# Always Use Welch's t Test to Compare the Means of Two Independent Groups

Joshua D. Wondra and Richard Gonzalez

University of Michigan

## Author Note

Joshua D. Wondra, Department of Psychology, University of Michigan.

Richard Gonzalez, Department of Psychology, University of Michigan.

Correspondence concerning this article should be addressed to Josh Wondra, Department of Psychology, University of Michigan, 530 Church St., Ann Arbor, MI 48109-1043.

Contact: jdwondra@umich.edu

## Abstract

This is an abstract

*Keywords:* t test, new statistics, Welch

# Always Use Welch's t Test to Compare the Means of Two Independent Groups

Data analysis involves a series of decisions on the part of the researcher about which statistical test answers the research question, whether the data fit the requirements of the test, and whether there are alternative options that will do a better job. Recent discussions of false positives in psychology research (e.g., Fiedler, Kutzner, & Krueger, 2012; Ioannidis, 2012; Nosek, Spies, & Motyl, 2012; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) highlight the tension between two valued outcomes of the decision process. On the one hand, researchers want to avoid mistakenly claiming that there is a true effect where none exists, which involves concerns about false positives. On the other hand, researchers want to find true effects where they do exist, which involves concerns about power. In addition to these two, there is a growing concern with estimating and reporting effect sizes (Cumming, 2014). Some have argued that due to some common research practices such as running underpowered studies, those effects that make it into published papers are spuriously large (e.g., Bakker, van Dijk, & Wicherts, 2012; Ioannidis, 2008).

One of the first decisions that many researchers learn is how to compare the means of two independent groups–they run a t test. But even this basic comparison presents a choice between the classic Student's t test (Student, 1908) or the alternative Welch-Satterthwaite test (Welch, 1938; Satterthwaite, 1946). Most researchers learn about Student's t test in the first statistics class that they ever take. When you use Student's t test to compare the means of independent groups, you make three assumptions:

1. Normality: The population for each group has a normal distribution.

2. Independence: All observations are independent of each other, meaning that the probability of one observation having a particular value does not depend on the probability of another observation having a particular value.

3. Equal variances: The population variances for the two groups are equal.

If these assumptions hold, then you can find the t-value by taking the difference in group means and dividing by the standard error of that difference:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}} \tag{1}$$

The p-value for the value of the t statistic depends on the degrees of freedom, $df = n_1 + n_2 - 2$. You are more likely to reject the null hypothesis and conclude that there is a difference in the group means as both the degrees of freedom and the t value get larger. This means that you are more likely to conclude that there is a difference when the sample size gets larger, when the difference in group means gets larger, or when the standard error gets smaller.

In order to compute the standard error for the t test, you need to find the variance of the difference in group means. When the population variances are equal, then the variance of the difference in means is equal to the variance of either group. Unfortunately, even if the population variances are equal, the group variances in the actual data are rarely identical due to sampling error, so you can't just use one of the observed group variances to find the standard error. Student's t test deals with this problem by pooling together the two group variances to estimate a single common variance. The group with the larger sample size is given more weight than the group with the smaller sample size. This means that when the larger group has the larger variance, the standard error is bigger, but when the larger group has the smaller variance, the standard error is smaller.

If either the data or the study design suggests that one or more of the assumptions has been violated, then Student's t test is not the right choice. Specifically, if the equal variances assumption has been violated, then the Welch-Satterthwaite t test (hereafter called the Welch t test for the sake of brevity) is a good alternative choice. Many researchers might not learn about the Welch-Satterthwaite test in their formal statistical training, though most have encountered it in their analyses. For those who use SPSS to analyze their data, the Welch t test is in the "Equal variances not assumed" row that appears by default whenever they run an independent samples t test. For those who use R, the Welch t test is the default when they use the `t.test()` function and they can only get Student's t test by setting the `var.equal` argument to `TRUE`.

As with Student’s t test, the Welch t test assumes normality and independence; however, it does not assume that the population variances are equal. The standard error is based on separate group variances instead of a common variance. Additionally, the Welch t test decreases the degrees of freedom to the extent that the group variances are unequal. Because of these differences, the two tests can disagree about whether there is a difference in group means. The penalty to the degrees of freedom pushes the Welch t test in the direction of being more conservative and less likely to reject the null. On the one hand, this might make the Welch t test a better choice if Student’s t test finds more false positives when variances are unequal. On the other hand, this might make the Welch t test a worse choice if it is not powerful enough to detect true effects.

However, the Welch t test is not necessarily always more conservative. The power of the two tests is not only based on the degrees of freedom, but also on the standard error. This means that the Welch t test could be more powerful than Student’s test if the separate variances standard error is smaller than the pooled variances standard error.

How do you decide which test to use? The typical approach is to use Student’s t test unless there is evidence that the two groups have unequal population variances. The challenge is how to find that evidence.

One option is to run another test of the null hypothesis that the variances are equal, such as Levene’s test for homogeneity which shows up by default in SPSS, and use the Welch t test if you reject the null. However, these tests of assumptions make their own assumptions that go unchecked and they are sensitive to sample size (Gonzalez, 2008). In addition, simulation studies find that a two-step process of running tests of equal variances to decide whether to use Student’s or Welch’s t test is not very effective (Zimmerman, 1996, 2004).

A second option is to visualize the data using boxplots and make a judgment about whether the variances appear to differ. With smaller sample sizes, you can tolerate larger apparent differences. This strategy can be enhanced by simulating data for two groups of sample sizes equal to those in your data, changing whether the variances are equal or unequal, and seeing if the boxplots of your data look like the boxplots of the simulations with equal variances.

A third option that has not been tested to our knowledge is to examine the ratio of the degrees of freedom between Student's t test and Welch's t test. If they differ to a large extent, then it might be a sign that the group variances differ.

A fourth option is to change the typical approach. Under ideal conditions, when variances and sample sizes are equal, Welch's t test is equivalent to Student's t test. If using the Welch t test generally leads to better decisions than Student's t test under both ideal and non-ideal conditions, then instead of using Student's t test by default it might be better to always use Welch's t test.

We examined these second, third, and fourth options in a Monte Carlo simulation study.

## Method

We ran Monte Carlo simulations of two independent groups with normally distributed data. We examined the type I error rate, power, and coverage probability for both Student's t test and Welch's t test under different conditions. We varied the ratio of population variances ($\sigma_1^2/\sigma_2^2 = 1/5$, 1/2, 1, 2, or 5; smallest $\sigma^2 = 2$), the sample sizes (smallest n = 20, 50, or 100), and the ratio of sample sizes ($n_1/n_2 = 1$, 2/3, or 1/2).

Additionally, we varied the size of the difference in group means based on Cohen's d values of 0, .2, .5, and .8 when variances were equal. Importantly, Cohen's d assumes that the population variances are equal and pools the group variances just like Student's t test. This means that there is no true Cohen's d when variances are unequal. Therefore, we used the same differences in group means when variances were unequal. Because we changed the variance ratio by increasing the variance of one group, the mean differences could be considered to represent smaller effects when variances are unequal.

For each condition, we set the seed to 2184 and ran 10,000 simulations. When we report conditions with equal sample sizes and variance ratios of 2 and 5, they are identical to the conditions with equal sample sizes and variance ratios of 1/2 and 1/5.

