ASSESSMENT OF ASSUMPTION VIOLATION OF TWO-SAMPLE T-TEST OF MEANS

František Pavlík – Ondřej Vozár

Abstract

The thesis examines how violations of normality and variance homogeneity affect tests on means, focusing on the one-sample and two-sample t-tests, Welch’s t-test, and permutation tests. After outlining theoretical assumptions, a Monte Carlo simulation study compares empirical Type I error rates and power across several data-generating processes (normal, truncated normal, skew-normal, and heavy-tailed) and a range of sample sizes and variance ratios. Results indicate that classical t-tests generally control Type I error for larger samples (n > 50) but show inflated error rates and loss of power in small samples drawn from skewed or heavy-tailed distributions. Welch’s test reliably handles pronounced variance heterogeneity and retains competitive power in many settings. The permutation test maintains nominal size under strong departures from normality and performs well for small samples, though it can be affected by variance imbalance. Usage of Welch’s test is generally recommended, when sample sizes are reasonably large, since it addresses unequal variances while preserving good power. For small samples with strong non-normality, consider permutation methods with caution due to sensitivity to heteroskedasticity. In all cases, conduct diagnostic checks of distributional shape and variance homogeneity before choosing a test.

Key words: assumption violation, permutation test, power of a test, tests of the mean, two-sample t-test, Welch’s test

JEL Code: C12

Introduction

Comparing the means of two independent populations is one of the most common inferential tasks in applied statistics. The classical Student’s two-sample t-test (pooled-variance form) is optimal under the assumptions of normally distributed populations with equal variances. However, real-world data often deviate from normality—exhibiting skewness, heavy tails, or truncation—and group variances may differ. Violations of these assumptions can lead to incorrect Type I error rates or loss of power, thus misleading scientific conclusions. Alternative procedures include Welch’s t-test, which adjusts degrees of freedom to accommodate unequal variances, and permutation (randomization) tests, which rely only on exchangeability under the null hypothesis and make no distributional assumptions. Yet practitioners lack clear quantitative guidance on when each method remains valid or loses accuracy. This paper focuses exclusively on tests comparing means from two independent samples. In this paper we summarize theoretical derivations and assumptions of the pooled two-sample t-test, Welch’s test, and the permutation test; design and run a Monte Carlo simulation study to evaluate empirical Type I error rates and power under a variety of nonideal conditions; and offer practical recommendations based on the results.

1 Test on means from two independent samples

When comparing the means of two independent samples, several statistical methods are available, each with specific assumptions and properties. The most widely used approaches are the two-sample t-test, Welch's t-test, and the permutation test. This section provides an overview of these methods, their assumptions, and their application in hypothesis testing.

1.1 Two-sample t-test and its assumptions

The two-sample t-test is a classical method for comparing the means of two independent groups. It assumes that both samples are drawn from normal distributions with equal variances and that the observations are independent. The test evaluates whether the difference in sample means is statistically significant, typically using a t-distribution with $m+n$-2 degrees of freedom. If the test statistic exceeds the critical value, the null hypothesis of equal means is rejected (Anděl, 2007; Malá, 2024).

1.2 Welch test

When the assumption of equal variances is not met, Welch's t-test offers a robust alternative. This test allows for unequal variances between groups and adjusts the degrees of freedom accordingly. Like the standard t-test, it assumes normality and independence, but is less sensitive to heterogeneity of variances (Welch, 1947; Anděl, 2007).

1.3 Permutation test

If the normality assumption is questionable, or a non-parametric approach is preferred, the permutation test provides a flexible alternative. This test compares the observed difference in means to the distribution of differences obtained by randomly reallocating the data between groups (Welch, 1990). The permutation test requires only independence of observations and does not assume normality or equal variances, making it especially useful for small samples or unknown distributions (Edington & Onghena, 2007; Collingridge, 2013).

Tab. 1: Assumptions of statistical tests

Test	Normality	Equal Variances	Independence
Two-sample t-test	Yes	Yes	Yes
Welch's test	Yes	No	Yes
Permutation test	No	No	Yes

Source: Author

1.4 Distributions used to study assumption on normal distribution

To comprehensively evaluate the performance of statistical tests under various conditions, the simulation study employs several probability distributions that represent both standard and non-standard data scenarios. The following distributions are considered: the normal distribution, the truncated normal distribution, Student's *t*-distribution, and the skew normal distribution.

