Revista Colombiana de Estadística - Applied Statistics
January 2026, volume 49, issue 1, pp. 213–230
http://doi.org/10.15446/rce.v49n1.118494

The Adaptive Baumgartner-Type Test Statistics for Two-Sample Independent Problem

Estadísticas de prueba tipo baumgartner adaptativas para el problema de dos muestras independientes

Zaheer Aslam[1,a], Syed Wajahat Ali Bokhari[2,b], Nasir Ali[2,c], Abid Hussain[2,d]

[1]Department of Statistics, Govt. Graduate College Asghar Mall, Rawalpindi, Pakistan

[2]Department of Statistics, PMAS-Arid Agriculture University, Rawalpindi, Pakistan

Abstract

The two-sample independent problem remains a persistent challenge in statistical analysis. Parametric tests, such as Student's t-test and Welch's t-test, are commonly employed to assess the significance of differences between the means of two groups. However, these methods rely on the assumption of normally distributed populations. When this assumption is violated, nonparametric alternatives like the Wilcoxon-Mann-Whitney, Yuen-Welch, Brunner-Munzel, and Baumgartner tests offer robust solutions. This study introduces an adaptive framework for nonparametric two-sample tests, building upon the foundation of Baumgartner-type tests. To enhance statistical power, we incorporate a recently proposed relative rank transformation method that is more resilient to scale differences between the two samples. The adaptive tests are suitable for both location and scale comparisons. Through extensive Monte Carlo simulations, we evaluate the power performance of our adaptive tests under diverse distributional scenarios. Our results demonstrate that adaptive tests offer a substantial advantage over traditional nonparametric methods. To illustrate the practical application of our approaches, we apply the adaptive tests along their competitors to six real-world biomedical datasets. These examples highlight the reliability and effectiveness of the proposed methodology in addressing the two-sample independent location-scale testing problem.

***Keywords*:** Baumgartner-type statistics; Location-scale shift; Nonparametric tests; Relative ranking; Simulations.

[a] Master's degree. E-mail: zaheeraslam_stat@gpgcam.edu.pk
[b] Master's degree. E-mail: wajahatbokhari2@gmail.com
[c] Ph.D. E-mail: nasir_stat@uaar.edu.pk
[d] Ph.D. E-mail: abid0100@gmail.com

Resumen

El problema de dos muestras independientes sigue siendo un desafío persistente en el análisis estadístico. Las pruebas paramétricas, como la prueba t de Student y la prueba t de Welch, se emplean comúnmente para evaluar la significancia de las diferencias entre las medias de dos grupos. Sin embargo, estos métodos se basan en el supuesto de poblaciones distribuidas normalmente. Cuando este supuesto se viola, alternativas no paramétricas como las pruebas de Wilcoxon-Mann-Whitney, Yuen-Welch, Brunner-Munzel y Baumgartner ofrecen soluciones robustas. Este estudio introduce un marco adaptativo para pruebas no paramétricas de dos muestras, basado en pruebas tipo Baumgartner. Para mejorar la potencia estadística, incorporamos un método de transformación de rangos relativos recientemente propuesto que es más resistente a las diferencias de escala entre las dos muestras. Las pruebas adaptativas son adecuadas para comparaciones de ubicación y escala. A través de extensas simulaciones de Monte Carlo, evaluamos el rendimiento de potencia de nuestras pruebas adaptativas bajo diversos escenarios distribucionales. Nuestros resultados demuestran que las pruebas adaptativas ofrecen una ventaja sustancial sobre los métodos no paramétricos tradicionales. Para ilustrar la aplicación práctica de nuestros enfoques, aplicamos las pruebas adaptativas junto con sus competidores a seis conjuntos de datos biomédicos del mundo real. Estos ejemplos destacan la confiabilidad y efectividad de la metodología propuesta para abordar el problema de prueba de ubicación-escala de dos muestras independientes.

***Palabras clave*:** Cambio de ubicación-escala; Estadísticas tipo Baumgartner; Pruebas no paramétricas; Ranking relativo; Simulaciones.

1. Introduction

A frequent challenge in data analysis, particularly within fields like psychology, medicine, and environmental science, is the limited sample size. Conventional parametric tests, such as the t-test, rely on assumptions of normality and homogeneity of variance, which can be difficult or impossible to verify with small sample sizes, see for example Siegel & Castellan (1988). When these assumptions are violated, the reliability of parametric tests diminishes, potentially leading to biased estimates and inferences, especially in cases of skewed data, ordinal data, or the presence of outliers. Nonparametric rank-based tests provide a valuable alternative to parametric methods, as they are less affected by departures from the assumptions of normality and homogeneity of variance. These tests are well-suited for small, skewed, or ordinal-scaled data, allowing for comparisons without strict distributional constraints. This flexibility expands the applicability of statistical analysis to situations with limited sample sizes or data that is not easily amenable to transformation or resampling techniques, see for example Conover (1999).

For small sample sizes, nonparametric tests are often implemented as permutation tests. This approach involves generating all possible permutations of the data under the null hypothesis and calculating a test statistic for each permutation. By comparing the observed test statistic to its permutation distribution, we can conduct exact inference. The resulting p-value indicates the probability of observing a test statistic as extreme or more extreme than the one calculated, under the assumption that the null hypothesis is correct, see Manly (2018). The rationale behind this approach is to assess the deviation of the test statistic from its expected value under the null hypothesis, see for example Casella & Berger (2002). However, this method becomes computationally expensive for large datasets, necessitating alternative approaches for inference. For sample sizes of eight or more in each group, asymptotic distributions, such as the asymptotic normality of the Wilcoxon rank-sum test, can provide a suitable and accurate approximation. However, the reliability of these approximations is contingent on the quantity and distribution of ties within the dataset, as demonstrated by Brunner & Munzel (2000).

