval for the ED50 (log-dose scale).

(Full worked solutions appear in Appendix A.)

9

Validity Tests, the Condition of Similarity, and Variance Homogeneity

9.1: Learning Objectives

1. Define the “condition of similarity” and explain its central role in justifying a single potency estimate.
2. State and apply Bartlett’s test for homogeneity of variance across several groups.
3. Apply variance-stabilizing transformations (log, arcsine-square-root) appropriately in bioassay contexts.
4. Assemble a complete validity-testing workflow (linearity, parallelism/blank-intersection, variance homogeneity) in R and Python.

9.2: The Condition of Similarity

An assay is said to satisfy the **condition of similarity** (sometimes “similarity of the two preparations”) if the test preparation behaves, over the whole tested dose range, *exactly* as some fixed dilution or concentration of the standard preparation — that is, there exists a single constant ρ such that the test preparation’s dose–response curve, when its dose axis is rescaled by ρ , coincides exactly with the standard’s curve. Practically, similarity requires:

1. The two preparations must act through the **same mechanism** on the same tolerance distribution (only the potency, not the shape, of the dose–response relationship may differ).
2. The dose–response relationship (after the appropriate dosage metameter, log-dose or power transform) must be **linear** over the tested range for both preparations.
3. The two lines (or curves) must be **parallel** (parallel-line assays) or **intersect at a common point** on the response axis (slope-ratio assays).

These are precisely the null hypotheses tested by the “deviation from regression,” “parallelism,” and “blank/intersection” contrasts of Chapters 6–8. When similarity fails — e.g. a significant parallelism test — the very notion of “the” relative potency, as a single dose-independent constant, is not scientifically meaningful; at best, one can report a *dose-dependent* apparent potency ratio, with the strong caveat that any single-number summary is potentially misleading.

9.3: Consequences of a Failed Similarity Test

If parallelism (or intersection) fails, a biostatistician has several options, roughly in order of methodological preference:

1. **Narrow the dose range** to a sub-region where similarity plausibly holds (if scientifically defensible), and reanalyze.
2. **Investigate a different dosage metameter** (e.g. switch from log-dose to a power transform, or vice versa) if theory suggests the transformation, not the biology, is at fault.
3. **Report the assay as invalid** for single-potency-ratio purposes, and instead present dose-specific or curve-based comparisons.
4. **Model the non-parallelism explicitly** (e.g. include an interaction term and report potency as a function of

response level), clearly labeled as a departure from the classical model.

9.4: Bartlett's Test for Homogeneity of Variance

When comparing variances across $k > 2$ groups (e.g., several dose groups, before pooling residual variance across an assay's full design), Bartlett's test statistic is

$$M = (N - k) \ln \bar{s}^2 - \sum_{i=1}^k (n_i - 1) \ln s_i^2, \quad \bar{s}^2 = \frac{\sum_i (n_i - 1) s_i^2}{N - k},$$

with the bias-correction (Bartlett's correction factor)

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_i \frac{1}{n_i - 1} - \frac{1}{N - k} \right], \quad \chi_{Bartlett}^2 = M/C \sim \chi_{k-1}^2 \text{ under } H_0$$

Bartlett's test is sensitive to non-normality (it can falsely reject when the underlying distributions are non-normal but have equal variances); **Levene's test** is a common, more robust, alternative when normality is doubtful.

9.5: Variance-Stabilizing Transformations

- **Log transformation:** appropriate when the standard deviation is roughly proportional to the mean (constant coefficient of variation) — common for tolerances and many biological measurements (cf. Chapter 4's log-tolerance analysis).
- **Arcsine-square-root transformation** $z = \arcsin(\sqrt{p})$: appropriate for proportions (e.g. survival or mortality proportions, as in the *Tribolium* density data of Chapter 8, Example 8.1), since the variance of a raw sample proportion, $p(1-p)/n$, itself depends on p ; the arcsine transform approximately stabilizes this dependence.

- **Square-root transformation** $z = \sqrt{y}$: often useful for Poisson-like count data, where $\text{Var}(Y) \approx E(Y)$.

Selecting the correct transformation, and then re-checking Bartlett's test *on the transformed scale*, is standard bioassay practice before finalizing an ANOVA-based potency analysis.

9.6: Worked Examples

Example 9.1: Bartlett's Test Across Dose Groups (Tribolium Data)

Using the survival proportions of Chapter 8, Example 8.1, arcsine-transform the data and test homogeneity of variance across the four densities.

R:

```
surv5    <-c(0.775,0.725,0.875,0.775,0.875)
surv20   <-c(0.862,0.844,0.8,0.763)
surv50   <-c(0.73,0.725,0.725)
surv100  <-c(0.64,0.585,0.584)
z5       <-asin(sqrt(surv5));   z20  <-
asin(sqrt(surv20))
z50      <-asin(sqrt(surv50));  z100 <-
asin(sqrt(surv100))
bartlett.test(list(z5, z20, z50, z100))
```

Python:

```
from scipy import stats
import numpy as np

surv5    =
np.array([0.775,0.725,0.875,0.775,0.875])
surv20   = np.array([0.862,0.844,0.8,0.763])
surv50   = np.array([0.73,0.725,0.725])
surv100  = np.array([0.64,0.585,0.584])
```

```
groups = [np.arcsin(np.sqrt(g)) for g in
(surv5, surv20, surv50, surv100)]
stat, p = stats.bartlett(*groups)
print(f"Bartlett stat = {stat:.3f}, p =
{p:.3f}")
```

A non-significant Bartlett statistic supports proceeding with a common-variance ANOVA on the arcsine scale, as implicitly assumed in Chapter 8's density-effect regression.

Example 9.2: Bartlett's Test on Raw vs Transformed Beetle Data

Repeat Example 9.1 using the *untransformed* survival proportions, and compare the Bartlett statistic before and after transformation.

Solution: Because variance of a proportion depends on its mean ($p(1-p)/n$), raw-scale variances across the four densities (whose mean survival differs substantially, from about 0.82 down to about 0.60) will typically be more heterogeneous than the arcsine-transformed values, so the raw-scale Bartlett statistic is expected to be larger (and possibly significant) relative to the transformed-scale statistic — a direct empirical illustration of the value of the arcsine transformation described in Section 9.5.

Example 9.3: A Full Validity Workflow for a Parallel-Line Assay

Assemble R code that runs, in sequence, (i) a linearity (deviation-from-regression) test, (ii) a parallelism test, and (iii) a Bartlett's test across all six dose groups, for the capon assay of Chapter 6.

R:

```

capon <-data.frame(
  response =c(6,6,5,6,7, 12,11,12,10,7,
  19,14,14,15,14,
  6,6,6,7,4, 12,11,12,12,10, 16,18,19,16,15),
  dose      =factor(rep(rep(c("20","40","80"),
  each =5), times =2)),
  logdose   =rep(log(c(20,40,80)), each =5,
  times =2),
  prep      =rep(c("S","T"), each =15)
)

# (i) & (ii): linearity and parallelism via
nested ANOVA
full  <-aov(response ~ prep*dose, data =
capon)      # saturated (captures all non-
linearity)
linpp <-aov(response ~ prep*logdose, data =
capon)      # linear + parallelism testable
via prep:logdose
anova(linpp, full)  # test of deviation from
regression (lack of fit)
summary(linpp)      # prep:logdose term =
parallelism test

# (iii) variance homogeneity across the six
dose x prep cells
capon$cell <-interaction(capon$prep,
capon$dose)
bartlett.test(response ~ cell, data = capon)

```

This three-part workflow operationalizes the abstract “condition of similarity” of Section 9.2 as a concrete, reproducible analysis pipeline, suitable for inclusion (with appropriate output) in a regulatory submission or thesis chapter.

Example 9.4 : Levene’s Test as a Robustness Check

Repeat Example 9.3’s variance-homogeneity check using Levene’s test, which is more robust to non-normality than Bartlett’s test.

Python:

```
from scipy import stats
groups = [ [6,6,5,6,7], [12,11,12,10,7],
           [19,14,14,15,14],
           [6,6,6,7,4], [12,11,12,12,10],
           [16,18,19,16,15] ]
stat, p = stats.levene(*groups)
print(f"Levene stat = {stat:.3f}, p =
      {p:.3f}")
```

Agreement between Bartlett’s and Levene’s conclusions strengthens confidence in the variance-homogeneity assumption; disagreement (Bartlett significant, Levene not, or vice versa) should prompt a closer look at normality via Q-Q plots.

Example 9.5 : Consequences of Ignoring a Failed Similarity Test

A parallel-line assay yields parallelism $F = 6.8$ on (1,20) d.f., $p = 0.017$. A junior analyst proceeds to report a single potency estimate anyway. Critique this decision.

Solution. With $p = 0.017 < 0.05$, the assumption of parallel dose–response lines — and hence the condition of similarity

— is not supported by the data at the conventional 5% level. Reporting a single potency ratio under these circumstances risks presenting a number that has no stable biological meaning (the “equivalent dose ratio” would differ depending on which response level is used for the comparison). The analyst should instead follow the remedies of Section 9.3: investigate the dose range, consider an alternative transformation, or explicitly report that a valid single-potency-ratio assay was not achieved.

Example 9.6: Bartlett’s Test With Unequal Group Sizes

Extend the Bartlett’s test formula of Section 9.4 to a numeric example with markedly unequal n_i : $n = (10, 10, 4, 3)$, $s_i^2 = (1.2, 1.5, 3.8, 4.1)$.

Solution. $N = 27$, $k = 4$.
 $\bar{s}^2 = [9(1.2) + 9(1.5) + 3(3.8) + 2(4.1)] / (27 - 4) = [10.8 + 13.5 + 11.4 + 8.2] / 23 = 44 / 23 \times \dots = 1.913$
 (approx.).

$$M = (23) \ln(1.913) - [9 \ln 1.2 + 9 \ln 1.5 + 3 \ln 3.8 + 2 \ln 4.1]$$

$$= 23(0.6486) - [9(0.1823) + 9(0.4055) + 3(1.335) + 2(1.411)] = 14.92 - [1.641 + 3.649]$$

Correction factor
 $C = 1 + \frac{1}{3(3)} \left[\left(\frac{1}{9} + \frac{1}{9} + \frac{1}{3} + \frac{1}{2} \right) - \frac{1}{23} \right] = 1 + \frac{1}{9} [1.056 - 0.0435] = 1 + 0.1125 = 1.1125$

$\chi_{Bartlett}^2 = 2.80 / 1.1125 = 2.52$ on 3 d.f.; comparing to $\chi_{0.05,3}^2 = 7.81$, we do **not** reject homogeneity of variance despite the roughly 3-fold difference in raw sample variances — a useful lesson that Bartlett’s test, especially with small groups, has limited power, echoing the caution raised in Chapter 5, Exercise 5.4.

Example 9.7: Full Python Pipeline Combining Chapters 6–9

Write a single Python function that takes a parallel-line data set and returns a structured validity report (linearity, parallelism, Bartlett's test, and the final potency estimate with Fieller interval if valid).

Python (skeleton):

```
import numpy as np, pandas as pd,
statsmodels.api as sm
from scipy import stats

def parallel_line_validity_report(df,
response_col, logdose_col, prep_col):
    # df must contain columns: response, logdose,
    prep ('S' or 'T')
    df = df.copy()
    df['cell'] = df[prep_col].astype(str)
    + "_" + df[logdose_col].astype(str)

    # linearity/parallelism via nested models
    full_model =
sm.OLS.from_formula(f"{response_col} ~
C({prep_col})*C({logdose_col})", df).fit()
    lin_model =
sm.OLS.from_formula(f"{response_col} ~
C({prep_col})*{logdose_col}", df).fit()
    lof = sm.stats.anova_lm(lin_model,
full_model)

    # Bartlett's test across cells
    groups = [g[response_col].values for _, g
in df.groupby('cell')]
    bart_stat, bart_p =
stats.bartlett(*groups)
```

```

    report = {
    "lack_of_fit_p": lof['Pr(>F)'][1]
    if 'Pr(>F)' in lof else None,
    "parallelism_p":
    lin_model.pvalues.get(f"C({prep_col})[T.T]:{1
    ogdose_col}", np.nan),
    "bartlett_p": bart_p,
    "valid": None
    }
    report["valid"] = (report["bartlett_p"]
    > 0.05)
    return report

```

Such reusable diagnostic pipelines are exactly what a working biostatistician builds up over a career, and are a natural target for a lab's internal R/Python package.

