Always Use the Separate Variances t Test for 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 the Separate Variances t Test for 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., ?, ?, ?, ?, ?, ?, ?) 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 (?, ?). Some have argued that due to some common research practices such as running underpowered studies, those that make it into published papers are spuriously large (e.g., ?, ?, ?).

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 (?, ?) or the alternative Welch-Satterthwaite test (?, ?, ?). 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:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}} \quad (1)$$

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 group variances are equal in the population, the group variances in the actual data are rarely identical due to sampling error, so you can't just use one of the group variances to find the standard error. Student's 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 own analyses. For those who use SPSS to analyze their data, the Welch 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 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 test in the direction of being more conservative and less likely to reject the null. On the one hand, this might make the Welch 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 test a worse choice if it is not powerful enough to detect true effects.

However, the Welch 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 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 (?, ?). 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 (?, ?, ?).

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 your own data, changing whether the variances are equal or unequal, and seeing if the boxplots of your data look like 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 test and Welch's 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 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 population variance ratio ($\sigma_1/\sigma_2 = 1/5$, 1/2, 1, 2, or 5; smallest $\sigma = 2$), the sample size ratio ($n_1/n_2 = 1$, 2/3, or 1/2), and sample sizes (smallest n = 20, 50, or 100).

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 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, these 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. We did not simulate the conditions with equal sample sizes and variance ratios of 2 and 5 because they were 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 for the first simulation each condition when sample sizes are equal. When variances are unequal, the larger variance is for the group to the right. When the sample sizes are smaller, it can be more difficult to tell whether the variances are unequal or not. When the sample sizes are 20, the variance of one group looks a bit larger than the other when they are really equal and the variances look approximately equal when one group has twice the variance of the other. As the sample size and increases, it becomes clearer to see whether there is a violation - at 100 subjects per condition, when variances are equal the boxes and whiskers are approximately the same length, and when variances are unequal both the box and whiskers appear wider in the group with the bigger variance. 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.

Figure 2 displays a sample of boxplots for simulations when the sample size ratio and variance ratio are both changing (NOTE: I'm not sure that this adds much so we might just drop it).

Does the df Ratio Help?

We examined whether looking at ratio of the Welch degrees of freedom to the classic t test degrees of freedom would provide a heuristic for deciding that equal variances does not hold. When both the sample sizes and variances are equal, As the sample sizes increase, the degrees of freedom ratio stays close to 1.

Figure 3 displays the ratio as the group variances become increasingly different when both the sample size and the sample size ratio stay the same. Under this condition, the distribution moves away from 1 as the difference in group variances increases.

In the first two plots, the degrees of freedom ratios are generally above 95% when variances are equal and below 95% when they differ. A useful heuristic might be to assume unequal variances when the ratio falls below 95%. But now look what happens when the variances are equal and the sample size ratio changes in Figure 4. Here too the ratio drops when the sample sizes become increasingly uneven, even though the variances stay the same. Our 95% heuristic would lead us astray and we would incorrectly conclude that the variances are unequal.

Finally, Figure 5 demonstrates one example of what happens to the degrees of freedom ratio when the variances become increasingly unequal and the sample sizes are unequal. In this case, the effect of different variances depends on whether the larger group has the larger variance or the smaller variance. In the first row, when the larger group has the larger variance, the move from equal to unequal variances actual counteracts the effect of the unequal sample sizes, and the degrees of freedom are generally higher at both unequal variance ratios that we sampled. However, in the second row, when the larger group has the smaller variance, the degrees of freedom ratio drops as the variances become increasingly unequal.

In short, the usefulness of a heuristic based on the ratio of degrees of freedom from the Welch test to degrees of freedom from Student's test is limited to cases when the sample sizes are equal.

One of the concerns about using the Welch test as an alternative to Student's t test is that the penalty on the degrees of freedom makes it difficult to find effects. The boxplots show that under ideal conditions, when the true population variances and the sample sizes are equal, the penalty is small. Even under non-ideal conditions, when both the population variances and the sample sizes are unequal, there is variability in the penalty depending on the specific configuration of sample sizes and variances.

When Does Each Test Perform Best?

The simple rule based on the degrees of freedom penalty did not work well, 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, and we report the observed Type I error rates, observed power, and coverage probability for the classic and separate variance t tests over the 10000 simulations.

Type I Error Rates

In this section we report type I error rates for the classic and separate variance t tests when the null hypothesis is true. Prior research has demonstrated that when either sample sizes or variances are equal, the type I error rates are preserved at .05 for both tests (CITATIONS). This prior work has typically examined sample sizes that are far smaller than the typical psychology experiment.

Figure 7 displays the type I error rate for the classic t test across the different conditions. Consistent with prior research, the type I error rate remained close to .05 when either the sample size or the population variances were equal, but it varied widely when both population variances and sample sizes were unequal. On the left side of Figure 1, when the group with the larger sample size had the larger variance, the type I error rate dropped as low as about .01, whereas on the right side, when the group with the larger sample size had the smaller variance, the type I error rate rose as high as .12, which is more than double the normally accepted false positive rate.

Figure 8 displays the type I error rate for the separate variance t test across the different conditions. In contrast to the classic t test, and consistent with prior research, the type I error rate remained close to .05 across all conditions.

Here we see the degrees of freedom penalty at work. Our boxplots showed that the penalty was greatest when the large group had the small variance, which reduced the false positive rate of the classic t test. In contrast, the penalty was less severe when the large group had the large variance, so that the Welch test did not reduce the type I error rate beyond the .05 level. Additionally, we can see the different effects of using pooled and separate variances to compute the standard errors. Because the pooled variance of the classic t test weighs the larger sample more heavily, it becomes more liberal when the associated variance is small and more conservative when the associated variance is large.

