

A New Transformed t -test for Skewed Data: A Goodness-of-fit Approach

Khairul Islam¹ & Tanweer J. Shapla²

^{1,2}Department of Mathematics and Statistics, Eastern Michigan University, USA.

Correspondence: Khairul Islam, Department of Mathematics and Statistics, Eastern Michigan University, Ypsilanti, MI 48197, USA

Received: June 7, 2020 Accepted: July 14, 2020 Online Published: July 27, 2020

doi:10.5539/ijsp.v9n5p30

URL: <https://doi.org/10.5539/ijsp.v9n5p30>

Abstract

A new transformed two-sample t -test has been proposed for testing equality of two population means for skewed distributions by means of a univariate normal goodness of fit to the combined sample. The small sample performance of the proposed test is compared with untransformed t -test and the non-parametric analogue of t -test via Wilcoxon rank sum test using real-life examples and simulation from skewed distributions with varying values of skewness, empirically. It reveals that the proposed new test is appropriate for estimating the level of significance and is more powerful than the untransformed t -test and the Wilcoxon rank sum test for skewed distributions.

Keywords: goodness-of-fit, power, transformation, two-sample t -test, Wilcoxon test

1. Introduction

Let $X = (X_1, X_2, \dots, X_m)$ and $Y = (Y_1, Y_2, \dots, Y_n)$ be two independent random samples from two populations having means $\mu_x = E(X_1)$ and $\mu_y = E(Y_1)$, respectively. We wish to test the null hypothesis

$$H_{01}: \mu_x = \mu_y$$

that is, the two populations from which the two samples are considered have the same mean. For testing H_{01} , the standard statistical models usually assume that the two population distributions are normal with the common unknown variance σ^2 . Under the assumption, a pooled estimator of σ^2 is given by

$$S_p^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$$

where S_x^2 and S_y^2 are sample variances of the two samples X and Y , respectively. Under H_{01} , the test statistic T given by

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

follows Student's t -distribution with $m+n-2$ degrees of freedom. This test is uniformly most powerful unbiased test (see, e.g., Lehmann 1994), and is omnipresent in statistical practice for making inference about the difference of the two population means.

In real life, the assumption of normality is often invalid or unmet. As such, one option is to use the nonparametric analog of t -test, namely, Wilcoxon rank sum test (Wilcoxon, 1945) or Mann Whitney U test (Mann & Whitney, 1947) which does not require the normality of the data for the validity of the inference. Alternately, one may use the t -test to transformed data following an appropriate transformation. With transformation an option, the common practice is to re-express the data to achieve the normality and then implement t -test (Mosteller & Tukey, 1977; Atkinson, 1985). In an oft-cited paper, Box and Cox (1964) suggested a power transformation for non-negative observations to achieve normality. Since then Box-Cox transformation has widely been used for of the problems of statistical inference.

In this article, a new method is proposed to estimate the Box-Cox transformation by means of the univariate normal goodness-of-fit approach. The idea is to combine the Box-Cox transformed data from two samples to fit into a normal distribution to estimate the transformation parameter, and then implement the t -test to the transformed data. The new transformed t -test outperforms existing transformed t -test, and the nonparametric Wilcoxon rank sum test or the Student's t -test in the violation of the normality.

2. Methods

In this section, we review some popular tests for comparing two groups with respect to their locations (means or medians). Section 2.1 presents a brief review of nonparametric Wilcoxon rank sum test for the completeness of the comparison. A Box-Cox transformed t -test achieved via a maximum likelihood method is discussed in section 2.2. The new transformation using the univariate normal goodness-of-fit is discussed in section 3. Examples from a real-life situation and a simulated data appear in section 4 to demonstrate the application and performance of the proposed test as compared with the other tests described. A simulation study is carried out in section 5 to compare the finite sample performance of all tests considered in this article. Results and discussion from examples and simulation study appear in section 6. The concluding remarks of the study appear in section 7.