Kruskal-Wallis, Wilcoxon, and Mann-Whitney U tests are nonparametric rank-based tests that assess whether two or more groups differ in central tendency without assuming normality or homogeneity of variance. These methods are robust to outliers and deviations from normality, maximizing statistical power, see Hollander et al. (2013). Consequently, these tests are widely used in fields that require precise inference from small sample sizes. They are particularly valuable in situations where large samples are unattainable due to logistical, financial, or other constraints, such as clinical trials or the analysis of rare events (Mittelstadt & Floridi, 2016).

The Wilcoxon-Mann-Whitney (WMW) test is a widely-used nonparametric statistical test that has been extensively studied. However, its sensitivity to departures from the pure location shift model and variance homogeneity remains a subject of ongoing debate. As shown by Moran (2006), these deviations can lead to significant levels that diverge from the nominal significance level. While some researchers advocate for the use of the WMW test only when variance ratios are below specific thresholds, such as 1.5, see, for example, Fagerland & Sandvik (2009), others adopt a more lenient approach. These contrasting perspectives underscore the limitations of traditional nonparametric methods in addressing diverse data scenarios.

To address the limitations of conventional location-based tests, Baumgartner et al. (1998) proposed a novel approach focusing on mean differences. However, this method's performance in comparing means of distributions has not been extensively explored, its applicability to scenarios with unequal variances remains uncertain, and a comparative analysis with other rank-based methods, particularly under varying sample sizes and outlier conditions, is lacking. Furthermore, the utilization of this test in extended information spaces and multi-attribute rankings has been relatively unexplored, suggesting potential avenues for further development. Yuen (1974) introduced the concept of trimming to enhance robustness against outliers. Brunner & Munzel (2000) proposed modifications to the Wilcoxon-Mann-Whitney test, extending its applicability to data environments with tied observations and unequal variances. Building upon these ideas, Murakami (2006) developed a modified Baumgartner test specifically designed for multi-sample evaluation, enabling simultaneous testing for location and scale parameters.

Despite significant advancements in statistical analysis, challenges remain in applying various interventions to diverse datasets. For example, the WMW-test, a widely used nonparametric method, has limitations when dealing with distributions that have differing variances. Its sensitivity to unequal variances and potential bias in significance levels are ongoing areas of research. Similarly, the Baumgartner and modified Baumgartner tests, which incorporate decision-maker input, may not be universally effective across various sample sizes, in the presence of outliers, or under conditions of unequal variance. This study aims to address these limitations by reviewing and improving the usability and robustness of existing nonparametric methods. Specifically, we focus on enhancing the performance of these methods under challenging conditions, such as small sample sizes, skewed distributions, and unequal variances. To achieve this goal, we propose new methods based on a modification of Baumgartner-type test statistics. By calculating ranks based on actual distances rather than uniform spacing, these new methods aim to better align with real-world data and improve their practical applicability.

The remaining part of this paper is organized as follows. In Section 2, we provide a background of notations and test statistics. Section 3 outlines the adaptive Baumgartner-type test statistics. Section 4 details the simulation study conducted. Section 5 presents an empirical study. To conclude, Section 6 provides a comprehensive summary of the primary results obtained in this research, followed by a detailed discussion of their significance and potential applications.

2. Notations and Test-Statistics

The t-test, a widely recognized statistical test proposed by Gosset (1908), is employed to assess the difference between the means of two independent normal populations, X and Y, particularly when sample sizes are limited. For this test, we arbitrarily select m observations from population X and n observations from population Y, and the test-statistic is defined as:

$$T_1 = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{(1/m + 1/n)}},$$

where, $\bar{X}$ and $\bar{Y}$ are the means of two samples, whereas the s_p is the pooled sample standard deviation, calculated as:

$$s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}.$$

Under the null hypothesis of equal population means, the statistic T_1 is assumed to follow a Student's t-distribution with $m + n - 2$ degrees of freedom (df).

Welch (1938) introduced a modified t-test, often referred to as the unequal variances t-test or Welch's t-test, to accommodate scenarios where the assumption of equal population variances cannot be reasonably upheld. This test is particularly robust when dealing with samples that exhibit unequal variances and potentially disparate sample sizes. The Welch statistic is computed as follows:

$$T_2 = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}},$$

where T_2 follows a t-distribution with the following df:

$$df = \frac{\dfrac{s_X^2}{m} + \dfrac{s_Y^2}{n}}{\dfrac{s_X^4}{m^3 - m^2} + \dfrac{s_Y^4}{n^3 - n^2}}.$$

The Wilcoxon-Mann-Whitney (WMW) test is a nonparametric statistical test used to compare two independent samples. This test involves calculating a test statistic and ranking the observations from both samples. The test statistic, T_3, is defined as:

$$T_3 = mn + \frac{m(m+1)}{2} - R_X,$$

where R_X denotes the sum of the ranks for the observations in sample X. Assuming the null hypothesis that the probability of an observation from sample X being smaller than an observation from sample Y is 0.5, the test statistic T_3 is approximately normally distributed with a mean of $\frac{mn}{2}$ and a variance of $\frac{mn(m+n+1)}{2}$, see for example Zaremba (1965). A standardized test statistic, T_3^*, can be computed as follows:

$$T_3^* = \frac{W_{MN} - mn/2}{\sqrt{mn(m+n+1)/12}},$$

which can be approximated by the standard normal distribution.

Yuen (1974) introduced a test statistic that builds upon the Welch statistic by incorporating trimming. This method involves discarding the lowest and highest 20% of values from each sample, resulting in a trimming proportion (tr) of 0.20, which is generally considered appropriate. The trimmed means of the samples, denoted as $\bar{X}_c$ and $\bar{Y}_c$, are subsequently used to compute the Yuen-Welch statistic:

$$T_4 = \frac{\bar{X}c - \bar{Y}c}{\sqrt{w_X + w_Y}},$$

where w_X and w_Y represent estimates of the squared standard errors. T_3 follows a t-distribution with the following degrees of freedom (df):

$$df = \frac{w_X + w_Y}{\frac{w_X^2}{h_X - 1} + \frac{w_Y^2}{h_Y - 1}},$$

where h_X and h_Y denote the number of observations remaining in samples X and Y, after the trimming process.