# Results

## Visualizing Data with Boxplots

One option for deciding whether the group variances are equal is to examine boxplots. Figure 1 displays boxplots from simulations of two groups with equal variances and Figure 2 displays boxplots from simulations of two groups when the variance of the group to the right is five times as large as the variance of the group to the left. The first row displays groups with $n$=20 in the second row displays groups with $n$=100. The population distributions are displayed at the top. When the sample sizes are smaller there is more variability in the boxplots. For example, in Figure 1 it appears as though the the third boxplot displays data from populations with unequal variances, whereas in Figure 2 it appears as though the third boxplot displays data from populations with equal variances. It would be difficult to decide that the boxplots provide evidence of unequal variances when $n$=20 unless the differences were quite extreme. In contrast, when sample sizes are larger there is more consistency in the boxplots. It would be easy to determine that differences in the visual variances of the boxplots point to different population variances differ when $n$=100. By examining several additional simulations it would be possible to see how much variability in the boxplots is normal when variances are equal or unequal.

## Does the df Ratio Help?

We examined whether looking at ratio of the Welch t test degrees of freedom to the Student t test degrees of freedom would provide a heuristic for deciding that the equal variances assumption does not hold. Rather than simulate the df ratio, we examined the analytical df ratio as a function of the variance ratio, as a function of the sample size ratio, and as a function of both.

Figure 3 displays the change in the df ratio as the variance ratio increases when sample sizes are equal. As expected, the ratio decreases as the difference in variances grows larger. When the variances are equal, the df ratio is equal to 1, though in real data the observed variances will rarely be exactly equal even if the population variances are equal. When once variance is twice the size of the other, the ratio drops to .9. A useful heuristic might to assume that the variances are unequal when the ratio falls below 96%.

But now look at what happens when the variances are equal and the sample size ratio changes (Figure 4). Here, too, the df ratio decreases as the difference in sample sizes grows larger, even though the variances stay the same. The 96% heuristic would lead us astray and we would incorrectly conclude that the variances are unequal in many cases when only the sample sizes are unqual.

The picture becomes even more complicated when both the sample sizes and variances are unequal (Figure 5). In this case, the effect of different variances depends on whether the larger group has the larger variance or the smaller variance. When the variance of the larger group is increasing, the move from equal to unequal variances actual counteracts the effect of the unequal sample sizes at first, and the df ratio initially begins to approach 1 before dropping again. Due to the difference in sample sizes, a 96% heuristic would mislead us into concluding that variances are equal when they are actually three times different from each other. However, when the variance of the smaller group is increasing the immediate drop in the df ratio is quite dramatic before it begins to level off.

In short, the usefulness of a heuristic based on the df ratio is limited to cases when the sample sizes are equal.

### When Does Each Test Perform Best?

There did not seem to be a simple rule based on the degrees of freedom penalty to detect whether variances are unequal, so we decided to examine when the Welch and Student t tests would perform best based on the sample size, variance ratio, and sample size ratio. We examined how well each test balances the concerns about false positives, power, and estimation. Some prior research has examined the Type I error rate for the two tests (Boneau, 1960; Zimmerman & Zumbo, 1993; Zimmerman, 2004, 1996; Zimmerman & Zumbo, 2009) and some has also examined the power of the two tests, though not always representing the complete configuration of conditions that we examined in our simulations (NeuhÃďuser, 2002; Zimmerman & Zumbo, 1993).

Nevertheless, we believe that it will be informative to display the false positives and power of the two tests here. We also discuss implications for effect size estimation, which follows from the false positive results but has not, to our knowledge, been discussed explicitly in past research.

NOTE: Zimmerman1993 only reports power when Student's t test looks bad; Neuhauser only reports power when sample sizes are equal

NOTE: I excluded here citations that looked at ANOVA vs. Welch for now (Overall et al., 1995a, 1995b), but I could add them if we want.

**Type I Error Rates.** The expected type I error rate is $\alpha = .05$. The observed type I error rate for Student's t test remained close to the expected .05 rate when either the sample size or the population variances were equal, but it varied widely when both population variances and sample sizes were unequal (see Figure 7). When the group with the larger sample size had the larger variance (the left side of Figure 7), the type I error rate dropped as low as about .01, but when the group with the larger sample size had the smaller variance (the right side of Figure 7), the type I error rate rose as high as .12, which is more than double the normally accepted false positive rate. In contrast, the observed type I error rate for Welch's t test remained close to the expected .05 rate across all conditions (see Figure 8). Overall, the Welch t test consistently behaved as expected when it comes to false positives. Student's t test did not.

**Power.** Is Welch's t test underpowered compared to Student's t test? Figure 9 displays the power of Student's t test and of Welch t tests to detect small, medium, and large effects under the different conditions. For both tests, power decreases as variances become unequal because we increased one group's variance; however, the power of each test decreases at a different rate depending on the sample size ratio. Figure 10 displays the difference in power between Student's t test and Welch's t test, with higher numbers indicating that Student's t test is more powerful. When the sample sizes or variances are equal, the power of the two tests is approximately equal. However, when both the sample sizes and variances are unequal, there are differences in power. Overall, Student's t test is more powerful when the large sample has the smaller variance, whereas the Welch t test is more powerful when the small sample has the smaller variance. These differences are the most dramatic when one sample is twice the size of the other.

The conditions in which Student's t test has the greatest power over Welch's t test, when one sample is twice the size of the other and the large sample has the small variance, are the same conditions in which Student's t test has a risk of doubling the false positive rate. In contrast, Welch's t test was more powerful than Student's t test under other conditions and never inflated the type I error rate beyond the expected rate.

Taken together, the type I error rates and power favor the Welch t test over Student's t test as better balancing researchers' concerns.

**Coverage Probability.** We examined the coverage probability of 95% confidence intervals, which is the proportion of confidence intervals that contain the true population value of the difference in group means, based on Student's t test and Welch's t test. Because the accuracy of a confidence interval is influenced by the variance and sample size, but not by the true effect size, we only show the coverage probability when the null hypothesis is true (the coverage probabilities are identical across all effect sizes). Additionally, when the null hypothesis is true and $\alpha = .05$, the coverage probability of 95% confidence intervals has a simple relationship with the type I error rate–the coverage probability is the complement of the type I error rate. Table 10 displays the coverage probabilities of the two t tests. The coverage probability for Student's t test varies dramatically, just as the type I error rates did. When Student's test is the least powerful, it is the most accurate at estimating the difference in means. When it is the most powerful to find an effect, it is the least accurate, and what you would believe is a 95% confidence interval drops as low as an 88% confidence interval in reality. In contrast, the Welch t test performs as expected and the confidence interval contains the true effect 95% of the time regardless of the variance and sample size ratios.

## Discussion

We set out to find a simple rule to help researchers decide when to use Student's t test and when to use Welch's t test. We believe that the simplest rule is to always use

Welch's t test to compare the means of independent groups.

The results of our simulations demonstrated that when the population variances or sample sizes were equal, using Welch's t test instead of Student's t test didn't hurt. Figures 3 and 4 show that the Welch degrees of freedom could drop below 70% the Student's t test degrees of freedom when only the variance ratio or the sample size ratio changed. Nevertheless, the false positive rates, power, and coverage probabilities of the two tests were almost identical under these conditions. It seems as though the difference in the degrees of freedom of the two tests was not very important at all. We illustrate this point in Figure 11, which shows two overlapping t distributions, one with half the degrees of freedom of the other. The tails of the distributions with the smaller degrees of freedom are a little larger, but not much. Given the same t-value, in the majority of cases these distributions would agree about whether or not to reject the null hypothesis.

More important than the degrees of freedom is the standard error, which affects the t-value itself. When either the variances or the sample sizes are equal, the pooled standard error of Student's t test and the separate variances standard error of Welch's t test are identical, and the two tests will generally agree with each other. However, when both the variances and the sample sizes are unequal, the pooled standard error of Student's t test gives more weight to the group with the larger sample size (CITATIONS); if that group has the larger variance, then Student's t test becomes more conservative, but if that group has the smaller variance, then Student's t test becomes more liberal. This can be seen in Figures 7 and 10, where the type I error rate and coverage probability of Student's t test vary widely. In contrast, Welch's t test was more stable, regardless of which sample had the larger variance. This can be seen in Figures 8 and 10, where the type I error rate and coverage probability of Welch's t test were exactly what you would expect.

