

# Testing Equivalence of Variances Using Hartley's Ratio

Jesse Frey  
Villanova University

# Outline for Talk:

- Motivating Example
- The ANOVA  $F$  Test
- The Standard Approach to Assumption-Checking
- An Alternate Approach to Assumption-Checking
- A Test for Equivalence of  $k$  Variances
- Comparing the Approaches
- Final Notes

## A Motivating Example:

- An experiment was run to examine how four different solvents affect the ability of a fungicide to destroy a certain fungus.
- The data were reported by Bishop and Dudewicz (1978). Each value is the percentage of the fungus destroyed.
- Is there a difference between the solvents?

Solvent 1		Solvent 2		Solvent 3		Solvent 4	
96.44	96.87	93.63	93.99	93.58	93.02	97.18	97.42
97.24	95.41	94.61	91.69	93.86	92.90	97.65	95.90
95.29	95.61	93.00	94.17	91.43	92.68	96.35	97.13
95.28	94.63	92.62	93.41	91.57	92.87	96.06	96.33
95.58	98.20	94.67	95.28	92.65	95.31	96.71	98.11
98.29	98.30	95.13	95.68	95.33	95.17	98.38	98.35
98.65	98.43	97.52	97.52	98.59	98.00	98.05	98.25
98.41		97.37		98.79		98.12	

## The ANOVA $F$ Test (I):

- The usual approach to testing whether  $k$  means are equal is the ANOVA  $F$  test.
- Data:  $y_{ij}$  for  $i = 1, \dots, k$  and  $j = 1, \dots, n_i$ . Here  $y_{ij}$  is the  $j$ th independent observation on the  $i$ th treatment.
- Model:  $y_{ij} = \mu_i + \epsilon_{ij}$ , where the error terms  $\{\epsilon_{ij}\}$  are independent  $N(0, \sigma^2)$  random variables.
- Null hypothesis:  $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ .
- Alternative hypothesis:  $H_1$  : The means are not all the same.
- Key assumptions: (1) The errors  $\{\epsilon_{ij}\}$  are normal. (2) The error variance  $\sigma^2$  is the same for all treatments.
- Note: In the solvent example, there are  $k = 4$  treatments, and there are  $n = 15$  independent observations on each treatment.

## The ANOVA $F$ Test (II):

- The test statistic is the ratio between two estimates of  $\sigma^2$ .
- When the sample sizes are all equal to  $n$ , the in-sample estimate of  $\sigma^2$  is  $(s_1^2 + \cdots + s_k^2) / k$ , where  $s_i^2$  is the sample variance for the observations on the  $i$ th treatment.
- When the sample sizes are all equal to  $n$ , the between-sample estimate of  $\sigma^2$  is  $ns_{\bar{y}}^2$ , where  $s_{\bar{y}}^2$  is the sample variance of the sample means  $\bar{y}_1, \dots, \bar{y}_k$ .
- The in-sample estimate is unbiased for  $\sigma^2$  whether  $H_0$  holds or not, but the between-sample estimate is unbiased only if  $H_0$  holds. It is unbiased under  $H_0$  since each value  $\bar{y}_i$  is distributed  $N(\mu, \sigma^2/n)$ , where  $\mu$  is the common mean.
- Thus, the test statistic is  $F = \frac{ns_{\bar{y}}^2}{(s_1^2 + \cdots + s_k^2) / k}$ , which has an  $F$  distribution with  $k - 1$  and  $n(k - 1)$  d.f. under  $H_0$ .
- Since any departure from  $H_0$  will tend to make the numerator of  $F$  bigger, we reject only when  $F$  is too big.

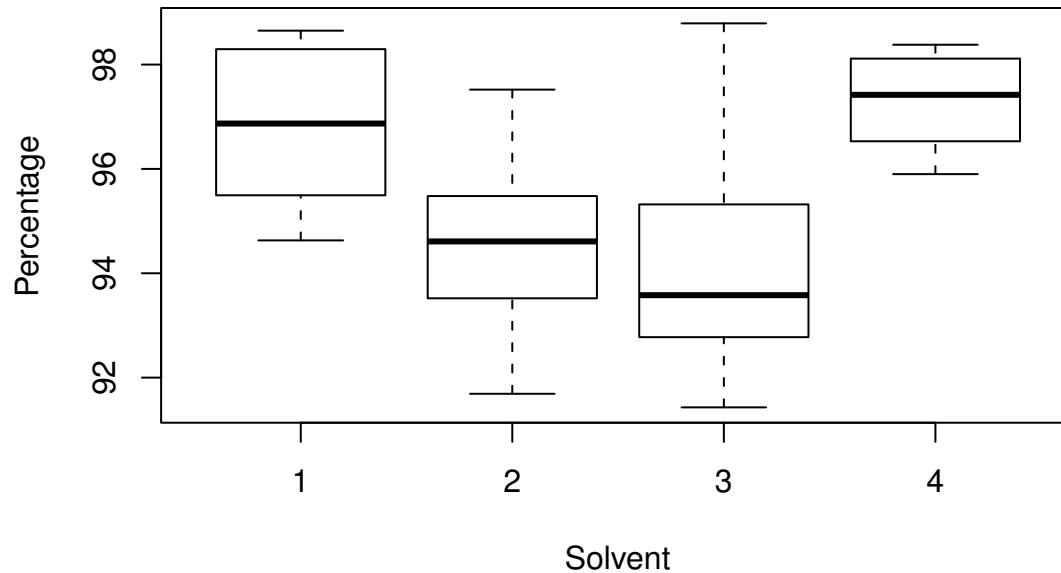
## What Happens if the Assumptions Fail?:

- Danger #1: Elevated Type I error rates. That is, the chance of falsely rejecting  $H_0$  may be much higher than we intended.
- Danger #2: Reduced power. That is, our ability to reject  $H_0$  when it isn't true may be compromised. The test may be less sensitive to differences between the  $k$  means than it should be.
- Thus, checking the assumptions is essential.
- Several alternate tests are available, such as Welch's  $F$  test. These tests tend to be slightly less powerful than the ANOVA  $F$  when all assumptions hold, but they are more robust to departures from the assumptions.