In contrast to the WMW-test, the Brunner & Munzel (2000) test is a nonparametric test capable of handling both unequal variances and tied observations. To accommodate these complexities, the test employs mid-ranks, which are the average ranks assigned to tied values. The Brunner-Munzel test statistic is calculated as:

$$T_5 = \frac{\bar{M}_X - \bar{M}_Y}{(m+n)\sqrt{SB_X^2/mn^2 + SB_Y^2/m^2n}}.$$

The distribution of T_5 can be approximated by a t-distribution with the following degrees of freedom (df):

$$df = \left(\frac{SB_X^2}{n} + \frac{SB_Y^2}{m}\right)^2 \Big/ \left(\frac{SB_X^4}{n^2(m-1)} + \frac{SB_Y^4}{m^4(n-1)}\right).$$

The nonparametric Baumgartner test, introduced by Baumgartner et al. (1998), is designed to test the null hypothesis H_0 that two samples originate from identical populations with a shared cumulative distribution function. The test statistic proposed by the authors is as follows:

$$T_6 = \frac{1}{2}(B_X + B_Y),$$

where,

$$B_X = \frac{1}{m}\sum_{i=1}^{m} \frac{\left(R_i - \frac{m+n}{m} i\right)^2}{\left(\frac{i}{m+1}\right)\left(1 - \frac{i}{m+1}\right)\frac{n(m+n)}{m}},$$

and

$$B_Y = \frac{1}{n}\sum_{j=1}^{n} \frac{\left(H_j - \frac{m+n}{n} j\right)^2}{\left(\frac{j}{n+1}\right)\left(1 - \frac{j}{n+1}\right)\cdot\frac{m(m+n)}{n}}.$$

where R_i and H_j denote the ranks of the samples from the first and second populations, respectively, when the samples are pooled.

Concerning the location parameter, Murakami (2006) proposed a modified Baumgartner nonparametric k-sample test, which exhibits nearly equivalent power to the Wilcoxon test. The author argues that in the k-sample setting, where $F(x) = G\left(\frac{y}{\sigma}\right)$ with $\sigma \neq 0$, the modified Baumgartner statistic can be used to assess both location and scale differences. The combined sample rankings of the X-values and Y-values in ascending order of magnitude are denoted by $R_1 < \cdots < R_m$ and $H_1 < \cdots < H_n$, respectively. The test statistic can be expressed as follows:

$$T_7 = \frac{1}{2}\left(B_X^* + B_Y^*\right),$$

where

$$B_X^* = \frac{1}{m}\sum_{i=1}^{m} \frac{(R_i - E(R_i))^2}{\operatorname{Var}(R_i)},$$

and

$$B_Y^* = \frac{1}{n}\sum_{j=1}^{n} \frac{(H_j - E(H_j))^2}{\operatorname{Var}(H_j)}.$$

3. Adaptive Baumgartner-Type Test-Statistics

Let $X = X_1, \dots, X_m$ and $Y = Y_1, \dots, Y_n$ be two random samples of sizes m and n, respectively, drawn independently from continuous distributions $F(x)$

and $G(y)$. Baumgartner's well-known test is designed to assess the equality of locations between these two distributions, under the null hypothesis $F(x) = G_(y-\delta)$. A modified version of Baumgartner's nonparametric two-sample test has been proposed, which exhibits nearly identical power to the WMW-test for the location parameter. Notably, this modified statistic can be extended to the k-sample setting to test for both location and scale differences, where the null hypothesis is $F(x) = G\left(\frac{y}{\sigma}\right)$ with $\sigma \neq 0$.

Given two distinct samples, we hypothesize that the smaller sample is derived from the distribution $F(x)$ and the larger sample from $G(y)$. These two samples are then merged into a consolidated set Θ_i as

$$\Theta_i = \{X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n\}.$$

Arrange the observations of Θ_i in ascending order as

$$\Theta_i = \{\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m+n)}\},$$

Each value of Θ_i is assigned to one of two samples, ordered such that $\theta_{(1)} < \theta_{(2)} < \cdots < \theta_{(m+n)}$. Given the continuity assumption of the cumulative distribution functions $F(x)$ and $G(y)$, the probability of ties within the set Θ_i is negligible. Therefore, all inequalities among the Θ_i values are straightforward. The adaptive Baumgartner-type test-statistic is formulated by applying a relative-rank transformation to the set Θ_i as follows:

1. Define $R_1^r = 1$.
2. For $i = 2, 3, \ldots, (m+n)$, we define the following expression
$$\nu_i = i - 1 + \frac{(n_1 + n_2 - 1)(\theta_{(i)} - \theta_{(i-1)})}{\theta_{(m+n)} - \theta_{(1)}}.$$
3. Define R_i^r as the $(i-1)$th smallest value of $\{\nu_2, \nu_3, \ldots, \nu_{m+n}\}$ for $i = 2, 3, \ldots, (m+n)$. This will ensure that
$$R_1^r < R_2^r < \cdots < R_{m+n}^r.$$

Let $R_1^{(r)} < \cdots < R_m^{(r)}$ and $H_1^{(r)} < \cdots < H_n^{(r)}$ denote the combined sample ranks of the X-values and Y-values in increasing order of magnitude, respectively.

$$T_8 \quad = \quad \frac{1}{2}(B_{M_X} + B_{M_Y}),$$

where

$$B_{M_X} = \frac{1}{m}\sum_{i=1}^{m} \frac{\left(R_i^{(r)} - \frac{n+m}{m} i\right)^2}{\left(\frac{i}{m+1}\right)\left(1 - \frac{i}{m+1}\right)\frac{n(m+n)}{m}},$$

and

$$B_{M_Y} = \frac{1}{n}\sum_{j=1}^{n} \frac{\left(H_j^{(r)} - \frac{m+n}{n} j\right)^2}{\left(\frac{j}{n+1}\right)\left(1 - \frac{j}{n+1}\right)\frac{m(m+n)}{n}}.$$