Power

Figure 9 displays the power of the classic and Welch tests to detect small, medium, and large effects under the different conditions. Overall, the classic t test is more powerful when the large sample has the smaller variance, whereas the Welch 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 the classic t test has the greatest power over the Welch 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 the classic t test had a risk of doubling the false positive rate. In contrast, the Welch test was more powerful than the classic test under other conditions and never inflated the type I error rate.

Taken together, the type I error rates and power favor the Welch test over the classic t test as better balancing researcher concerns.

Coverage Probability

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). Table 10 displays the coverage probability, which is how often the 95% confidence interval contains the true mean difference in groups, for the two tests under the different conditions. The coverage probability for the classic t test varies dramatically. When it is the least powerful, it is the most accurate. When it is the most powerful, the effect size estimation is the least accurate, and what seems to be a 95% confidence interval drops as low as an 88% confidence interval. Once again, the Welch test retains the expected 95% rate and turns out to best meet a researcher's concerns.

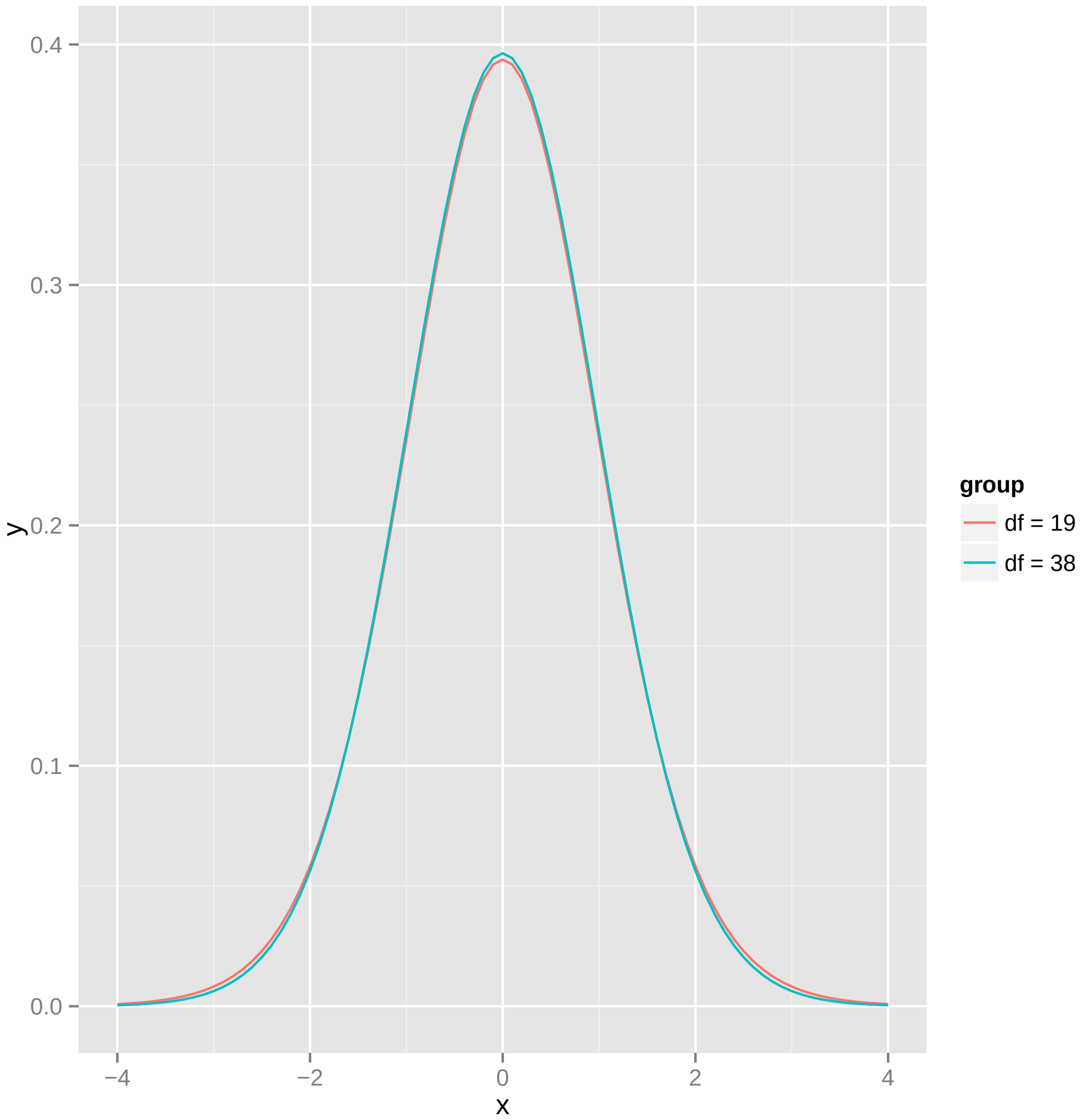
Discussion

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 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



Possible Cuts

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

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 $\alpha - errorcontroltovalidityproper : Problemswithashort - sightedfalse - positivedebate.$ *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

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

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi: http://dx.doi.org/10.1177/1745691612463401

