

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.

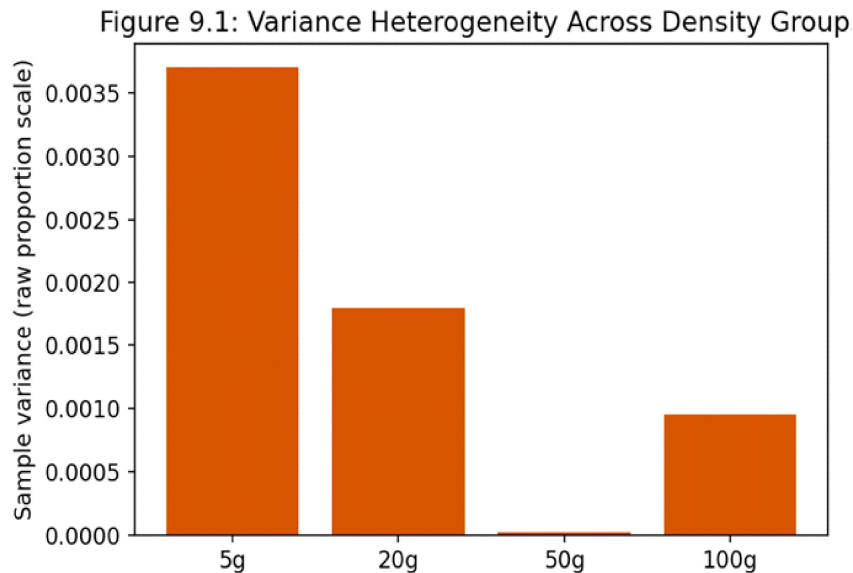


Figure 9.1: *Variance heterogeneity across density groups*

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.