2.1 Wilcoxon Rank Sum Test

The nonparametric Wilcoxon rank-sum test, also known as the Mann-Whitney U test, is well-known and preferable to the two-sample t -test when the two populations the samples come from depart from normality. Let $\{X_1, \dots, X_m\}$ and $\{Y_1, \dots, Y_n\}$ be two independent samples from two populations with continuous cdfs F_X and F_Y and location parameters μ_x and μ_y , respectively. Then, the basic null hypothesis of the Wilcoxon rank sum test is that the two populations have an identical distribution (Gibbons & Chakraborti, 2014; Kvam & Vidakovic, 2007; Desu & Raghavarao, 2004). That is $H_{02}: F_X(x) = F_Y(x)$. Note that when the two random variables X and Y have the identical distribution, they will have the same median or mean, say, $\mu_x = \mu_y$. Then, one can test the equality of two location parameters using the test $H_{02}: F_X = F_Y$ or equivalently $H_{02}: \mu_x = \mu_y$. In order to test H_{02} , the Mann-Whitney (U) test compares each $X_i, i = 1, 2, \dots, m$ with each $Y_j, j = 1, 2, \dots, n$ and is defined as follows:

$$U_{yx} = \# \text{ of pairs } (X_i, Y_j) \text{ for which } X_i > Y_j$$

It follows that $U_{yx} = \sum_{i=1}^m R_i - \frac{m(m+1)}{2}$, where $R_1 < R_2 < \dots < R_m$ are the ordered ranks of "m" x -observations in the combined sample. On the other hand, the Wilcoxon rank sum test (W) is defined in terms of "sum of X ranks in the combined sample": $W_x = \sum_{i=1}^m R_i$. It is easy to verify that W_x and U_{yx} are connected by the equation $W_x = U_{yx} + \frac{m(m+1)}{2}$ (e.g., see Gibbons & Chakraborti, 2014; Kvam & Vidakovic, 2007; Desu & Raghavarao, 2004). In view of this relationship, one can use either of the statistics W_x or U_{yx} or similarly defined W_y or U_{xy} for testing H_{02} . For an example, given a level of significance α , the inference procedure using Wilcoxon rank sum statistics can be made as follows:

- 1) Reject H_{02} against $H_{2a}: \mu_x > \mu_y$ if W_x is larger i.e., p - value = $P(W \geq W_x) \leq \alpha$.
- 2) Reject H_{02} against $H_{2b}: \mu_x < \mu_y$ if W_y is larger i.e., p - value = $P(W \geq W_y) \leq \alpha$.
- 3) Reject H_{02} against $H_{2c}: \mu_x \neq \mu_y$ using $W_{min} = \min(W_x, W_y)$ or $W_{max} = \max(W_x, W_y)$ if the p - value = $2P(W \leq W_{min}) = 2P(W \geq W_{max}) \leq \alpha$.

We implement this test using the statistical software R.

2.2 Box-Cox Transformed Test

An alternative to Wilcoxon rank sum test, one can use the Box-Cox transformation (Box & Cox, 1964) to achieve normality before applying t -test when the data deviate from normality. For simplicity of presentation, let $X = (X_1, \dots, X_m)$ and $Y = (Y_1, \dots, Y_n)$ be non-negative random variables having a positive skewed distribution or deviating from normality. Given a scalar λ , the Box-Cox power transformation to the sample X , $X(\lambda)$, is defined by

$$X_i(\lambda) = \begin{cases} (X_i^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0 \\ \log(X_i), & \text{if } \lambda = 0 \end{cases} \quad (1)$$

The transformation to Y_j , $Y_j(\lambda)$ is defined in a similar way.

Let $\bar{X}(\lambda) = m^{-1} \sum_{i=1}^m X_i(\lambda)$ be the mean of the transformed sample $X(\lambda)$. Let $\bar{Y}(\lambda)$ be defined similarly. Let $S^2(\lambda)$ be the pooled maximum likelihood estimate of the variance to the transformed data given by

$$S^2(\lambda) = \{1/(m+n)\} \left[\sum_{i=1}^m \{X_i(\lambda) - \bar{X}(\lambda)\}^2 + \sum_{j=1}^n \{Y_j(\lambda) - \bar{Y}(\lambda)\}^2 \right]$$