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

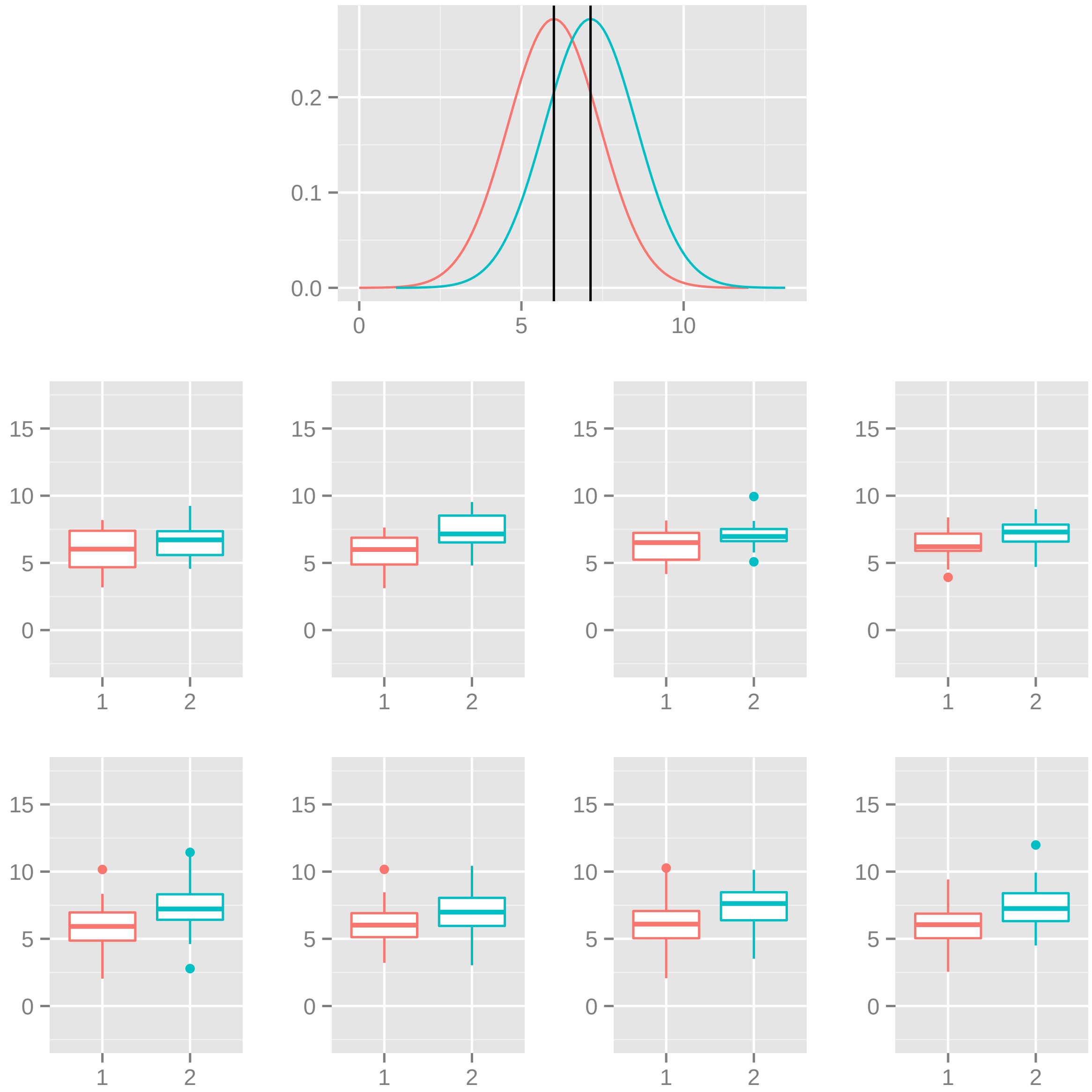


Figure 1. Boxplots for equal variances when sample sizes equal 10 and 100.

```
## Warning:  Removed 1 rows containing non-finite values (stat_boxplot).

## Warning:  Removed 1 rows containing non-finite values (stat_boxplot).
```