Normal distribution serves as the baseline for most simulations. It is widely used in statistics due to its theoretical properties and frequent occurrence in analyzed data. In the simulations, samples were generated from the standard normal distribution to represent ideal conditions where the assumptions of parametric tests are fully met.

To explore the impact of deviations from normality, the study also includes the truncated normal distribution. This distribution is created by restricting a normal variable to a finite interval, effectively removing extreme values (outliers). In the simulations, truncation was applied to the standard normal distribution, with the interval chosen to reflect practical data limitations. This allows for the evaluation of test performance when the data are more concentrated and lack heavy tails.

Student's *t*-distribution was used to generate data with heavier tails than the normal distribution, which is common in real datasets with more extreme values. In the simulation study, samples from the *t*-distribution with various degrees of freedom were used to assess how well statistical tests maintain their reliability when the data are prone to outliers or have higher variability. Additionally, the *t*-distribution is central to the construction of the *t*-test itself, as the test statistic under the null hypothesis follows this distribution when the population variance is unknown.

To investigate the effect of asymmetry, the skew normal distribution was included. This distribution introduces a skewness parameter to the normal distribution, allowing the simulation of data that are not symmetric. By varying the skewness parameter, the study examines how increasing levels of skewness influence the accuracy and reliability of statistical tests.

By generating samples from these distributions, the simulation study systematically evaluates the sensitivity of statistical methods to violations of their underlying assumptions. This approach provides practical insights into the conditions under which standard tests remain valid and when alternative methods may be necessary.

2 Results of the simulation study and discussion

2.1 Design of the simulation study

The simulation study was designed to systematically evaluate the reliability of various statistical tests under different conditions of data normality and sample size. In all scenarios, the null hypothesis (H_0) tested was the equality of means between generated samples. The primary objective was to assess the robustness of the one-sample t-test, paired t-test, two-sample t-test, Welch's test, and the permutation test when applied to samples drawn from both normal and non-normal distributions.

To achieve this, several types of distributions were considered: mildly, moderately, and strongly skewed distributions, distributions with heavy tails, truncated normal distributions, and standard normal distribution. For each distribution, the theoretical mean was determined, ensuring that the null hypothesis held true in all simulated scenarios. In the case of skewed and truncated distributions, the theoretical mean was derived analytically or referenced from relevant literature.

The simulation procedure involved generating random samples of varying sizes (5, 10, 25, 50, and 100 observations) from each distribution. For each combination of distribution type and sample size, 1000 independent replications were performed. The relevant statistical test was then applied to each sample, with the significance level (α) set at 0.05, 0.01, and 0.10 in separate experiments. For the permutation test, 1000 permutations were used to compute the p-value.

The main outcome measure was the proportion of confidence intervals (or p-values) that correctly included the theoretical mean, reflecting the empirical Type I error rate of each test. Results were summarized in tabular form, displaying the percentage of intervals containing the true mean for each scenario.

This design allows for a comprehensive comparison of test reliability across a range of realistic data conditions, highlighting the sensitivity of classical and nonparametric tests to violations of normality and the presence of outliers. The findings provide practical guidance for the selection of appropriate statistical methods in empirical research where the assumption of normality may not hold.

In addition to evaluating the empirical Type I error rate, the simulation study also investigated the power of the tests, that is, the probability of correctly rejecting the null hypothesis when it is false. For this purpose, the simulations were repeated under scenarios where the true mean (or difference in means) differed from the value specified in the null hypothesis by a parameter δ. By systematically varying δ across a range of values, the power function of each test was estimated for different sample sizes and distribution types. This approach allowed for a detailed assessment of how the ability to detect true effects depends on the underlying distribution, sample size, and the magnitude of the effect. The results provide insight into the practical sensitivity of each test and inform recommendations for their use in empirical research.

2.2 Effects of assumptions violation on the size of the tests

The simulation study revealed several important findings regarding the impact of violating key assumptions of statistical tests, particularly normality and homogeneity of variances. For one-sample tests, the results demonstrated that the classical t-test maintains reasonable robustness to violations of normality when sample sizes exceed 50 observations. The empirical Type I error rates remained close to the nominal significance levels $(\alpha = 0.01, 0.05, 0.10)$ under these conditions. However, for smaller sample sizes $(n < 30)$, departures from normality led to notable distortions in test size, particularly when data exhibited substantial skewness or heavy tails.

