2025 ■ Vol. 21 ■ no. 1

# Same or different? Comparing the coverage rate of five different approaches for testing the difference of two groups means

Felix Bittmann [a]

[a]Leibniz Institute for Educational Trajectories, Wilhelmsplatz 4, 96047 Bamberg, Germany

**Abstract** ■ testing for statistically significant differences between two group means is one of the most common requirements in psychological research, for example, after an experiment has been conducted. While the classical t-test is probably the most popular approach, its deficiencies under violated assumptions have been acknowledged, and various alternative tests have been developed. In this research paper, five widely available methods are compared to investigate the coverage of the generated 95% confidence intervals. We utilize the coverage of confidence intervals as it corresponds to nominal type-I-error rates (Alpha), yet is more adequate since confidence intervals are preferred in contrast to p-values, which often facilitate binary conclusions. The approaches tested are the classical t-test, Welch's t-test, OLS regressions with robust standard errors, and two flavors of bootstrapping (normal and bias-corrected). Three different outcome distributions are generated (normal, uniform, skewed), and 75,000 simulations with a wide range of sample sizes (15 to 200 per group) and standard deviations are conducted for each. The results outline that Welch's t-test and the regression approach perform best. The bootstrap approaches tend to consistent undercoverage. The regular t-test produces larger deviations when its assumptions, especially the equality of variances, are violated. When distributions are skewed, all approaches result in undercoverage.

**Keywords** ■ mean comparison, t-test Welch test, robust standard error, bootstrapping, simulation, coverage.

✉ felix.bittmann@lifbi.de

10.20982/tqmp.21.1.p001

**Acting Editor** ■ Denis Cousineau (Université d'Ottawa)

**Reviewers**
■ One anonymous reviewer

## Introduction

Testing for differences between group means is highly relevant in psychology and related disciplines. Especially when an experiment has been conducted where a randomization procedure guarantees that participants are similar to each other regarding all background variables, a simple comparison of group means is adequate to estimate the causal effect of treatments (Morgan & Winship, 2015). In such scenarios, the t-test has been firmly established for a very long time to compute differences between group means and provide not only p-values but also confidence intervals for inference (Cressie & Whitford, 1986). The standard t-test makes specific assumptions on the data, which, if violated, can lead to incorrect conclusions (Boneau, 1960; Skaik, 2015). Unfortunately, these violations happen regularly in applied research, and numerous alternatives have been suggested that relax assumptions on the data (Wilcox, 2017, p, 162f). Some recent results also call for a general abandonment of the classical t-test in favor of Welch's approach, which controls type-I-error rates better when the variance homogeneity assumption is not met and has, in general, only few disadvantages (Delacre et al., 2017). In this investigation, we focus on confidence intervals, which have been suggested as an alternative and amendment to classical p-values (Gardner & Altman, 1986; Greenland et al., 2016). As they provide more information, they help researchers avoid the binary-conclusion trap and facilitate a more nuanced interpretation of the outcome (Wasserstein et al., 2019).

When evaluating confidence intervals, the coverage rate of the interval is of special interest. A 95% confidence

interval should contain the true population parameter in 95% of all independent samples. If the actual coverage rate is higher, this is referred to as overcoverage, if the rate is lower, it is undercoverage. If a method consistently produces deviations from the nominal coverage rate, this can lead to wrong conclusions (Hazra, 2017). If overcoverage is present, this leads to conservative results. This error implies that the null hypothesis (type-II-error) is accepted, even if it should be rejected. If undercoverage is present, the opposite error arises (type-I-error). There, the null is rejected incorrectly. Both errors can be critical, and a good statistical test should consistently produce confidence intervals very close to the nominal coverage rate. In general, the nominal coverage rate is 1 minus the Alpha level. Therefore, a test with a nominal coverage rate of 95% corresponds to an Alpha level of 5%. While in classical hypothesis testing frameworks, Alpha is the classical definition to control the type-I-error, in this paper we refer to coverage rates instead as they are more general and correspond to confidence intervals, which are preferable over reporting p-values.

Our research uses a simulation approach and thoroughly tests the empirical coverage rate of five widespread statistical approaches to test for group differences. These are the standard t-test, the t-test with Welch's correction, an OLS regression approach with robust standard errors, and two flavors of bootstrapping (standard and bias-corrected). We have chosen these approaches since they are either popular or widely used, well-known to many researchers, and implemented in most modern statistical software packages, ready to use "out of the box".

In this paper, we would like to be able to give practical advice to researchers who face various data constellations that often violate the assumptions of some tests. We chose a range of distributions common in empirical research (normal, uniform, skewed) and selected a wide range of sample sizes and variances. Since most research is constrained by financial means, relatively small sample sizes (below 200 per group) are common in psychology (Marszalek et al., 2011). Hence, the focus on these sample sizes as larger sample sizes usually produce more robust statistical results. The variances are also related to the sample sizes. It is well known that even when probability sampling is carried out correctly, sample characteristics can deviate strongly from the population due to random sampling error (Tipton et al., 2017). Variances can also be diverse in small samples, which is why we implement a wide range of potential variances. Our specifications reflect the situations many researchers face in their daily work and are hence able to give practical advice.

## Statistical approaches to compare group means

In this section, we give a short overview of all five methods and explain their design and the assumptions they make. Note that the point estimate is always and for all methods generated by computing the arithmetic mean for each group and forming the difference between these two values. The question remains how to generate confidence intervals for this difference.

### *Student's t-test*

The first step is to compute the t-statistic as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{2}/\sqrt{\tilde{n}}} \tag{1}$$

where $\overline{X}_1$ and $\overline{X}_2$ are group means. The value

$$\frac{\sqrt{2}}{\sqrt{\tilde{n}}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is related to the harmonic mean of the sample sizes. Here, $s_p$ is the pooled standard deviation, computed as follows:

$$s_p = \sqrt{\frac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n_1+n_2-2}} \tag{2}$$

This formula applies when group sizes are unequal and variances are similar (yet there is no requirement for identical variances) and is therefore the general form. Based on the t-statistic, inference can be done. The degrees of freedom for this test are $n_1 + n_2 - 2$. Based on the degrees of freedom one can compute the critical value with the inverse cumulative Student's t distribution and generate the CI as follows:

$$CI_{\text{lower}} = (\overline{X}_1 - \overline{X}_2) - T_{\text{crit}} \times s_p \tag{3}$$

and

$$CI_{\text{upper}} = (\overline{X}_1 - \overline{X}_2) + T_{\text{crit}} \times s_p \tag{4}$$

The general form of computing CIs is the same for the Welch tests as well for the regression approach yet computing standard errors and the critical value are different. Note that assumptions on the data for this basic test are rather strict. The sample sizes must be close to equal and the distributions of the outcome variable are assumed to have a similar variance (a factor of 2 is the largest difference generally accepted).