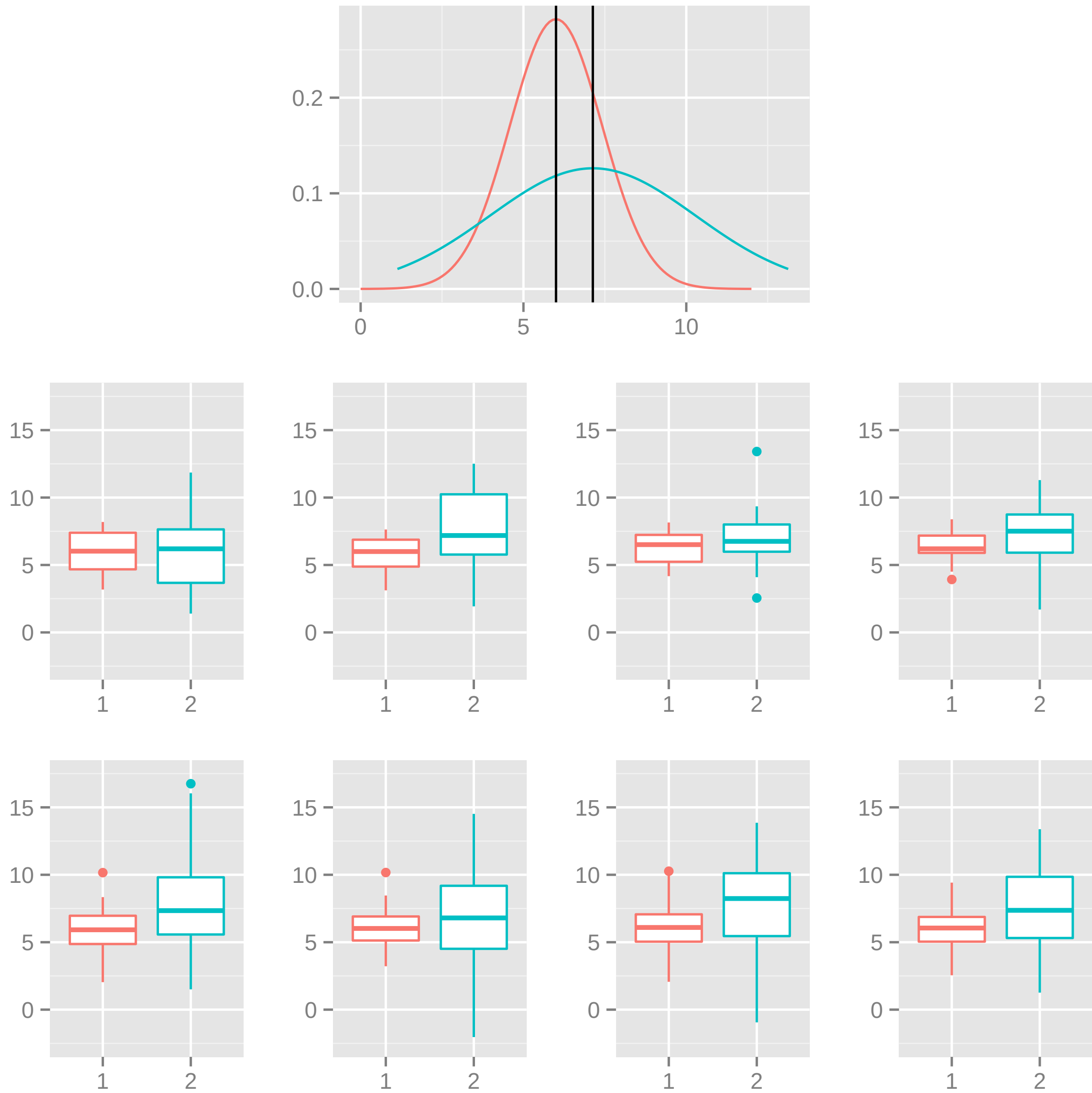


Figure 2. Boxplots for unequal variances when sample sizes equal 10 and 100.

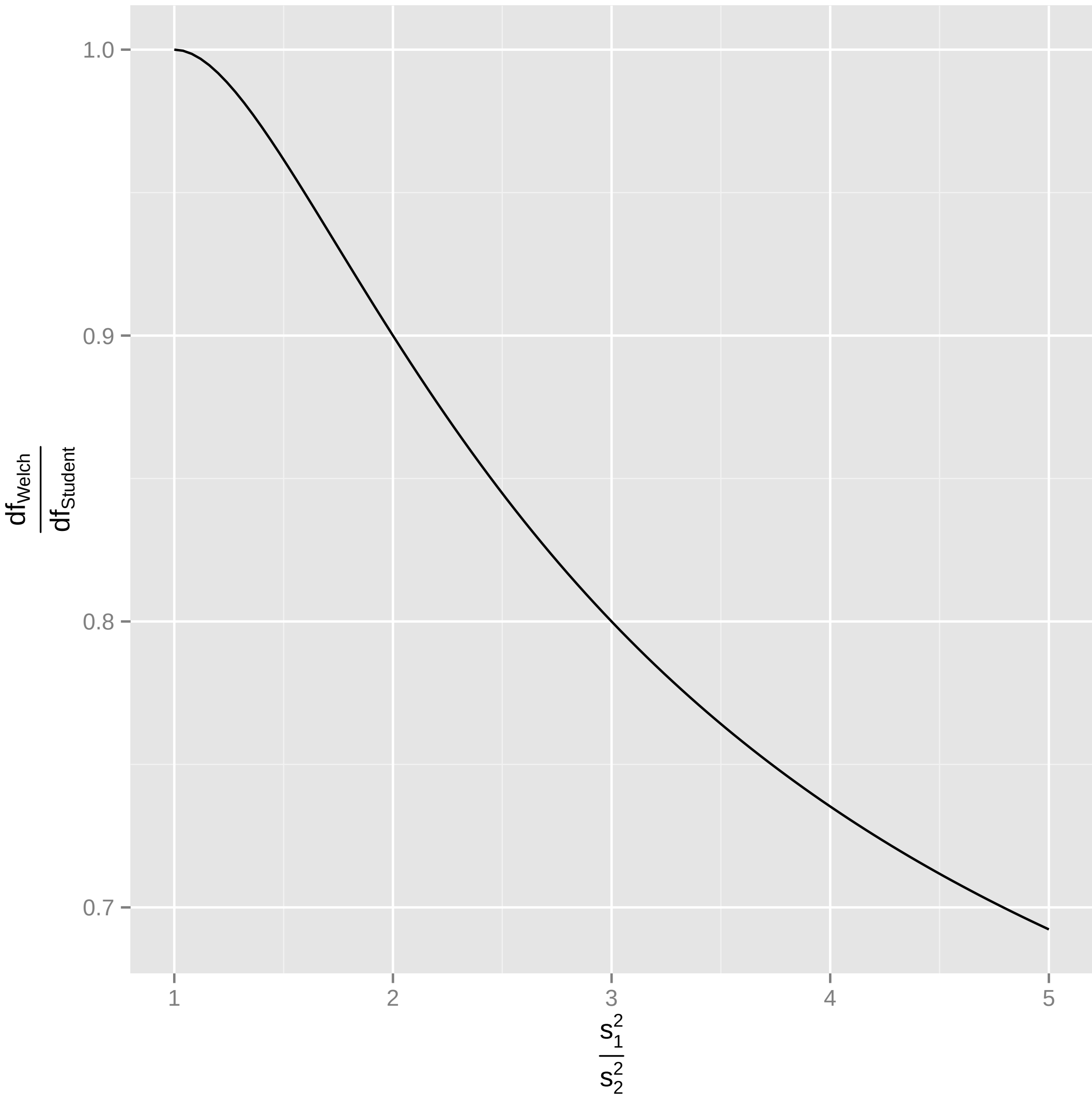


Figure 3. Degrees of freedom ratio when sample sizes are equal and variances are unequal.