The biggest benefit of Student's t test was that it had more power than Welch's t test to detect true effects when the larger sample had the smaller variance–yet under these same conditions, it had an inflated type I error rate and the lowest coverage probability. Far from being underpowered, Welch's t test was more powerful than

Student's t test when the larger sample had the larger variance, but under these same conditions it retained the expected type I error rate and coverage probability. Overall, Welch's t test did a better job of balancing researcher concerns about false positives, power, and estimation.

We believe that researchers are more likely to use simple decision rules, and so we echo the recommendations of previous researchers to always use the Welch t test (CITATIONS). For researchers who have a reason to prefer Student's t test, there is some good news for those who are doing experimental work. With experiments, subjects are usually assigned evenly to conditions. In this case, the two tests will perform equally well. However, for researchers who are interested in comparing pre-existing groups, it might not be possible to have equal sample sizes. In this latter case, the Welch t test is likely to outperform Student's t test because the observed variances are unlikely to be precisely equal. If a researcher still prefers a complex rule where using either test is still possible, then one option is to look at boxplots to determine whether you can reasonably assume that the group variances are equal. This judgment will be easier if the sample sizes are large. Nevertheless, researchers should rest assured that they won't suffer from using the simple rule to always go with Welch's t test.

We examined whether the ratio of the degrees of freedom from Welch's t test to the degrees of freedom from Student's t test could provide a good heuristic for determining whether or not the variances of the two groups are unequal. The rule had the greatest potential when it mattered the least–when the sample sizes of the two groups were equal. When the sample sizes were unequal the degrees of freedom ratio was not a reliable indicator of unequal variances. It could become very high as variances became unequal and it could drop very low when the variances were equal.

Make a note that our df ratio rule works best when it doesn't matter - when sample sizes are equal.

In much experimental work, the choice between the tests is probably fine if the experimenter ensures that sample sizes are equal. This is generally not an option with pre-existing groups.

Notably, the two standard errors are equal when either the sample sizes or the variances of the two groups are identical, so the Welch t test could only be more powerful when both the sample sizes and variances are unequal.

Benefits of just using welch rule: simplifies the decision which makes it easier, puts researcher motivation in line with false positive preservation

One of the concerns about using the Welch t test as an alternative to Student's t test is that the penalty on the degrees of freedom makes it difficult to find effects. Figures 3 and 4 show that the penalty is small when the variances and sample sizes are approximately equal, which is within the range that one might expect from normal sampling error when the true population variances are equal. However, the penalty to the degrees of freedom might not be very important in general. Figure 11 displays two t-distributions where one has half the degrees of freeom of the other. Despite the drastic difference in the degrees of freedom, the two distributions largely overlap and so the differences in standard errors of the two tests, which will affect the value of the t-test, might be more important.

Cite Zimmerman 1996 when talking about how the pooled variance is the major cause of problems with Student's t test

Cohen's d is a population standardized mean difference based on the standard deviation of the population - but what if there is no standard deviation of the population? This is exactly the situation we're faced with when lifting the equal variances assumption.

# References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi: http://dx.doi.org/10.1177/1745691612459060

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*, 49–64. doi: http://dx.doi.org/10.1037/h0041412

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi: http://dx.doi.org/10.1177/0956797613504966

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669. doi: http://dx.doi.org/10.1177/1745691612462587

Gonzalez, R. (2008). Data analysis for experimental design.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. doi: http://dx.doi.org/10.1097/EDE.ObO

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654. doi: http://dx.doi.org/10.1177/1745691612464056

NeuhÃďuser, M. (2002). Two-sample tests when variances are unequal. *Animal Behaviour*, *63*, 823–825. doi: http://dx.doi.org/10.1006/anbe.2002.1993

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. doi: http://dx.doi.org/10.1177/1745691612459058

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114. doi: http://dx.doi.org/10.2307/3002019

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi: http://dx.doi.org/10.1177/0956797611417632

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25. doi: http://dx.doi.org/10.1093/biomet/6.1.1

Wagenmakers, E.-J., Wetzels, R. W., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi: http://dx.doi.org/10.1177/1745691612463078

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362. doi: http://dx.doi.org/10.1093/biomet/29.3-4.350

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *The Journal of General Psychology*, *123*, 217–231. doi: http://dx.doi.org/10.1080/00221309.1996.9921274

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173–181. doi: http://dx.doi.org/10.1348/000711004849222

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the student t test and welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*, *47*, 523-539. doi: http://dx.doi.org/10.1037/h0078850

Zimmerman, D. W., & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t tests. *PsicolÃşgica*, *30*, 371-390.

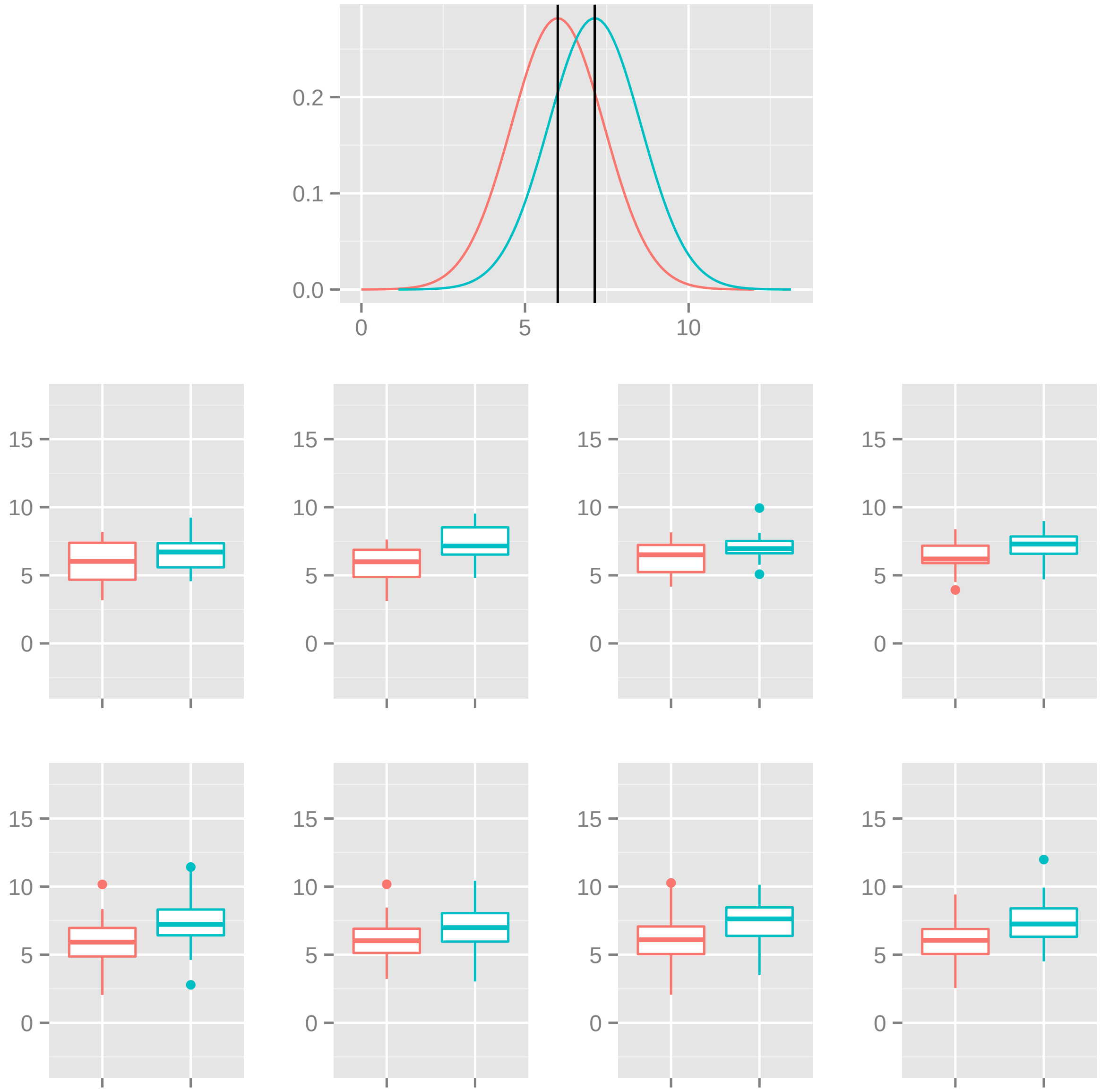

*Figure 1.* Boxplots for groups with equal variances. The first row displays groups with $n$=20 and the second row displays groups with $n$=100.