For two-sample comparisons, the interplay between unequal variances and sample sizes emerged as a critical factor. When sample sizes were equal, both Student's t-test and Welch's test maintained adequate control of Type I error rates, even under moderate violations of the homoscedasticity assumption. However, under unequal sample sizes combined with unequal variances, the classical Student's t-test showed significant size distortions, with empirical error rates exceeding the nominal level by up to 8 percentage points in extreme cases.

The permutation test demonstrated stability in maintaining the nominal significance level across various distributional shapes and variance structures. Even in small samples *(n* = 10), the empirical Type I error rates remained within 2 percentage points of the nominal level, regardless of the underlying distribution. This robustness, however, comes at the cost of increased computational complexity, particularly for larger sample sizes.

Tab. 2: Verification of the reliability of tests at a 5% significance level

	Severely skewed			Heavy tails			Truncated		
Sample size	Two	Welch.	Perm.	Two	Welch.	Perm.	Two	Welch.	Perm.
5	94.1	96.7	95.8	**97.9**	**99.2**	95.6	95.0	95.0	95.0
10	95.0	95.6	94.1	**97.5**	**97.8**	95.7	94.9	95.6	95.1
25	94.6	94.9	94.2	**96.7**	**96.2**	95.8	94.6	94.2	96.4
50	95.9	95.8	96.0	96.6	95.7	94.9	95.1	94.8	94.8
100	95.8	94.8	95.0	94.6	95.2	94.4	95.2	95.6	95.4

Source: Author

Tab. 3: Verification of the reliability of tests at a 5% significance level

Sample size		**Ratio between variances**														
		1 / 4			**1 / 2**			**1**			**2**			**4**		
n1	n2	Two	Welch	Perm.	Two	Welch	Perm.	Two	Welch	Perm.	Two	Welch	Perm.	Two	Welch	Perm.
10	10	94.0	95.0	93.2	93.9	94.2	95.1	95.9	95.7	95.3	95.6	95.8	93.7	94.5	95.0	95.0
10	20	**83.5**	94.7	**84.0**	**88.7**	96.2	**90.0**	95.3	95.2	94.4	**98.2**	94.1	**98.2**	**99.4**	95.1	**99.0**
10	40	**72.7**	95.7	**70.5**	**83.7**	94.5	**82.6**	94.5	94.6	94.6	**99.7**	95.4	**99.6**	**99.9**	94.7	**100.0**
100	100	96.0	95.4	94.9	95.0	94.6	95.2	96.1	95.8	95.4	95.4	95.5	93.9	94.0	95.5	94.6
100	200	**85.3**	95.1	**84.2**	**87.7**	94.1	**90.1**	93.8	95.5	95.5	**99.1**	96.3	**98.7**	**99.2**	95.0	**99.5**
100	400	**71.6**	94.8	**73.9**	**82.7**	95.3	**82.4**	95.1	94.9	94.5	**99.7**	94.5	**99.3**	**100.0**	94.5	**99.9**

Source: Author

2.3 Effects of assumptions violation on the power of the tests

The investigation of test power revealed complex interactions between sample size, effect size, and violations of assumptions. Results presented in this section represent only a part of a larger simulation study conducted for the bachelor thesis. Specifically, these results correspond to the most extreme scenarios, chosen to illustrate the general behavior of the statistical tests. Under normality and homoscedasticity, all three tests (Student's, Welch's, and permutation) showed similar power characteristics, with power increasing predictably with both sample size and effect size. However, notable differences emerged under various violations of assumptions.

The results on the power of tests are presented for the most extreme cases observed in the simulation study. The heavy-tailed distribution is chosen because it best represents the effects of violations of assumptions, and these effects are most prominent in this scenario. Specifically, for the heavy-tailed distribution, results are shown for sample size $n = 5$, and for the scenario with unequal variances, results are given for $n = 10$ and $n = 40$ with variance ratios of 1/4.