Given the transformation (1) is successful to transform the data to fit a normal model, the profiled log-likelihood function for the transformation parameter λ is

$$l(\lambda) = -\{(m + n)/2\} \log S^2(\lambda) + \lambda \{ \sum_{i=1}^m \log X_i + \sum_{j=1}^n \log Y_j \} \tag{2}$$

Box and Cox (1964) proposed to estimate λ by the maximum likelihood estimator (MLE), $\hat{\lambda}_l$ over a pre-specified set I of values of λ so as to the equation (2) is maximized. Then, the two-sided transformed t -test is to reject $H_{01}: \mu_x = \mu_y$ if $|T(\hat{\lambda}_l)|$ is greater than the Student's t critical value $t_{\alpha/2, m+n-2}$, where $T(\hat{\lambda}_l) = \frac{\bar{X}(\hat{\lambda}_l) - \bar{Y}(\hat{\lambda}_l)}{S(\hat{\lambda}_l) \sqrt{1/m + 1/n}}$.

The theoretical aspects of the Box-Cox transformed data analysis described above have been reported in literature. For examples, Hinkley (1975) and Hernandez and Johnson (1980) investigated the asymptotic properties of the parameter estimates. Bickel and Doksum (1981) critically examined the behavior of the asymptotic variances of the parameter estimates for regression and analysis of variance situations. Chen and Loh (1992) and Chen (1995) proved that the Box-Cox transformed t -test is typically more efficient asymptotically than the t -test without transformation. Islam and Chen (2007) justified the use of transformed t -test by fitting a t distribution to transformed data.

3. The New Proposed Transformed T -Test

In this article, we propose a new transformed t -test by applying a univariate normal goodness-of-fit to the transformed combined sample data. This method is easy to implement using any standard statistical software, and it outperforms other tests considered in this study while applied in real-life problems and simulations. Below we describe the new method along with an algorithm to implement it.

Given the transformation $X(\lambda)$ is successful or nearly successful in achieving normality, it is expected that $Z_x(\lambda) = \frac{X(\lambda) - \mu_x(\lambda)}{\sigma_x(\lambda)} = (Z_{1,x}(\lambda), Z_{2,x}(\lambda), \dots, Z_{m,x}(\lambda))$ represents a random sample from a $N(0,1)$ distribution. With the similar argument, $Z_y(\lambda) = \frac{Y(\lambda) - \mu_y(\lambda)}{\sigma_y(\lambda)} = (Z_{1,y}(\lambda), Z_{2,y}(\lambda), \dots, Z_{n,y}(\lambda))$ represents a random sample from a $N(0,1)$ distribution.

Then, the combined $Z_{x,y}(\lambda) = (Z_x(\lambda), Z_y(\lambda))$ represents a sample

$$(Z_{1,x}(\lambda), Z_{2,x}(\lambda), \dots, Z_{m,x}(\lambda), Z_{1,y}(\lambda), Z_{2,y}(\lambda), \dots, Z_{n,y}(\lambda))$$

of size $N = m + n$ from a $N(0,1)$ distribution, which for the simplicity of the presentation we write as:

$$Z(\lambda) = (Z_1(\lambda), Z_2(\lambda), \dots, Z_N(\lambda))$$

We propose to estimate λ by $\hat{\lambda}_n$ in a way that $Z(\hat{\lambda}_n)$ is as close as possible to the true $N(0,1)$ distribution. Viewing this problem as a goodness-of-fit to normal distribution, we test the hypothesis:

$H_0: Z_1(\lambda), Z_2(\lambda), \dots, Z_N(\lambda)$ is coming from a $N(0,1)$ distribution, against

$H_1: Z_1(\lambda), Z_2(\lambda), \dots, Z_N(\lambda)$ is not a $N(0,1)$ distribution.