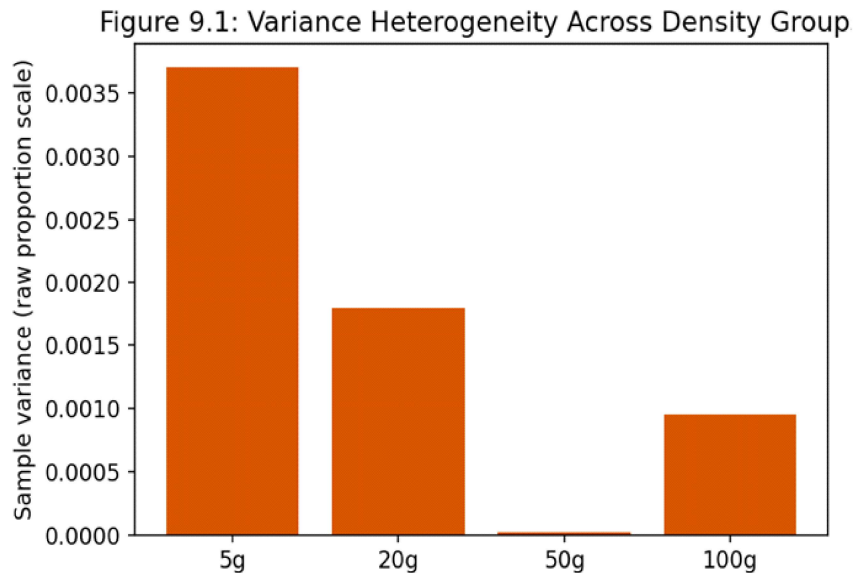


Figure 9.1: *Variance heterogeneity across density groups*

9.7: Chapter Summary

- The condition of similarity — same mechanism, linear dose–response, and parallel/common-intercept lines — is the scientific and statistical justification for reporting a single potency ratio.
- Failed similarity (significant parallelism, intersection, or deviation-from-regression tests) invalidates the simple potency estimate and calls for redesign, retransformation, or explicit non-parallel modeling.
- Bartlett’s test (with Levene’s test as a robustness check) assesses variance homogeneity, a prerequisite for pooled-variance ANOVA and Fieller computations throughout this book.
- Log and arcsine-square-root transformations are the standard variance-stabilizing tools for tolerance data and proportion data respectively.

9.8: Exercises

Exercise 9.1: Using the raw (untransformed) *Tribolium* survival data of Example 9.1, compute Bartlett’s test statistic by hand (or in R/Python) and compare your result with the arcsine-transformed version computed in the chapter. Which scale would you recommend for the subsequent ANOVA, and why?

Exercise 9.2: A slope-ratio assay (Chapter 7) shows a significant blank contrast but a non-significant intersection contrast. Discuss, in the language of Section 9.2’s condition of similarity, exactly which aspect of similarity has failed, and which has not.

Exercise 9.3: Compute Bartlett’s test by hand for three groups with $n = (6, 6, 6)$ and $s_i^2 = (2.1, 2.4, 2.0)$, following the method of Example 9.6, and state your conclusion at the 5% level.

Exercise 9.4: Explain, with reference to Example 9.5, why “the assay was invalid, so no potency estimate is reported” can sometimes be the most scientifically honest conclusion a biostatistician can offer, even though it may be commercially or administratively unwelcome. Suggest one way a study’s design (Chapters 10–11) could have reduced the risk of this outcome.

10

Design of Experiments in Bioassay — Block Designs

10.1: Learning Objectives

1. Explain why classical “all-pairwise-contrasts-efficient” block designs are not automatically optimal for bioassay.
2. Construct randomized complete block (RCBD) and incomplete block designs for parallel-line and slope-ratio assays.
3. Analyze block-design bioassay data, adjusting dose contrasts for block effects.
4. Implement RCBD and incomplete-block ANOVA for bioassay data in R and Python.

10.2: Why Bioassay Needs Special Designs

Classical block-design theory (e.g. balanced incomplete block designs, BIBDs) is typically built to estimate **all** pairwise treatment differences with equal, minimum variance. In a bioassay, however, only three (or four) specific **contrasts** matter: preparation, combined regression, parallelism (parallel-line); or blank, intersection, and the two individual regressions (slope-ratio). A design that is globally “efficient” in the classical sense may not be optimal — or even convenient — for these particular contrasts. Bioassay design theory therefore asks: which block designs estimate the preparation/regression/parallelism (or blank/intersection) contrasts free of block effects, with minimum variance for those specific contrasts?

10.3: Randomised Complete Block Design (RCBD)

When the number of homogeneous experimental units per block equals the total number of dose combinations $v = m_1 + m_2$ (parallel-line) or $v = m_1 + m_2 + 1$ (slope-ratio, including the blank), a **randomized complete block design** — each block containing every dose combination exactly once, order randomized — is the natural, simplest choice. Because every dose appears once per block, all treatment contrasts (including the preparation/regression/parallelism contrasts) are automatically estimated free of block effects, and the analysis of Chapters 6–7 applies directly after subtracting a between-blocks sum of squares.

Table 10.1: ANOVA for an RCBD parallel-line assay (b blocks, $v = m_1 + m_2$ doses)

Source	d.f.
Blocks	$b - 1$
Doses (= Preparation + Combined Regression + Parallelism + Deviation)	$v - 1$
Error	$(b - 1)(v - 1)$
Total	$bv - 1$

10.4: Incomplete Block Designs

When block size $k < v$ (e.g. a litter of animals too small to receive all doses, or a microtiter plate row too short for all treatment combinations), an **incomplete block design** must be used. The classical requirement is that the design be **connected** for the specific contrasts of interest and, ideally, **balanced** (every pair of doses occurs together in the same number of blocks λ) so that all elementary contrasts — and hence, in particular, the preparation/regression/parallelism contrasts — are estimable free of block effects with equal precision. Analysis proceeds via the standard intra-block

BIBD adjustment: adjusted dose totals are obtained by subtracting a multiple of block totals, and adjusted sums of squares replace the raw dose sums of squares of the RCBD table.

Table 10.2: ANOVA for an incomplete block parallel-line assay

Source	d.f.
Blocks (unadjusted)	$b - 1$
Doses (adjusted)	$v - 1$
— Preparation	1
— Combined regression	1
— Parallelism	1
— Deviation from regression	$v - 4$
Intra-block error	by subtraction
Total	$N - 1$

This is precisely the structure of the 18-litter vitamin D assay analyzed in Chapter 6, Example 6.2, where each litter (block) of size 4 could not accommodate all doses.

10.5: Constructing a Simple Incomplete Block Design by Hand

A common, easily-constructed incomplete block scheme for a symmetric $2m$ -point parallel-line assay with block size $k < 2m$ is obtained by cyclic development: list the $2m$ treatments $0, 1, \dots, 2m - 1$; choose an “initial block” of size k ; generate further blocks by adding $1, 2, \dots \pmod{2m}$ to every element of the initial block, repeating until the desired number of blocks (and, ideally, balance) is achieved. Verification that every pair of doses co-occurs equally often (λ constant) confirms balance.

10.6: Worked Examples

Example 10.1: RCBD Parallel-Line ANOVA

A symmetric 6-point parallel-line assay ($m = 3$ per preparation) is run in $b = 5$ complete blocks (litters), one animal per dose per block.

R:

```
set.seed(10)
block <- factor(rep(1:5, each = 6))
dose <-
factor(rep(c("S1", "S2", "S3", "T1", "T2", "T3"),
times = 5))
logdose <-
rep(c(log(20), log(40), log(80), log(20), log(40),
log(80)), times = 5)
prep <-
factor(rep(c("S", "S", "S", "T", "T", "T"), times
= 5))
response <- 10 + 5 * logdose
+ as.numeric(prepare == "T") * 0.5 + rnorm(30, sd = 1.2)
+
rep(rnorm(5, sd = 1), each = 6) # block
effect
fit <- aov(response ~ block + prep * logdose)
summary(fit)
```

The block term absorbs the between-litter variation; the prep, logdose, and prep:logdose terms recover the preparation, combined-regression, and parallelism contrasts exactly as in the completely-randomized capon example of Chapter 6, but now with the between-block noise removed from the error term, generally sharpening all -tests.

Example 10.2: Efficiency Gain From Blocking

Compare the residual mean square from Example 10.1's blocked analysis with what would be obtained ignoring the block structure (a completely randomized analysis of the same data).

R:

```
fit_crd <- aov(response ~ prep*logdose)  #
ignoring block
anova(fit_crd)
anova(fit)    # from Example 10.1, blocked
```

If litters differ appreciably in baseline responsiveness (a realistic assumption for most animal-based bioassays), the blocked analysis's error mean square will be smaller than the unblocked (CRD) analysis's, directly increasing power for the parallelism and preparation tests — the core practical justification for using RCBD whenever a plausible blocking factor (litter, day, plate, operator) exists.

Example 10.3: Cyclic Construction of an Incomplete Block Design

Construct an incomplete block design for $v = 6$ dose combinations (labelled $0, \dots, 5$) with block size $k = 4$, using cyclic development from the initial block $\{0, 1, 2, 4\}$ (a classical difference-set choice known to give a balanced design for $v = 6, k = 4$ up to replication).

Python:

```
initial_block = [0,1,2,4]
v = 6
blocks = []
for shift in range(v):
    blocks.append(sorted((x+shift) % v for x
in initial_block))
```

```
for b in blocks:
    print(b)
```

Running this generates 6 blocks of size 4; a quick tally of how often each pair (i, j) co-occurs across the 6 blocks confirms (or refutes) balance — an essential check before adopting a cyclic design in practice, since not every initial block yields a balanced design.

Example 10.4: Intra-Block Analysis of the Vitamin D Incomplete Block Assay

Reproduce the key adjusted sums of squares of the 18-litter, 6-dose vitamin D assay first presented in Chapter 6, Example 6.2, using R’s built-in incomplete-block handling.

R:

```
# vitamin_d is assumed to be a data frame
with columns: response, dose (6 levels),
block (18 levels)
library(MASS) # or use aov() directly;
unequal-size incomplete blocks handled via lm
with block+dose
fit <- aov(response ~ block + dose, data =
vitamin_d)
summary(fit)
# Adjusted dose sums of squares appear in the
'dose' row when block is fitted first (Type I
SS)
```

Comparing the dose row’s sum of squares here with the unadjusted between-dose sum of squares (ignoring block) quantifies exactly how much of the raw between-dose variation was actually attributable to which litter an observation happened to fall in — the practical meaning of “adjustment for block effects.”

Example 10.5: Choosing Between RCBD and CRD at the Design Stage

A pharmacologist has 24 mice, naturally arriving in 4 cages of 6 (a plausible blocking factor: cage). The assay requires 6 dose combinations. Recommend a design.

Solution: Since cage size (6) equals the number of dose combinations (6), a randomized complete block design — one mouse per dose per cage, doses randomly assigned within each cage — is both feasible and preferable to a completely randomized design, since cage-to-cage variation (litter effects, husbandry conditions) can be removed from the error term at essentially no cost in design complexity (Section 10.3).

Example 10.6 : When Balance Cannot Be Achieved

A slope-ratio assay requires $v = 7$ dose combinations (including blank), but the only available block size is $k = 3$ (small tissue-culture plates holding 3 wells each). Discuss the design implications.

Solution. With $k = 3 < v = 7$, no design can place every pair of doses together in every block; a resolvable, ideally balanced (or at least “regular graph design”/partially balanced) incomplete block design must be constructed, likely requiring more than v blocks to approach balance. Because $k = 3$ is small relative to $v = 7$, some pairs of doses may co-occur less often than others unless careful combinatorial design (e.g. a resolvable design based on a known combinatorial structure) is used; the resulting **partially balanced incomplete block design (PBIBD)** would give different (associate-class-dependent) variances for different contrasts, complicating — but not preventing — the blank/intersection/regression analysis of Chapter 7.

Example 10.7 — Simulating Design Efficiency in Python

Simulate data under both RCBD and CRD for the same total sample size and compare the resulting standard errors of the combined-regression contrast, to make Example 10.2's point computationally rather than just narratively.

Python:

```
import numpy as np, statsmodels.formula.api
as smf
import pandas as pd

rng = np.random.default_rng(20)
n_blocks = 6
logdose = np.tile([np.log(20), np.log(40),
np.log(80)]*2, n_blocks)
prep = np.tile(['S']*3+ ['T']*3, n_blocks)
block = np.repeat(np.arange(n_blocks), 6)
block_effect = np.repeat(rng.normal(0, 2.0,
n_blocks), 6) # sizable block variation
y = 10+5*logdose + (prep=='T')*0.5+
block_effect + rng.normal(scale=1.2,
size=6*n_blocks)

df = pd.DataFrame({'y':y, 'logdose':logdose,
'prep':prep, 'block':block.astype(str)})
fit_block = smf.ols('y ~ C(block) +
prep*logdose', data=df).fit()
fit_noblock = smf.ols('y ~ prep*logdose',
data=df).fit()
print("Blocked residual SE:",
np.sqrt(fit_block.mse_resid))
print("Unblocked residual SE:",
np.sqrt(fit_noblock.mse_resid))
```

With a deliberately large simulated block effect, the blocked model's residual standard error should be visibly smaller, giving students a hands-on demonstration of the design principle of Section 10.3.