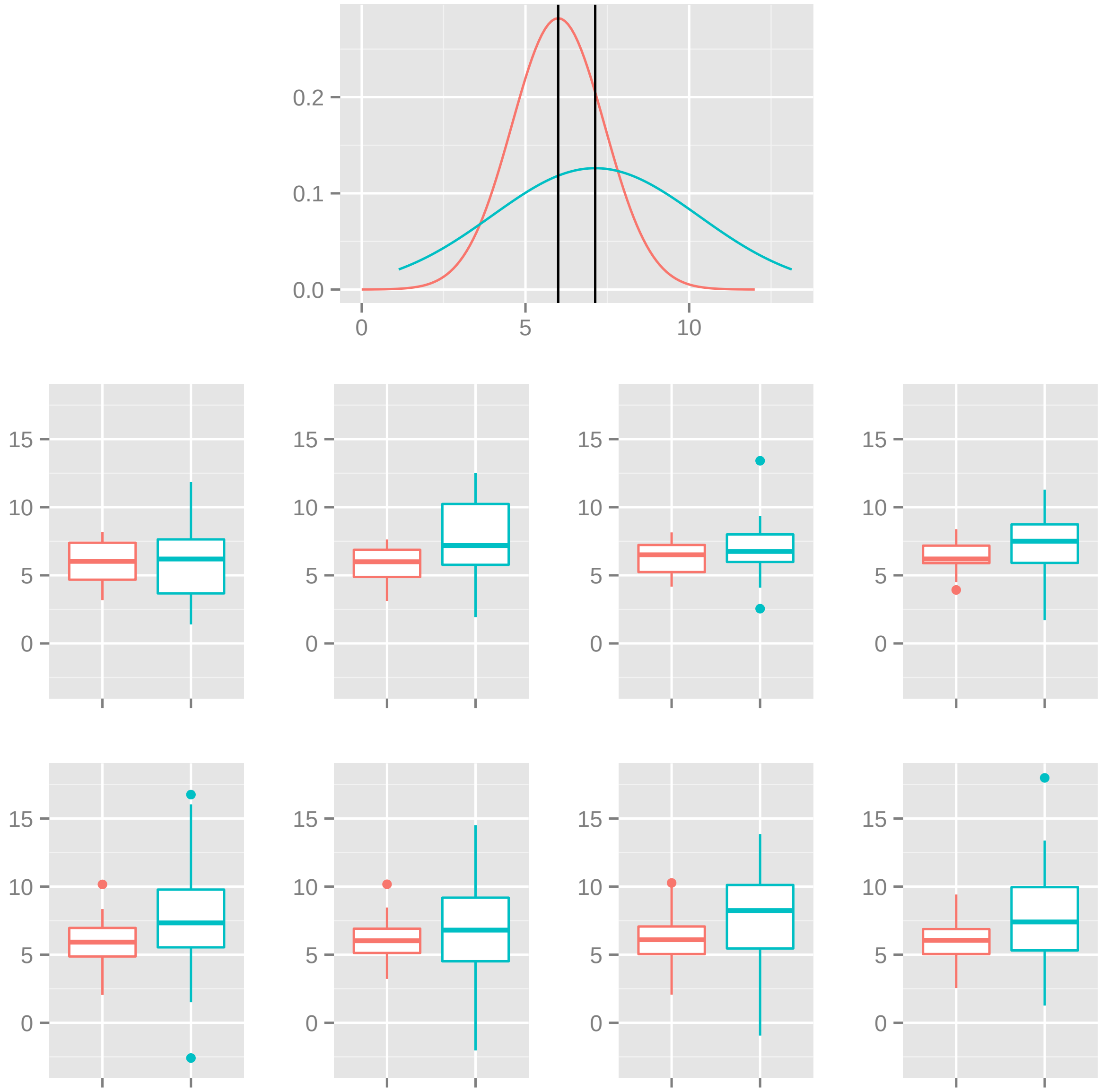

*Figure 2.* Boxplots for groups with unequal variances. The first row displays groups with $n$=20 and the second row displays groups with $n$=100.