We also provided an adaptive test-statistic for Murakami (2006), which can be defined as:

$$T_9 \quad = \quad \frac{1}{2}\left(B^*_{M_X} + B^*_{M_Y}\right),$$

where

$$B^*_{M_X} = \frac{1}{m}\sum_{i=1}^{m}\frac{(R_i^{(r)} - E(R_i^{(r)}))^2}{\mathrm{Var}(R_i^{(r)})},$$

and

$$B^*_{M_Y} = \frac{1}{n}\sum_{j=1}^{n}\frac{(H_j^{(r)} - E(H_j^{(r)}))^2}{\mathrm{Var}(H_j^{(r)})}.$$

The expected values and variances can be computed as:

$$E(R_i) = \frac{m+n+1}{n+1}i,$$

$$V(R_i) = \frac{i}{n+1}(1 - \frac{i}{n+1})\frac{m(n+m+1)}{n+2},$$

$$E(H_j) = \frac{n+m+1}{m+1}j,$$

and

$$V(H_j) = \frac{j}{m+1}(1 - \frac{j}{m+1})\frac{n(n+m+1)}{m+2}.$$

The mean and variance of Baumgartner-type statistics are indeterminate and can only be approximated through rigorous simulation. In this study, we utilized several distributions, like normal, uniform, exponential and Laplace with 10 000 replications to derive these values, which are presented in Table 1. These values are closely aligned with those of both adaptive and original Baumgartner-type statistics. Additionally, Figure 1 illustrates the distributional properties of Baumgartner-type statistics within the simulation framework. The presented statistics exhibit enhanced robustness without introducing bias or altering distributional assumptions, making them valuable additions to non-parametric approaches.

Table 1: Mean and variance of Baumgartner-type test-statistics.

	T_6	T_7	T_8	T_9
μ	0.9880	1.0181	1.2672	1.2769
σ^2	0.6488	0.7744	0.7331	0.7884

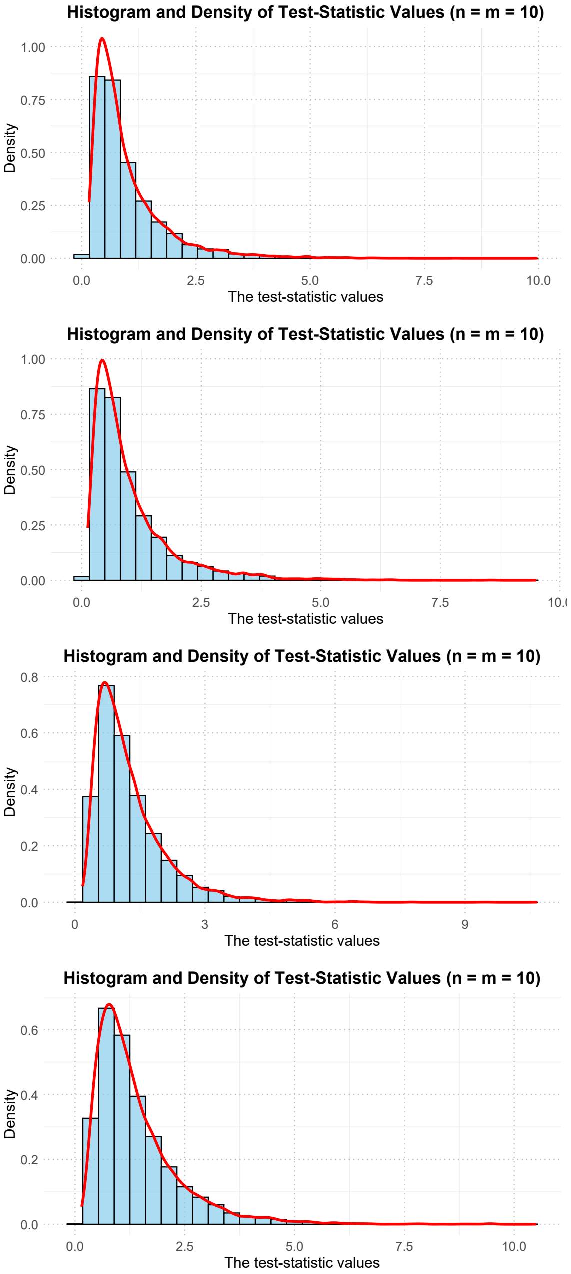

FIGURE 1: The distributional behaviors of Baumgartner-type statistics: (a) T_6, (b) T_7, (c) T_8, and (d) T_9.

This study demonstrates the effectiveness of adaptive methodological enhancements in improving the power of nonparametric statistical tests. By strategically adjusting rankings, we align our findings with previous research on power optimization in ranking-based methods. Our approaches highlight the potential for enhancing hypothesis testing through targeted modifications to test statistic calculations, offering valuable insights for both theoretical and applied statistical research.

4. Simulations Study

Tables 2-6 present a comprehensive comparison of nine test statistics across diverse scenarios, encompassing shifts in location, scale, or both, along varying sample sizes. This study employed a range of distributions with dynamic characteristics, including normal, exponential, uniform, and Laplace. All simulations were executed using R software (version 4.4.2), and statistical tests were conducted at the 5% significance level. To assess the performance of each test statistic, 10 000 replications were performed. For precise comparisons, simulation results were rounded to four decimal places.

Table 2: The probability of rejecting $H_0 : X \sim N(0, 1)$ vs. $H_1 : Y \sim N(\mu, 1)$.