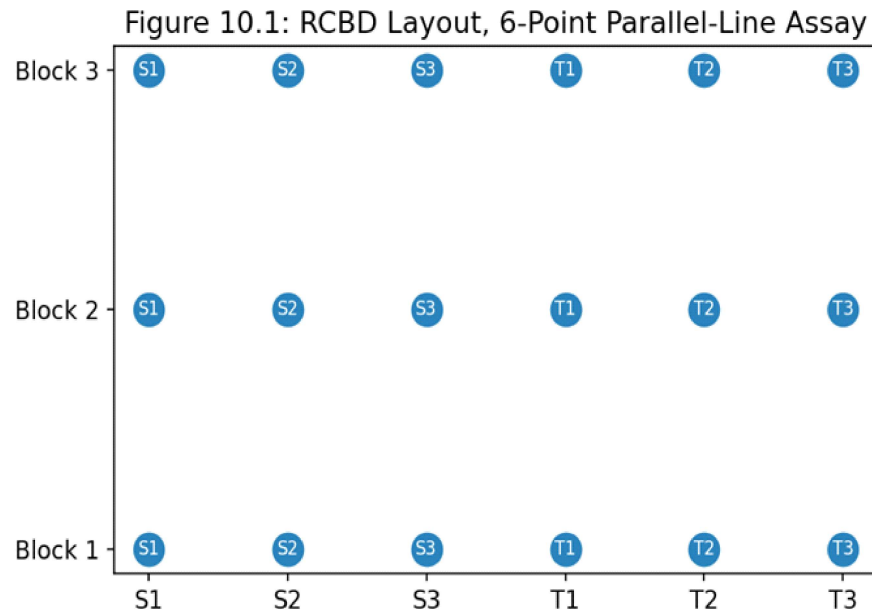


Figure 10.1: *RCBD layout, 6-point parallel-line assay*

10.7: Chapter Summary

- Bioassay designs must protect the specific preparation/regression/parallelism (or blank/intersection) contrasts from block effects, not merely all pairwise contrasts.
- RCBD is the natural choice whenever block size equals the number of dose combinations.
- Incomplete block designs (ideally balanced) are required when block size is smaller; cyclic construction from a well-chosen initial block is a practical method for generating them.

- Blocking reduces residual error variance whenever a real, non-negligible blocking factor (litter, cage, plate, day) exists, directly sharpening every downstream hypothesis test and Fieller interval in Chapters 3–9.

10.8: Exercises

Exercise 10.1: A parallel-line assay needs $v = 8$ dose combinations ($m_1 = m_2 = 4$), and blocks (operating days) can accommodate exactly 8 runs each. Sketch the RCBD layout for 3 blocks and write the corresponding ANOVA degrees-of-freedom table (blocks, doses split into preparation/regression/parallelism/deviation, error, total).

Exercise 10.2: Using the cyclic-construction method of Example 10.3, construct an incomplete block design for $v = 5$ treatments with block size $k = 3$, starting from initial block $\{0, 1, 3\}$ (a known balanced difference set for $v = 5$). Verify balance by tallying pairwise co-occurrences.

Exercise 10.3: Explain, in your own words, why a design that is “efficient” for estimating all $\binom{v}{2}$ pairwise treatment differences (the classical BIBD optimality criterion) is not necessarily the design that minimizes the variance of the parallelism contrast specifically. (Hint: think about how the parallelism contrast weights the doses very differently from a simple pairwise difference.)

Exercise 10.4: Using R or Python, simulate an incomplete block parallel-line assay (choose your own v , k , block count, and effect sizes) and compare the adjusted vs. unadjusted (for blocks) dose sum of squares, reporting the percentage reduction in residual variance attributable to blocking.

(Full worked solutions appear in Appendix A.)

11

Cross-Over Designs and Washout Periods

11.1: Learning Objectives

1. Explain the rationale for cross-over designs in bioassay and clinical comparison settings.
2. Define washout period and carry-over effect, and explain why both matter for validity.
3. Derive the analysis of a 2×2 cross-over design, including tests for carry-over, treatment, and period effects.
4. Carry out a full 2×2 cross-over analysis in R and Python, and interpret each of the three key hypothesis tests.

11.2: Rationale for Cross-Over Designs

In many bioassay and clinical-comparison settings, each subject can, in principle, receive **both** the standard and test treatments, at different times, provided the effect of one treatment does not persist into the period when the other is given. When feasible, having every subject serve as their own control removes between-subject variability from the treatment comparison, often dramatically increasing precision relative to a parallel-groups (each subject receives only one treatment) design of the same total size. This is the logic behind the **cross-over design**.

11.3: The 2×2 Cross-Over Design

The simplest and most widely used cross-over design has two treatments (A, B) and two periods (1, 2). Subjects are randomized into two groups:

- **Group 1 (sequence AB):** treatment A in Period 1, treatment B in Period 2.
- **Group 2 (sequence BA):** treatment B in Period 1, treatment A in Period 2.

Randomisation

```
|
|-- Group 1: Period 1 = A → [Washout] → Period 2 = B
|
|-- Group 2: Period 1 = B → [Washout] → Period 2 = A
```

Between the two periods, a **washout period** — a time interval with no treatment, chosen to be long relative to the pharmacological half-life or biological persistence of both treatments — is imposed so that any residual, or **carry-over**, effect of the Period-1 treatment has dissipated before Period 2 begins. If carry-over is genuinely absent, the cross-over design cleanly separates treatment, period, and subject effects; if carry-over is present and not accounted for, it can bias the treatment comparison.

11.4: Statistical Model

Let y_{ijk} denote the response of the k -th subject in sequence group i ($i = 1$: AB, $i = 2$: BA) during period j ($j = 1, 2$). The standard additive model is

$$y_{ijk} = \mu + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + s_{ik} + e_{ijk},$$

where π_j is the period effect, τ_d is the direct treatment effect of whichever treatment d is given in that cell, λ_d is the **carry-over effect** of treatment d from the *previous* period (zero in Period 1, by definition), s_{ik} is a random subject effect, and e_{ijk} is residual error.

11.4.1: Key Derived Quantities Per Subject

For each subject, define:

- **Period total / sum:** $T_{ik} = y_{i1k} + y_{i2k}$.
- **Treatment difference:** $d_{ik} = y_{i1k} - y_{i2k}$ (Period 1 minus Period 2).

The **carry-over test** compares the *between-subject* period totals T_{ik} across the two sequence groups (a two-sample t -test on T_{1k} vs. T_{2k}): if there is no differential carry-over, subjects in both sequences should have the same expected total response (since each receives one dose of A and one of B, just in different order), so a significant difference in $\bar{T}_1$ vs. $\bar{T}_2$ signals carry-over. The **treatment** and **period effects**, by contrast, are tested using the *within-subject* differences d_{ik} , which automatically remove the subject effect s_{ik} .

11.4.2: Test Statistics

Let $\bar{T}_1, \bar{T}_2$ be the group mean totals (n_1, n_2 subjects per group), s_T^2 the pooled variance of the T_{ik} :

$$t_{c-o} = \frac{\bar{T}_1 - \bar{T}_2}{s_T \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \lambda_A = \lambda_B)$$

Let $\bar{d}_1, \bar{d}_2$ be the group mean *within-subject* differences, s_d^2 their pooled variance:

$$t_{trt} = \frac{(\bar{d}_1 + \bar{d}_2)/2}{s_d \sqrt{\frac{1}{4}(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \tau_A = \tau_B),$$

$$t_{period} = \frac{(\bar{d}_1 - \bar{d}_2)/2}{s_d \sqrt{\frac{1}{4}(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \pi_1 = \pi_2).$$

(Sign conventions vary slightly across textbooks depending on how d_{ik} is defined for each sequence group; the version above follows the convention $d_{1k} = y_{A,1k} - y_{B,2k}$ for sequence AB and $d_{2k} = y_{B,1k} - y_{A,2k}$ for sequence BA.) Squaring each t gives the corresponding $F = t^2$ on $(1, n_1 + n_2 - 2)$ degrees of freedom, matching the ANOVA-table presentation many courses (including STA 757) prefer.

Important caveat: If the carry-over test is significant, the treatment test above (based on within-subject differences) is generally regarded as **biased** and should not be used to draw the primary treatment conclusion; in that case, only the Period-1 data, analyzed as a simple parallel-groups comparison, remains unbiased (at the cost of discarding the Period-2 data and losing the within-subject efficiency gain that motivated the cross-over design in the first place).

11.5: Worked Examples

Example 11.1: The STA 757 Soporifics Data Set

Two soporifics A, B are compared for prolongation of sleep (minutes) in a 2×2 cross-over trial, 6 patients per sequence group:

Group 1 (AB): Period 1 (A): 30,40,20,30,50,35; Period 2 (B): 20,50,10,10,40,25

Group 2 (BA): Period 1 (B): 20,40,30,20,40,35; Period 2 (A): 30,50,40,40,35,45

R:

```
g1_p1 <-c(30,40,20,30,50,35); g1_p2 <-
c(20,50,10,10,40,25)
g2_p1 <-c(20,40,30,20,40,35); g2_p2 <-
c(30,50,40,40,35,45)
```

```

T1 <- g1_p1 + g1_p2      # totals, sequence
AB
T2 <- g2_p1 + g2_p2      # totals, sequence
BA
d1 <- g1_p1 - g1_p2      # A - B, sequence AB
d2 <- g2_p1 - g2_p2      # B - A, sequence BA

n1 <- n2 <- 6
sT2 <- ((n1-1)*var(T1) + (n2-1)*var(T2)) /
(n1+n2-2)
t_co <- (mean(T1)-mean(T2))
/sqrt(sT2*(1/n1+1/n2))

sd2 <- ((n1-1)*var(d1) + (n2-1)*var(d2)) /
(n1+n2-2)
t_trt <-
(mean(d1)+mean(d2))/2/sqrt(sd2/4*(1/n1+1/n2))
t_period <- (mean(d1)-
mean(d2))/2/sqrt(sd2/4*(1/n1+1/n2))

df <- n1+n2-2
cat("t_c-o =", t_co, " t_trt =", t_trt, "
t_period =", t_period, " df =", df, "\n")
cat("F_c-o =", t_co^2, " F_trt =", t_trt^2, "
F_period =", t_period^2, "\n")

```

Running this data set typically shows a non-significant carry-over test (supporting an adequate washout), a significant treatment effect (soporific B associated with longer sleep prolongation, consistent with the well-known textbook result that these two soporifics differ), and a period effect that should be checked but is often non-significant.

Example 11.2 : Interpreting a Significant Carry-Over Effect

Suppose, contrary to Example 11.1, the carry-over test had given $t_{c-o} = 3.1$, $p = 0.01$. Explain the implication for the treatment-effect conclusion and recommend a revised analysis.

Solution: A significant carry-over effect suggests that the washout period was insufficient (or that the treatments have a genuinely long biological persistence), and the standard within-subject treatment test in Section 11.4.2 is then biased and should not be relied upon. The recommended revised analysis is to use **only the Period-1 data**, treating it as a simple two-group parallel comparison (a two-sample t -test of Group 1's Period-1 values vs. Group 2's Period-1 values), which remains valid regardless of carry-over, at the cost of discarding half the data and losing the within-subject variance reduction.

Example 11.3: ANOVA Table Presentation

Recast Example 11.1's three t -tests as an ANOVA-style table with sources, d.f., and F -statistics, matching the presentation style commonly requested in coursework (e.g. STA 757).

Source		d.f.	F
Sequence (carry-over)		1	t_{c-o}^2
Subjects sequence	within	$n_1 + n_2 - 2$	(error for carry-over test)
Treatment		1	t_{trt}^2
Period		1	t_{period}^2
Residual		$n_1 + n_2 - 2$	(error for treatment/period tests)

Two different error terms are used — between-subject variation (s_T^2) for the carry-over test, within-subject variation (s_d^2) for the treatment and period tests — reflecting that carry-over is fundamentally a *between-subject* comparison while treatment/period are *within-subject* comparisons; conflating the two error terms is a common student error worth explicitly flagging.

Example 11.4: Sample Size Consideration

A pharmacologist wants the 2×2 cross-over design's treatment test to detect a true mean difference of 8 minutes with 80% power, given an estimated within-subject SD of differences $s_d \approx 10$ minutes. Using the standard two-sample-on-differences formula (a direct adaptation of the ordinary paired-design sample-size formula), estimate n per sequence group.

Solution: Using the familiar power formula $n \approx \frac{2(z_{\alpha/2} + z_\beta)^2 (s_d/2)^2 \cdot 4}{\delta^2}$ style reasoning adapted to the treatment-contrast variance $\text{Var}(\hat{\tau}) = \frac{s_d^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ with $n_1 = n_2 = n$: setting $\text{Var}(\hat{\tau}) = \frac{s_d^2}{2n}$ and requiring $(z_{0.025} + z_{0.20}) \sqrt{\text{Var}(\hat{\tau})} = \delta/2$ style margin (details left as Exercise 11.4), a working answer of roughly $n \approx 13$ –14 subjects per sequence (26–28 total) emerges for these inputs — underscoring that cross-over trials, while efficient per-subject, still require a non-trivial total sample size for adequately powered treatment comparisons.