### *Welch's t-test*

The assumptions of equal sample size and equal variances are relaxed by Welch's approach. While the general computation is very similar, the degrees of freedom are computed

differently. The degrees of freedom are computed using the following equation:

$$df \approx \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{s_1^4/(n_1^2 \times \nu_1) + s_2^4/(n_2^2 \times \nu_2)} \quad (5)$$

where $\nu_i = n_i - 1$ are computed according to the Welch–Satterthwaite equation (Kirkup & Frenkel, 2006).

### *OLS with robust standard errors*

Regression models are a highly popular method in empirical research since they are flexible, easily adapt to a wide range of outcome distributions, and allow the inclusion of control variables. We chose Ordinary Least Square (OLS) regression models in this example since the outcome variables are continuous. In this regression, there is exactly one dependent (outcome) variable and one independent (explanatory) variable, which is binary and indicates the group. To produce robust results, we compute robust standard errors.[1] The literature describes various options for robust standard errors. Usually, the Huber/White sandwich estimator is meant when "robust" standard errors in regressions are discussed. We test an even more conservative version, sometimes referred to as HC3. For a discussion of the question, also refer to the literature (Angrist & Pischke, 2009; Davidson & MacKinnon, 1993).

### *Bootstrapping*

Bootstrapping is a conceptually simple yet powerful approach to generating standard errors or confidence intervals for virtual any statistic (Bittmann, 2024; Efron & Tibshirani, 1994). It is a nonparametric approach that makes little assumptions and relies on resampling. First, the point estimator is computed as described above (the difference between the two group means). Afterward, the data is resampled many times; random samples are drawn with replacement. By doing so, many new differences are computed and stored. After the resampling, there are two common ways to generate a confidence interval. The first is to compute the standard deviation of all generated bootstrap results and use this as the standard error of interest (normal approach). The second approach is to use the empirical bootstrap distribution and compute the quantiles of interest (for a 95% confidence interval, quantiles 2.5 and 97.5; Diciccio & Romano, 1988). Additionally, if the arithmetic mean of the bootstrap distribution deviates from the empirical difference, one can apply a bias-correction factor (bias-corrected bootstrap CI).

### Simulation study

The point estimate is the difference between the two groups' means when comparing two independent samples. Reporting a confidence interval for this statistic is recommended as it directly corresponds to a classical p-value. For example, if the confidence level is set to 95% and zero is not included in this confidence interval, the p-value of the test is smaller than 5%. However, in contrast to p-values, confidence intervals report more information and are recommended to evaluate statistical significance not only in a binary fashion but also to assess the potential effect size. We conduct a simulation study to evaluate the coverage rate of the five methods. In a simulation study, the researchers know the true parameters and various methods can be compared to these known values. In this study, we vary the following parameters:

- the distribution of the outcome variable (normal, uniform, skewed)
- the sizes of both groups
- the standard deviations of the outcome variable of both groups

We conduct the simulations separately by type of distribution. We start with a baseline simulation that serves as a benchmark. Next, for simulation one, the outcome variables in both groups are normally distributed with their means fixed to zero. The standard deviation is drawn uniformly from the interval (0.4, 1.6; meaning that the nominal factor is at most 4), and the group sizes are drawn uniformly from the interval (15, 200).

For simulation two, the outcome variables in both groups are uniformly distributed. The distributions are transformed to have standard deviations from the interval (0.4, 1.6), just as in the first simulation. The group sizes have the same specifications.

For simulation three, we would like to have a skewed outcome variable. To this end, we utilize the beta distribution and vary the $p$ and $q$ parameters. $p$ is drawn from the interval (1.5, 4), and $q$ from the interval (5, 9).

Note that in each simulation, the null hypothesis (no difference between the two group means) is specified to be correct, and, on average, one would not expect to see a group difference. Under these specifications, the confidence intervals are assumed to include zero in 95% of all simulations. A good approach should generate confidence intervals that are, on average, close to this nominal level. For each scenario, we conduct a substantial number of 75,000 simulations. Our empirical tests have shown that this number of simulations is sufficient to produce high-quality and precise results. For the bootstrapping ap-

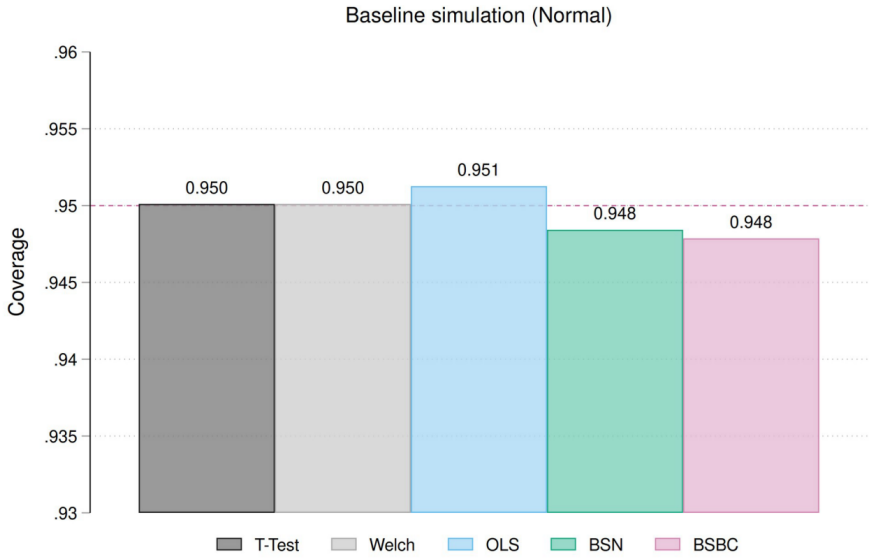
[1] To produce these standard errors in Stata, use the option `vce(hc3)`; not `vce(robust)`. For a discussion of this pitfall in Stata, refer to https://datacolada.org/99 [last accessed: 2024-06-12] and https://blog.stata.com/2022/10/06/heteroskedasticity-robust-standard-errors-some-practical-considerations/

**Table 1** ■ Summary statistics of the baseline simulation

| | Mean | Median | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| $Mean_1$ | 0.00034 | 0.00012 | 0.100 | -0.39 | 0.49 | 0.011 | 3.00 |
| $Mean_2$ | -0.000093 | -0.000060 | 0.10 | -0.48 | 0.38 | 0.012 | 2.98 |
| $SD_1$ | 1.00 | 1.00 | 0.071 | 0.71 | 1.28 | 0.067 | 3.01 |
| $SD_2$ | 1.00 | 1.00 | 0.071 | 0.69 | 1.31 | 0.069 | 3.00 |
| $N_1$ | 100 | 100 | 0 | 100 | 100 | . | . |
| $N_2$ | 100 | 100 | 0 | 100 | 100 | . | . |
| Simulations | | | | 75000 | | | |

*Note.* Source: own computations. Some statistics are undefined since the values are constant.