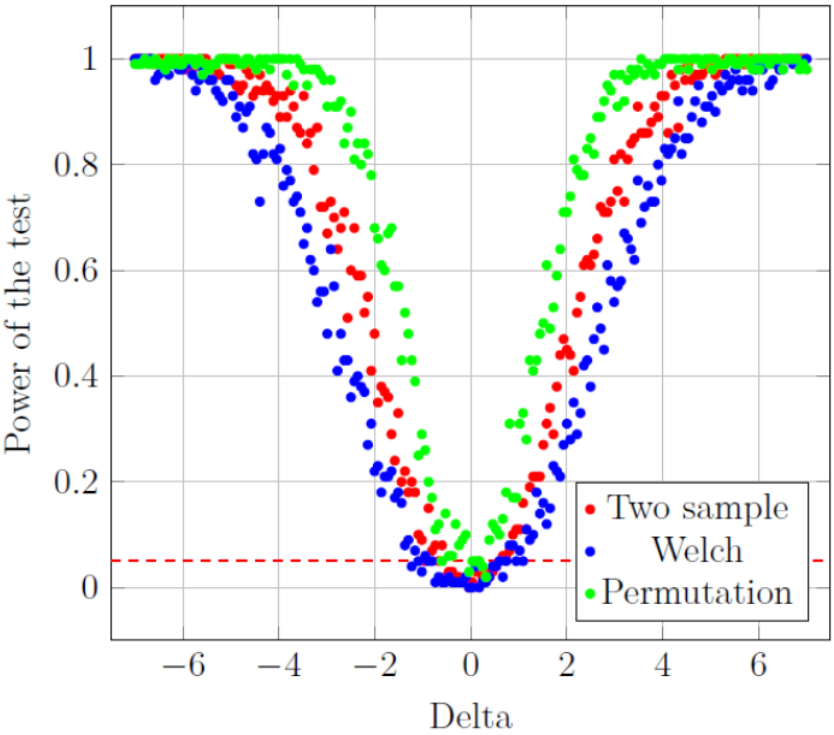
For skewed distributions, the power of the classical t-test was substantially reduced in small samples ($n < 30$), particularly for detecting small to moderate effects. The Welch's test maintained better power characteristics under these conditions, showing only modest reductions in power compared to the ideal normal case. The permutation test demonstrated particularly strong performance in detecting differences under skewed conditions, often achieving higher power than both parametric alternatives.

Under heavy-tailed distributions, all tests showed reduced power compared to the normal case, but the magnitude of this reduction varied considerably. The classical t-test suffered the largest power loss, while the permutation test showed the smallest reduction in power, particularly for moderate to large effect sizes. This pattern was consistent across different sample sizes, though the differences became less pronounced with larger samples ($n > 50$).

The presence of unequal variances had a particularly notable impact on power when combined with unequal sample sizes. When the larger sample was paired with the larger variance, both parametric tests showed reduced power compared to the homoscedastic case. However, when the larger sample was paired with the smaller variance, the Welch's test maintained better power characteristics than the Student's t-test, especially for small to moderate effect sizes.

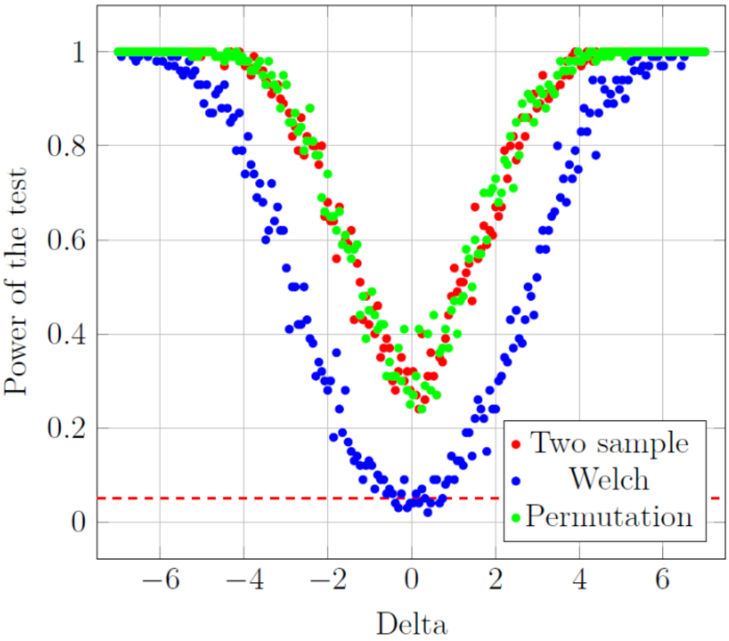
The simulation results suggest that for practical applications, researchers should carefully consider sample size and potential violations of assumptions when selecting a testing procedure. For small samples ($n < 30$) or when normality is questionable, the permutation test offers a robust alternative that maintains both size and power. For larger samples, Welch's test provides a good compromise between computational simplicity and robustness to violations of assumptions, particularly when variances may be unequal. The classical Student's t-test should be used with caution when sample sizes are small or assumptions are violated, as it may lead to both inflated Type I error rates and reduced power to detect true effects.