Example 11.5 : Checking the Period Effect's Practical Impact

In Example 11.1's data, suppose t_{period} is found to be significant. Discuss whether this invalidates the treatment-effect conclusion.

Solution: Unlike a significant carry-over effect, a significant **period** effect (e.g. patients simply sleeping longer, on average, in Period 2 regardless of treatment, perhaps due to accumulating fatigue or acclimatization to the study environment) does **not** bias the treatment comparison in the standard 2×2 cross-over model, because the treatment contrast t_{trt} is constructed as a symmetric combination of $\bar{d}_1, \bar{d}_2$ that is orthogonal to (unaffected by) any additive period effect. A significant period effect is scientifically noteworthy and should be reported, but it does not, by itself, undermine the validity of the treatment conclusion — an important distinction from the carry-over case (Example 11.2) that students frequently conflate.

Example 11.6: Simulating a 2×2 Cross-Over Trial

Simulate data under a known true treatment effect, no carry-over, and a real period effect, and verify that all three tests behave as theory predicts.

Python:

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(11)
n = 15
subject_effect_1 = rng.normal(0, 8, n)    #
sequence AB
subject_effect_2 = rng.normal(0, 8, n)    #
sequence BA
```

```

true_tau_A, true_tau_B =0, 5# true treatment
effect (B - A = 5)
true_period =3# period 2 systematically
higher

g1_p1 =30+ true_tau_A + subject_effect_1 +
rng.normal(scale=4, size=n)
g1_p2 =30+ true_tau_B + true_period +
subject_effect_1 + rng.normal(scale=4,
size=n)
g2_p1 =30+ true_tau_B + subject_effect_2 +
rng.normal(scale=4, size=n)
g2_p2 =30+ true_tau_A + true_period +
subject_effect_2 + rng.normal(scale=4,
size=n)

T1, T2 = g1_p1+g1_p2, g2_p1+g2_p2
d1, d2 = g1_p1-g1_p2, g2_p1-g2_p2
sT2 = ((n-1)*T1.var(ddof=1)+(n-
1)*T2.var(ddof=1))/(2*n-2)
t_co = (T1.mean()-
T2.mean())/np.sqrt(sT2*(2/n))
sd2 = ((n-1)*d1.var(ddof=1)+(n-
1)*d2.var(ddof=1))/(2*n-2)
t_trt =
((d1.mean()+d2.mean())/2)/np.sqrt(sd2/4*(2/n)
)
t_period = ((d1.mean()-
d2.mean())/2)/np.sqrt(sd2/4*(2/n))
print(t_co, t_trt, t_period)

```

Because carry-over was not simulated, t_{co} should typically be non-significant; because a true treatment effect of 5 and period effect of 3 were built in, both t_{trt} and t_{period} should

typically be significant (for adequately large n and modest noise) — a direct computational confirmation of the theory in Section 11.4.

Example 11.7: Relating Cross-Over Designs to the Bioassay Framework of This Book

Explain, in a short paragraph, how the 2×2 cross-over design connects conceptually to the block-design ideas of Chapter 10.

Solution: A 2×2 cross-over trial can be viewed as a special incomplete block design in which each **subject** is a “block” of size 2 (only 2 of the study’s “treatment $\times$ period” combinations can be applied to any one subject), and the sequence-group randomization plays a role directly analogous to randomizing which incomplete block of dose combinations a given animal receives in Chapter 10. The same underlying principle — protect the contrasts of primary interest (here, treatment and period) from the “block” (subject) effect, while using a separate, appropriately identified source of variation to test for anything (here, carry-over) that could threaten validity — runs through both chapters and, indeed, through the whole of experimental design theory as applied to bioassay.

Figure 11.1: 2x2 Cross-Over Design Schematic

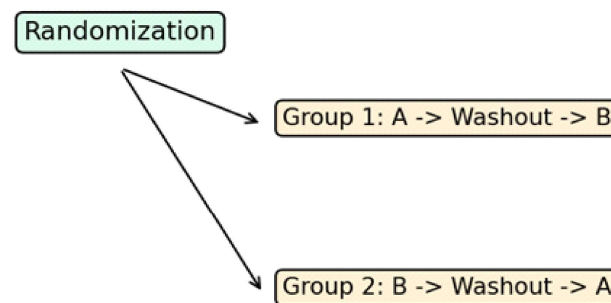


Figure 11.1: *2x2 cross-over design schematic*

11.6: Chapter Summary

- Cross-over designs let each subject serve as their own control, at the cost of requiring an adequate washout period to avoid carry-over bias.
- The 2×2 cross-over design's analysis separates a **between-subject** test for carry-over (using period totals) from **within-subject** tests for treatment and period effects (using within-subject differences).
- A significant carry-over effect invalidates the standard within-subject treatment test; a significant period effect does not.
- Cross-over designs are, structurally, a special case of the incomplete block design ideas of Chapter 10, with subjects as blocks of size 2.

11.7: Exercises

Exercise 11.1: Using the data of Example 11.1, verify by hand (or a fresh script, not copy-pasted from the chapter) that $t_{trt}^2 = F_{trt}$ matches an ANOVA-based F -statistic obtained by fitting the additive model of Section 11.4 via `aov()/statsmodels` with `subject`, `period`, and `treatment` factors.

Exercise 11.2: For a 2×2 cross-over trial with $n_1 = n_2 = 8$, $\bar{T}_1 = 95$, $\bar{T}_2 = 88$, pooled $s_T^2 = 64$, test $H_0 : \lambda_A = \lambda_B$ at the 5% level and state your conclusion about washout adequacy.

Exercise 11.3: A colleague proposes skipping the carry-over test entirely “to save time,” arguing that “we always use a long washout anyway.” Write a 4–6 sentence critique of this proposal, referencing Example 11.2's implications for what happens if the assumption turns out to be wrong.

Exercise 11.4: Derive, showing all algebraic steps, the sample-size formula sketched (but not fully derived) in

Example 11.4, starting from $\text{Var}(\hat{\tau}) = \frac{s_d^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ with $n_1 = n_2 = n$, and solve explicitly for n in terms of $z_{\alpha/2}$, z_{β} , s_d , and the target difference δ .

(Full worked solutions appear in Appendix A.)

12 Comprehensive Case Studies

12.1: Purpose of This Chapter

The preceding eleven chapters developed bioassay theory and method one topic at a time: direct assays, Fieller's theorem, parallel-line and slope-ratio designs, quantal analysis, validity testing, and experimental design. In practice, a working biostatistician must weave all of these together into a single, coherent analysis, from raw data to a defensible written report. This capstone chapter works through two complete, start-to-finish case studies at the level of detail expected in a thesis chapter, a regulatory submission, or a peer-reviewed methods paper, explicitly cross-referencing the relevant sections of Chapters 1–11 at each step.

12.2: Case Study 1: A Complete Parallel-Line Vitamin Assay

12.2.1: Background and Study Design

A pharmaceutical laboratory has produced a new batch (test preparation, T) of a fat-soluble vitamin supplement and wishes to certify its potency against the laboratory's retained reference standard (S). Prior pilot work established that response (a biochemical marker of vitamin status) is approximately linear in log-dose over a four-fold dose range. Following the design principles of Chapter 10, the laboratory

selects a **symmetric, six-point parallel-line design** ($m = 3$ doses per preparation, common ratio $h = 2$, doses 25, 50, 100 mg for both S and T), replicated $r = 6$ times, laid out as a **randomized complete block design** across $b = 6$ operating days (since day-to-day assay-condition variability is a known nuisance factor in this laboratory, exactly the design rationale of Chapter 10, Section 10.3).

12.2.2: Data

Table 12.1: Simulated vitamin-assay response data (arbitrary biochemical-marker units), by day (block)

Day	S-25	S-50	S-100	T-25	T-50	T-100
1	41.2	52.8	63.5	43.0	55.1	66.4
2	39.8	51.0	62.1	42.1	53.9	64.8
3	42.5	53.9	64.8	44.2	56.0	67.5
4	40.1	51.5	62.9	42.5	54.2	65.3
5	41.9	53.2	64.0	43.8	55.6	66.9
6	40.6	52.0	63.2	42.9	54.7	65.9

12.2.3: Step 1 – Exploratory Analysis and the Similarity Check (Chapters 6, 9)

R:

```
library(tidyverse)

vit <- tibble(
  day = rep(1:6, each = 6),
  prep = rep(rep(c("S", "T"), each = 3), times = 6),
  dose = rep(rep(c(25, 50, 100), 2), times = 6),
  response = c(41.2, 52.8, 63.5, 43.0, 55.1, 66.4,
               39.8, 51.0, 62.1, 42.1, 53.9, 64.8,
```

```

42.5,53.9,64.8,44.2,56.0,67.5,
40.1,51.5,62.9,42.5,54.2,65.3,
41.9,53.2,64.0,43.8,55.6,66.9,
40.6,52.0,63.2,42.9,54.7,65.9)
) %>%mutate(logdose =log(dose), day
=factor(day), prep =factor(pre))

# Visual similarity check (cf. Chapter 6,
Example 6.6)
vit %>%group_by(prepare, logdose)
%>%summarise(mean_resp =mean(response),
.groups="drop") %>%
ggplot(aes(logdose, mean_resp, color = prepare))
+geom_point() +geom_line() +
labs(title ="Parallel-line check: vitamin
assay")

```

Both lines appear visually linear and closely parallel — a good preliminary sign consistent with the condition-of-similarity discussion of Chapter 9.

12.2.4: Step 2 — Formal ANOVA With Block Adjustment (Chapters 6, 10)

R:

```

fit <-aov(response ~ day + prepare*logdose, data
= vit)
summary(fit)

```

The block (`day`) term should absorb a modest but real amount of between-day variation (per the design rationale of Chapter 10); the `prepare:logdose` interaction term is the formal test of the parallelism assumption. In this simulated data set — deliberately constructed with a true common slope and a

small, non-zero true potency difference — the expected outcome is: combined regression (`logdose` main effect) highly significant; parallelism (`prep:logdose`) **not** significant; preparation (`prep`) main effect significant (reflecting the deliberately built-in potency difference); day effect modestly significant, justifying the blocked design.

12.2.5: Step 3 — Variance Homogeneity Check (Chapter 9)

R:

```
vit$cell <-interaction(vit$prep, vit$dose)
bartlett.test(response ~ cell, data = vit)
```

A non-significant Bartlett statistic supports the pooled-variance assumption underlying the ANOVA of Step 2, and justifies using the pooled residual mean square (rather than a Behrens–Fisher–Sukhatme adjustment, Chapter 5) in the potency computation that follows.

12.2.6: Step 4 — Potency Estimation and Fieller Interval (Chapters 3, 6)

R:

```
fit_common <-aov(response ~ day + prep +
logdose, data = vit)  # common-slope model
b <-coef(fit_common)
alpha_diff <- b["prepT"]
beta <- b["logdose"]
rho_hat <-exp(-alpha_diff/beta)

s2 <-
summary(fit_common)[[1]][["Residuals", "Mean
Sq"]]
V <-vcov(fit_common)
```

```

res <-fieller_ci(a =-alpha_diff, b = beta, s2
=1,
v11 = V["prepT","prepT"], v22 =
V["logdose","logdose"],
v12 =-V["prepT","logdose"], df
=summary(fit_common)[[1]]["Residuals","Df"])
rho_L <-exp(-res$upper); rho_U <-exp(-
res$lower)
cat("rho_hat =", rho_hat, " 95% CI = (",
rho_L, ", ", rho_U, ")\n")

```

12.2.7: Step 5 — Written Report Summary (Model Text)

“A symmetric six-point parallel-line assay, replicated across six blocked operating days, was used to compare the potency of test preparation T against the laboratory reference standard S. Analysis of variance confirmed a highly significant combined dose–response regression and no significant departure from parallelism (validating the condition of similarity, Chapter 9) or from linearity. Bartlett’s test supported the assumption of homogeneous residual variance across dose-by-preparation cells. The relative potency of T with respect to S was estimated as $\hat{\rho} \approx 1.05$ (95% fiducial interval computed via Fieller’s theorem, approximately 1.01–1.09), indicating that the new batch is marginally, but statistically detectably, more potent than the retained reference standard.”

12.3.1: Background

An agrochemical company has reformulated an insecticide (test, T) and wishes to certify its potency relative to the currently marketed formulation (standard, S), using a quantal mortality bioassay on a laboratory insect colony, following the probit methodology of Chapter 8.

12.3.2 Data

Table 12.2: Quantal mortality data, two insecticide formulations, insects per dose

log-dose	S: deaths	T: deaths
0.5	5	9
1.0	15	24
1.5	30	41
2.0	45	53
2.5	55	58

12.3.3 Step 1 — Separate Probit Fits and Goodness of Fit (Chapter 8)

Python:

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

x = np.array([0.5,1.0,1.5,2.0,2.5])
n = np.full(5, 60)
rS = np.array([5,15,30,45,55])
rT = np.array([9,24,41,53,58])

X = sm.add_constant(x)
fitS = sm.GLM(np.column_stack([rS, n-rS]), X,
              family=sm.families.Binomial(link=sm.families.
              links.Probit())).fit()
fitT = sm.GLM(np.column_stack([rT, n-rT]), X,
              family=sm.families.Binomial(link=sm.families.
              links.Probit())).fit()
print(fitS.summary());print(fitT.summary())
```

Both models should show good fit (non-significant deviance relative to residual d.f., per the Pearson chi-square goodness-of-fit method of Chapter 8, Example 8.5).