# Does Our Data Meet the Assumptions for the ANOVA $F$ Test?:

Summary Statistics:

Statistic	Solvent 1	Solvent 2	Solvent 3	Solvent 4
Sample mean	96.84	94.68	94.38	97.33
Sample variance	2.11	3.17	5.88	0.78



## The Standard Approach to Assumption-Checking:

- The standard approach to assessing the equal-variance assumption is to test  $H_0 : \sigma_1^2 = \dots = \sigma_k^2$  against the general alternative  $H_1$  : Not all variances are the same.
- If we reject  $H_0$ , then we conclude that the equal-variances assumption fails, and we would not use the ANOVA  $F$  test.
- If we fail to reject  $H_0$ , then we conclude that the assumption holds well enough and use the ANOVA  $F$  test
- Possible tests: Bartlett's test, Levene's test, Hartley's test, and many others. Each of these tests assumes normality, but Levene's test is more robust to departures from normality than Hartley's test and Bartlett's test.

## Hartley's Test:

- Hartley's test: Find the smallest and largest values among the  $k$  sample variances. Compute the ratio  $R = s_{max}^2/s_{min}^2$ . If this ratio is too large, then reject  $H_0 : \sigma_1^2 = \dots = \sigma_k^2$ .
- Our data: Here  $s_{max}^2 = 5.88$ , and  $s_{min}^2 = 0.78$ . Thus,  $F_{max} = 5.88/0.78 \approx 7.54$ . The critical value for a level-0.05 test with  $k = 4$  and  $n = 15$  is 4.22. Thus, we reject  $H_0$ , and the standard approach would lead us to use some test other than the ANOVA  $F$  test.
- Academic genealogy: H. N. Nagaraja (PhD 1980) was a student of H. A. David (PhD 1953), who was a student of H. O. Hartley (PhD 1940).

## Problem:

- The logic of the standard approach is backwards.
- When we do a hypothesis test, rejecting  $H_0$  allows us to conclude that  $H_0$  is false since we controlled the chance of falsely rejecting  $H_0$ .
- Failing to reject  $H_0$  doesn't allow us to conclude  $H_0$  is true.
- Thus, not rejecting  $H_0 : \sigma_1^2 = \dots = \sigma_k^2$  doesn't allow us to conclude that the variances are equal.
- Danger #1: When the sample size is small, a test for homogeneity of variances may not pick up even major differences among the variances.
- Danger #2: When the sample size is large, even small differences in the variances will be detected. Thus, we may end up deciding not to use the ANOVA  $F$  test when using the test would actually be fine.

# An Alternate Approach to Assumption-Checking:

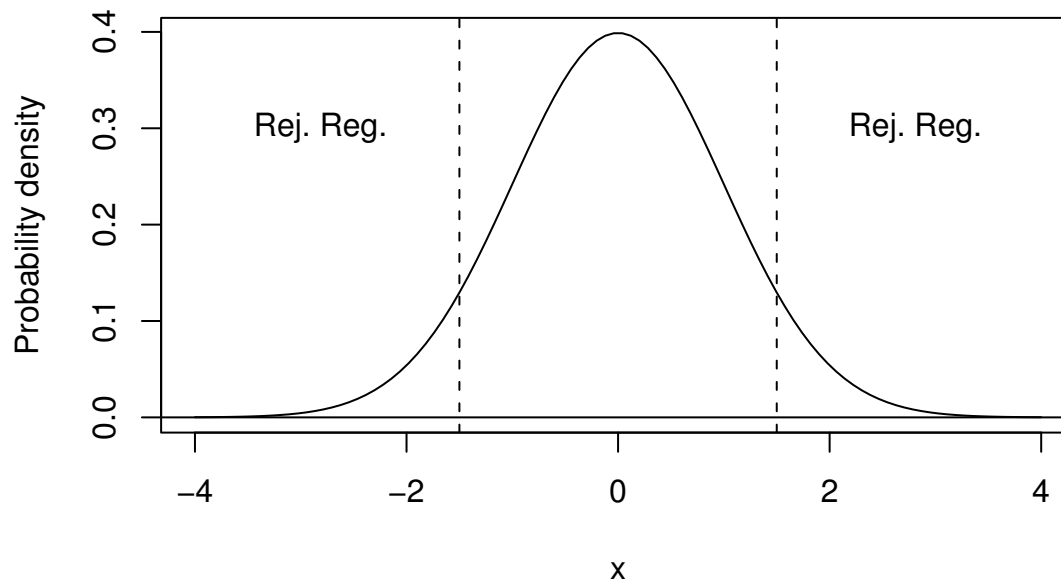
- Instead of looking for evidence that the equal-variances assumption fails, we look for evidence that it holds in an approximate sense.
- This can be done using equivalence testing.
- We reverse the roles of  $H_0$  and  $H_1$  in the test for homogeneity of variances, while introducing a zone of indifference around the original point null hypothesis.
- New hypotheses:  $H_0$  : The variances are not roughly the same and  $H_1$  : The variances are roughly the same.
- If we can reject the new  $H_0$ , then we may conclude that the variances are nearly the same. Provided that the other key assumptions hold, it is then reasonable to use the ANOVA  $F$  test.