**Figure 1** ■ Overall coverage probabilities by approach / baseline. Source: own computations, 75K simulations. Welch: t-test with Welch's correction; OLS: OLS regression with robust standard errors; BSN: bootstrapping with normal CIs, BSBC: bootstrapping with percentile and bias-corrected CIs. The nominal (target) coverage is indicated by the dashed line.




proach, we specify 7,000 random resamples, which should provide a high degree of precision. We evaluate 95% confidence intervals of the group mean differences. All analyses are conducted in `Stata` 16.1 (Bittmann, 2019), and replication files are available from the author. We have used the user-written program `parallel` (Vega Yon & Quistorff, 2019).

## Results

### *Baseline simulation*

To ensure correct results, we conduct a baseline simulations first where all assumptions of the classical t-test are met. This has multiple purposes. First, it serves as a general test of the simulations and ensures that the programming is correct. Second, it is a comparison point for all other methods and makes it easier to check whether some methods are conservative or liberal. In this simulation, the sample sizes are rather large ($N = 100$ per group), the means are fixed to zero, the standard deviations are fixed to 1. This ensures homogeneity of variances. The distribution of the outcome variable is normal. We conduct 75,000 simulations. The descriptive results are as follows, showing that the specifications are done correctly (Table 1).

As these results from Table 1 look fine, we present the results graphically in Figure 1.

As these first results indicate, all methods are very close to the nominal coverage rate of 95% under ideal conditions. The point estimates are reported in Figure 1, 95% confidence intervals of these estimates are provided in

**Table 2** ■ Summary statistics of simulation 1 specifications (normal)

| | Mean | Median | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| $Mean_1$ | 0.00066 | 0.00056 | 0.13 | -1.24 | 1.07 | -0.037 | 7.07 |
| $Mean_2$ | -0.00046 | -0.00060 | 0.13 | -1.21 | 1.17 | -0.053 | 7.20 |
| $SD_1$ | 1.00 | 0.99 | 0.36 | 0.26 | 2.34 | 0.12 | 2.00 |
| $SD_2$ | 1.00 | 0.99 | 0.36 | 0.23 | 2.19 | 0.11 | 1.98 |
| $N_1$ | 107.3 | 108 | 53.6 | 15 | 200 | -0.0037 | 1.80 |
| $N_2$ | 107.6 | 108 | 53.9 | 15 | 200 | -0.0046 | 1.79 |
| Ratio $SD_1/SD_2$ | 1.65 | 1.46 | 0.61 | 1.00 | 6.14 | 1.42 | 5.05 |
| Ratio $N_1/N_2$ | 2.48 | 1.76 | 1.86 | 1 | 13.3 | 2.24 | 8.65 |
| Unequal $SD$ | 0.68 | | | 0 | 1 | | |
| Simulations | | | | 75000 | | | |

*Note*. Source: own computations. $Mean$: Empirical arithmetic mean; $SD$: Empirical standard deviation; $N$: Number of observations.

parentheses. The classical t-test (94.85%; 95.17%) and the Welch-test (94.85%; 95.17%) deliver an ideal result as expected. The OLS approach shows a very small overcoverage (94.97%; 95.27%), yet the nominal value is included in the confidence interval, so it is unlikely that the overcoverage is substantial, given the large number of simulations. The bootstrap approaches tend to slight undercoverage in both the BS normal (94.68%; 95.00%) and the BS BC approach (94.62%; 94.94%). However, in any case, these deviates are rather small. We can conclude from these results that some methods are not more conservative in general but reach the nominal coverage under ideal conditions. Furthermore, these baseline results demonstrate that the simulations are specified correctly and all methods deliver fine results when the data conditions are ideal. In the following simulations, we can gauge how the quality of the methods change as soon as ideal conditions are no longer present.

### *Normal distribution*

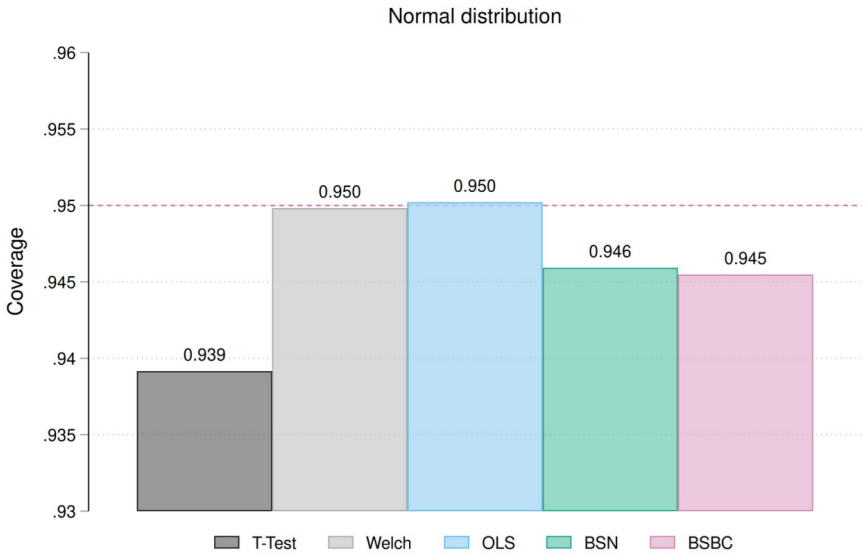
We start by presenting some descriptive statistics of the simulation study with a normal distribution (Table 2). This is relevant to show that the simulations come close to what has been specified before. Due to generating variables with random numbers, small deviations are always expected yet uncritical. A Levene test has also been conducted to test whether variances in the two groups are unequal. A p-value below 0.05 indicates a difference of variances, which has been recoded into a binary variable (1 = unequal variances, 0 = equal variances).