Test	m	n	$\mu = 0.0$	0.5	1.0	1.5	2.0	2.5	3.0
T_1	5	10	0.0533	0.1344	0.3605	0.6636	0.8757	0.9723	0.9961
T_2			0.0402	0.1100	0.3423	0.6456	0.8737	0.9742	0.9962
T_3			0.0484	0.1282	0.3614	0.6565	0.8733	0.9727	0.9952
T_4			0.0685	0.1630	0.4131	0.7059	0.8995	0.9723	0.9900
T_5			0.0638	0.1157	0.2892	0.5145	0.6993	0.8314	0.9076
T_6			0.0501	0.1023	0.3137	0.6117	0.8543	0.9640	0.9946
T_7			0.0536	0.1276	0.3684	0.6668	0.8965	0.9761	0.9974
T_8			0.0432	0.0853	0.2591	0.5508	0.8166	0.9488	0.9906
T_9			0.0554	0.1190	0.3315	0.6433	0.8758	0.9732	0.9960
T_1	10	10	0.0485	0.1861	0.5528	0.8860	0.9888	0.9993	1.0000
T_2			0.0416	0.1603	0.5179	0.8525	0.9808	0.9990	1.0000
T_3			0.0463	0.1825	0.5500	0.8830	0.9890	0.9994	1.0000
T_4			0.0552	0.1948	0.5631	0.8828	0.9846	0.9997	1.0000
T_5			0.0500	0.1487	0.4617	0.7888	0.9537	0.9950	1.0000
T_6			0.0482	0.1741	0.5215	0.8599	0.9818	0.9983	1.0000
T_7			0.0484	0.1767	0.5279	0.8523	0.9798	0.9992	1.0000
T_8			0.0391	0.1374	0.4486	0.8101	0.9677	0.9974	1.0000
T_9			0.0444	0.1481	0.4666	0.8116	0.9708	0.9980	1.0000
T_1	20	20	0.0503	0.3353	0.8751	0.9961	0.9990	1.0000	1.0000
T_2			0.0505	0.3300	0.8489	0.9945	1.0000	1.0000	1.0000
T_3			0.0480	0.3302	0.8723	0.9960	1.0000	1.0000	1.0000
T_4			0.0531	0.3297	0.8567	0.9945	1.0000	1.0000	1.0000
T_5			0.0468	0.2966	0.8026	0.9891	1.0000	1.0000	1.0000
T_6			0.0497	0.3176	0.8464	0.9929	1.0000	1.0000	1.0000
T_7			0.0486	0.3023	0.8335	0.9933	0.9999	1.0000	1.0000
T_8			0.0470	0.2759	0.8063	0.9903	0.9999	1.0000	1.0000
T_9			0.0423	0.2694	0.7960	0.9903	1.0000	1.0000	1.0000

Table 3: The probability of rejecting $H_0 : X \sim N(0,1)$ vs. $H_1 : Y \sim N(\mu, \sigma^2)$, for the case of $(m = n = 10)$.

Test	μ	$\sigma^2 = 2.0$	3.0	4.0	5.0	6.0
T_1	0.0	0.0547	0.0644	0.0667	0.0656	0.0661
T_2		0.0672	0.0873	0.0921	0.0988	0.1019
T_3		0.0485	0.0523	0.0509	0.0515	0.0523
T_4		0.0497	0.0347	0.0335	0.0329	0.0299
T_5		0.0619	0.0691	0.0561	0.0603	0.0702
T_6		0.0960	0.2630	0.2657	0.3409	0.4109
T_7		0.0499	0.0736	0.1141	0.2202	0.2743
T_8		0.0863	0.1785	0.2744	0.3597	0.4345
T_9		0.6509	0.8878	0.9492	0.9686	0.9767
T_1	1.0	0.2749	0.1667	0.1289	0.1010	0.0884
T_2		0.2486	0.1652	0.1212	0.1121	0.1059
T_3		0.2616	0.1567	0.1055	0.0858	0.0761
T_4		0.2593	0.1467	0.0990	0.0797	0.0640
T_5		0.2119	0.1374	0.1093	0.0878	0.0833
T_6		0.3443	0.3360	0.3655	0.4003	0.4518
T_7		0.3240	0.3075	0.3315	0.3870	0.4283
T_8		0.3423	0.3621	0.4129	0.4701	0.5211
T_9		0.5213	0.5850	0.6710	0.7485	0.7952
T_1	2.0	0.6538	0.4821	0.3682	0.3364	0.2401
T_2		0.7090	0.4346	0.2963	0.2263	0.1888
T_3		0.7439	0.4485	0.2774	0.2052	0.1519
T_4		0.7069	0.4050	0.2452	0.1635	0.1271
T_5		0.6253	0.3625	0.2402	0.1895	0.1408
T_6		0.7936	0.6399	0.5772	0.5568	0.5542
T_7		0.7811	0.6177	0.5397	0.5139	0.5286
T_8		0.6960	0.8160	0.9140	0.9678	0.9865
T_9		0.8956	0.8293	0.8212	0.8417	0.8610

Table 2 presents the simulated results, which rejected the null hypothesis of normal distribution for various sample sizes: $(m, n) = (5, 10), (10, 10), (20, 20)$. The Type I error rates for all tests are approximately 0.05, as indicated by the first column in each panel. In terms of power, the test statistic T_4 exhibits superior performance for small location shifts. The proposed adaptive tests consistently outperform their original counterparts across all table entries. Table 3 presents the simulated results for location-scale shifts with the sample size $(m, n) = (10, 10)$. For all shifts, it is evident that the proposed adaptive tests outperform the original tests across all table entries.

Table 4 presents the results of the uniform distribution experiment for two sample size pairs: $(m, n) = (5, 10), (10, 10)$. The observed significance levels closely approximate the nominal error rate for all employed test statistics. In terms of power, the adaptive tests (T_8 and T_9) demonstrated superior performance when sample sizes were unequal. However, these tests did not meet expectations for the uniform distribution for equal sample size. This discrepancy can be attributed to the fact that these tests are designed to detect outliers within a sample, while the uniform distribution assigns equal weight to all sample units.

Table 4: The probability of rejecting $H_0 : X \sim U(0,1)$ vs. $H_1 : Y \sim U(0,1) + \delta$.