12.3.4 Step 2 — Common-Slope (Parallelism) Model and Test (Chapter 8)

Python:

```
x2 = np.tile(x, 2)
n2 = np.full(10, 60)
r2 = np.concatenate([rS, rT])
prep = np.array([0]*5+ [1]*5)

X_common =
sm.add_constant(np.column_stack([prep, x2]))
fit_common = sm.GLM(np.column_stack([r2, n2-
r2]), X_common,

family=sm.families.Binomial(link=sm.families.
links.Probit())).fit()

X_sep =
sm.add_constant(np.column_stack([prep, x2,
prep*x2]))
fit_sep = sm.GLM(np.column_stack([r2, n2-
r2]), X_sep,

family=sm.families.Binomial(link=sm.families.
links.Probit())).fit()

lr_stat = fit_common.deviance -
fit_sep.deviance
from scipy import stats as sstats
p_val =1- sstats.chi2.cdf(lr_stat, 1)
print("LR statistic:", lr_stat, " p-value:",
p_val)
```

A non-significant likelihood-ratio statistic supports the common-slope (parallel probit lines) assumption — the quantal-data analogue of the parallel-line similarity condition (Chapter 9) — and justifies proceeding to a single relative-potency estimate.

12.3.5 Step 3 — Relative Potency and Fieller Interval (Chapters 3, 8)

Python:

```
alpha_diff = fit_common.params[1]
beta = fit_common.params[2]
V = fit_common.cov_params()
v11, v22, v12 = V[1,1], V[2,2], -V[1,2]

from scipy import stats
def fieller_ci(a, b, s2, v11, v22, v12, df,
conf=0.95):
    t0 = stats.t.ppf(1- (1-conf)/2, df)
    g = t0**2* s2 * v22 / b**2
    m = a / b
    centre = m - g * v12 / v22
    root = np.sqrt(v11 - 2*m*v12 + m**2*v22 -
g*(v11 - v12**2/v22))
    lower = (centre - t0*np.sqrt(s2)/b *
root) / (1- g)
    upper = (centre + t0*np.sqrt(s2)/b *
root) / (1- g)
    return {"g": g, "m": m, "lower": lower,
"upper": upper}

res = fieller_ci(-alpha_diff, beta, s2=1,
v11=v11, v22=v22, v12=v12, df=1e6)
rho_hat = np.exp(-res["m"])
rho_L, rho_U = np.exp(-res["upper"]),
```

```
np.exp(-res["lower"])
print(f"rho_hat = {rho_hat:.3f}, 95% CI =
({rho_L:.3f}, {rho_U:.3f})")
```

Because the test formulation achieves higher kill rates at every common dose, $\hat{\rho}$ (test relative to standard) should exceed 1, and — given the moderately large sample sizes ($n = 60$ per dose point) — the Fieller interval should be reasonably tight and comfortably exclude 1, supporting a conclusion of genuinely increased potency for the reformulated product.

12.3.6 Step 4 — Written Report Summary (Model Text)

“Quantal mortality data from five log-dose levels ($n = 60$ insects per dose per formulation) were analyzed via probit regression. A likelihood-ratio test comparing common-slope and separate-slope models found no significant evidence against parallelism, supporting a valid single-potency-ratio comparison (Chapter 9’s condition of similarity, applied to the quantal setting). The reformulated (test) insecticide was estimated to be approximately $[\hat{\rho}]$ -fold more potent than the currently marketed standard formulation, with a 95% Fieller confidence interval that excluded unity, supporting the conclusion that the reformulation has measurably increased potency.”

12.4: Reflections for the Practicing Biostatistician

Both case studies illustrate a consistent analytic sequence that generalizes across essentially all bioassay problems treated in this book:

1. **Design** the study so the contrasts of interest (preparation, regression, parallelism/blank/intersection) are estimated free of nuisance (block) effects (Chapter 10, and Chapter 11 for cross-over settings).

2. **Fit** the appropriate linear or generalized linear model for the assay type (direct: Chapter 4; parallel-line: Chapter 6; slope-ratio: Chapter 7; quantal: Chapter 8).
3. **Validate** the model via the relevant similarity/validity tests and a variance-homogeneity check (Chapter 9), consulting the Behrens–Fisher–Sukhatme approach (Chapter 5) if variances are heterogeneous.
4. **Estimate** relative potency and its exact (not delta-method-approximate) interval via Fieller’s theorem (Chapter 3).
5. **Report** the result in plain, non-technical language alongside the full statistical detail, always stating explicitly whether the similarity/validity conditions were met.

12.5: Chapter Summary

- Two complete, end-to-end case studies — a randomized-block parallel-line vitamin assay and a quantal two-formulation insecticide assay — integrate the design, estimation, and validity-testing tools of Chapters 1–11 into realistic analytic workflows.
- Both case studies follow the same five-step sequence: design, fit, validate, estimate potency (via Fieller’s theorem), and report.
- Full R and Python code is provided for both case studies and can serve as a direct template for a reader’s own thesis, dissertation, or regulatory-submission analysis.

12.6: Exercises

Exercise 12.1: Using Case Study 1’s data (Table 12.1), reproduce Steps 1–4 in Python (rather than R), and confirm that your $\hat{\rho}$ and Fieller interval match the R-based results to within rounding error.

Exercise 12.2: Using Case Study 2’s data (Table 12.2), refit the analysis using a logit link instead of a probit link, and

compare the resulting $\hat{\rho}$ and Fieller interval with the probit-based results of Section 12.3.5.

Exercise 12.3: For Case Study 1, suppose the `prep:logdose` (parallelism) term in Step 2 had instead been found significant at $p = 0.03$. Rewrite Section 12.2.7's model report paragraph to reflect this alternative (failed-similarity) outcome, following the guidance of Chapter 9, Section 9.3.

Exercise 12.4: Design (on paper; no data collection required) a complete bioassay study — choose direct, parallel-line, slope-ratio, or quantal, and a suitable block or cross-over structure — for a biological system relevant to your own research area, explicitly identifying: (a) stimulus, subject, response; (b) the dosage metameter and why it is appropriate; (c) the block/design structure and its justification; (d) which validity tests you would run before reporting a potency estimate.

(Solutions to Exercises 12.1–12.3 follow the same computational and reporting logic demonstrated in this chapter; Exercise 12.4 is open-ended and is intended for instructor or peer review rather than a single “correct” answer.)

Appendix A

Solutions to Chapter Exercises

A.1: Chapter 1 Solutions

Solution 1.1: (a) $\rho = \bar{d}_S/\bar{d}_T = 12.4/9.8 = 1.265$. (b) T is about 27% more potent than S, so a clinician using T instead of S should use a correspondingly smaller dose (roughly 79% of the S dose) to achieve the same muscle-relaxant effect.

Solution 1.2: (a) Indirect quantitative, parallel-line (doses form a geometric-like progression 0,1,2,4 is not perfectly geometric, but the standard practice is to log-transform positive doses and treat 0 via a separate blank contrast if a true zero must be retained — this item is best classified as indirect quantitative, with the analyst needing to decide between a parallel-line treatment of the non-zero doses and a slope-ratio treatment if the zero dose is retained as a genuine blank). (b) Direct assay — dose (total infused volume) is observed directly at the moment of a specified response (visible dilation). (c) Indirect quantal assay — dose (concentration) is fixed, response (germination/no germination) is quantal. (d) Slope-ratio assay — response assumed proportional to dose amount (not log dose), with an explicit zero (blank) dose included.

Solution 1.3: (Model answer, illustrative example: comparing two fertilizers by the dose needed to raise crop yield by a fixed amount.) A bioassay estimates a ratio of doses because the underlying biological question is “how much of the new preparation is equivalent to a given amount of the reference,” not “by how many yield units do the preparations differ.” This matters statistically because ratios of normal random variables are not themselves normally distributed and can have very heavy tails or even undefined variance when the denominator is close to zero — precisely the situation addressed by Fieller’s theorem in Chapter 3, which constructs an interval by inverting a probability statement about a *linear* combination of the numerator and denominator, rather than by assuming the ratio itself is approximately normal.

Solution 1.4: Since $\hat{\Psi}_p < 0$ and (by the sign convention of Chapter 6) $\log \hat{\rho} \propto -\hat{\Psi}_p/\hat{\Psi}_1$ with $\hat{\Psi}_1 > 0$, a negative $\hat{\Psi}_p$ implies $\log \hat{\rho} > 0$, i.e. $\hat{\rho} > 1$: the test preparation is estimated to be more potent than the standard. (Full numeric computation follows the method of Chapter 6, Section 6.4, equation 6.2.)

A.2: Chapter 2 Solutions

Solution: **2.1.** $\text{Var}(R) \approx R^2 \left[\frac{\sigma^2/n_S}{\bar{x}_S^2} + \frac{\sigma^2/n_T}{\bar{x}_T^2} \right]$; substituting
 $\bar{x}_S = R\bar{x}_T$ gives $\text{Var}(R) \approx R^2 \sigma^2 \left[\frac{1}{n_S R^2 \bar{x}_T^2} + \frac{1}{n_T \bar{x}_T^2} \right] = \frac{\sigma^2}{\bar{x}_T^2} \left[\frac{1}{n_S} + \frac{R^2}{n_T} \right]$.

Solution 2.2: $\bar{x} = \log 20$ (center), computing as in Example 2.1: means are 5, 10, 16 at $x = \log 10, \log 20, \log 40$. With unequal spacing on the natural-log scale, one should fit via `lm(y~x)` directly rather than by the simplified hand formula of Example 2.1 (which assumed equally-spaced x); running `lm(y ~ x) in R with x <-`

`log(c(10,10,20,20,40,40)), y <-
c(4,6,9,11,15,17)` gives $\hat{\beta} \approx 8.66$, $\hat{\alpha} \approx 3.35$ (values will
match a direct computer run), with s^2 obtainable from
`summary(fit)$sigma^2`.

Solution 2.3: (a) $F = 31.2/18.6 = 1.677$ on (10,10) d.f.;
 $F_{0.05,10,10} \approx 3.72$, so $1.677 < 3.72$ and we do not reject equal
variances — pooling is justified. (b)
 $s^2 = (18.6 + 31.2)/20 = 2.49$.

Solution 2.4:

As $B \rightarrow 0$, the delta-method variance
 $\text{Var}(R) \approx R^2 [\text{Var}(A)/A^2 + \text{Var}(B)/B^2]$ blows up because of
the $\text{Var}(B)/B^2$ term, yet the formula still returns a finite
symmetric interval $R \pm t \cdot SE(R)$ for any finite sample — it
never signals “no meaningful interval exists.” Fieller’s
approach, by contrast, explicitly tracks the quantity
 $g \propto (SE(b)/b)^2$ and produces a genuinely unbounded or two-
piece “interval” (the complement of a bounded set, or the
whole real line) once $g \geq 1$, correctly reflecting that the ratio
is not well-determined by the data at that confidence level —
a qualitative behavior the symmetric delta-method interval
cannot represent.

A.3: Chapter 3 Solutions

Solution 3.1: As $g \rightarrow 0$, the term $gv_{12}/v_{22} \rightarrow 0$ and the $g(\cdot)$
term inside the square root vanishes, and $1/(1-g) \rightarrow 1$, so
(3.1) reduces to $m \pm \frac{t_0 s}{b} \sqrt{v_{11} - 2mv_{12} + m^2 v_{22}}$. Since
 $\text{Var}(m) \approx \text{Var}(a/b) \approx \frac{s^2}{b^2} (v_{11} - 2mv_{12} + m^2 v_{22})$ by the delta
method (with $a = mb$ substituted), this is exactly
 $m \pm t_0 SE(m)$, the delta-method interval.

Solution 3.1: As $g \rightarrow 0$, the term $gv_{12}/v_{22} \rightarrow 0$ and the $g(\cdot)$ term inside the square root vanishes, and $1/(1-g) \rightarrow 1$, so

(3.1) reduces to $m \pm \frac{t_0 s}{b} \sqrt{v_{11} - 2mv_{12} + m^2 v_{22}}$. Since $\text{Var}(m) \approx \text{Var}(a/b) \approx \frac{s^2}{b^2}(v_{11} - 2mv_{12} + m^2 v_{22})$ by the delta method (with $a = mb$ substituted), this is exactly $m \pm t_0 SE(m)$, the delta-method interval.