For a more convenient interpretation, we have recoded three key parameters. First, the size of the smaller of the two groups was recoded into three categories (15 to 30 / 31 to 75 / 76 to 200). The ratio of standard deviations has been recoded into three categories (1 to 1.3 / 1.3 to 2.2 / 2.2 to 6.5). The ratio of the two group sizes has been recoded into three categories (1 to 1.5 / 1.5 to 2.5 / 2.5 to 15). These categories were chosen to have approximately similar case numbers in each category. For more insight, we also formed all combinations between sd-ratio and size-ratio groups, resulting in nine distinct groups. We start with an overview of all simulations in Figure 2.

We see here that Welch's approach and OLS regressions with robust standard errors are very close to the nominal level of 95%. The bootstrap approaches are slightly off, while the standard t-test has the largest deviations. However, this is a general summary of all simulations. Next, we present the detailed results in table form (Table 4 at the end).

Starting with the overall sample and all specifications, we note that the Welch t-test and the OLS approach are very close to the nominal coverage rate of 95%. This means that, on average, these two tests are recommended when all dataset details are unknown. Moving to the first distinction of interest, we divide the dataset by constellations where the assumption of equal variances has been violated, indicated by a statistically significant Levene test ($p < 0.05$). When the assumption is met, all tests produce rather fine results, yet the best approach is the regression approach with robust standard errors. When the assumption is violated, the Welch approach is slightly closer. Next, we test different groups for the SD and N ratios of both groups, and since the overall number of comparisons is large, we do not discuss them in detail. However, the result is that the Welch tests and the regression approach always perform best. The standard t-test often shows a strong undercoverage, sometimes with coverage rates below 90%. On average, the bootstrap approaches are better than the standard t-test yet have consistent under-coverage. Based on these findings, we conclude that either the Welch t-test or

**Figure 2** ■ Overall coverage probabilities by approach / simulation 1 (normal) Source: own computations, 75K simulations. Welch = t-test with Welch's correction; OLS = OLS regression with robust standard errors; BSN: bootstrapping with normal CIs, BSBC: bootstrapping with percentile and bias-corrected CIs. The nominal (target) coverage is indicated by the dashed line.




the regression approach deliver the best results in terms of nominal coverage under a wide range of simulation specifications when the baseline distributions are normal.

### *Uniform distribution*

The simulation setup is exactly the same as before, but the outcome variables are now uniformly distributed in both groups. The descriptive statistics are presented in Table 3.

The overall summary is graphically presented (Figure 3). The conclusion is that for this kind of distribution, the Welch approach and the regression approach yield the best coverage results. However, all approaches tend to slight un dercoverage. The basic t-test performs worst. Numerical results are reported in Table 5.

Overall, the Welch approach and the regression approach are usually the closest to the nominal coverage rate.

### *Skewed distribution (beta)*

Finally, we repeat the simulation with an outcome variable skewed to the right (e.g., income distribution). In this simulation, we systematically varied the beta distribution parameters $p$ and $q$. When analyzing this dataset, we focus on the skewness and kurtosis of the variable and not the standard deviation, as this has already been done in the two other simulations. The descriptive statistics are presented in Table 6. The absolute difference between the skewness and kurtosis for the two outcome variables must be compared. The classifications are as follows. For the skewness (0 to 0.10 / 0.10 to 0.36 / 0.36 to 3) and the kurtosis (0 to 0.20 / 0.20 to 0.85 / 0.85 to 11) to generate groups for a more convenient comparison.

A graphical summary of the results is as follows (Figure 4). The standard t-test works better with skewed distributions than others, as the results are quite close to the nominal coverage rate. The numerical results are presented in Table 7.

Interestingly, in the simulation with a skewed outcome variable, where the variances of the groups are highly similar, the classical t-test performs fine. However, Welch's test and the regression approach are also quite good; only the bootstrap approach results in inconsistent undercoverage. However, overall, all approaches result in undercoverage.

## Summary and conclusions

In this paper, we have tested which approach is the most advantageous when computing confidence intervals for the difference of two independent group means. We suggest two methods that consistently deliver the best results by conducting many simulations with a great variety of input to mimic realistic conditions. We have seen that Welch's t-test and OLS regressions with robust standard errors come closest to the nominal coverage rate of 95% with-

**Table 3** ■ Summary statistics of simulation 2 specifications (uniform)

| | Mean | Median | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| $Mean_1$ | 1.73 | 1.73 | 0.61 | 0.46 | 3.74 | 0.071 | 1.93 |
| $Mean_2$ | 1.73 | 1.73 | 0.61 | 0.015 | 3.51 | 0.0016 | 1.92 |
| $SD_1$ | 1.00 | 1.00 | 0.35 | 0.30 | 2.04 | 0.051 | 1.89 |
| $SD_2$ | 1.00 | 1.00 | 0.35 | 0.25 | 2.01 | 0.039 | 1.87 |
| $N_1$ | 107.7 | 108 | 53.7 | 15 | 200 | -0.0033 | 1.80 |
| $N_2$ | 107.8 | 108 | 53.6 | 15 | 200 | -0.0063 | 1.80 |
| Ratio $SD_1/SD_2$ | 1.64 | 1.45 | 0.60 | 1.00 | 5.26 | 1.34 | 4.47 |
| Ratio $N_1/N_2$ | 2.46 | 1.75 | 1.83 | 1 | 13.3 | 2.25 | 8.74 |
| Unequal $SD$ | 0.67 | 1 | | 0 | 1 | | |
| Simulations | | | | 75000 | | | |

*Note*. Source: own computations. Mean: Empirical arithmetic mean; SD: Empirical standard deviation; N: Number of observations.

out a true treatment effect. As the baseline simulation has shown, this is not due to the fact that these methods are overly conservative as they reach the nominal coverage of 95% under ideal conditions. The classical t-test creates larger deviations from the norm as soon as the assumptions it makes are violated. Interestingly, the bootstrap approach results in consistent undercoverage, already in the baseline specification. These findings are stable and independent of distribution type or other simulation specifications. What recommendation can be given to researchers?

In contrast to our simulation study, a researcher usually only has a single sample at hand, and it is unclear whether one method will produce a "good" confidence interval. However, as we have seen, Welch's test and OLS regressions (with robust HC3 standard errors) usually result in fine outcomes; hence, we recommend these approaches. These methods are easy to use and implemented in most statistical software packages, making them available to many researchers. Given the well-known shortcomings of the classical t-test, we see little reason to use it. Bootstrapping, which can be highly relevant if standard errors cannot be derived analytically, is probably unnecessary to test for group differences. However, if the distribution of the outcomes is highly skewed, researchers should be alerted. Our simulation has shown that all approaches result in undercoverage. As a remedy, researchers can apply the log-transformation to make their data normal; they can compute the Mann–Whitney U test (McKnight & Najab, 2010), or they should adjust their confidence intervals to another level to account for this, which is sometimes referred to as calibration. In any case, special attention is required for such distributions.