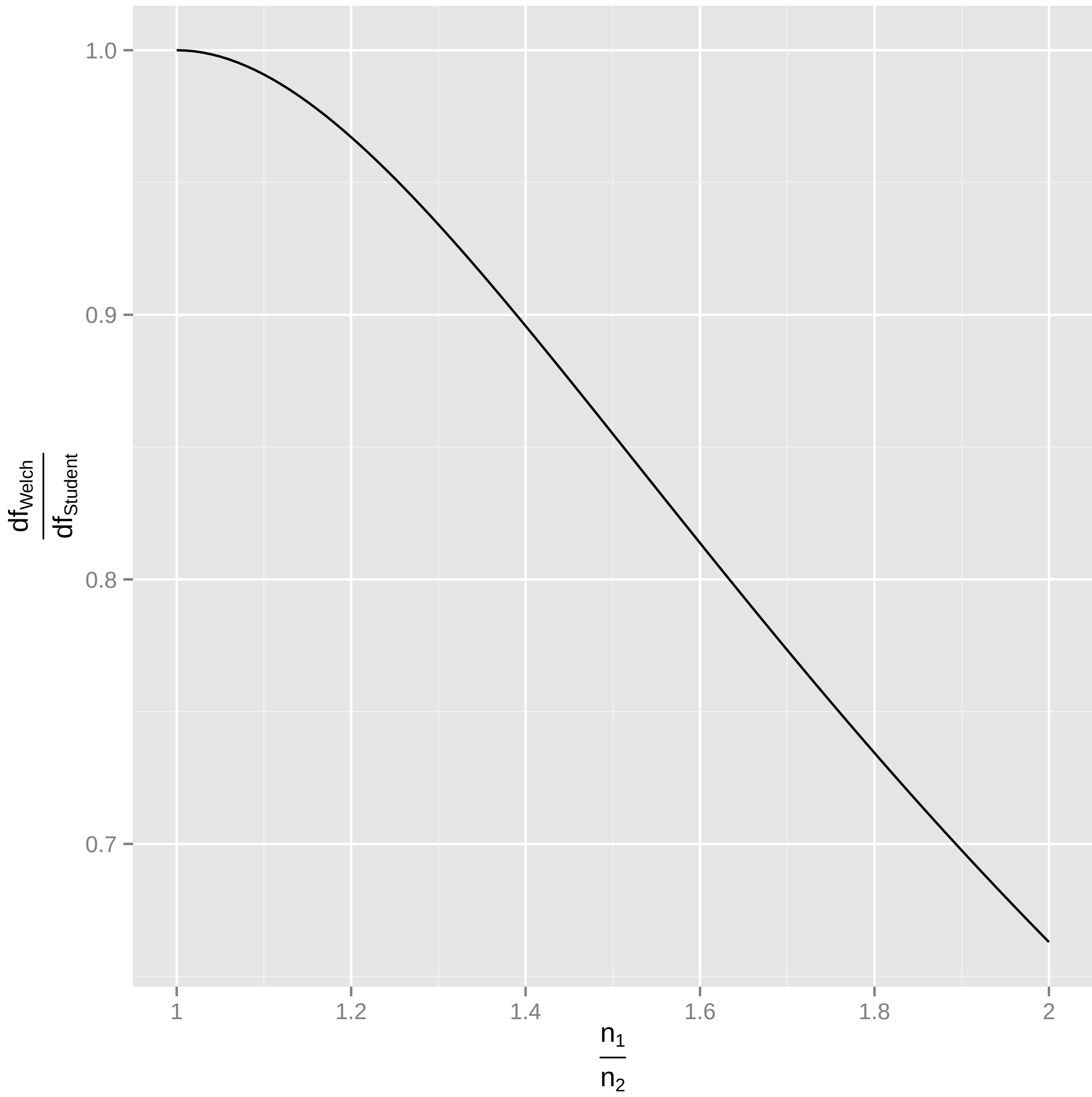


Figure 4. Degrees of freedom ratio when sample sizes are unequal and variances are equal.