Following Shapiro and Wilk (1965), we use the test statistic $W_Z(\lambda)$ to test H_0 , which is given by

$$W_Z(\lambda) = \frac{[\sum_{i=1}^N a_i Z_{(i)}(\lambda)]^2}{\sum_{i=1}^N (Z_i(\lambda) - \bar{Z}_i(\lambda))^2}, \text{ where}$$

$Z_{(i)}(\lambda), i = 1, \dots, N$ represents the i th order statistic of the sample $Z(\lambda)$,

$\bar{Z}_i(\lambda) = (\sum_{i=1}^N Z_i(\lambda))/N$ is the sample mean,

$$(a_1, \dots, a_N) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

$$m = (m_1, \dots, m_N)^T,$$

$m_i = E(Z_{(i)}(\lambda)), i = 1, \dots, N$, is the expected value of the i th order statistic $Z_{(i)}(\lambda)$,

$V = (v_{i,j})$ is the variance-covariance matrix of order $N \times N$, and

$v_{i,j} = Cov(Z_{(i)}(\lambda), Z_{(j)}(\lambda)), i, j = 1, \dots, N$, is the covariance between i th and j th order statistics.

While the value of $W_Z(\lambda)$ lies between zero and one, the small value of $W_Z(\lambda)$ leads to the rejection of normality, whereas a value close to one indicates normality. Given a level of significance α , one may reject the null hypothesis if p -value $p(\lambda) = P(W \leq w_Z(\lambda)) \leq \alpha$ and accept otherwise. We propose to estimate λ by observing the maximum p -value associated with $W_Z(\lambda)$ over all possible values of λ to achieve the desired normality of the transformed data. In other words, the new estimate $\hat{\lambda}_n$ using the univariate goodness-of-fit to $N(0,1)$ distribution satisfies the equation

$$p(\hat{\lambda}_n) = \max_{\lambda \in I} P(W \leq w_Z(\lambda)) \tag{3}$$

where I is a pre-specified set of values of λ considered in the search. In this article, the search for $\hat{\lambda}_n$ is made over the interval $[-1,1]$ with an increment of 0.1, and therefore, hereafter, we express it by $I = \{-1:0.1:1\}$. Once $\hat{\lambda}_n$ is obtained, we re-express the original samples and apply Student's t -test to the transformed data. We employ the software R in all examples and simulation to obtain the optimum $\hat{\lambda}_n$ and other computational purposes.

An algorithm for the estimate $\hat{\lambda}_n$ and the transformed test using $\hat{\lambda}_n$ is as follows:

Given X and Y and a fixed $\lambda \in I = \{-1:0.1:1\}$,

- i. Obtain the transformations to X and Y , $X(\lambda)$ and $Y(\lambda)$ using equation (1).
- ii. Estimate $Z_x(\lambda) = \frac{X(\lambda) - \bar{X}(\lambda)}{S_x(\lambda)}$ and $Z_y(\lambda) = \frac{Y(\lambda) - \bar{Y}(\lambda)}{S_y(\lambda)}$, where $S_x(\lambda)$ and $S_y(\lambda)$ are estimated using the transformed data by $S_x(\lambda) = \sqrt{\sum_{i=1}^m (x_i(\lambda) - \bar{x}(\lambda))^2 / m}$ and $S_y(\lambda) = \sqrt{\sum_{j=1}^n (y_j(\lambda) - \bar{y}(\lambda))^2 / n}$. The term $X(\lambda) - \bar{X}(\lambda)$ refers to element-wise subtraction of sample mean $\bar{X}(\lambda)$ from the vector $X(\lambda)$ and similar operation applies to $Y(\lambda) - \bar{Y}(\lambda)$, which are allowed by any standard statistical software.
- iii. Combine the two samples together to form $Z(\lambda) = (Z_1(\lambda), Z_2(\lambda), \dots, Z_N(\lambda))$, where $N = m + n$.
- iv. Compare $Z(\lambda)$ with the $N(0,1)$ distribution using the Shapiro-Wilk goodness-of-fit and find the p -value.
- v. Repeat steps (i)-(iv) for all $\lambda \in I = \{-1:0.1:1\}$.
- vi. Select the maximum p -value among all p -values from steps (i)-(v).
- vii. $\hat{\lambda}_n$ is the value of λ corresponding to the maximum p -value in step (vi).
- viii. Obtain transformations $X(\hat{\lambda}_n)$ and $Y(\hat{\lambda}_n)$.
- ix. Perform usual t -test based on transformed data in step (viii) and decide about the acceptance and rejection of the null hypothesis comparing with critical value of t distribution.