Lastly, we would like to acknowledge the limitations of our study. First, even if a wide range of simulation specifications have been chosen, no simulation can account for all potential data constellations. Researchers can easily check the distribution and variances of their samples and see whether they think our simulation is similar. Usually, if sample sizes are larger, the results are generally more robust, which is why they were not included in this simulation. Second, this study focused on confidence intervals, related to the statistical testing of the alpha error. The question of statistical power has yet to be investigated (De Winter, 2019). This means that a good statistical test should have the power to detect a true effect if present. Whether the approaches we have investigated in this simulation also have high power remains unanswered, and future studies could focus on that question.

## References


Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. doi: 10.1515/9781400829828.

Bittmann, F. (2019). *Stata: A really short introduction* (1st). De Gruyter Oldenbourg. doi: 10.1515/9783110617160.

Bittmann, F. (2024). *Applied bootstrap analysis with imputed data in stata*. doi: 10.20944/preprints202401.0813.v1.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49–64. doi: 10.1037/h0041412.

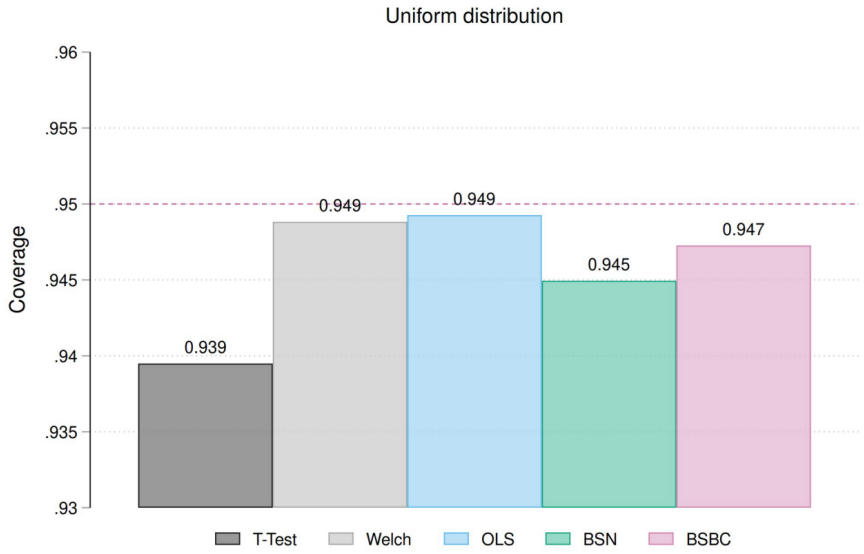
Cressie, N., & Whitford, H. (1986). How to use the two sample t-test. *Biometrical Journal*, *28*(2), 131–148.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. Oxford University Press. doi: 10.1017/S0266466600009452.

De Winter, J. C. (2019). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, *18*(1), 10–19.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's

**Figure 3** ■ Overall coverage probabilities by approach / simulation 2 (uniform) Source: own computations, 75K simulations. Welch: t-test with Welch's correction; OLS: OLS regression with robust standard errors; BSN: bootstrapping with normal CIs, BSBC: bootstrapping with percentile and bias-corrected CIs. The nominal (target) coverage is indicated by the dashed line.




t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi: 10.5334/irsp.82.

Diciccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *50*(3), 338–354. doi: 10.1111/j.2517-6161.1988.tb01732.x.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press. doi: 10.1201/9780429246593.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *BMJ*, *292*(6522), 746–750. doi: 10.1136/bmj.292.6522.746.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi: 10.1007/s10654-016-0149-3.

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4124–4129. doi: 10.21037/jtd.2017.09.14.

Kirkup, L., & Frenkel, R. B. (2006). *An introduction to uncertainty in measurement: Using the gum (guide to the expression of uncertainty in measurement)* (1st). Cambridge University Press. doi: 10.1017/CBO9780511755538.

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. doi: 10.2466/03.11.PMS.112.2.331-348.

McKnight, P. E., & Najab, J. (2010). Mann-whitney u test. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology (1st ed.)* (pp. 1–1). Wiley. doi: 10.1002/9780470479216.corpsy0524.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (Second). Cambridge University Press. doi: 10.1017/CBO9781107587991.

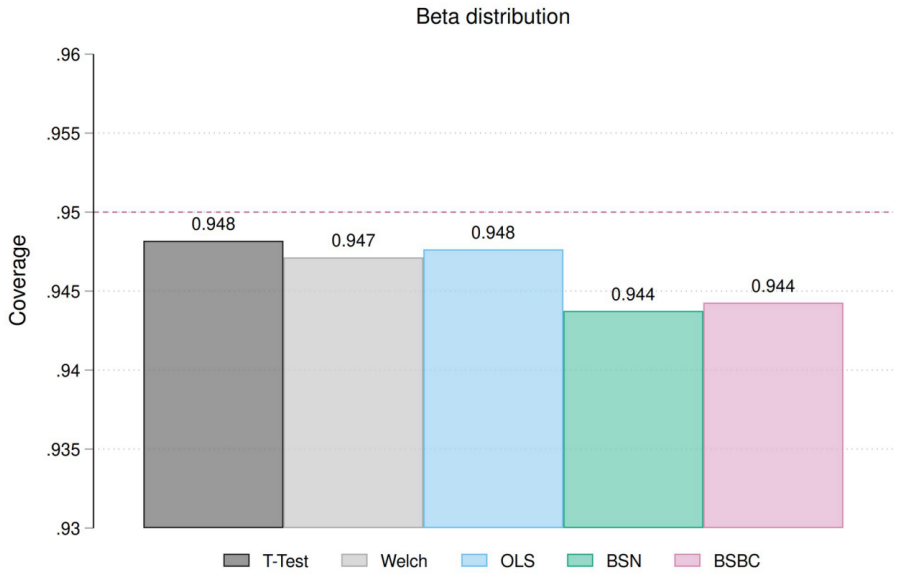
Skaik, Y. A. (2015). The bread and butter of statistical analysis "t-test": Uses and misuses. *Pakistan Journal of Medical Sciences*, *31*(6), 1558–1559. doi: 10.12669/pjms.316.8984.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472–505. doi: 10.1177/0193841X16655665.