## Equivalence Testing:

- Failing to reject  $H_0$  does not allow us to conclude  $H_0$  is true.
- However, we often wish to conclude that  $H_0$  is true or approximately true. Example: Generic drugs.
- To create a test that allows us to draw these kinds of conclusions, we reverse the roles of  $H_0$  and  $H_1$ , while also introducing a zone of indifference around the original point null hypothesis.
- Example: The equivalence testing analogues of  $H_0 : \mu = 0$  and  $H_1 : \mu \neq 0$  would be the hypotheses  $H_0 : |\mu| \geq \Delta$  and  $H_1 : |\mu| < \Delta$ , where  $\Delta$  is a value small enough that distances of  $\Delta$  or less have little practical significance.
- If we use the hypotheses in the example, then rejecting  $H_0$  gives us positive evidence that  $\mu$  is close to 0.

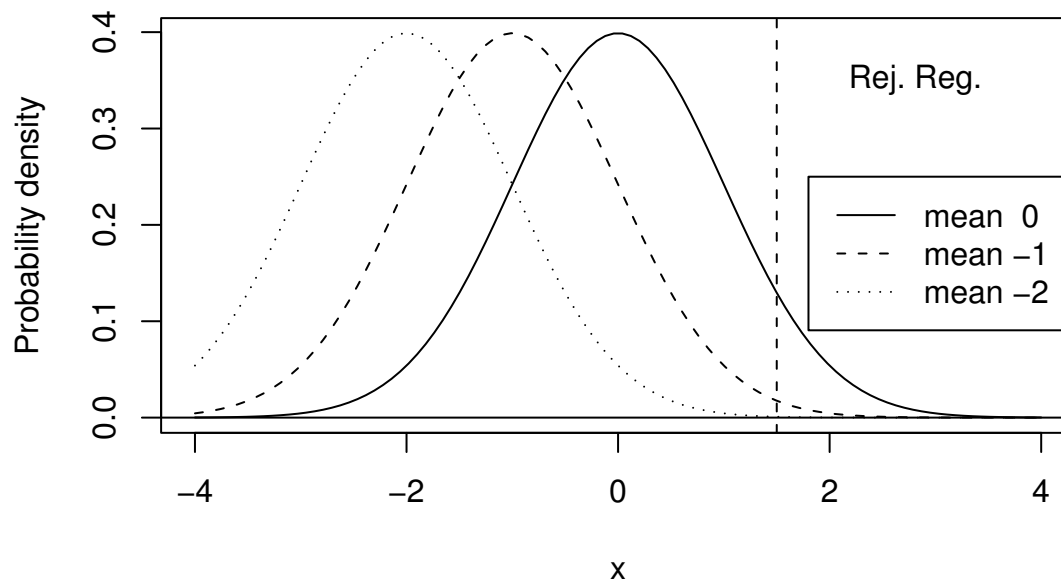
## Detailed Example (I):

- Test  $H_0 : \mu = 0$  against  $H_1 : \mu \neq 0$  using one observation  $X \sim N(\mu, 1)$  and a rejection region  $\{X : |X| > 1.5\}$ .
- The Type I error rate  $\alpha$  is  $P(|X| > 1.5 | \mu = 0)$ . This gives  $\alpha = 2(1 - \Phi(1.5)) = 2(0.0668) = 0.1336$ .



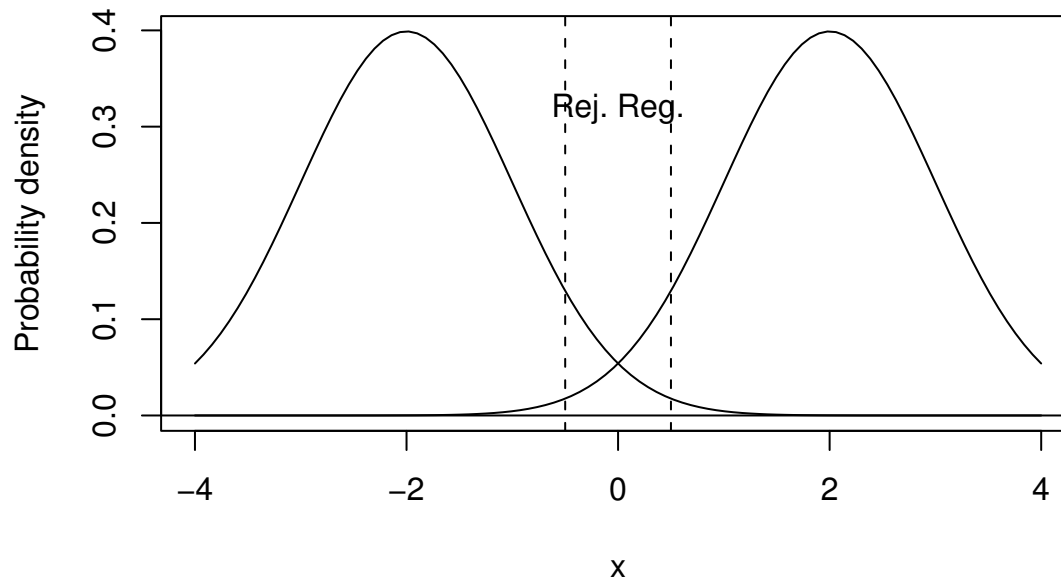
## Detailed Example (II):

- Test  $H_0 : \mu \leq 0$  against  $H_1 : \mu > 0$  using one observation  $X \sim N(\mu, 1)$  and a rejection region  $\{X : X > 1.5\}$ .
- To find  $\alpha$ , we find the supremum of  $P(X > 1.5|\mu)$  over all  $\mu \leq 0$ . The supremum is achieved when  $\mu = 0$ , and the  $\alpha$  level is  $P(X > 1.5|\mu = 0) = 0.0668$ .