Test	m	n	$\delta = 0.0$	0.1	0.2	0.3	0.4	0.5	0.6
T_1	5	10	0.0609	0.0888	0.1845	0.3385	0.5619	0.7801	0.9297
T_2			0.0436	0.0724	0.1608	0.3191	0.5225	0.7308	0.8880
T_3			0.0679	0.0882	0.1868	0.3482	0.5697	0.7769	0.9247
T_4			0.0719	0.1046	0.2115	0.3877	0.6105	0.7992	0.9321
T_5			0.0719	0.0906	0.1409	0.2235	0.3398	0.4675	0.5982
T_6			0.0480	0.0671	0.1451	0.2839	0.4775	0.6885	0.8601
T_7			0.0507	0.0779	0.1776	0.3400	0.5432	0.7456	0.8967
T_8			0.0325	0.0868	0.1809	0.3522	0.5825	0.7904	0.9269
T_9			0.0437	0.1078	0.2304	0.4287	0.6620	0.8500	0.9561
T_1	10	10	0.0514	0.1060	0.2868	0.5776	0.8316	0.9615	0.9970
T_2			0.0442	0.0439	0.2534	0.5016	0.7476	0.9119	0.9823
T_3			0.0508	0.1058	0.2948	0.5718	0.8305	0.9631	0.9971
T_4			0.0512	0.1182	0.3009	0.5619	0.7952	0.9363	0.9881
T_5			0.0582	0.0840	0.1807	0.3471	0.5581	0.7595	0.9008
T_6			0.0459	0.1017	0.2588	0.5213	0.7603	0.9243	0.9865
T_7			0.0528	0.1024	0.2519	0.5148	0.7656	0.9274	0.9841
T_8			0.0344	0.0650	0.1944	0.4229	0.6893	0.9001	0.9820
T_9			0.0371	0.0628	0.1922	0.4369	0.7007	0.8991	0.9801

For location-scale problems involving Laplace distributions, we compared the performance of our proposed adaptive tests with their competitors. The results, summarized in Table 5, demonstrate the superior performance of our suggested tests across all scenarios. A similar pattern of superiority was observed for the exponential distribution, as shown in Table 6. In conclusion, our simulation study indicates that the proposed adaptive test statistics, particularly T_9, exhibit higher power than the competing methods.

Upon examination of Tables 2-6, the following observations can be made:

1. Statistical test performance: The T_1 and T_4 consistently outperforms the reference test under normal and T_1 perform better in the case of uniform distribution assumptions.

2. Location-scale problem: Adaptive procedures exhibit superior performance compared to traditional two-sample tests, regardless of distributional assumptions, with the exception of uniform distributions.

3. Nonnormal distributions: In the context of nonnormal distributions, adaptive procedures demonstrate superior performance to both the Baumgartner statistic and the modified Baumgartner statistic across all study conditions.

Table 5: The probability of rejecting $H_0 : X \sim L(0,1)$ vs. $H_1 : Y \sim L(\mu, \sigma)$, for the case of $(m = n = 10)$.

Test	μ	$\sigma = 1.00$	1.25	1.50	1.75	2.00	2.25
T_1	0.0	0.0424	0.0442	0.0457	0.0430	0.0419	0.0455
T_2		0.0435	0.0481	0.0580	0.0638	0.0614	0.0705
T_3		0.0513	0.0513	0.0417	0.0435	0.0403	0.0422
T_4		0.0407	0.0548	0.0560	0.0561	0.0516	0.0498
T_5		0.0447	0.0425	0.0407	0.0396	0.0400	0.0454
T_6		0.0482	0.0779	0.1216	0.1774	0.2280	0.2785
T_7		0.0470	0.0710	0.1238	0.1673	0.2199	0.2711
T_8		0.0669	0.0684	0.0839	0.1015	0.1154	0.1275
T_9		0.0895	0.0909	0.1031	0.1220	0.1417	0.1671
T_1	1.0	0.3590	0.1747	0.1049	0.0830	0.0718	0.0602
T_2		0.3791	0.2072	0.1472	0.1235	0.1026	0.1002
T_3		0.3471	0.1771	0.1159	0.0822	0.0678	0.0623
T_4		0.4122	0.2224	0.1387	0.1016	0.0861	0.0712
T_5		0.4041	0.2035	0.1250	0.0902	0.0766	0.0631
T_6		0.4105	0.2826	0.2647	0.2872	0.3098	0.3430
T_7		0.4158	0.2672	0.2521	0.2588	0.2963	0.3235
T_8		0.3827	0.3516	0.3333	0.3307	0.3318	0.3425
T_9		0.4043	0.3791	0.3587	0.3568	0.3546	0.3639
T_1	2.0	0.8406	0.5129	0.2988	0.1972	0.1479	0.1171
T_2		0.8642	0.5756	0.3782	0.2791	0.2254	0.1939
T_3		0.8423	0.5101	0.3000	0.1961	0.1406	0.1152
T_4		0.8691	0.5691	0.3553	0.2403	0.1848	0.1383
T_5		0.8814	0.5628	0.3444	0.2393	0.1640	0.1353
T_6		0.3633	0.6733	0.5547	0.5082	0.4795	0.4782
T_7		0.8971	0.6628	0.5209	0.4812	0.4586	0.4556
T_8		0.8588	0.7919	0.7483	0.7384	0.7553	0.7667
T_9		0.9260	0.8063	0.7608	0.7563	0.7703	0.7918

Table 6: The probability of rejecting $H_0 : X \sim Exp(0)$ vs. $H_1 : Y \sim Exp(\lambda)$, for the case of $(m = n = 10)$.

Test	$\lambda = 2.0$	3.0	4.0	5.0	6.0	7.0
T_1	0.2148	0.4542	0.6283	0.7311	0.7956	0.8376
T_2	0.2129	0.4693	0.6524	0.7827	0.8569	0.8989
T_3	0.2077	0.4543	0.6318	0.7353	0.7978	0.8351
T_4	0.2536	0.5168	0.6897	0.8013	0.8606	0.9052
T_5	0.1311	0.2659	0.3506	0.4087	0.4585	0.4972
T_6	0.2370	0.5029	0.6928	0.8036	0.8836	0.9177
T_7	0.2383	0.5078	0.6933	0.8088	0.8723	0.9180
T_8	0.5827	0.8210	0.9325	0.9694	0.9869	0.9948
T_9	0.6673	0.8758	0.9486	0.9810	0.9926	0.9962

5. Empirical Data

This section demonstrates the applicability of adaptive tests using six real-world data examples presented in Table 7.