Vega Yon, G. G., & Quistorff, B. (2019). Parallel: A command for parallel computing. *The Stata Journal: Promoting Communications on Statistics and Stata*, *19*(3), 667–684. doi: 10.1177/1536867X19874242.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond " p < 0.05." the American statis-

**Figure 4** ■ Overall coverage probabilities by approach / simulation 3 (beta) Source: own computations, 75k simulations. Welch: t-test with Welch's correction; OLS: OLS regression with robust standard errors; BSN: bootstrapping with normal CIs, BSBC: bootstrapping with percentile and bias-corrected CIs. The nominal (target) coverage is indicated by the dashed line.




tician. *73(sup1)*, 1–19. doi: 10.1080/00031305.2019.1583913.

Wilcox, R. R. (2017). *Modern statistics for the social and behavioral sciences: A practical introduction* (Second). CRC Press. doi: 10.1201/9781315154480.

## Citation

Bittmann, F. (2025). Same or different? Comparing the coverage rate of five different approaches for testing the difference of two groups means. *The Quantitative Methods for Psychology*, *21*(1), 1–12. doi: 10.20982/tqmp.21.1.p001.

Copyright © 2025, *Bittmann*. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 19/07/2024 ~ Accepted: 07/10/2024

Tables 4 to 7 follow.

**Table 4** ■ Coverage probabilities by approach / simulation 1 (normal)

| | T-Test | | Welch | | OLS | | BSnormal | | BSBC | |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall | 93.92 | | 94.98 | | 95.02 | | 94.59 | | 94.55 | |
| $N = 75000$ | 93.74 | 94.09 | 94.82 | 95.14 | 94.86 | 95.18 | 94.43 | 94.75 | 94.38 | 94.71 |
| Equal SD | 94.64 | | 94.88 | | 94.93 | | 94.43 | | 94.41 | |
| $N = 24153$ | 94.35 | 94.92 | 94.59 | 95.15 | 94.64 | 95.20 | 94.13 | 94.71 | 94.11 | 94.70 |
| Unequal SD | 93.57 | | 95.03 | | 95.07 | | 94.67 | | 94.61 | |
| $N = 50847$ | 93.35 | 93.78 | 94.84 | 95.22 | 94.87 | 95.25 | 94.47 | 94.86 | 94.41 | 94.81 |
| Minsize = Small | 92.40 | | 95.12 | | 95.05 | | 94.28 | | 94.13 | |
| $N = 12497$ | 91.92 | 92.86 | 94.73 | 95.49 | 94.65 | 95.42 | 93.86 | 94.68 | 93.71 | 94.54 |
| Minsize = Medium | 93.77 | | 95.03 | | 95.08 | | 94.57 | | 94.55 | |
| $N = 28617$ | 93.49 | 94.05 | 94.77 | 95.28 | 94.82 | 95.33 | 94.30 | 94.83 | 94.28 | 94.81 |
| Minsize = Large | 94.60 | | 94.89 | | 94.96 | | 94.73 | | 94.70 | |
| $N = 33886$ | 94.35 | 94.83 | 94.65 | 95.13 | 94.72 | 95.19 | 94.48 | 94.96 | 94.46 | 94.94 |
| Ratio N = Small | 94.82 | | 94.97 | | 95.07 | | 94.73 | | 94.68 | |
| $N = 29092$ | 94.56 | 95.07 | 94.71 | 95.22 | 94.81 | 95.31 | 94.46 | 94.98 | 94.41 | 94.93 |
| Ratio N = Medium | 94.14 | | 94.92 | | 95.00 | | 94.60 | | 94.56 | |
| $N = 22552$ | 93.82 | 94.44 | 94.63 | 95.21 | 94.71 | 95.28 | 94.30 | 94.89 | 94.26 | 94.85 |
| Ratio N = Large | 92.58 | | 95.06 | | 94.99 | | 94.41 | | 94.37 | |
| $N = 23356$ | 92.23 | 92.91 | 94.77 | 95.33 | 94.70 | 95.26 | 94.11 | 94.70 | 94.07 | 94.67 |
| Ratio SD = Small | 94.62 | | 94.93 | | 95.01 | | 94.60 | | 94.56 | |
| $N = 27447$ | 94.34 | 94.88 | 94.67 | 95.19 | 94.74 | 95.26 | 94.32 | 94.86 | 94.29 | 94.83 |
| Ratio SD = Medium | 93.81 | | 94.92 | | 94.93 | | 94.48 | | 94.44 | |
| $N = 34797$ | 93.55 | 94.06 | 94.68 | 95.14 | 94.69 | 95.16 | 94.23 | 94.72 | 94.19 | 94.67 |
| Ratio SD = Large | 92.71 | | 95.27 | | 95.30 | | 94.89 | | 94.82 | |
| $N = 12756$ | 92.24 | 93.15 | 94.89 | 95.63 | 94.91 | 95.66 | 94.49 | 95.26 | 94.42 | 95.20 |
| Ratio N + SD = SS | 94.87 | | 94.86 | | 95.01 | | 94.70 | | 94.70 | |
| $N = 10819$ | 94.44 | 95.28 | 94.43 | 95.27 | 94.58 | 95.41 | 94.26 | 95.12 | 94.26 | 95.12 |
| Ratio N + SD = SM | 94.80 | | 95.04 | | 95.12 | | 94.77 | | 94.69 | |
| $N = 13435$ | 94.41 | 95.17 | 94.65 | 95.40 | 94.74 | 95.48 | 94.39 | 95.15 | 94.30 | 95.07 |
| Ratio N + SD = SL | 94.77 | | 95.02 | | 95.06 | | 94.65 | | 94.58 | |
| $N = 4838$ | 94.11 | 95.38 | 94.37 | 95.61 | 94.41 | 95.65 | 93.97 | 95.26 | 93.91 | 95.21 |
| Ratio N + SD = MS | 94.37 | | 94.69 | | 94.77 | | 94.35 | | 94.27 | |
| $N = 8248$ | 93.86 | 94.86 | 94.18 | 95.16 | 94.27 | 95.24 | 93.83 | 94.84 | 93.74 | 94.76 |
| Ratio N + SD = MM | 94.17 | | 94.84 | | 94.92 | | 94.54 | | 94.50 | |
| $N = 10426$ | 93.70 | 94.61 | 94.40 | 95.26 | 94.48 | 95.33 | 94.09 | 94.97 | 94.05 | 94.93 |
| Ratio N + SD = ML | 93.55 | | 95.64 | | 95.69 | | 95.31 | | 95.33 | |
| $N = 3878$ | 92.73 | 94.31 | 94.95 | 96.26 | 95.01 | 96.31 | 94.59 | 95.95 | 94.62 | 95.98 |
| Ratio N + SD = LS | 94.52 | | 95.26 | | 95.24 | | 94.70 | | 94.68 | |
| $N = 8380$ | 94.01 | 95.00 | 94.79 | 95.71 | 94.76 | 95.68 | 94.20 | 95.17 | 94.18 | 95.15 |
| Ratio N + SD = LM | 92.25 | | 94.84 | | 94.71 | | 94.06 | | 94.06 | |
| $N = 10936$ | 91.73 | 92.74 | 94.41 | 95.25 | 94.28 | 95.13 | 93.60 | 94.49 | 93.60 | 94.49 |
| Ratio N + SD = LL | 89.43 | | 95.22 | | 95.20 | | 94.78 | | 94.60 | |
| $N = 4040$ | 88.44 | 90.36 | 94.52 | 95.86 | 94.49 | 95.84 | 94.05 | 95.44 | 93.86 | 95.28 |