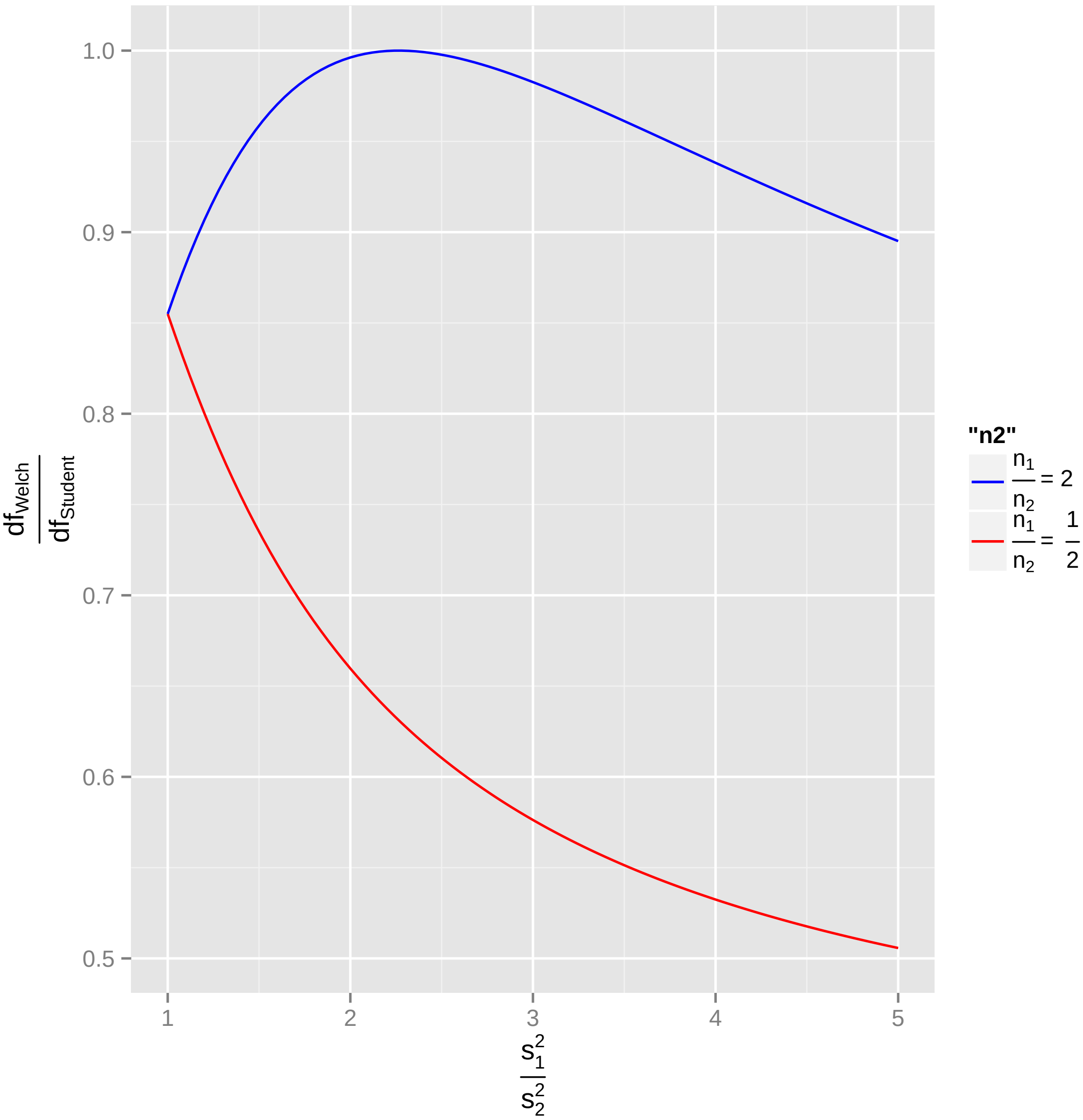


Figure 5. Degrees of freedom ratio when sample sizes are unequal and variances are unequal.

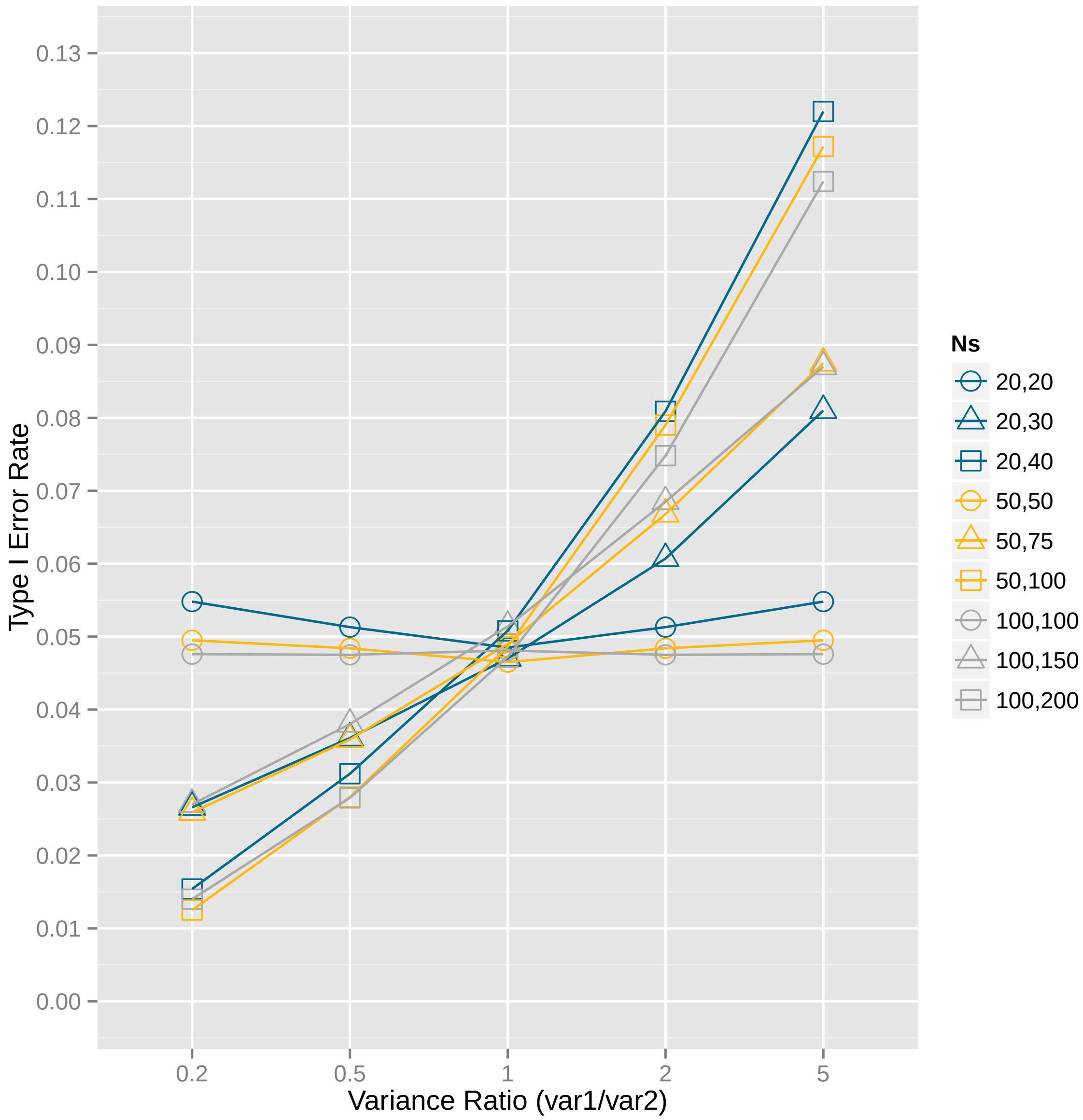


Figure 7. Type I error rates for Student's t test.