## Detailed Example (III):

- Test  $H_0 : |\mu| \geq 2$  against  $H_1 : |\mu| < 2$  using one observation  $X \sim N(\mu, 1)$  and a rejection region  $\{X : |X| < 0.5\}$ .
- To find  $\alpha$ , we find the supremum of  $P(|X| < 0.5 | \mu)$  over all  $\mu$  such that  $|\mu| \geq 2$ . This supremum is achieved at  $\mu = 2$  and at  $\mu = -2$ . The  $\alpha$  level is  $P(|X| < 0.5 | \mu = 2) = 0.0606$ .



## A Notion of Equivalence for Variances:

- Equivalence of  $k$  population means is usually defined in terms of differences.
- Since variances are scale parameters rather than location parameters, it makes sense to define equivalence in terms of ratios rather than differences.
- Thus, we say that  $\sigma_1^2, \dots, \sigma_k^2$  are equivalent if  $\sigma_{max}^2 / \sigma_{min}^2 < c$ , where  $c > 1$  is a user-chosen value.
- Example: If  $c = 4$ , then the variances 1, 2, 3 are equivalent, but the variances 1, 5, 1 are not.
- Wellek (2003) defined equivalence of variance in terms of the Euclidean distance between the  $k$ -vector of log variances and the  $k$ -vector in which each entry is the average log variance. This can be used to obtain an asymptotic test for equivalence.

## The Equivalence Test:

- We test  $H_0 : \sigma_{max}^2 / \sigma_{min}^2 \geq c$  against  $H_1 : \sigma_{max}^2 / \sigma_{min}^2 < c$  by using exactly the statistic  $R = s_{max}^2 / s_{min}^2$  that Hartley uses.
- Hartley concludes that the variances are different if  $R$  is *larger* than a critical value  $r_H$  that depends on  $k$ , the sample sizes, and  $\alpha$ .
- We conclude that the variances are equivalent if  $R$  is *smaller* than a critical value  $r$  that depends on  $k$ , the sample sizes, and  $\alpha$ .
- The critical values for the two tests are different since the null hypotheses are different.
- For the equivalence test with critical value  $r$ , the  $\alpha$  level is

$$\alpha = \max_{(\sigma_1^2, \dots, \sigma_k^2) \in C} \{P(R \leq r | \sigma_1^2, \dots, \sigma_k^2)\},$$

where  $C = \{(\sigma_1^2, \dots, \sigma_k^2) : \sigma_{max}^2 / \sigma_{min}^2 \geq c\}$ .

## Key Theoretical Results:

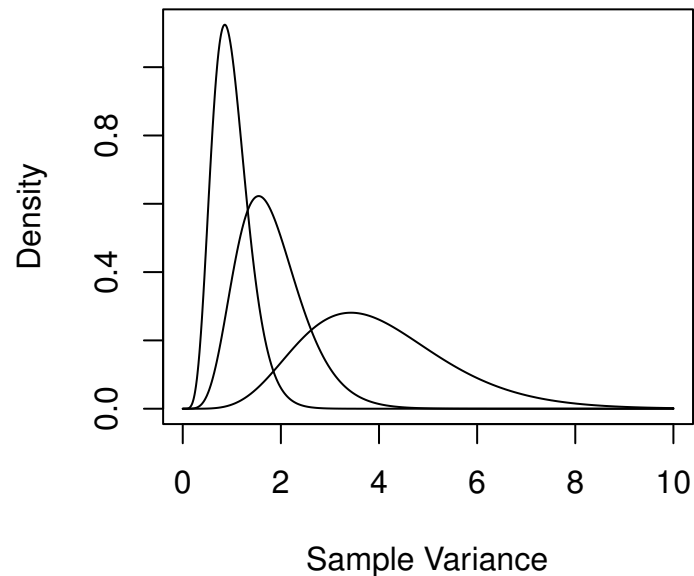
- Finding critical values for the equivalence test required some results about the function  $G(\theta_1, \dots, \theta_k) = P(R \leq r | \sigma_1^2, \dots, \sigma_k^2)$ , where  $\theta_i = \log \sigma_i^2$ .
- Result #1: I derived a one-dimensional integral representation for  $G$ . This made it possible to compute values of  $G$  using numerical integration.
- Result #2: I proved that  $G$  is a unimodal function.
- Result #3: I proved a majorization result which shows that if we take any collection of  $\theta$  values corresponding to treatments with the same sample size, then replacing each  $\theta_i$  with a convex combination  $\lambda\theta_i + (1 - \lambda)\bar{\theta}$  can only increase  $G(\theta_1, \dots, \theta_k)$ .
- Putting these results together led to an algorithm for finding critical values for the test.

## Finding $\alpha$ Given $c$ and $r$ :

- Recall that  $\alpha = \max_{(\sigma_1^2, \dots, \sigma_k^2) \in C} \{P(R \leq r | \sigma_1^2, \dots, \sigma_k^2)\}$ , and assume equal sample sizes.
- To be in the set  $C = \{(\sigma_1^2, \dots, \sigma_k^2) : \sigma_{max}^2 / \sigma_{min}^2 \geq c\}$ , we must have  $\theta_{max} - \theta_{min} \geq \log c$ .
- By Result #3, the intermediate  $\theta$  values must all be the same.
- Also by Result #3,  $\theta_{max} - \theta_{min}$  must equal  $\log c$ .
- Thus, the  $\alpha$  level for a given choice of  $c$  and  $r$  must occur when  $(\theta_1, \dots, \theta_k)$  has the form  $(0, \nu, \dots, \nu, \log c)$  for  $\nu \in (0, \log c)$ .
- Thus, the worst-case configuration for  $(\sigma_1^2, \dots, \sigma_k^2)$  has the form  $(1, v, \dots, v, c)$  for  $v \in (1, c)$ .
- When  $n$  is very large, the  $v$  that leads to  $\alpha$  is roughly  $\sqrt{c}$ .