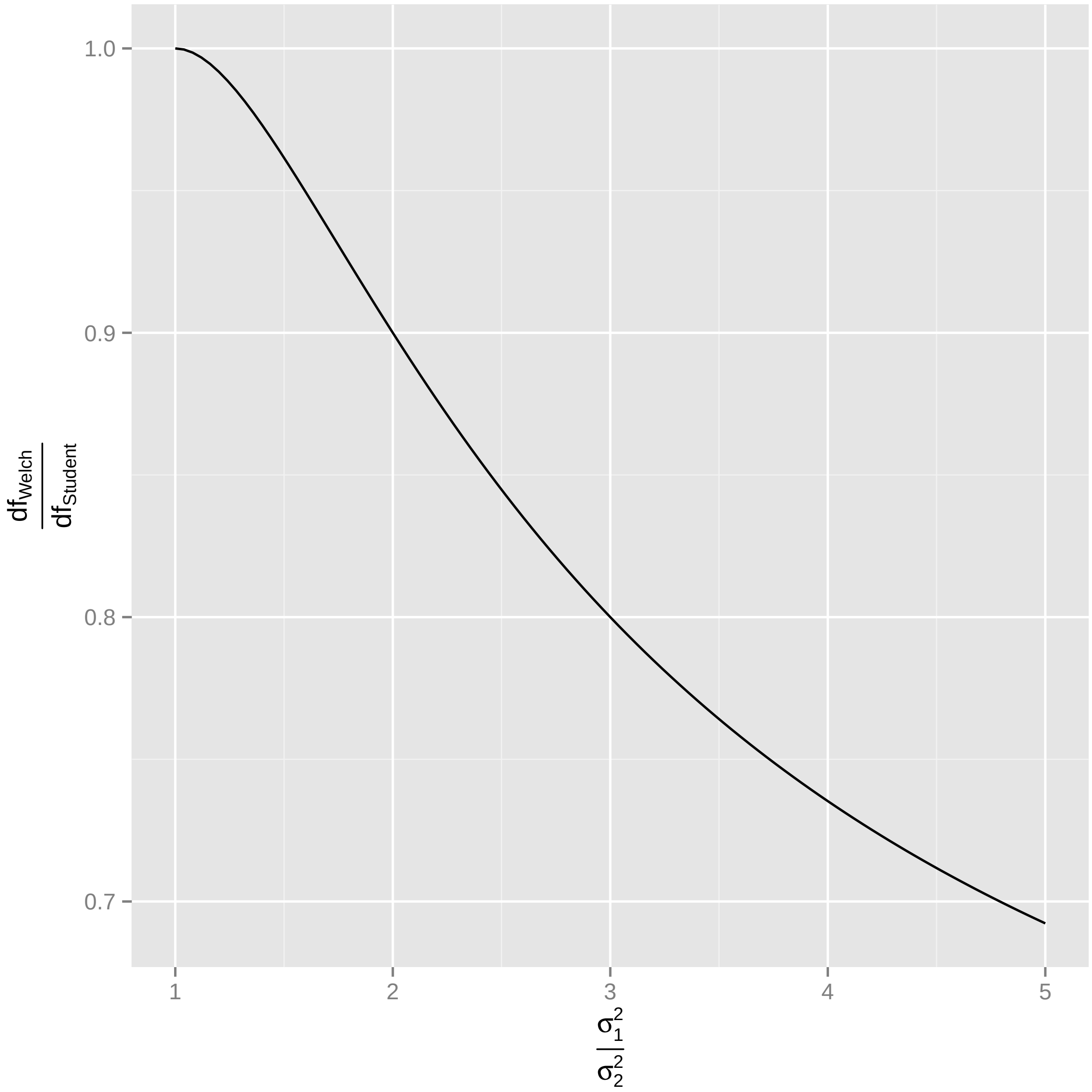

*Figure 3.* Degrees of freedom ratio when sample sizes are equal and variances are unequal.

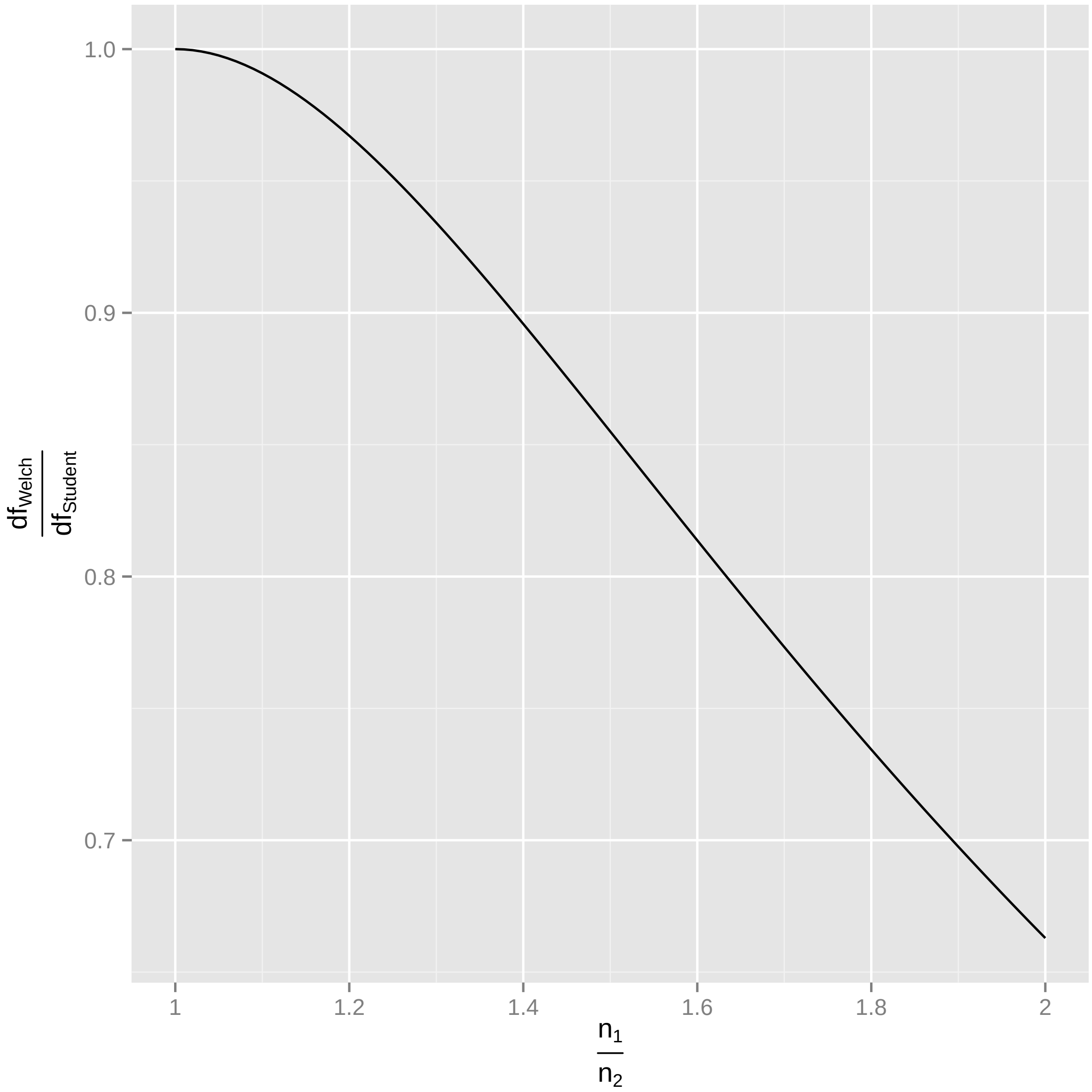

*Figure 4.* Degrees of freedom ratio when sample sizes are unequal and variances are equal.