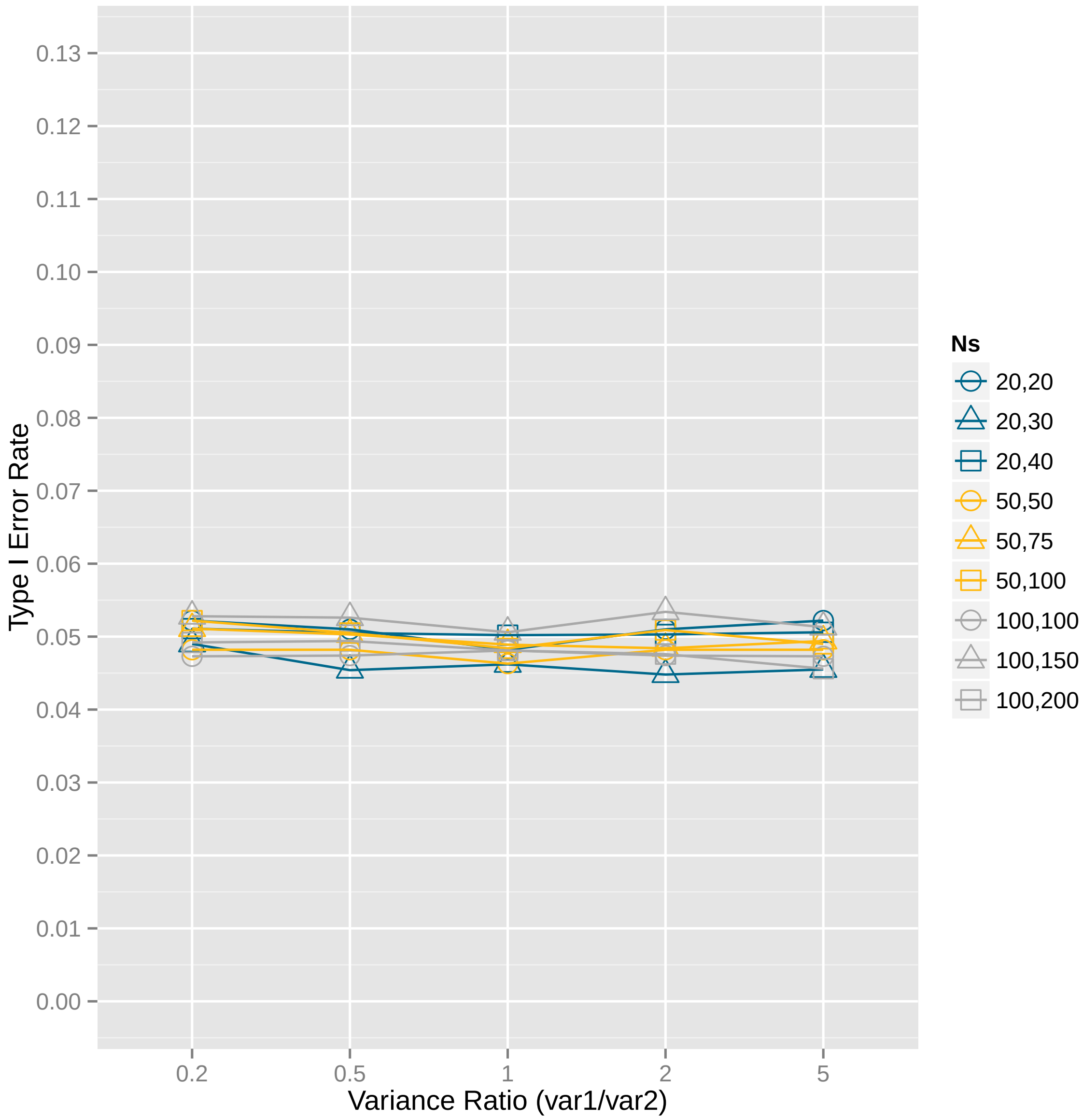


Figure 8. Type I error rates for Welch's t test.