- Dataset 1: Sourced from the UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets/HCV+data`, this dataset examines Cholinesterase (CHE), a liver enzyme used to assess liver function. Lower CHE levels indicate poor liver protein synthesis capacity, while higher levels may suggest conditions such as nephrotic syndrome, hyperthyroidism, or fatty liver.
- Dataset 2: This dataset, as described in Wild & Seber (1999), investigates genetic inheritance by analyzing the mean sister chromatid exchange (MSCE) of Native American and Caucasian individuals from diverse ethnic backgrounds.
- Dataset 3: This dataset, initially presented in Tasdan & Sievers (2009), examines the effect of thyroxine on young mice. It compares a control group of 7 mice with the remaining mice used for thyroxine treatment.
- Datasets 4 and 5: These datasets, respectively from Hettmansperger & McKean (2011) and Lindsey et al. (1987), explore the following: Plasma LDL levels in quails fed a special diet containing a drug compared to a non-drug-fed control group. The width of the first tarsal joint in two different Chaetocnema insect species.
- Dataset 6: This dataset, initially studied by Karpatkin et al. (1981) and further discussed by Hollander & Wolfe (1999), investigates the influence of maternal steroid medication on newborn platelet counts. It compares platelet count parameters in infants born to mothers with autoimmune thrombocytopenia purpura (ATP) who received prednisone treatment with those born to mothers without ATP.

To visually represent the data, we constructed violin plots as shown in Figure 2. To analyze these datasets, we employed two newly developed adaptive tests alongside seven well-established tests. The results, summarized in Table 8, are presented as p-values, whereas the significance level is indicated at the 0.05. The outcomes of the adaptive tests strongly support the distributional patterns observed in Figure 2. Therefore, when dealing with outliers, adaptive test-statistics can be considered reliable and appropriate alternatives. Based on the observed data and at the given significance level, we conclude that the results are consistent. These findings demonstrate the potential of the suggested adaptive tests as valuable tools, comparable to other well-established tests in the literature.

Table 7: The datasets used in this study.

Dataset	Values
Dataset 1: Cholinesterase levels (hepatitis and fibrosis)	
Fibrosis:	11.49, 9.64, 6.97, 7.76, 7.28, 10.43, 8.74, 8.77, 8.59, 6.60 9.45, 7.10, 9.92, 9.24, 8.55, 8.61, 7.29, 10.21, 3.99, 7.75, 6.65
Hepatitis:	9.58, 7.55, 7.09, 6.00, 8.77, 8.79, 12.16, 10.11, 13.80, 9.71 10.30, 11.42, 10.23, 9.67, 8.91, 16.41, 9.54, 10.12, 5.75, 5.95 6.88, 7.08, 7.51, 9.48
Dataset 2: Wild and Seber data (MSCE values)	
Native American:	8.50, 9.48, 8.65, 8.16, 8.83, 7.76, 8.63
Caucasian:	8.27, 8.20, 8.25, 8.14, 9.00, 8.10, 7.20, 8.32, 7.70
Dataset 3: Thyroid weights (grams)	
Control group:	0.7, 1.2, 1.4, 2.3, 1.6, 0.9, 1.3
Treatment group:	4.1, 4.4, 3.3, 2.1, 3.5, 2.9, 2.8, 4.3
Dataset 4: Hettmansperger and McKean (LDL levels)	
Treatment group:	10, 1, 20, 18, 122, 14, 44, 8, 51, 34
Control group:	34, 19, 24, 34, 67, 36, 46, 14, 41, 59, 40, 42 41, 25, 30, 32, 16, 47, 56, 41
Dataset 5: Lindsey et al. (tarsal joint widths)	
Species A:	31, 34, 37, 27, 28, 18, 34, 29, 31, 15
Species B:	7, 22, 44, 31, 8, 18, 22, 27, 25, 24
Dataset 6: Newborn platelet counts	
Control subjects:	12, 20, 112, 32, 60, 40
Case subjects:	120, 124, 215, 90, 67, 95, 190, 180, 135, 399

Table 8: The p-values for each test statistic across datasets.

Dataset	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
1	0.1483	0.2026	0.1385	0.2034	0.2440	0.1900	0.1980	0.2122	0.1112
2	0.1066	0.1142	0.1141	0.1028	0.0606	0.1562	0.1135	0.0670	0.0912
3	0.0001	0.0006	0.0001	0.0000	0.0003	0.0001	0.0000	0.0000	0.0000
4	0.5814	0.1343	0.6760	0.2110	0.1725	0.0612	0.0712	0.0600	0.0612
5	0.1834	0.1031	0.1860	0.1012	0.0258	0.0912	0.0601	0.0000	0.0000
6	0.0141	0.0017	0.0047	0.0001	0.0051	0.0000	0.0010	0.0000	0.0000