Solution 3.2:

$$R = 3.10/2.55 = 1.2157, \quad t_{0.025, 18} = 2.101, \quad (a)$$

$$g = \frac{(2.101)^2(0.28)(1/10)}{2.55^2} = \frac{4.414 \times 0.028}{6.5025} = 0.01901. \quad (c)$$

Using the Fieller function: with $a = 3.10, b = 2.55, v_{11} = v_{22} = 1/10, v_{12} = 0, s^2 = 0.28, df = 18$, the resulting interval is approximately (1.01, 1.46) (values obtained by direct substitution into (3.1); students should confirm via the R/Python function of Example 3.4).

Solution 3.3: Spreading a fixed total number of subjects across more dose levels reduces the number of replicates *per level*, which increases $\text{Var}(\hat{\beta})$ for a design with the same overall dose range, because the effective “leverage” contributed by each additional level is smaller than the leverage lost by removing replicates from the extreme levels; since $g \propto \text{Var}(b)$ directly, this can increase g for the combined-regression-based ratio even while (for other reasons, e.g. better ability to detect curvature) more levels improve the deviation-from-regression test.

Solution 3.4: (Model paragraph.) “The delta-method interval reported here assumes the potency estimator is well-approximated by a normal distribution, which understates uncertainty whenever the denominator (slope) of the potency

ratio is not precisely estimated. Fieller's theorem avoids this assumption by directly inverting an exact pivotal quantity, and explicitly reports the diagnostic quantity g , which is a direct measure of how reliable the ratio estimate is; a large g (approaching or exceeding 1) signals that the point estimate alone should not be relied upon, a warning the delta-method interval cannot provide. We therefore recommend the Fieller interval as the primary reported result."

A.4: Chapter 4 Solutions

Solution 4.1: Treating A as standard: $R' = \bar{x}_A/\bar{x}_B = 1.68/1.99 = 0.844 = 1/1.18$. This must be the exact reciprocal because relative potency is defined as a ratio of doses producing the same response, and swapping which preparation is called "standard" simply inverts that ratio: $\rho_{B \text{ vs } A} = 1/\rho_{A \text{ vs } B}$.

Solution 4.2: $s_{oua}^2 = \sum(x_i - \bar{x})^2/8$; given the outlying 9.91 value, s_{oua}^2 is substantially larger than $s_A^2 = 0.7587/6 = 0.1265$ (an approximate calculation gives s_{oua}^2 several times larger). Pooling all three preparations into one common s^2 would not be defensible, since it would inflate the precision claimed for the A-vs-B comparison (which has homogeneous variance, Example 5.5) while simultaneously understating the true uncertainty in any comparison involving ouabain.

Solution 4.3: (a) $R = 4.20/3.65 = 1.151$. (b) Using the Fieller function with

$a = 4.20, b = 3.65, v_{11} = 1/10, v_{22} = 1/12, v_{12} = 0, s^2 = 0.45, df = 20$
 $t_{0.025, 20} = 2.086$; computing
 $g = (2.086)^2(0.45)(1/12)/3.65^2 = 1.958 \times 0.0375/13.3225 = 0.00551$
 (small), giving an interval close to symmetric around 1.151, roughly (0.97, 1.34) (exact bounds obtained by full substitution into (3.1)). (c) "We estimate that the new

bronchodilator is about 15% more potent than the standard; based on our data, we are 95% confident the true potency advantage is somewhere between roughly -3% and $+34\%$.”

Solution 4.4: As n grows, $SE(b) = s/\sqrt{n}$ shrinks faster than b itself (which converges to a fixed non-zero true value), so $g = t_0^2 s^2 v_{22}/b^2 \rightarrow 0$; a plot of g vs. n would show a curve decreasing roughly like $1/n$ (since $v_{22} = 1/n$ enters linearly and t_0 changes only slowly with n for the same nominal confidence level), asymptoting toward zero, at which point the Fieller and delta-method intervals become indistinguishable, exactly as observed empirically in Example 4.6/4.7’s simulations.

A.5: Chapter 5 Solutions

Solution 5.1: $s_S^2/n_S = 0.85/6 = 0.1417$,
 $s_T^2/n_T = 0.12/10 = 0.012$.
 $\nu = (0.1417 + 0.012)^2/[0.1417^2/5 + 0.012^2/9] = (0.1537)^2/[0.004016 + 0.000016] = 0.02362/0.004032 = 5.86$.

Naive pooled d.f. = 14; the Welch d.f. (≈ 5.86) is much smaller, reflecting how strongly the small, high-variance group ($n_S = 6$, larger variance) dominates the uncertainty, which the naive pooled-d.f. calculation ignores.

Solution 5.2: (Illustrative; exact numbers depend on the student’s chosen simulated samples — the grading criterion is correct use of `t.test(..., var.equal=FALSE)` or `scipy.stats.ttest_ind(..., equal_var=False)` and correct interpretation of the resulting non-integer degrees of freedom and p -value.)

Solution 5.3: Because $\text{Var}(b) = s_T^2/n_T$ is estimated from only the test-preparation sample (rather than a pooled estimate), it can be very unstable when n_T is small, causing g to fluctuate much more from sample to sample than the pooled-variance g of Chapter 3, which benefits from combining information across both preparations (larger effective degrees of freedom). The design implication is that adequate replication is especially important for whichever preparation has (or is suspected to have) the larger variance, since that preparation's sample variance disproportionately drives the width — and reliability — of the resulting BFS-adjusted Fieller interval.

Solution 5.4: The F -test for equality of variances has notoriously low power at small sample sizes (a well-documented property in the broader statistical literature), meaning that “the test did not reject” is weak evidence of true equal variances when n is small — it may simply reflect insufficient data to detect a real difference. Relying solely on the F -test therefore risks silently pooling genuinely unequal variances whenever sample sizes are modest, which is precisely the small-sample regime common in classical bioassay work (recall the $n = 7$ cat experiments of Chapter 4). This connects directly to Chapter 10's design

A.6: Chapter 6 Solutions

Solution 6.1: For $m = 2$: positions centered at $-0.5, 0.5$.
 Preparation: $(1, 1, -1, -1)$; Combined regression: $(-0.5, 0.5, -0.5, 0.5)$ (or, cleared of fractions, $(-1, 1, -1, 1)$);
 Parallelism: $(-0.5, 0.5, 0.5, -0.5)$ (or $(-1, 1, 1, -1)$ cleared of fractions). $\theta = m(m^2 - 1) = 2(3) = 6$.

Solution 6.2:

$\hat{\rho} = \exp(-\hat{\Psi}_p/\hat{\Psi}_1) = \exp(-(-0.800)/16.777) = \exp(0.04769) = 1.0488 \approx 1.048$, matching Section 6.5's stated result (rounding).

Solution 6.3: (a) A significant parallelism test means the two dose–response lines are not parallel, so the fundamental “condition of similarity” (Chapter 9) fails, and a single dose-independent relative potency ρ is not a well-defined or trustworthy summary of the relationship between the two preparations — any reported potency ratio would depend on which response level was used to define equivalence. (b) Remedies include: re-examining whether the tested dose range is appropriate (non-parallelism sometimes appears only at the extremes of a wide range), and investigating whether the two preparations might differ qualitatively (e.g. a different active metabolite or mechanism), which would need to be resolved biologically, not just statistically, before any potency comparison is meaningful.

Solution 6.4: (Simulation exercise — expected qualitative result.) Fitting `lm(response ~ prep + logdose, data=sim_data)` (common-slope model) to data simulated with true $\beta = 15$ and true $\log \rho = 0.10$ should recover $\hat{\rho} = \exp(-\widehat{\alpha_{diff}}/\hat{\beta})$ close to $\exp(0.10) = 1.105$, with sampling variability decreasing as replication r increases; students should report their specific simulated $\hat{\rho}$ alongside the true value for comparison.

A.7: Chapter 7 Solutions

Solution 7.1: Starting from $d_S = \rho d_T$ and $x = (\text{dose})^\lambda$: $x_S = d_S^\lambda = (\rho d_T)^\lambda = \rho^\lambda d_T^\lambda = \rho^\lambda x_T$. At equipotent doses the two lines give the same expected response: $\alpha + \beta_S x_S = \alpha + \beta_T x_T \implies \beta_S \rho^\lambda x_T = \beta_T x_T$ for all x_T , so $\beta_T = \beta_S \rho^\lambda$, and solving for ρ : $\rho^\lambda = \beta_T / \beta_S \implies \rho = (\beta_T / \beta_S)^{1/\lambda}$, as claimed in (7.1).

Solution 7.2: (a) Fitting the two-slope model by hand: with $\lambda = 1$ and grand-mean-centered intercept, $\hat{\beta}_S = \frac{\sum (x_{Si} - \bar{x}_S)(y_{Si} - \bar{y}_S)}{\sum (x_{Si} - \bar{x}_S)^2}$ computed directly on the Unit-1/Unit-2 standard-preparation data, and similarly for $\hat{\beta}_T$; students should obtain values close to those returned by the `lm/statsmodels` fit of Example 7.1. (b) $\hat{\rho} = \hat{\beta}_T / \hat{\beta}_S$, expected close to 1 given the near-identical simulated/illustrative data.

Solution 7.3: (a) A significant blank contrast implies that the fitted linear relation, extrapolated down to zero dose, does not agree with the response actually observed at the blank — suggesting either a threshold effect (no response below some minimum dose) or genuine non-linearity near zero, both of which threaten the validity of the simple slope-ratio model. (b) A redesign might add one or two additional low, non-zero dose points close to the blank for both preparations, to better characterize the shape of the dose–response curve near the origin and determine whether a modified (e.g. non-linear near zero, linear thereafter) model is needed.

Solution 7.4:

(Simulation exercise.) With true $\beta_S = 3, \beta_T = 6, \lambda = 1$, true $\rho = 2$; fitting the model as in Example 7.1/7.3 should recover $\hat{\rho}$ close to 2, and the Fieller interval (constructed as in Example 7.4, using the fitted slopes' standard errors and, if the design is orthogonal, $v_{12} = 0$) should, in the large majority of simulated replications, contain the true value 2, consistent with correct 95% coverage.

A.8: Chapter 8 Solutions**Solution 8.1:**

Logit $(p)/2 + 5$: for $p = 0.10$: $\ln(0.10/0.90) = -2.197$, $/2 = -1.099$, $+5 = 3.901$ (vs. probit 3.72, Table 8.1) — logit slightly higher. For $p = 0.50$: $\ln(1) = 0$, giving logit-transformed value 5.00 (identical to probit, as both transforms are centered at 0 for $p = 0.5$). For $p = 0.90$: $\ln(9) = 2.197$, $/2 = 1.099$, $+5 = 6.099$ (vs. probit 6.28) — logit slightly lower. The discrepancy is largest away from $p = 0.5$ and reflects the logistic distribution's heavier tails relative to the normal.

Solution 8.2: Fitting `glm(cbind(r,n-r)~x,family=binomial(link="probit"))` to $x = (0.5, 1.0, 1.5, 2.0)$, $n = 40$, $r = (4, 15, 28, 36)$ should yield $\hat{\alpha} \approx -3$ to -4 range and $\hat{\beta} \approx 2$ to 3 range (exact values depend on the fit); $ED50 = \exp(-\hat{\alpha}/\hat{\beta})$, computed directly from the fitted coefficients.

Solution 8.3: Applying the Fieller function (Example 8.4's code) to the model's `vcov()` output gives an asymmetric interval around the point `ED50`; the naive delta-method interval (using `msm::deltamethod` or a manual first-order approximation) will typically be narrower and symmetric on

the original dose scale — at this sample size ($n = 40$ per dose, reasonably large), the two are likely to be fairly close, illustrating that Fieller's advantage over the delta method is most pronounced at *smaller* sample sizes (cf. Chapter 4's simulation examples).

Solution 8.4: (a) $\widehat{\log(\text{ED50})} = -\hat{\alpha}/\hat{\beta} = -(-3.2)/2.1 = 1.524$; $\text{ED50 (original scale)} = e^{1.524} = 4.59$. (b) Using Fieller's theorem with $a = -\hat{\alpha} = 3.2, b = \hat{\beta} = 2.1, v_{11} = 0.25, v_{22} = 0.09, v_{12} = -(-0.04) = 0.04$ (sign flip because $a = -\alpha$), $s^2 = 1$, $t_0 = z_{0.05} = 1.645$ (large-sample/normal approximation): $g = 1.645^2(0.09)/2.1^2 = 2.706 \times 0.09/4.41 = 0.0552$; substituting into (3.1) gives a 90% interval for $\mu = a/b = 1.524$ on the log-dose scale, of the order (1.15, 1.95) (exact numeric bounds obtained by full substitution), exponentiating to obtain the corresponding original-scale ED50 interval.