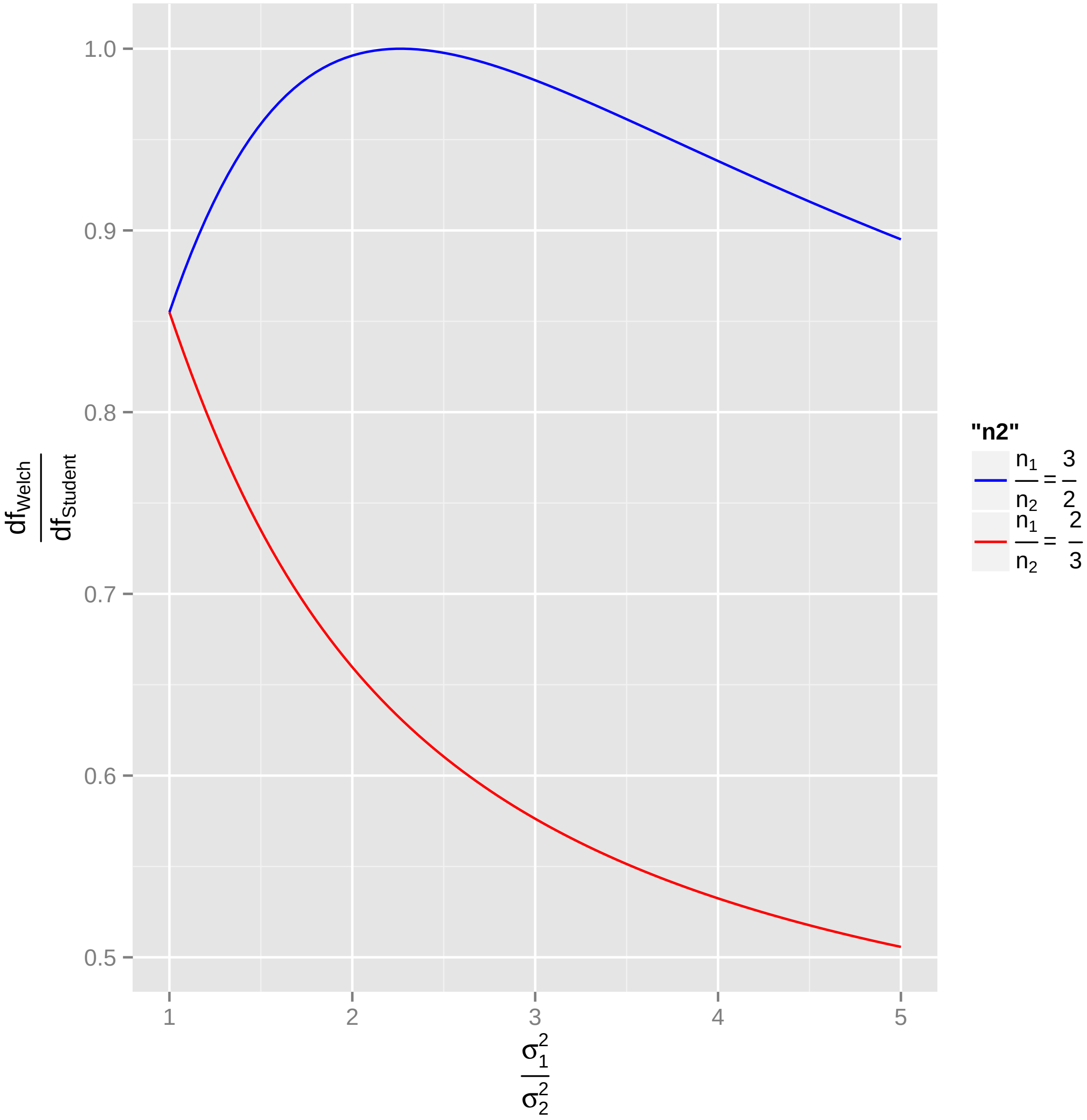

*Figure 5.* Degrees of freedom ratio when sample sizes are unequal and variances are unequal.

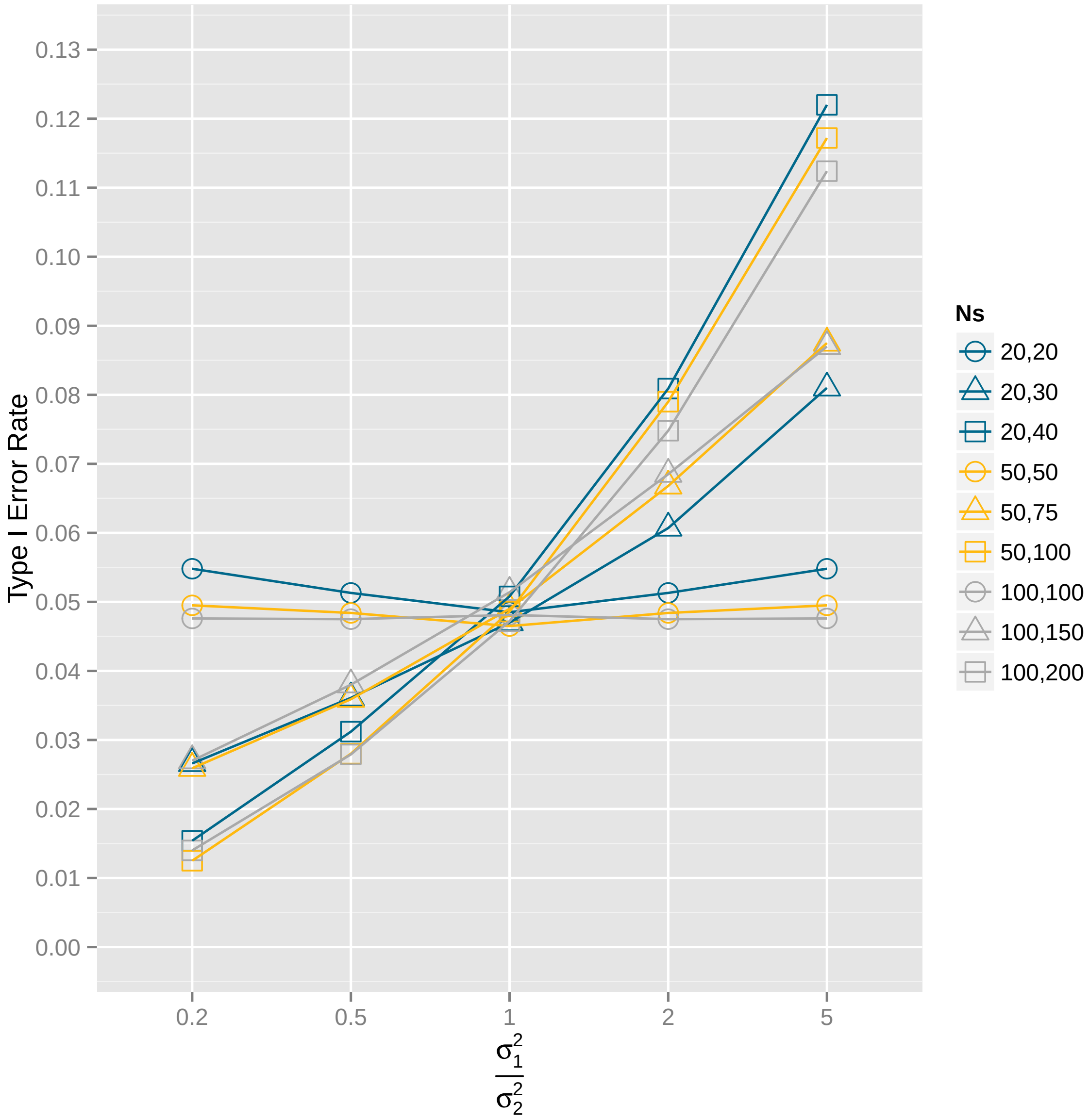

*Figure 7.* Type I error rates for Student's t test.