*Note*. Source: own computations. The first line indicates the point estimate, the second line the 95% confidence interval. The results closest to the nominal coverage of 95% are highlighted in green. S: Small, M: Medium, L: Large. OLS: OLS regression with robust standard errors.

**Table 5** ■ Coverage probabilities by approach / simulation 2 (uniform)

| | T-Test | | Welch | | OLS | | BSnormal | | BSBC | |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall | 93.95 | | 94.88 | | 94.93 | | 94.49 | | 94.73 | |
| $N = 75000$ | 93.78 | 94.12 | 94.72 | 95.04 | 94.77 | 95.08 | 94.33 | 94.66 | 94.56 | 94.89 |
| Equal SD | 94.69 | | 94.69 | | 94.75 | | 94.32 | | 94.59 | |
| $N = 24601$ | 94.40 | 94.97 | 94.40 | 94.97 | 94.47 | 95.03 | 94.02 | 94.60 | 94.30 | 94.87 |
| Unequal SD | 93.59 | | 94.98 | | 95.01 | | 94.58 | | 94.80 | |
| $N = 50399$ | 93.37 | 93.80 | 94.78 | 95.17 | 94.82 | 95.20 | 94.38 | 94.78 | 94.60 | 94.99 |
| Minsize = Small | 92.65 | | 94.73 | | 94.77 | | 93.95 | | 94.56 | |
| $N = 12191$ | 92.17 | 93.11 | 94.31 | 95.12 | 94.36 | 95.16 | 93.52 | 94.37 | 94.14 | 94.96 |
| Minsize = Medium | 93.77 | | 94.99 | | 95.04 | | 94.58 | | 94.83 | |
| $N = 28567$ | 93.48 | 94.05 | 94.73 | 95.24 | 94.78 | 95.29 | 94.32 | 94.84 | 94.57 | 95.09 |
| Minsize = Large | 94.56 | | 94.85 | | 94.89 | | 94.61 | | 94.70 | |
| $N = 34242$ | 94.32 | 94.80 | 94.61 | 95.08 | 94.65 | 95.12 | 94.37 | 94.85 | 94.45 | 94.93 |
| Ratio N = Small | 94.83 | | 95.01 | | 95.09 | | 94.73 | | 94.86 | |
| $N = 29350$ | 94.57 | 95.08 | 94.76 | 95.26 | 94.84 | 95.34 | 94.47 | 94.99 | 94.60 | 95.11 |
| Ratio N = Medium | 94.12 | | 94.70 | | 94.76 | | 94.36 | | 94.50 | |
| $N = 22519$ | 93.81 | 94.43 | 94.40 | 94.99 | 94.46 | 95.04 | 94.05 | 94.65 | 94.19 | 94.79 |
| Ratio N = Large | 92.66 | | 94.89 | | 94.88 | | 94.33 | | 94.79 | |
| $N = 23131$ | 92.32 | 92.99 | 94.60 | 95.17 | 94.59 | 95.16 | 94.02 | 94.62 | 94.49 | 95.07 |
| Ratio SD = Small | 94.60 | | 94.68 | | 94.72 | | 94.34 | | 94.56 | |
| $N = 27980$ | 94.33 | 94.86 | 94.41 | 94.94 | 94.46 | 94.98 | 94.06 | 94.60 | 94.28 | 94.82 |
| Ratio SD = Medium | 93.88 | | 94.90 | | 94.97 | | 94.50 | | 94.75 | |
| $N = 34443$ | 93.62 | 94.13 | 94.67 | 95.13 | 94.73 | 95.20 | 94.26 | 94.74 | 94.51 | 94.98 |
| Ratio SD = Large | 92.69 | | 95.28 | | 95.26 | | 94.82 | | 95.04 | |
| $N = 12577$ | 92.22 | 93.14 | 94.89 | 95.64 | 94.88 | 95.63 | 94.42 | 95.20 | 94.64 | 95.41 |
| Ratio N + SD = SS | 94.81 | | 94.80 | | 94.90 | | 94.58 | | 94.67 | |
| $N = 10993$ | 94.37 | 95.21 | 94.36 | 95.20 | 94.47 | 95.30 | 94.14 | 94.99 | 94.23 | 95.08 |
| Ratio N + SD = SM | 94.76 | | 94.96 | | 95.05 | | 94.65 | | 94.81 | |
| $N = 13551$ | 94.37 | 95.13 | 94.58 | 95.32 | 94.67 | 95.41 | 94.26 | 95.02 | 94.43 | 95.18 |
| Ratio N + SD = SL | 95.09 | | 95.65 | | 95.67 | | 95.32 | | 95.40 | |
| $N = 4806$ | 94.44 | 95.68 | 95.04 | 96.21 | 95.06 | 96.23 | 94.68 | 95.90 | 94.77 | 95.98 |
| Ratio N + SD = MS | 94.49 | | 94.45 | | 94.55 | | 94.14 | | 94.29 | |
| $N = 8402$ | 93.98 | 94.97 | 93.94 | 94.93 | 94.04 | 95.02 | 93.62 | 94.64 | 93.77 | 94.77 |
| Ratio N + SD = MM | 94.21 | | 94.99 | | 95.04 | | 94.63 | | 94.74 | |
| $N = 10313$ | 93.74 | 94.65 | 94.55 | 95.40 | 94.60 | 95.45 | 94.18 | 95.06 | 94.30 | 95.17 |
| Ratio N + SD = ML | 93.09 | | 94.48 | | 94.45 | | 94.09 | | 94.30 | |
| $N = 3804$ | 92.23 | 93.87 | 93.71 | 95.18 | 93.68 | 95.16 | 93.29 | 94.81 | 93.51 | 95.01 |
| Ratio N + SD = LS | 94.44 | | 94.75 | | 94.68 | | 94.21 | | 94.68 | |
| $N = 8585$ | 93.94 | 94.92 | 94.25 | 95.21 | 94.18 | 95.14 | 93.70 | 94.70 | 94.18 | 95.14 |
| Ratio N + SD = LM | 92.43 | | 94.75 | | 94.80 | | 94.20 | | 94.68 | |
| $N = 10579$ | 91.91 | 92.93 | 94.31 | 95.17 | 94.36 | 95.22 | 93.73 | 94.63 | 94.23 | 95.10 |
| Ratio N + SD = LL | 89.41 | | 95.59 | | 95.54 | | 94.93 | | 95.31 | |
| $N = 3967$ | 88.41 | 90.35 | 94.90 | 96.21 | 94.85 | 96.16 | 94.20 | 95.59 | 94.61 | 95.95 |

*Note*. Source: own computations. The first line indicates the point estimate, the second line the 95% confidence interval. The result closest to the nominal coverage of 95% is marked in green. S: Small, M: Medium, L: Large. OLS: OLS regression with robust standard errors.