## Worst-Case Configuration for $k = 4$ , $n = 15$ , $c = 4$ , and $\alpha = 0.05$ :

If we test for equivalence using  $c = 4$  in the context of the motivating example, the critical value is 2.033. One worst-case configuration for the variances in this case is  $(1, 1.806, 1.806, 4)$ . The plot below shows the distribution of the sample variance for  $\sigma^2 = 1$ ,  $\sigma^2 = 1.806$ , and  $\sigma^2 = 4$ .

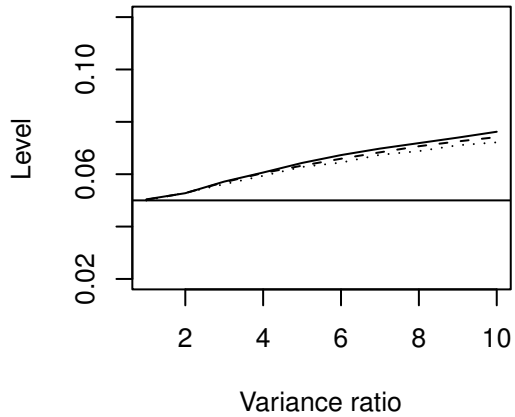


## Choosing the Zone of Indifference:

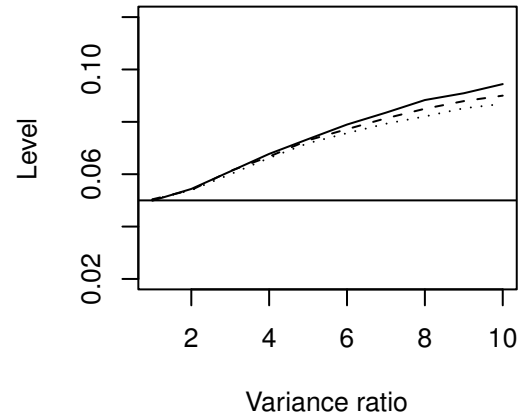
- To use the equivalence test in checking the equal-variances assumption, we must choose the value  $c$  that determines the zone of indifference.
- One way to make this choice is to examine how the ANOVA  $F$  test behaves under violations of the equal-variances assumption.
- I did this via a simulation study with different number of treatments, different common sample sizes, and different choices of the variance ratio  $\eta = \sigma_{max}^2 / \sigma_{min}^2$ .
- For each choice of  $\eta$ , I considered four different configurations of the variances.
- Configuration #1:  $(\sigma_1^2, \dots, \sigma_k^2) = (1, \sqrt{\eta}, \dots, \sqrt{\eta}, \eta)$ .
- Configuration #3:  $(\sigma_1^2, \dots, \sigma_k^2) = (1, \eta, \dots, \eta)$ .