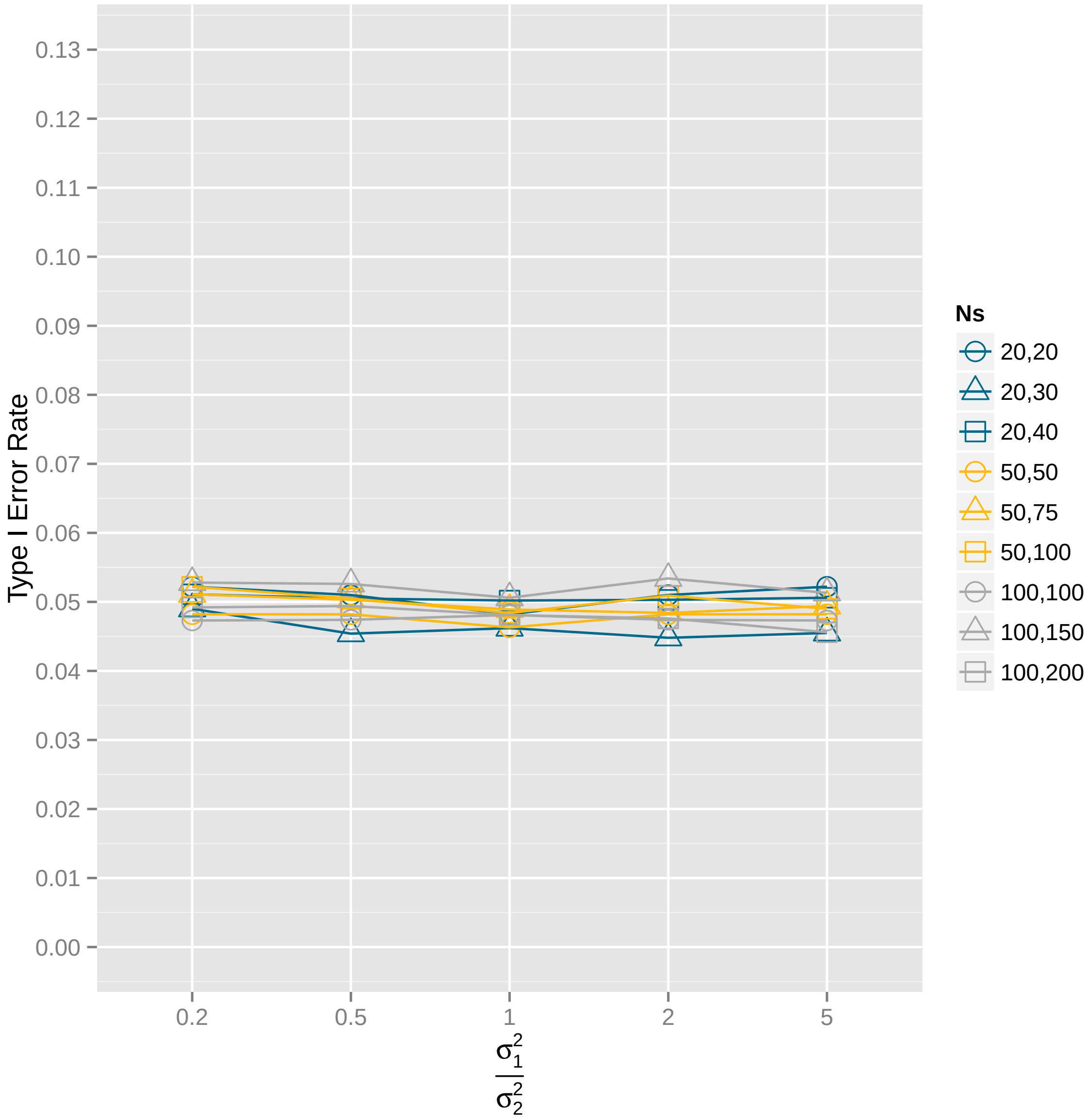

*Figure 8.* Type I error rates for Welch's t test.

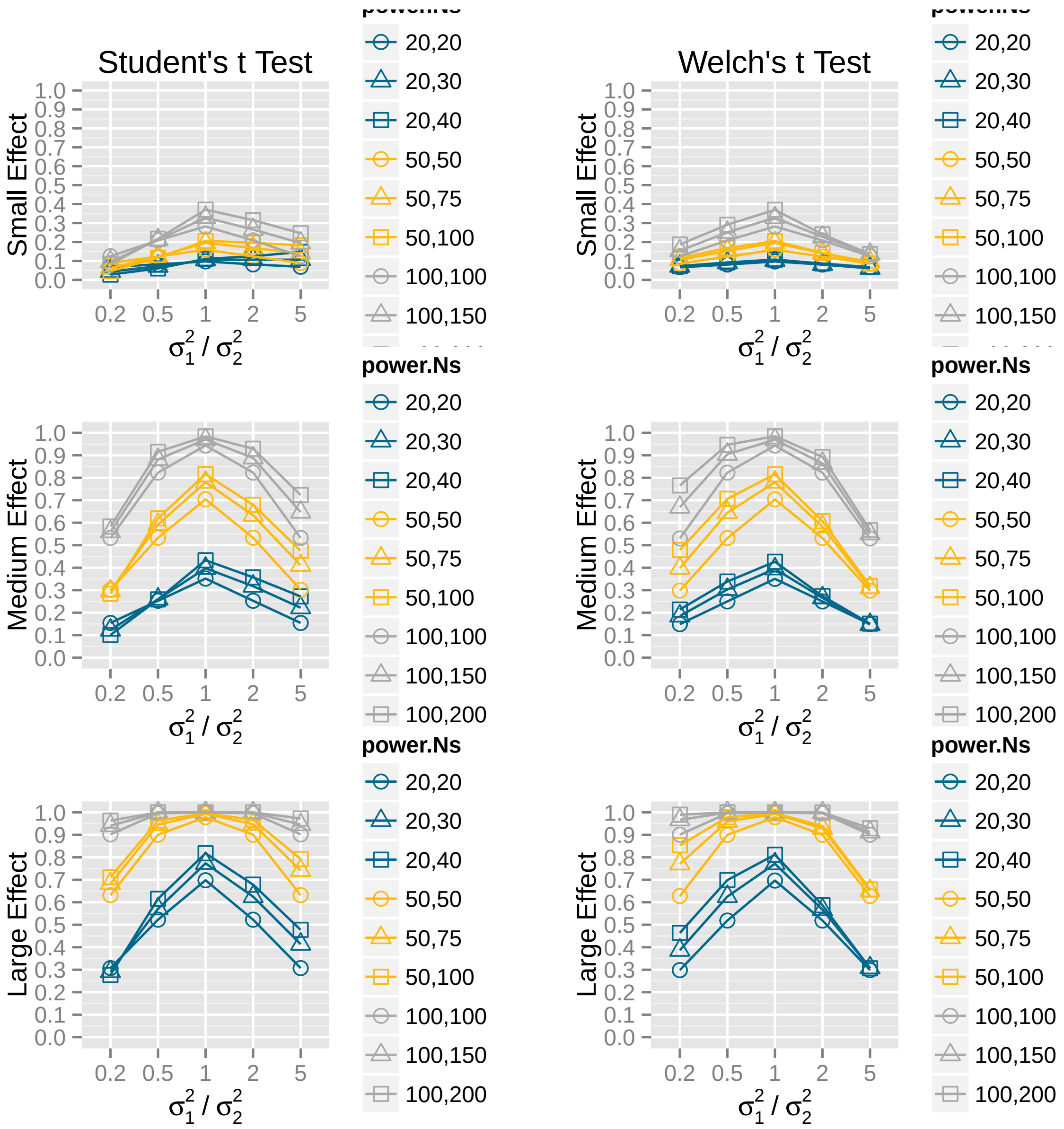

*Figure 9.* Power of Student's and Welch's t tests.

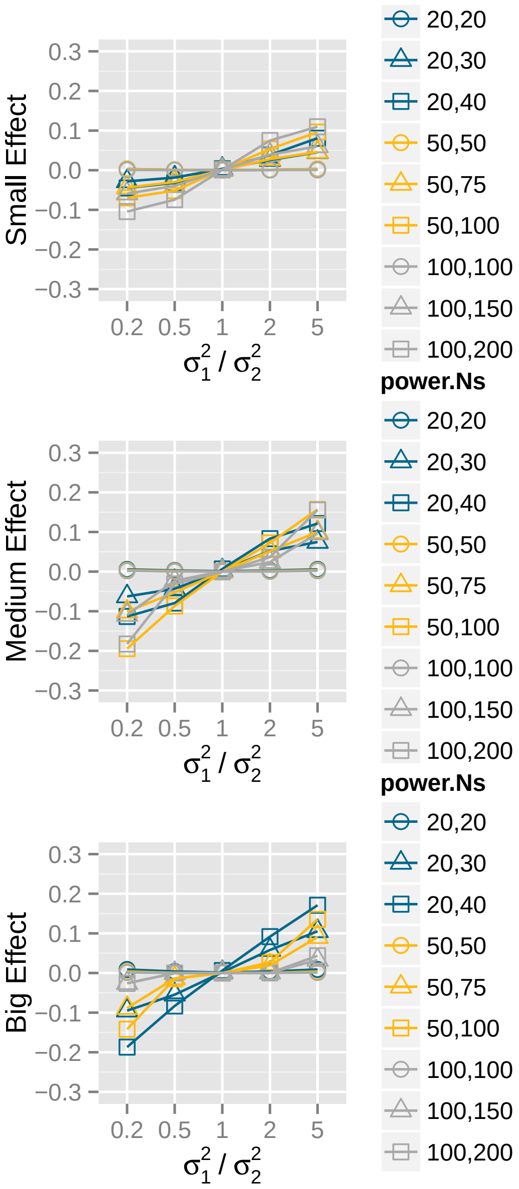

*Figure 10.* Difference in the Power of Student's and Welch's t tests (Student-Welch).