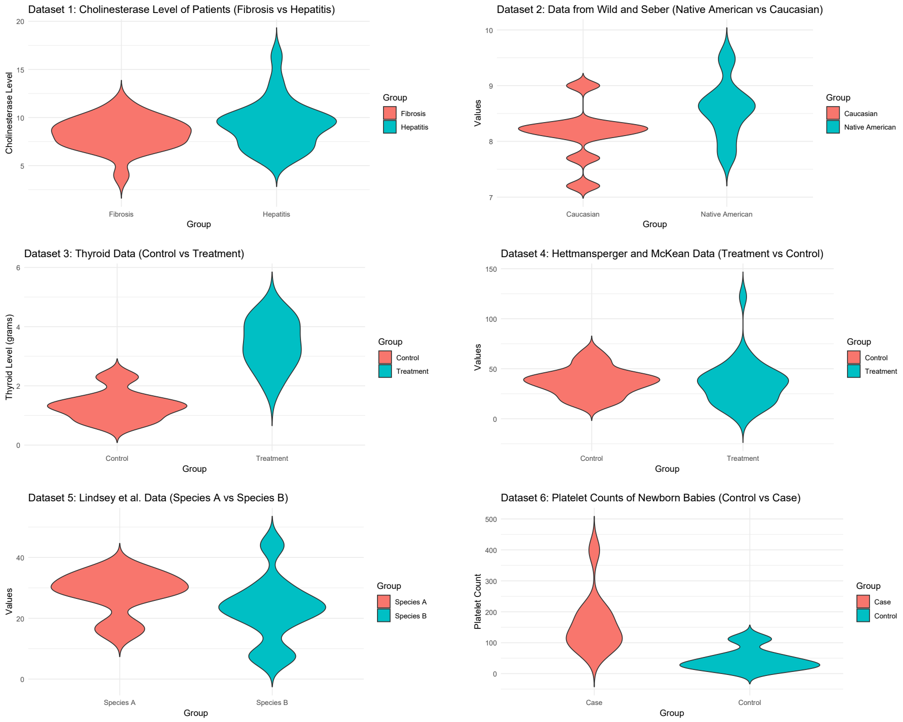

Figure 2: Violin plots illustrating the distribution of various real-life datasets: (a) Dataset 1, (b) Dataset 2, (c) Dataset 3, (d) Dataset 4, (e) Dataset 5, and (f) Dataset 6.

6. Conclusions

The application of rank transformations in nonparametric statistical inference has been a subject of ongoing debate, see for example Zimmerman (2012). To address the limitations of traditional rank transformations, Hussain et al. (2024) proposed a novel approach, the relative rank transformation, which preserves finer distinctions between values. This study investigates the utility of relative ranks in the context of Baumgartner-type statistics. We evaluate the performance of our adaptive approaches against widely used test statistics for the two-sample independent problem. Simulation results indicate that our adaptive tests, employing relative ranks, outperform all other well-known tests when both location and scale parameters shift under various normal and non-normal distributions. Real-world medical case studies further demonstrate the significant improvement in performance achieved by the proposed adaptive techniques. Overall, our investigation suggests that employing relative ranks can yield more reliable results for the two-sample independent problem, especially when both location and scale parameters shift simultaneously.

[Received: January 2025 — Accepted: November 2025]

References

Baumgartner, W., Weiß, P. & Schindler, H. (1998), 'A nonparametric test for the general two-sample problem', *Biometrics* pp. 1129–1135.

Brunner, E. & Munzel, U. (2000), 'The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation', *Biom J* **42**, 17–25.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, Duxbury Press, Pacific Grove, CA.

Conover, W. J. (1999), *Practical nonparametric statistics*, Vol. 350, John Wiley & Sons, New York.

Fagerland, M. W. & Sandvik, L. (2009), 'Performance of five two-sample location tests for skewed distributions with unequal variances', *Contemporary Clinical Trials* **30**(5), 490–496.

Gosset, W. S. (1908), William sealy gosset, *in* 'Biographical Encyclopedia of Mathematicians', Vol. 1, p. 239.

Hettmansperger, T. P. & McKean, J. W. (2011), *Robust Nonparametric Statistical Methods*, 2nd edn, CRC Press, Boca Raton, FL.

Hollander, M. & Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, Wiley, New York.

Hollander, M., Wolfe, D. A. & Chicken, E. (2013), *Nonparametric Statistical Methods*, 3rd edn, John Wiley & Sons.

Hussain, A., Drekic, S. & Cheema, S. A. (2024), 'A relative-rank measure for the rank transformation', *Statistics & Probability Letters* **204**, 109932.

Karpatkin, M., Porges, R. F. & Karpatkin, S. (1981), 'Platelet counts in infants of women with autoimmune thrombocytopenia: Effects of steroid administration to the mother', *The New England Journal of Medicine* **305**(16), 936–939.

Lindsey, D. C., Herzberg, A. M. & Watts, D. G. (1987), 'A method of cluster analysis based on projections and quantile-quantile plots', *Biometrics* **43**, 327–341.

Manly, B. F. (2018), *Randomization, bootstrap and Monte Carlo methods in biology*, Chapman and Hall/CRC.

Mittelstadt, B. D. & Floridi, L. (2016), The ethics of big data: current and foreseeable issues in biomedical contexts, *in* 'The Ethics of Biomedical Big Data', pp. 445–480.

Moran, J. L. (2006), Statistical issues in the analysis of outcomes in critical care medicine, PhD thesis, Doctoral dissertation.

Murakami, H. (2006), 'A k-sample rank test based on modified Baumgartner statistic and its power comparison', *Journal of the Japanese Society of Computational Statistics* **19**(1), 1–13.

Siegel, S. & Castellan, N. J. (1988), *Nonparametric statistics for the behavioral sciences*, 2nd edn, McGraw-Hill, New York.

Tasdan, F. & Sievers, G. (2009), 'Smoothed Mann-Whitney-Wilcoxon procedure for the two-sample location problem', *Communications in Statistics - Theory and Methods* **38**(6), 856–870.

Welch, B. L. (1938), 'The significance of the difference between two means when the population variances are unequal', *Biometrika* **29**(3/4), 350–362.

Wild, C. J. & Seber, G. A. F. (1999), *Chance Encounters: A First Course in Data Analysis and Inference*, John Wiley & Sons, New York.

Yuen, K. K. (1974), 'The two-sample trimmed t for unequal population variances', *Biometrika* **61**, 165–170.

Zaremba, S. K. (1965), 'Note on the wilcoxon-mann-whitney statistic', *The Annals of Mathematical Statistics* **36**(3), 1058–1060.

Zimmerman, D. W. (2012), 'A note on consistency of non-parametric rank tests and related rank transformations', *British Journal of Mathematical and Statistical Psychology* **65**(1), 122–144.