Fig. 1: Power of tests for samples (n = 5) originating from heavy-tailed distributions and a variance ratio of 1, at 5% significance level



Source: Author

Fig. 2: Power of tests for sample sizes 10 and 40 and a variance ratio of ¼, at 5% significance level



Source: Author

Conclusion

This study systematically investigated the robustness of classical mean comparison tests. One-sample and two-sample Student’s t-tests, Welch’s test, and the permutation test, under violations of normality and homogeneity of variance assumptions. Through extensive Monte Carlo simulations, the analysis quantified the sensitivity of these methods to various degrees of distributional asymmetry, heavy tails, and unequal variances, providing valuable practical insights.

The results demonstrate that for sufficiently large samples (n > 50), traditional t-tests generally maintain acceptable Type I error rates even in the presence of moderate non-normality. However, in small samples drawn from highly skewed or heavy-tailed distributions, the one-sample t-test exhibits a notable inflation of error rates, while the paired and two-sample t-tests, as well as Welch’s test, display greater robustness due to their reliance on difference statistics. Nevertheless, when dealing with heavy-tailed distributions and limited sample sizes, none of the classical tests consistently maintain nominal error rates, and only the permutation test provides reliable performance across these challenging scenarios.

In cases of pronounced variance heterogeneity, the classical two-sample t-test and the permutation test fail to adequately control Type I error rates, whereas Welch’s test consistently demonstrates robustness and superior power. Based on these findings, Welch’s test is recommended as the most generally reliable procedure, particularly when sample sizes exceed 100 observations. For small samples with substantial departures from normality, the permutation test remains a viable alternative, provided that its computational demands and sensitivity to variance differences are properly accounted for.

Overall, the study emphasizes the importance of conducting diagnostic checks and considering potential assumption violations when selecting appropriate statistical tests. Careful test selection is essential to ensure the validity of inferential conclusions in empirical research.

Acknowledgment

The work of the third author was prepared under Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague.

References

Anděl, J. (2007). *Statistické metody* (4., upr. vyd.). Matfyzpress.

Azzalini, A. (1985). A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics, 12*, 171–178.

Azzalini, A. (2023). *The R package : The skew-normal and related distributions such as the skew-t and the SUN (version 2.1.1).* Università degli Studi di Padova, Italia. Načteno z https://cran.r-project.org/package=sn

Collingridge, D.S. (2013). A Primer on Quantitized Data Analysis and Permutation Testing. *Journal of Mixed Methods Research*. **7** (1): 79–95. doi:10.1177/1558689812454457

Crump, M. J. (2024). The randomization test (permutation test). *LibreTexts*. https://stats.libretexts.org/@go/page/7917

Edgington, E., & Onghena, P. (2007). *Randomization Tests* (4th Edition. vyd.). Chapman and Hall/CRC.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1* (2. vyd., Sv. 1). New York: Wiley.

Malá, I. (2024). *Statistické úsudky* (Druhé vydání. vyd.). Professional Publishing.

Marek, L. (2024). *Pravděpodobnost* (Druhé vydání ed.). Praha: Professional Publishing.

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2023). *truncnorm: Truncated Normal Distribution.* Načteno z https://CRAN.R-project.org/package=truncnorm

Pavlík, F. (2025). Porušení předpokladů testů o střední hodnotě z normálního rozdělení. [Bachelor thesis], Prague university of economics and business

R Core Team. (2024). *R: A Language and Environment for Statistical Computing.* Vienna. https://www.R-project.org/

Ramsey, F. L., & Schafer, D. W. (2009). *The statistical sleuth* (2. ed.. vyd.). Belmont, Calif.: Brooks/Cole.

Welch, B. L. (January 1947). The generalization of 'Student's' problem when several different population varlances are involved. *Biometrika, 34*, 28-35.

Welch, W. J. (1990). Construction of permutation tests. *Journal of American Statistical Association, 85 (411)*, 693-698. doi:10.1080/01621459.1990.10474929

Contact

Bc. František Pavlík
Prague University of Economics and Business
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
pavf05@vse.cz

Ing. Ondřej Vozár, Ph.D.
Department of Statistics and Probability, Prague University of Economics and Business
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
vozo01@vse.cz