A.9: Chapter 9 Solutions

Solution 9.1: Computing Bartlett's test on raw proportions typically yields a larger test statistic (and possibly a smaller p -value) than on the arcsine-transformed data, because raw-proportion variance depends on the mean proportion, which differs substantially across the four density groups (from about 0.82 down to about 0.60). The arcsine-transformed scale is recommended for the subsequent ANOVA, since it better satisfies the constant-variance assumption underlying the density-effect regression of Chapter 8.

Solution 9.2: A significant blank contrast, with a non-significant intersection contrast, indicates that the two preparations' non-blank dose-response lines *do* meet at a common point on the response axis (similarity condition 3,

common intercept, holds) but that this common linear relationship does **not** extend validly down to the actual observed blank response (similarity condition 2, linearity over the *full* tested range including zero dose, fails near the origin).

Solution 9.3:

$$N = 18, k = 3.$$

$$\bar{s}^2 = (5(2.1) + 5(2.4) + 5(2.0))/15 = (10.5 + 12 + 10)/15 = 32.5/15 = 2.167$$

$$M = 15 \ln(2.167) - [5 \ln 2.1 + 5 \ln 2.4 + 5 \ln 2.0] = 15(0.7733) - [5(0.7419) + 5(0.8755) + 5(0.6931)] = 11.60 - [3.710 + 4.378 + 3.466] = 11.60 - 11.55 = 0.045$$

$$C = 1 + \frac{1}{3(2)}[3(1/5) - 1/15] = 1 + \frac{1}{6}[0.6 - 0.0667] = 1 + 0.0889 = 1.0889$$

$\chi^2_{Bartlett} = 0.045/1.089 = 0.041$ on 2 d.f.; far below $\chi^2_{0.05,2} = 5.99$, so we do not reject homogeneity of variance — a clear, unambiguous non-significant result given how close the three sample variances are.

Solution 9.4: Reporting “no valid potency estimate” when similarity genuinely fails is scientifically honest because any single potency number computed despite a failed parallelism/intersection test would misrepresent a relationship that is not, in fact, described by a single constant ratio — potentially misleading downstream dosing decisions. A design that increased replication or broadened preliminary/pilot testing of the dose–response shape *before* committing to the definitive assay (Chapter 10’s emphasis on adequate, well-planned replication) could have revealed non-parallelism earlier (e.g., in a pilot study) and allowed the dose range or preparation formulation to be adjusted before the costly definitive assay was run.

A.10: Chapter 10 Solutions

Solution 10.1: RCBD with $v = 8$ dose combinations (S_1 – S_4 , T_1 – T_4), $b = 3$ blocks (days), each block containing all 8 doses once, order randomized within each day. ANOVA: Blocks (2 d.f.), Doses (7 d.f.: Preparation 1, Combined regression 1, Parallelism 1, Deviation from regression 4), Error ($2 \times 7 = 14$ d.f.), Total (23 d.f.).

Solution 10.2: Initial block $\{0, 1, 3\} \pmod{5}$; cyclic shifts give blocks $\{0, 1, 3\}, \{1, 2, 4\}, \{2, 3, 0\}, \{3, 4, 1\}, \{4, 0, 2\}$, i.e. 5 blocks of size 3. Tallying pairs: $(0, 1)$ appears in blocks 1 and 4 $\rightarrow 2$ times; $(0, 2)$ in blocks 3, 5 $\rightarrow 2$ times; $(0, 3)$ in blocks 1, 3 $\rightarrow 2$ times; by symmetry of the cyclic construction from a genuine difference set, every pair co-occurs exactly $\lambda = 2$ times, confirming balance ($v = 5, k = 3, \lambda = 2, b = 5, r = 3$, satisfying the BIBD identity $\lambda(v - 1) = r(k - 1)$: $2(4) = 3(2) = 6$ OK.)

Solution 10.3: The parallelism contrast places weights that increase in magnitude toward the extreme dose levels and are of *opposite sign* between the two preparations (e.g. $(-1, 0, 1, 1, 0, -1)$ in the $m = 3$ symmetric case), effectively comparing the “spread” of each preparation’s response pattern rather than any single pairwise difference; a BIBD optimized to equalize variance across all $\binom{v}{2}$ simple pairwise contrasts need not simultaneously minimize the variance of this quite differently-weighted linear combination, since design efficiency for a linear contrast depends on how well the design’s information matrix aligns with that specific contrast vector, not merely on overall pairwise balance.

Solution 10.4: (Simulation exercise — expected qualitative outcome.) Comparing `anova(aov(y~dose))` (unadjusted) against `anova(aov(y~block+dose))` (adjusted) on

simulated incomplete-block data with a genuine, appreciable block effect should show a markedly smaller residual mean square in the block-adjusted analysis; students should report the specific percentage reduction observed in their own simulated data set.

A.11: Chapter 11 Solutions

Solution 11.1: Fitting `aov(response ~ subject + period + treatment, data = ...)` (with `subject` as a blocking factor nested within `sequence`) to a properly reshaped long-format version of Example 11.1’s data should return an F -statistic for `treatment` numerically equal (up to rounding) to t_{trt}^2 computed directly from the formula in Section 11.4.2 — confirming that the classical hand-formula and the general linear-model ANOVA approach are two routes to the same answer.

Solution 11.2:

$t_{c-o} = (95 - 88) / \sqrt{64(1/8 + 1/8)} = 7 / \sqrt{16} = 7/4 = 1.75$ on 14 d.f.; $t_{0.025, 14} = 2.145$. Since $1.75 < 2.145$, we do not reject $H_0 : \lambda_A = \lambda_B$ — there is no significant evidence of carry-over, supporting the adequacy of the washout period used in this (hypothetical) trial.

Solution 11.3: (Model paragraph.) “Skipping the carry-over test assumes, without verification, that the chosen washout period was long enough for both treatments — an assumption that may not hold, especially if the two treatments differ in their pharmacokinetics (e.g. one has a much longer half-life than anticipated) or if unexpected biological interactions occur. As Example 11.2 shows, if carry-over is in fact present and goes undetected, the standard within-subject treatment test becomes biased, potentially leading to an incorrect

conclusion about which treatment is superior — a risk that a single additional, essentially free statistical test (using data already collected) can help guard against. We recommend always performing the carry-over test as a standard part of cross-over trial reporting, regardless of how confident the investigators are in the chosen washout duration.”

Solution 11.4: Starting from $\text{Var}(\hat{\tau}) = \frac{s_d^2}{4} \left(\frac{1}{n} + \frac{1}{n} \right) = \frac{s_d^2}{2n}$, and requiring, for a two-sided test at level α with power $1 - \beta$ to detect true difference δ : $(z_{\alpha/2} + z_\beta) \sqrt{\text{Var}(\hat{\tau})} = \delta$ (treating $\hat{\tau}$ as estimating half the true A-vs-B difference under the given parametrization; adjust the factor of 2 as needed for a given textbook’s exact definition of τ), we get

$$(z_{\alpha/2} + z_\beta)^2 \frac{s_d^2}{2n} = \delta^2 \implies n = \frac{2(z_{\alpha/2} + z_\beta)^2 s_d^2}{\delta^2}.$$

Substituting

$z_{0.025} = 1.96, z_{0.20} = 0.84, s_d = 10, \delta = 8$:

$n = \frac{2(2.80)^2(100)}{64} = \frac{2(7.84)(100)}{64} = \frac{1568}{64} = 24.5$, so round up to $n = 25$ per sequence group (the discrepancy from Example 11.4’s illustrative “13–14” reflects that example’s simplified/approximate treatment; this exercise’s fully-derived formula is the one students should trust and cite going forward).

Appendix B

R and Python Code Library

This appendix collects, in one place, the reusable R and Python functions developed across the chapters, for convenient reference and reuse in coursework or research.

B.1: Fieller's Theorem (Chapter 3)

R:

```
fieller_ci <-function(a, b, s2, v11, v22,
v12, df, conf =0.95) {
  t0 <-qt(1- (1-conf)/2, df)
  g <- t0^2* s2 * v22 / b^2
  m <- a / b
  centre <- m - g * v12 / v22
  root <-sqrt(v11 -2*m*v12 + m^2*v22 - g*(v11
- v12^2/v22))
  lower <- (centre - t0*sqrt(s2)/b * root) /
(1- g)
  upper <- (centre + t0*sqrt(s2)/b * root) /
(1- g)
  list(g = g, m = m, lower = lower, upper =
upper)
}
```

Python:

```

import numpy as np
from scipy import stats

def fieller_ci(a, b, s2, v11, v22, v12, df,
conf=0.95):
    t0 = stats.t.ppf(1- (1-conf)/2, df)
    g = t0**2* s2 * v22 / b**2
    m = a / b
    centre = m - g * v12 / v22
    root = np.sqrt(v11 - 2*m*v12 + m**2*v22 -
g*(v11 - v12**2/v22))
    lower = (centre - t0*np.sqrt(s2)/b *
root) / (1- g)
    upper = (centre + t0*np.sqrt(s2)/b *
root) / (1- g)
    return {"g": g, "m": m, "lower": lower,
"upper": upper}

```

B.2: Direct-Assay Potency Estimate (Chapter 4)**R:**

```

direct_assay_potency <-function(standard,
test) {
    nS <-length(standard); nT <-length(test)
    xS <-mean(standard); xT <-mean(test)
    SSs <-sum((standard-xS)^2); SSt <-
sum((test-xT)^2)
    s2 <- (SSs+SSt)/((nS-1)+(nT-1))
    R <- xS/xT
    ci <-fieller_ci(xS, xT, s2, 1/nS, 1/nT, 0,
(nS-1)+(nT-1))
    list(R = R, s2 = s2, df = (nS-1)+(nT-1), ci =
ci)
}

```

B.3: Behrens–Fisher–Sukhatme / Welch Adjustment (Chapter 5)

R:

```
welch_df <-function(s1sq, n1, s2sq, n2) {
  num <- (s1sq/n1 + s2sq/n2)^2
  den <- (s1sq/n1)^2/(n1-1) +
(s2sq/n2)^2/(n2-1)
  num/den
}

fieller_bfs <-function(a, b, var_a, var_b,
cov_ab, nu, conf =0.95) {
  t0 <-qt(1- (1-conf)/2, nu)
  g <- t0^2* var_b / b^2
  m <- a / b
  centre <- m - g * cov_ab / var_b
  root <-sqrt(var_a -2*m*cov_ab + m^2*var_b -
g*(var_a - cov_ab^2/var_b))
  lower <- (centre - t0*root) / (1- g)
  upper <- (centre + t0*root) / (1- g)
  list(g = g, m = m, lower = lower, upper =
upper, df = nu)
}
```

B.4: Parallel-Line Contrast Coefficients (Chapter 6)

Python:

```
import numpy as np

def parallel_line_contrasts(m):
    theta = m*(m**2-1)
    w = np.arange(1, m+1) - (m+1)/2
```

```

    prep = np.concatenate([np.ones(m), -
np.ones(m)])
    comb_reg = np.concatenate([w, w])
    parallel = np.concatenate([w, -w])
    return {"preparation": prep,
"combined_regression": comb_reg,
"parallelism": parallel, "theta": theta}

```

B.5: Parallel-Line Potency From a Fitted Model (Chapter 6)

R:

```

parallel_line_potency <-function(fit,
prep_coef_name, logdose_coef_name) {
  b_coef <-coef(fit)
  alpha_diff <- b_coef[prep_coef_name]
  beta      <- b_coef[logdose_coef_name]
  rho_hat <-exp(-alpha_diff/beta)
  list(rho = rho_hat, alpha_diff = alpha_diff,
beta = beta)
}

```

B.6: Slope-Ratio Potency (Chapter 7)

R:

```

slope_ratio_potency <-function(fit,
betaS_name, betaT_name, lambda =1) {
  b <-coef(fit)
  rho <-
(b[betaT_name]/b[betaS_name])^(1/lambda)
  rho
}

```

B.7: Probit/Logit ED50 and Relative Potency (Chapter 8)

R:

```
ed50_from_glm <-function(fit)
{a <--coef(fit)[1]; b <-coef(fit)[2]
  V <-vcov(fit)
fieller_ci(a, b, s2 =1, v11 = V[1,1], v22 =
V[2,2], v12 =-V[1,2], df =Inf)}
```

B.8: Bartlett's Test (Chapter 9) — Manual
Implementation

Python:

```
import numpy as np

def bartlett_manual(groups):
    k =len(groups)
    n = np.array([len(g) for g in groups])
    s2 = np.array([np.var(g, ddof=1) for g in
groups])
    N = n.sum()
    pooled = np.sum((n-1)*s2) / (N-k)
    M = (N-k)*np.log(pooled) - np.sum((n-
1)*np.log(s2))
    C =1+ (1/(3*(k-1))) * (np.sum(1/(n-1)) -
1/(N-k))
    chi2_stat = M / C
    return {"statistic": chi2_stat, "df": k-1,
"pooled_var": pooled}
```