# Simulation Study Results (I):

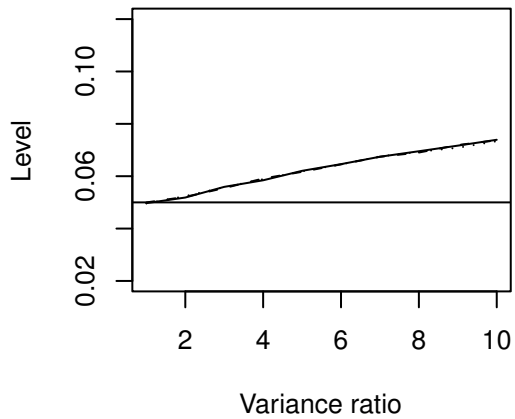
**k = 3, Configuration #1**



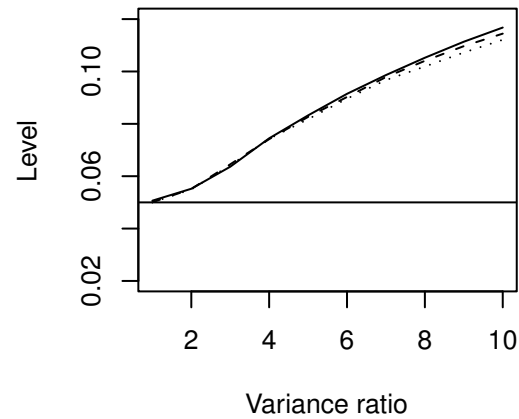
**k = 3, Configuration #3**



**k = 6, Configuration #1**

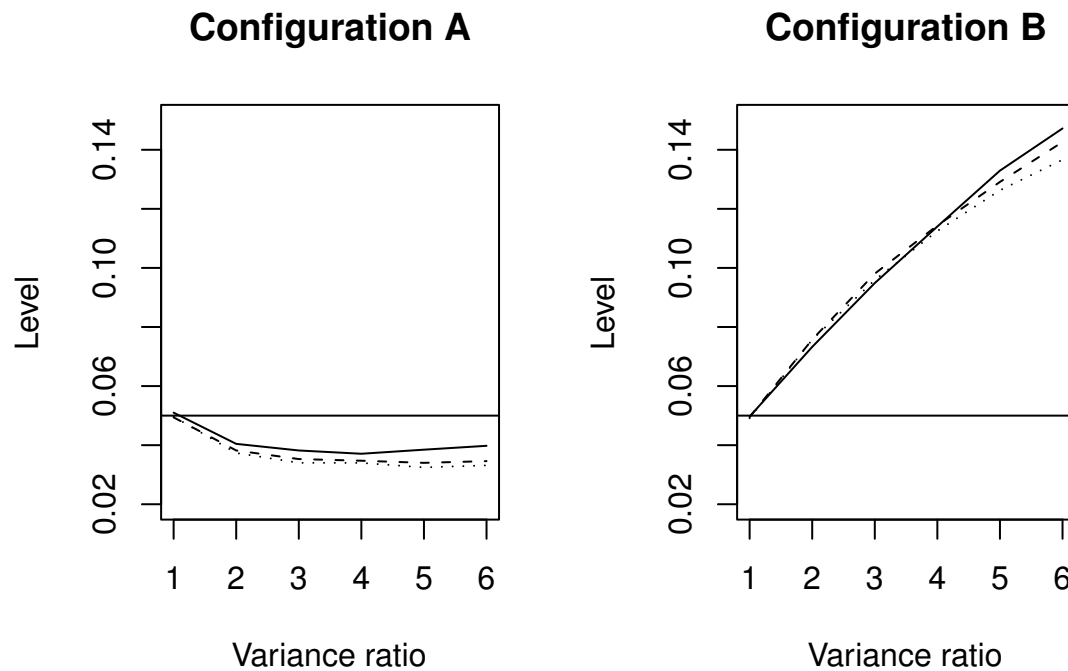


**k = 6, Configuration #3**



## Simulation Study Results (II):

Here we consider unequal sample sizes in a 2 : 3 : 4 ratio. In Configuration A, big variances go with big sample sizes. In Configuration B, big variances go with small sample sizes.



## Simulation Study Results (III):

- With equal sample sizes, choosing  $c = 4$  will keep the true  $\alpha$  level for a nominal level-0.05 test from exceeding 0.075.
- Choosing  $c = 8$  will keep the true  $\alpha$  level for a nominal level-0.05 test from exceeding 0.110.
- Small sample sizes and Configuration #3 provide the biggest challenge to the ANOVA  $F$  test.
- A smaller  $c$  may be needed if the sample sizes are unequal.
- I computed tables of critical values for the  $c = 4$  and  $c = 8$  cases.

# Critical Values for Level-0.05 Tests With $c = 4$ :

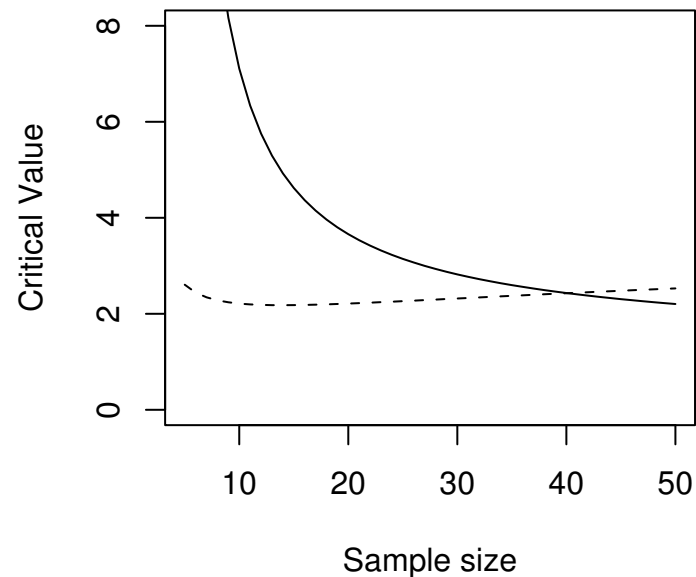
$n$	Number of treatments $k$							
	3	4	5	6	7	8	9	10
4	1.696	2.295	2.908	3.522	4.132	4.738	5.337	5.932
5	1.649	2.136	2.606	3.056	3.490	3.907	4.310	4.702
6	1.635	2.054	2.442	2.803	3.142	3.462	3.767	4.059
7	1.638	2.010	2.343	2.647	2.926	3.187	3.433	3.666
8	1.653	1.988	2.281	2.542	2.782	3.001	3.208	3.401
9	1.673	1.979	2.240	2.470	2.678	2.870	3.046	3.211
10	1.700	1.978	2.212	2.418	2.603	2.771	2.925	3.070
12	1.758	1.992	2.186	2.353	2.502	2.637	2.760	2.875
15	1.852	2.033	2.181	2.309	2.422	2.524	2.617	2.703
20	1.996	2.114	2.213	2.299	2.375	2.444	2.506	2.565
30	2.216	2.271	2.319	2.361	2.399	2.433	2.465	2.495
60	2.605	2.611	2.618	2.624	2.628	2.634	2.639	2.644
120	2.956	2.956	2.956	2.956	2.956	2.956	2.956	2.956

# Critical Values for Level-0.05 Tests With $c = 8$ :

$n$	Number of treatments $k$							
	3	4	5	6	7	8	9	10
4	2.108	2.850	3.572	4.274	4.958	5.626	6.282	6.924
5	2.179	2.780	3.329	3.838	4.318	4.773	5.280	5.626
6	2.293	2.795	3.235	3.633	4.000	4.343	4.666	4.972
7	2.427	2.848	3.209	3.533	3.827	4.098	4.351	4.590
8	2.566	2.918	3.219	3.487	3.728	3.951	4.157	4.350
9	2.700	2.997	3.250	3.473	3.675	3.861	4.032	4.192
10	2.827	3.078	3.292	3.481	3.652	3.808	3.953	4.087
12	3.058	3.239	3.394	3.532	3.656	3.770	3.876	3.975
15	3.355	3.467	3.565	3.653	3.733	3.806	3.875	3.940
20	3.750	3.802	3.849	3.892	3.931	3.968	4.003	4.036
30	4.312	4.323	4.334	4.344	4.354	4.364	4.373	4.381
60	5.195	5.195	5.195	5.196	5.196	5.196	5.196	5.196
120	5.910	5.910	5.910	5.910	5.910	5.910	5.910	5.910