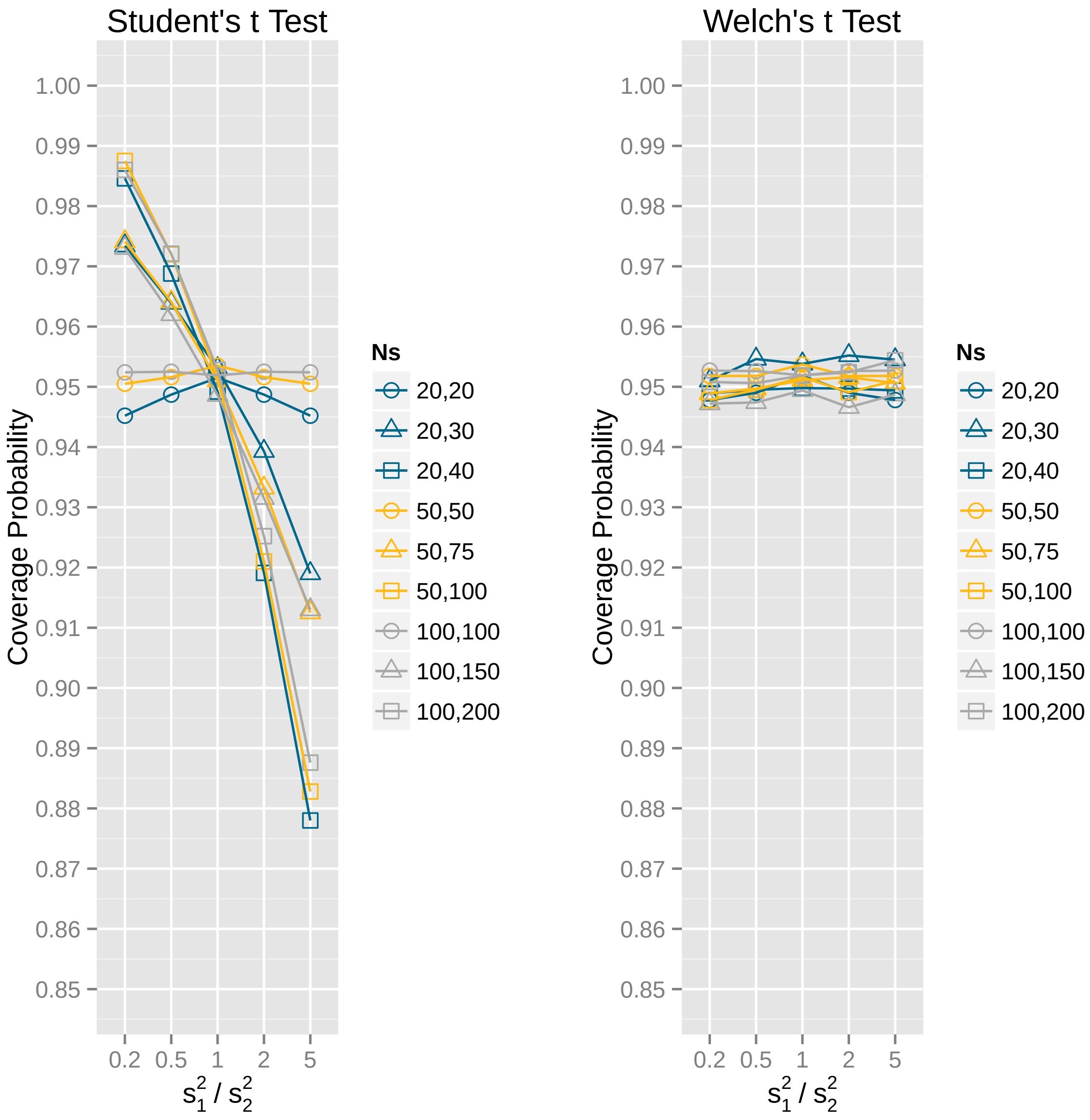

*Figure 10.* Coverage probabilities for Student's and Welch's t tests.

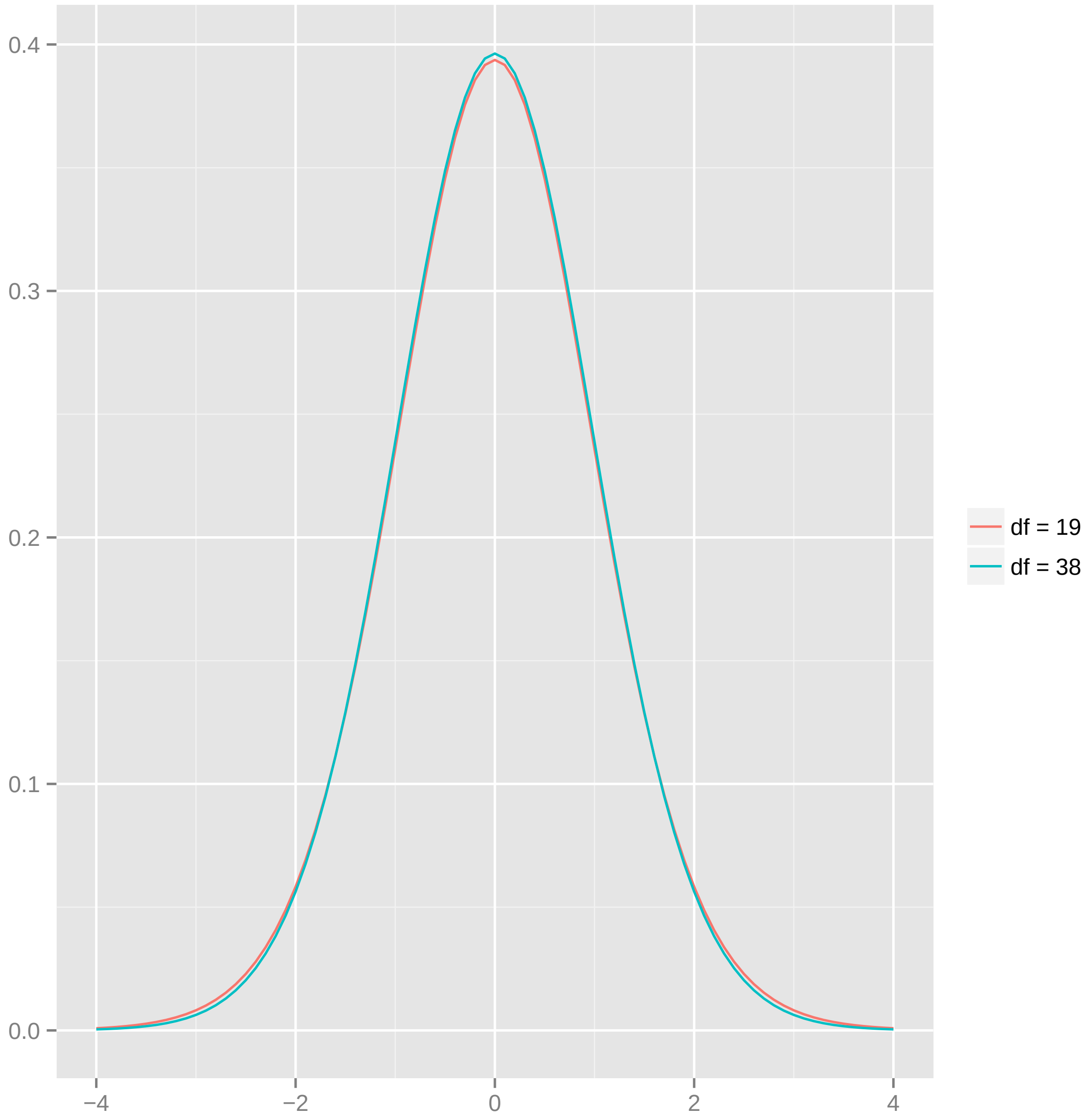

*Figure 11.* *Two t distributions with different degrees of freedom.*