**Table 6** ■ Summary statistics of simulation 3 specifications (beta)

| | Mean | Median | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| $Mean_1$ | 0.28 | 0.28 | 0.066 | 0.10 | 0.52 | 0.11 | 2.40 |
| $Mean_2$ | 0.28 | 0.28 | 0.066 | 0.077 | 0.55 | 0.12 | 2.40 |
| $SD_1$ | 0.14 | 0.14 | 0.017 | 0.033 | 0.24 | 0.063 | 3.17 |
| $SD_2$ | 0.14 | 0.14 | 0.017 | 0.053 | 0.25 | 0.078 | 3.17 |
| $N_1$ | 107.5 | 108 | 53.7 | 15 | 200 | 0.0022 | 1.80 |
| $N_2$ | 107.5 | 108 | 53.8 | 15 | 200 | -0.0017 | 1.80 |
| $Skewness1$ | 0.53 | 0.51 | 0.31 | -0.90 | 2.99 | 0.35 | 4.10 |
| $Skewness2$ | 0.53 | 0.51 | 0.31 | -1.29 | 2.48 | 0.32 | 3.89 |
| $Kurtosis1$ | 2.93 | 2.79 | 0.69 | 1.30 | 13.2 | 2.15 | 13.6 |
| $Kurtosis2$ | 2.93 | 2.79 | 0.69 | 1.32 | 13.4 | 1.97 | 11.5 |
| Ratio $SD_1/SD_2$ | 1.10 | 1.08 | 0.096 | 1.00 | 2.57 | 2.31 | 13.4 |
| Ratio $N_1/N_2$ | 2.47 | 1.77 | 1.84 | 1 | 13.3 | 2.25 | 8.76 |
| Unequal $SD$ | 0.053 | 0 | | 0 | 1 | | |
| Abs. diff. skewness | 0.26 | 0.21 | 0.22 | 0.0000054 | 2.79 | 1.53 | 6.67 |
| Abs. diff. kurtosis | 0.63 | 0.46 | 0.62 | 0.0000091 | 10.8 | 2.68 | 16.7 |
| Simulations | | | | 75000 | | | |

*Note.* Source: own computations. Mean: Empirical arithmetic mean; SD: Empirical standard deviation; N: Number of observations.

**Table 7** ■ Coverage probabilities by approach / simulation 3 (beta)

| | T-Test | | Welch | | OLS | | BSnormal | | BSBC | |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall | 94.82 | | 94.71 | | 94.76 | | 94.37 | | 94.43 | |
| $N = 75000$ | 94.66 | 94.98 | 94.55 | 94.87 | 94.60 | 94.92 | 94.21 | 94.54 | 94.26 | 94.59 |
| Minsize = Small | 94.87 | | 94.49 | | 94.41 | | 93.66 | | 93.84 | |
| $N = 12331$ | 94.47 | 95.26 | 94.07 | 94.88 | 93.99 | 94.81 | 93.21 | 94.08 | 93.41 | 94.26 |
| Minsize = Medium | 94.78 | | 94.70 | | 94.78 | | 94.33 | | 94.41 | |
| $N = 28886$ | 94.52 | 95.04 | 94.44 | 94.96 | 94.51 | 95.03 | 94.06 | 94.60 | 94.14 | 94.67 |
| Minsize = Large | 94.83 | | 94.81 | | 94.88 | | 94.67 | | 94.65 | |
| $N = 33783$ | 94.59 | 95.06 | 94.56 | 95.04 | 94.64 | 95.11 | 94.43 | 94.91 | 94.41 | 94.89 |
| Abs. Diff. Skew = Small | 95.56 | | 95.50 | | 95.54 | | 95.18 | | 95.15 | |
| $N = 19364$ | 95.26 | 95.85 | 95.20 | 95.78 | 95.24 | 95.83 | 94.87 | 95.48 | 94.84 | 95.45 |
| Abs. Diff. Skew = Med | 95.06 | | 94.97 | | 95.01 | | 94.69 | | 94.70 | |
| $N = 36276$ | 94.83 | 95.28 | 94.74 | 95.19 | 94.78 | 95.23 | 94.46 | 94.92 | 94.47 | 94.93 |
| Abs. Diff. Skew = Large | 93.62 | | 93.45 | | 93.53 | | 92.97 | | 93.18 | |
| $N = 19360$ | 93.27 | 93.96 | 93.09 | 93.79 | 93.17 | 93.87 | 92.60 | 93.33 | 92.82 | 93.53 |
| Abs. Diff. Kurt = Small | 95.38 | | 95.13 | | 95.17 | | 94.87 | | 94.83 | |
| $N = 18001$ | 95.07 | 95.69 | 94.80 | 95.44 | 94.84 | 95.48 | 94.54 | 95.19 | 94.50 | 95.15 |
| Abs. Diff. Kurt = Med | 94.74 | | 94.67 | | 94.72 | | 94.32 | | 94.38 | |
| $N = 38407$ | 94.51 | 94.96 | 94.44 | 94.89 | 94.49 | 94.94 | 94.09 | 94.55 | 94.15 | 94.61 |
| Abs. Diff. Kurt = Large | 94.43 | | 94.40 | | 94.46 | | 94.00 | | 94.13 | |
| $N = 18592$ | 94.09 | 94.76 | 94.06 | 94.73 | 94.12 | 94.78 | 93.65 | 94.34 | 93.78 | 94.46 |

*Note.* Source: own computations. The first line indicates the point estimate, the second line the 95% confidence interval. The result closest to the nominal coverage of 95% is marked in green. S: Small, M: Medium, L: Large. OLS: OLS regression with robust standard errors.