## Comparison of Critical Values:

- The plot shows critical values for Hartley's test and for the equivalence test when  $k = 5$ ,  $c = 4$ , and  $\alpha = 0.05$ .
- The critical values for Hartley's test (solid line) converge to 1 as  $n$  goes to infinity, while those for the equivalence test (dashed line) converge to 4.
- The cross-over occurs as  $n$  goes from 40 to 41.

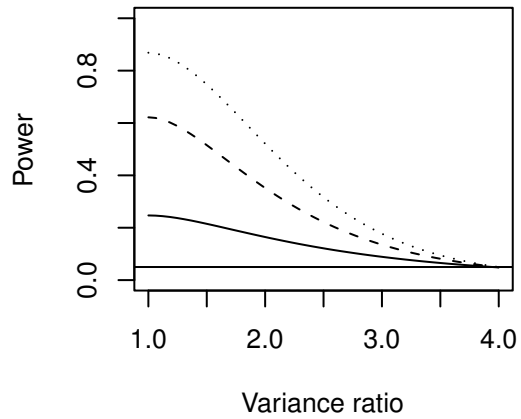


## Power of the Equivalence Test:

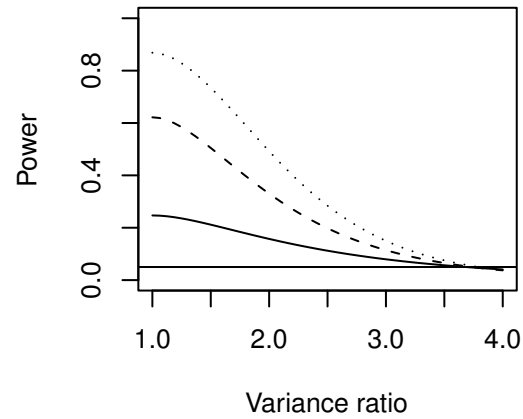
- I assessed the power of the equivalence test via another simulation study. I used the same configurations of variances used in the ANOVA  $F$  test simulation.
- I chose  $k = 4$ , sample sizes  $n = 10$ ,  $n = 20$ , and  $n = 30$ , and zones of indifference defined by  $c = 4$  and  $c = 8$ .
- The power study suggests that power is not high for small sample sizes.
- However, the tests for homogeneity of variances used in the standard approach also have low power for small sample sizes.
- When sample sizes are small, the data simply don't tell us much about whether the equal-variances assumption holds or not.