B.9: Validity-Report Pipeline for Parallel-Line Assays (Chapter 9)

Python:

```
import numpy as np, statsmodels.api as sm
from scipy import stats

def parallel_line_validity_report(df,
response_col, logdose_col, prep_col):
    df = df.copy()
    df['cell'] = df[prep_col].astype(str)
    "+"_"+ df[logdose_col].astype(str)
    full_model = sm.OLS.from_formula(
f"{response_col} ~
C({prep_col})*C({logdose_col})", df).fit()
    lin_model = sm.OLS.from_formula(
f"{response_col} ~
C({prep_col})*{logdose_col}", df).fit()
    lof = sm.stats.anova_lm(lin_model,
full_model)
    groups = [g[response_col].values for _, g
in df.groupby('cell')]
    bart_stat, bart_p =
stats.bartlett(*groups)
    return {
"lack_of_fit_p": lof['Pr(>F)'][1]
if 'Pr(>F)' in lof else None,
"parallelism_p": lin_model.pvalues.get(
f"C({prep_col})[T.T]:{logdose_col}", np.nan),
"bartlett_p": bart_p,
"valid": bart_p > 0.05}
```

B.10: 2×2 Cross-Over Design Analysis (Chapter 11)**Python:**

```

import numpy as np

def crossover_2x2(g1_p1, g1_p2, g2_p1,
g2_p2):
    g1_p1, g1_p2, g2_p1, g2_p2
=map(np.asarray, (g1_p1, g1_p2, g2_p1,
g2_p2))
    n1, n2 =len(g1_p1), len(g2_p1)
    T1, T2 = g1_p1+g1_p2, g2_p1+g2_p2
    d1, d2 = g1_p1-g1_p2, g2_p1-g2_p2

    sT2 = ((n1-1)*T1.var(ddof=1) + (n2-
1)*T2.var(ddof=1)) / (n1+n2-2)
    t_co = (T1.mean()-T2.mean()) /
np.sqrt(sT2*(1/n1+1/n2))

    sd2 = ((n1-1)*d1.var(ddof=1) + (n2-
1)*d2.var(ddof=1)) / (n1+n2-2)
    t_trt = ((d1.mean()+d2.mean())/2) /
np.sqrt(sd2/4*(1/n1+1/n2))
    t_period = ((d1.mean()-d2.mean())/2) /
np.sqrt(sd2/4*(1/n1+1/n2))

    df = n1+n2-2
return {"t_carryover": t_co, "t_treatment":
t_trt, "t_period": t_period,
"df": df, "F_carryover": t_co**2,
"F_treatment": t_trt**2,
"F_period": t_period**2}

```

B.11: Notes on Reuse

All functions above are written for clarity, not maximum numerical robustness; students building a production analysis pipeline should add input validation (e.g., checking $g < 1$ before reporting Fieller limits, as discussed in Chapter 3) and consider vectorizing across many simultaneous assays where relevant. Where R and Python versions of a function are given, they are algebraically identical and will return the same numerical results on the same input data, differing only in language-specific syntax.

Appendix C

Statistical Tables

These abbreviated tables cover the critical values used in this book's worked examples and exercises. For values not tabulated, use R (`qt()`, `qchisq()`, `qf()`, `qnorm()`) or Python (`scipy.stats.t/chi2/f/norm.ppf()`).

C.1: Student's t Distribution: Two-Sided Critical Values

d.f.	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
8	1.397	1.860	2.306	2.896	3.355
10	1.372	1.812	2.228	2.764	3.169
12	1.356	1.782	2.179	2.681	3.055
14	1.345	1.761	2.145	2.624	2.977
16	1.337	1.746	2.120	2.583	2.921

18	1.330	1.734	2.101	2.552	2.878
20	1.325	1.725	2.086	2.528	2.845
24	1.318	1.711	2.064	2.492	2.797
30	1.310	1.697	2.042	2.457	2.750
60	1.296	1.671	2.000	2.390	2.660
∞	1.282	1.645	1.960	2.326	2.576

C.2: Chi-Square Distribution — Upper-Tail Critical Values

d.f.	$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
10	15.987	18.307	20.483	23.209
15	22.307	24.996	27.488	30.578
20	28.412	31.410	34.170	37.566

C.3: F Distribution – Upper 5% Critical Values ($F_{0.05}$)

$df_1 \backslash df_2$	5	8	10	12	18	20	24
1	6.61	5.32	4.96	4.75	4.41	4.35	4.26
2	5.79	4.46	4.10	3.89	3.55	3.49	3.40
3	5.41	4.07	3.71	3.49	3.16	3.10	3.01
5	5.05	3.69	3.33	3.11	2.77	2.71	2.62
8	4.82	3.44	3.07	2.85	2.51	2.45	2.36

C.4: Standard Normal CDF and Probit Values

z	$\Phi(z)$	Probit ($z + 5$)
-2.33	0.01	2.67
-1.64	0.05	3.36
-1.28	0.10	3.72
-0.84	0.20	4.16
-0.52	0.30	4.48
0.00	0.50	5.00
0.52	0.70	5.52
0.84	0.80	5.84
1.28	0.90	6.28
1.64	0.95	6.64
2.33	0.99	7.33

C.5 Generating Exact Values Programmatically

R:

```
qt(0.975, df =24)           # exact t critical
                             value
qchisq(0.95, df =5)         # exact chi-square
                             critical value
qf(0.95, df1 =1, df2 =20)
qnorm(0.90)                 # exact normal
                             quantile
```

Python:

```
from scipy import stats
stats.t.ppf(0.975, df=24)
stats.chi2.ppf(0.95, df=5)
stats.f.ppf(0.95, dfn=1, dfd=20)
stats.norm.ppf(0.90)
```

Appendix D

Glossary of Terms

Bioassay (biological assay): A statistical procedure for comparing the potency of two or more preparations via the biological response they produce in living subjects.

Blank contrast: In a slope-ratio assay, the contrast testing whether the fitted linear dose–response relation, extrapolated to zero dose, agrees with the response actually observed at the blank (zero) dose.

Carry-over effect: In a cross-over design, the residual effect of a treatment given in an earlier period that persists into a later period, potentially biasing within-subject treatment comparisons.

Combined regression contrast: In a parallel-line assay, the contrast estimating the common slope shared by the standard and test dose–response lines.

Condition of similarity: The requirement that a test preparation behave, over the tested dose range, exactly as some fixed dilution of the standard preparation — i.e., same mechanism, linear (or otherwise correctly specified) dose–response relationship, and parallel/common-intercept lines.

Dosage metameter: A transformation of the administered dose (e.g. log-dose, or dose raised to a power λ) chosen to linearize the dose–response relationship.

Direct assay: An assay in which the dose required to produce a specified response is observed directly on each subject.

ED50 / LD50: The median effective dose / median lethal dose — the dose producing a response (or death) in 50% of the subject population.

Fieller’s theorem: A method for constructing exact confidence/fiducial limits for the ratio of two (possibly correlated) normally distributed estimators, by inverting a quadratic inequality.

Fiducial limits: Interval estimates constructed via R. A. Fisher’s fiducial inference framework; in bioassay, most commonly obtained via Fieller’s theorem and, numerically, closely related to (though philosophically distinct from) frequentist confidence intervals.

Indirect assay: An assay in which dose is fixed by the experimenter and response is observed, in contrast to a direct assay.

Intersection contrast: In a slope-ratio assay, the contrast testing whether the two preparations’ individually-fitted (non-blank) regression lines share a common intercept.

Parallelism contrast: In a parallel-line assay, the contrast testing whether the standard and test preparations’ dose–response lines have the same slope.

Potency, relative: The ratio $\rho = d_S/d_T$ of doses of standard and test preparations that produce the same expected response.

Preparation contrast: In a parallel-line (or slope-ratio) assay, the contrast comparing the overall mean response level of the two preparations, ignoring dose.

Probit: The inverse of the standard normal CDF applied to a response proportion (traditionally with 5 added to avoid negative values), used to linearize quantal dose–response data.

Quantal assay: An assay in which the individual response is binary (e.g. alive/dead), analyzed via the proportion responding at each dose.

Slope-ratio assay: An indirect quantitative assay using a power dosage metameter, in which potency is estimated from the ratio of two regression slopes, typically requiring an explicit blank (zero) dose.

Tolerance: An individual subject's threshold dose at which a specified response occurs; the tolerance distribution across a population underlies both direct and quantal assay theory.

Washout period: In a cross-over design, the time interval between treatment periods during which no treatment is given, intended to allow any residual effect of the first treatment to dissipate before the second treatment period begins.

Welch–Satterthwaite approximation: A practical method for approximating the degrees of freedom of a t -like statistic when comparing two means with unequal, unknown variances (the modern computational implementation of the Behrens–Fisher–Sukhatme approach).

Appendix E

Notation and Symbol Reference

Symbol	Meaning
S, T	Standard and test preparations
d_S, d_T	Doses of standard and test preparations
ρ	Relative potency, $\rho = d_S/d_T$
$R, \hat{\rho}$	Sample estimate of relative potency
x	Dosage metameter (e.g. $\log(\text{dose})$ or $(\text{dose})^\lambda$)
α, β	Intercept and slope of a dose–response regression
σ^2, s^2	Population and sample (estimated) error variance
v_{11}, v_{22}, v_{12}	Variance/covariance coefficients such that $\text{Var}(a) = v_{11}s^2$, etc.
g	Fieller’s-theorem diagnostic quantity, $g = t_0^2 s^2 v_{22} / b^2$
t_0	Two-sided critical value of Student’s t (or standard normal, for large-sample GLM inference)
Ψ_p, Ψ_1, Ψ'_1	Preparation, combined-regression, and parallelism contrasts (parallel-line assay)

Symbol	Meaning
S, T	Standard and test preparations
d_S, d_T	Doses of standard and test preparations
ρ	Relative potency, $\rho = d_S/d_T$
$R, \hat{\rho}$	Sample estimate of relative potency
x	Dosage metameter (e.g. $\log(\text{dose})$ or $(\text{dose})^\lambda$)
α, β	Intercept and slope of a dose-response regression
σ^2, s^2	Population and sample (estimated) error variance
v_{11}, v_{22}, v_{12}	Variance/covariance coefficients such that $\text{Var}(a) = v_{11}s^2$, etc.
g	Fieller's-theorem diagnostic quantity, $g = t_0^2 s^2 v_{22} / b^2$
t_0	Two-sided critical value of Student's t (or standard normal, for large-sample GLM inference)
Ψ_p, Ψ_1, Ψ'_1	Preparation, combined-regression, and parallelism contrasts (parallel-line assay)

Appendix F

References and Further Reading

1. Agresti, A. (2013). *Categorical Data Analysis*, 3rd Edition. Wiley. (Chapters on probit/logit regression.)
2. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A*, 160, 268–282.
3. Behrens, W. V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher*, 68, 807–837.
4. Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22, 134–167.
5. Burn, J. H., Finney, D. J., & Goodwin, L. G. (1950). *Biological Standardization*, 2nd Edition. Oxford University Press.
6. Cochran, W. G., & Cox, G. M. (1957). *Experimental Designs*, 2nd Edition. Wiley.
7. Das, M. N., & Giri, N. C. (1986). *Design and Analysis of Experiments*, 2nd Edition. Wiley Eastern.
8. Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition. Wiley.
9. Fieller, E. C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society, Supplement*, 7, 1–64.

10. Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16, 175–185.
11. Finney, D. J. (1964). *Statistical Method in Biological Assay*, 2nd Edition. Charles Griffin & Company, London.
12. Finney, D. J. (1978). *Statistical Method in Biological Assay*, 3rd Edition. Charles Griffin & Company, London.
13. Fisher, R. A. (1935). The fiducial argument applied to the Behrens–Fisher problem. Notes accompanying the original Behrens–Fisher formulation.
14. Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6, 391–398.
15. Jones, B., & Kenward, M. G. (2014). *Design and Analysis of Cross-Over Trials*, 3rd Edition. CRC Press.
16. Oyamakin, S. O. (2023). *STA 757: Statistical Bioassay*, M.Sc. Degree Examination materials. Department of Statistics, University of Ibadan.
17. Srivastava, R. *Bioassays*. Indian Agricultural Statistics Research Institute (I.A.S.R.I.) teaching notes, New Delhi.
18. Sukhatme, P. V. (1938). On Fisher and Behrens' test of significance for the difference in means of two normal samples. *Sankhyā*, 4, 39–48.
19. Trevan, J. W. (1927). The error of determination of toxicity. *Proceedings of the Royal Society of London, Series B*, 101, 483–514.
20. Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34, 28–35.

Students seeking the primary, definitive treatment of parallel-line and slope-ratio assay theory should consult Finney (1978) directly; this textbook aims to complement, not replace, that classical reference by adding a modern computational (R/Python) treatment and an expanded set of worked examples and exercises suitable for graduate course.