

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.”

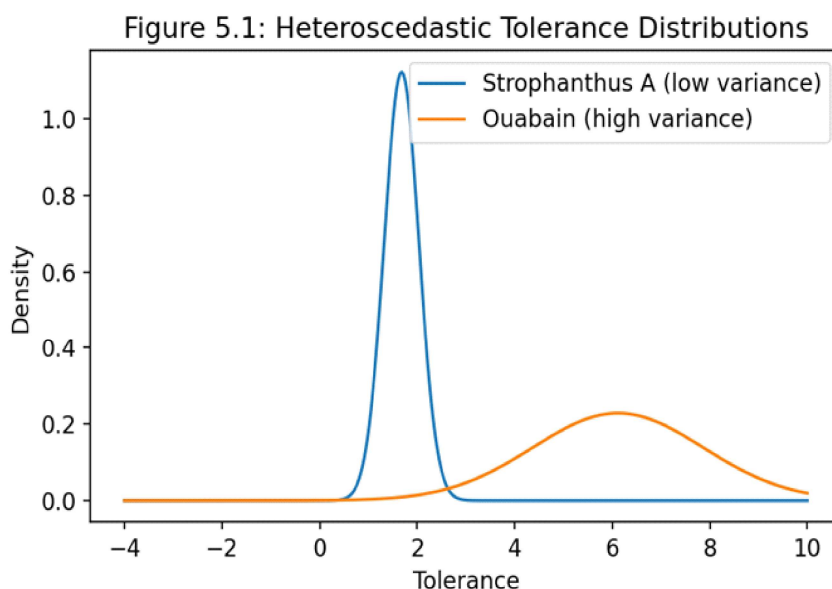


Figure 5.1: *Heteroscedastic tolerance distributions*

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.)