# Power of the Equivalence Test with $c = 4$ :

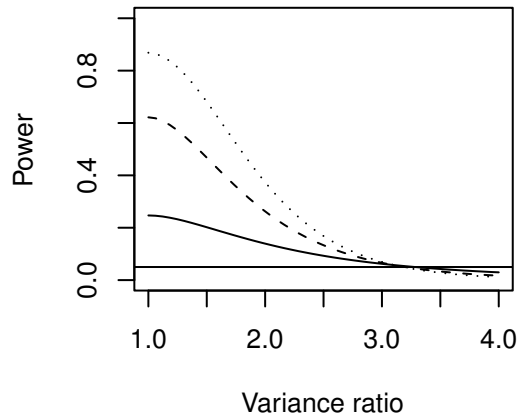
**Configuration #1**



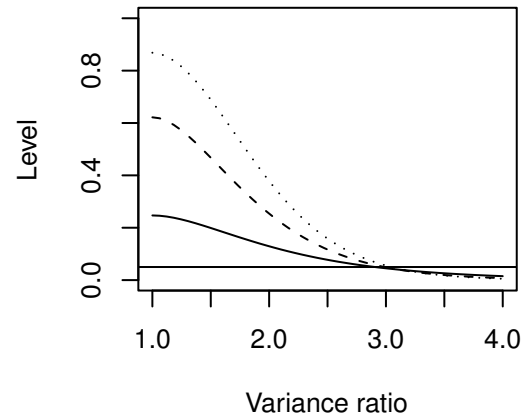
**Configuration #2**



**Configuration #3**

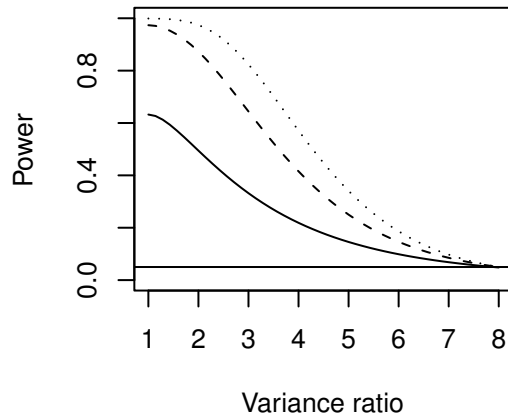


**Configuration #4**

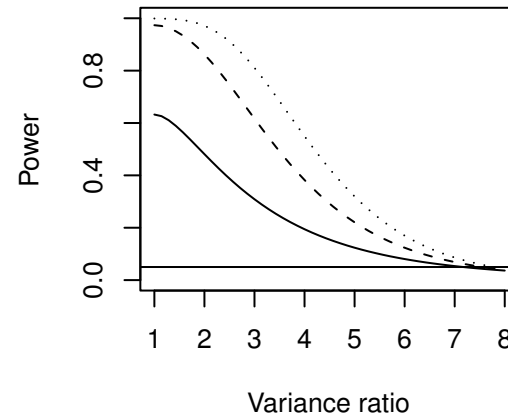


# Power of the Equivalence Test with $c = 8$ :

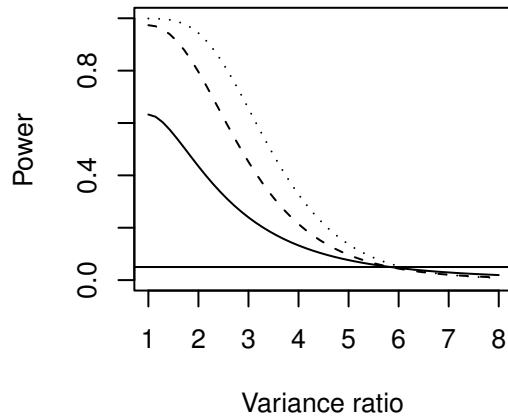
**Configuration #1**



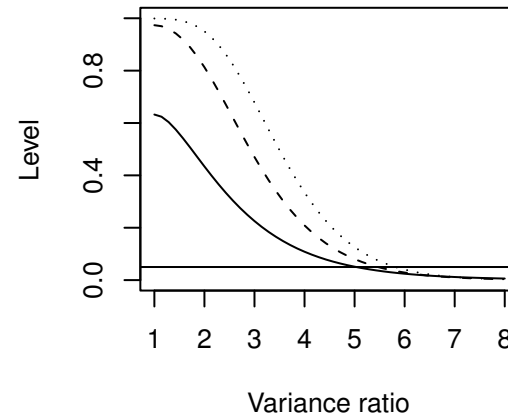
**Configuration #2**



**Configuration #3**



**Configuration #4**



## Back to the Motivating Example:

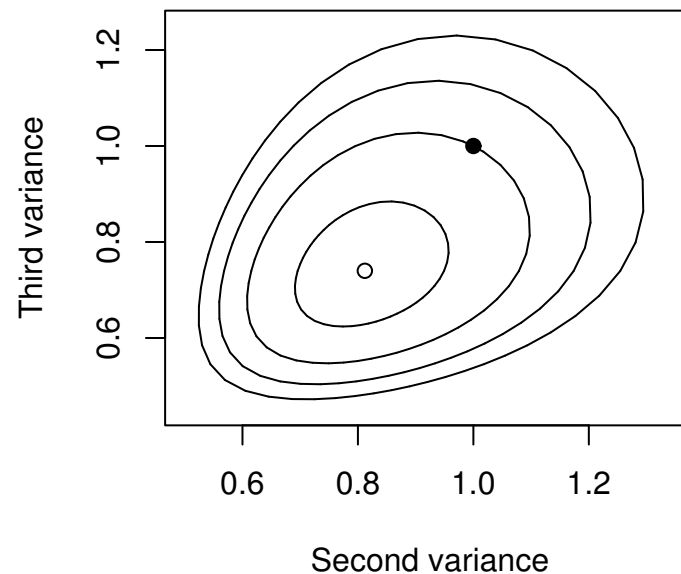
- The ratio between the maximum and minimum sample variances is  $R = s_{max}^2/s_{min}^2 = 7.55$ .
- If we test  $H_0 : \sigma_{max}^2/\sigma_{min}^2 \geq 4$  against  $H_1 : \sigma_{max}^2/\sigma_{min}^2 < 4$ , the  $p$ -value is 0.862. Thus, there is no evidence that the variances are equivalent.
- If we relax our standard for equivalence by using  $c = 8$ , the  $p$ -value is 0.444. Again, there is no evidence of equivalence.
- Thus, our conclusion would match that obtained using the standard approach: we should choose an alternate test.
- Welch's  $F$  test:  $F = 12.93$  and  $p < 0.0001$ . Thus, we conclude that there is a difference between the solvents.

## Second Example:

- Dean and Voss (1999) report data on an experiment run to compare the average lifetimes (in minutes) for four types of batteries.
- Here  $k = 4$ , and the common sample size was  $n = 4$ .
- Sample variances:  $(s_1^2, \dots, s_4^2) = (1333.7, 3152.3, 557.7, 601.6)$ .
- Test statistic:  $R = 3152.3/557.7 \approx 5.65$ .
- Using  $c = 4$  and  $c = 8$ , we obtain  $p$ -values of 0.323 and 0.200. Thus, we would want to use some test other than the ANOVA  $F$  test.
- Standard approach: Hartley's test gives  $p$ -value 0.4041, which would lead us to use the ANOVA  $F$  test.

# Hartley's Test With Unequal Sample Sizes is Biased:

- This plot shows power contours for Hartley's test when  $k = 3$ ,  $\alpha = 0.05$ ,  $\sigma_1^2 = 1$ , and  $(n_1, n_2, n_3) = (5, 10, 15)$ .
- The contours are for powers of 0.045, 0.050, 0.055, and 0.060.
- The minimum power occurs at the open dot, not the solid dot. Thus, the test is biased.



## Final Notes:

- Equivalence testing provides an alternate way to check the assumptions for the ANOVA  $F$  test.
- The new equivalence test is also useful if it is the variances themselves that are of interest.
- The test is easy to run, and the notion of equivalence used in the test is easy to understand.
- Like Hartley's test, the test is likely to be fairly sensitive to departures from normality.