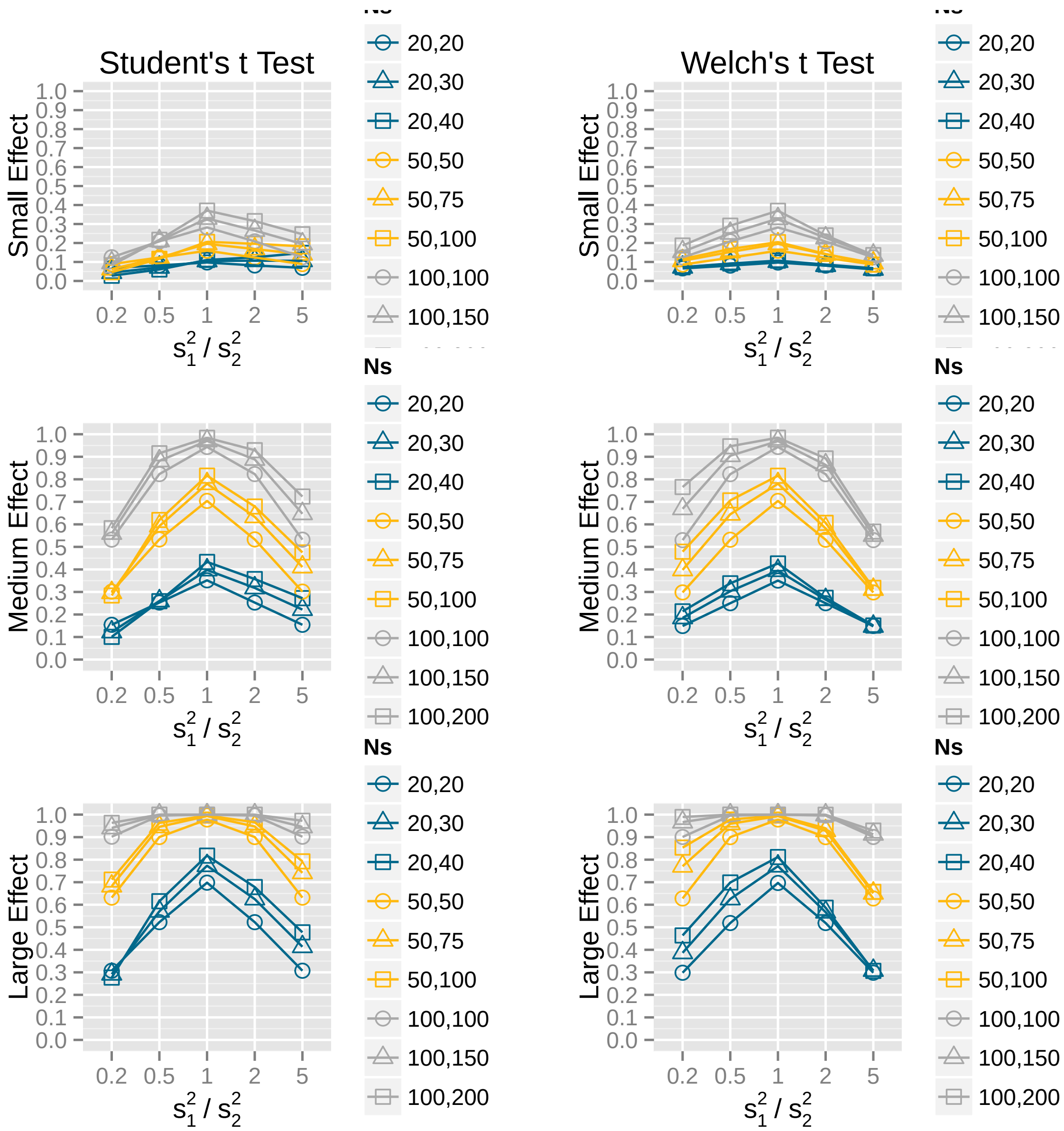


Figure 9. Power of Student's and Welch's t tests.

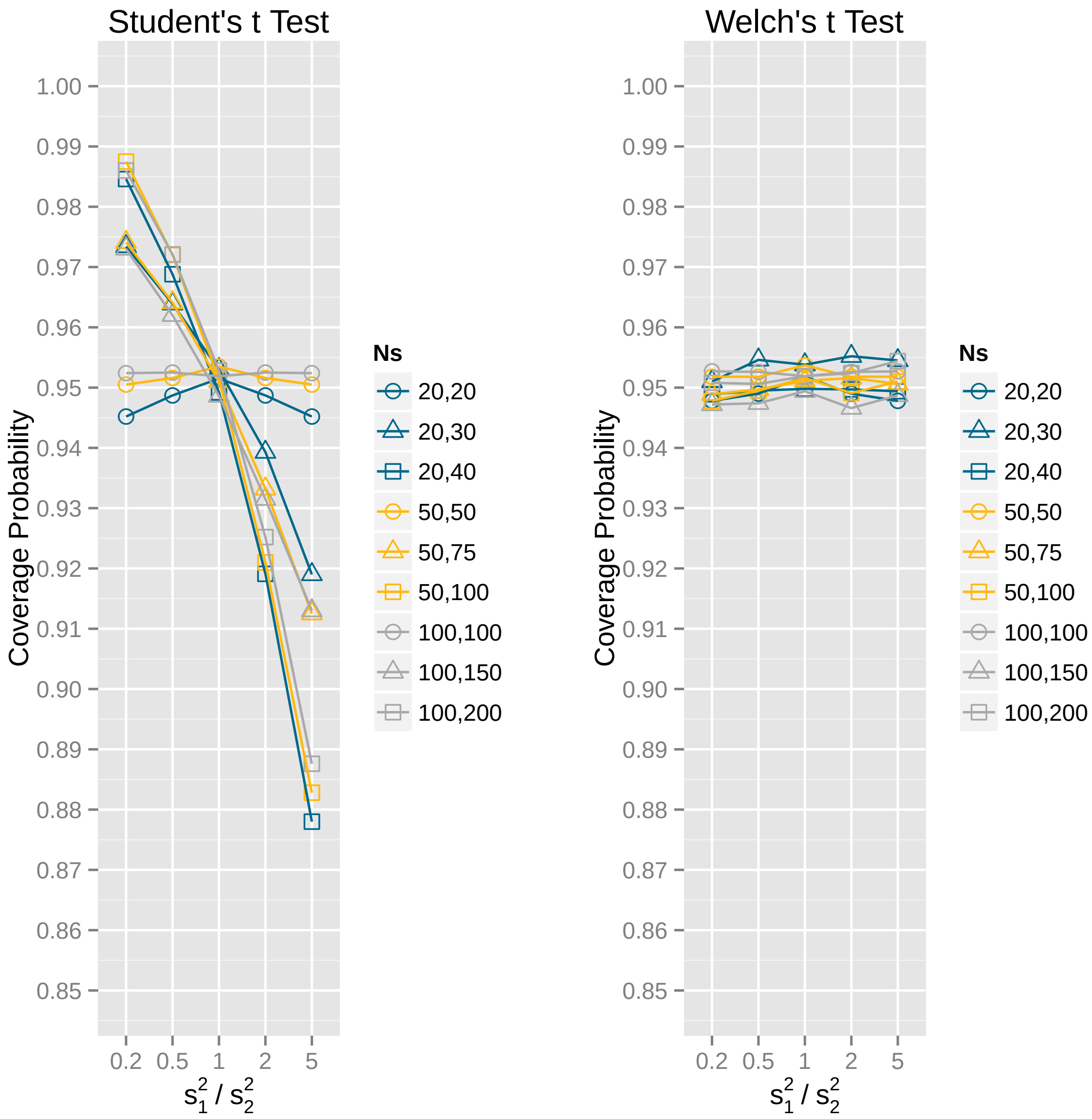


Figure 10. Coverage probabilities for Student's and Welch's t tests.