4. Applications

In this section, we will present two examples, one with real life data and the other with simulated data from a skewed distribution to show application and performance of various tests in making inference about acceptance or rejection of the equality of two population means.

Example 1

In this example, X and Y refer to sample data of checkout times, in minutes, of two grocery checkers. This data is due to Verzani (2005).

X	5.8	1.0	1.1	2.1	2.5	1.1	1.0	1.2	3.2	2.7
Y	1.5	2.7	6.6	4.6	1.1	1.2	5.7	3.2	1.2	1.3

The summary statistics for X and Y are, mean=2.17 and skewness=1.71 (for X); mean=2.91 and skewness=0.86 (for Y). The histograms and boxplots of X and Y are presented in Figure 1 to understand the shape of the simulated data.

Based on values of skewness and the shape from the histograms and boxplots, both samples X and Y seem to have positively skewed distributions. Let the population mean difference $\Delta x = \mu_x - \mu_y$. The results of the test $H_0: \Delta x = 0$ against the two-sided alternative $H_1: \Delta x \neq 0$ using various tests discussed in this article appear in Table 1.

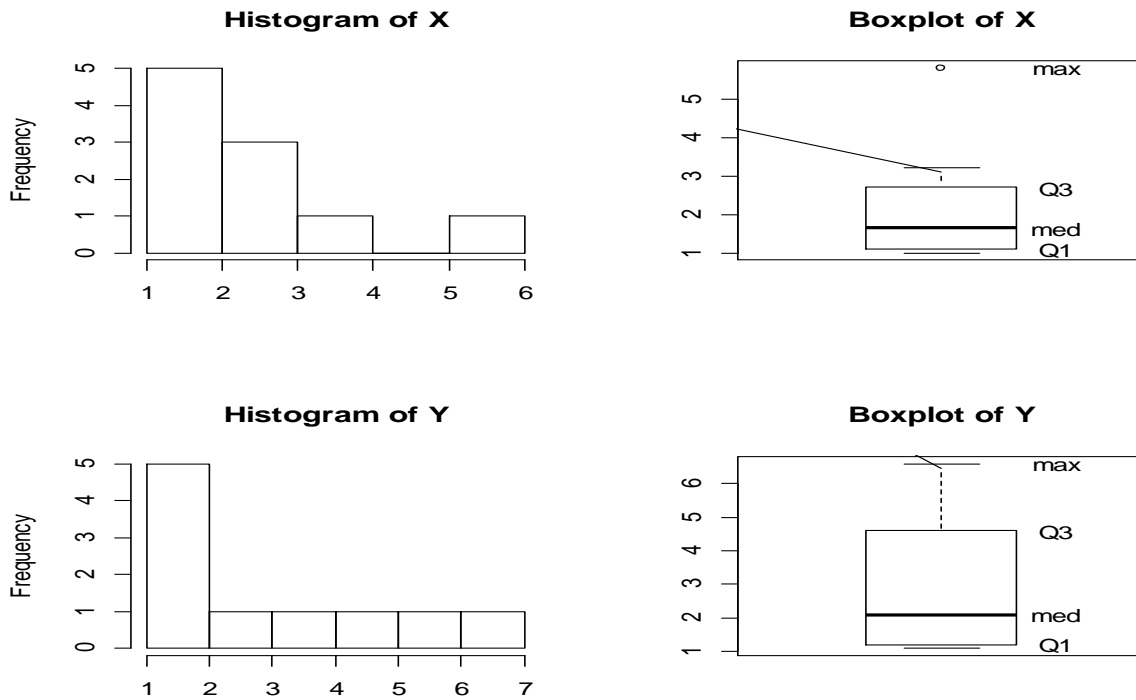


Figure 1. Histograms and boxplots of samples X and Y in Example 1 with their shapes

Table 1. Test statistics and p -values for various tests for data in Example 1

Tests**	Test statistic	p -value	$\hat{\lambda}$
t	-0.9164	0.3716	-
w	34.000	0.2394	-
$t(l)$	-0.8201	0.4229	-0.6
$t(n)$	-0.8415	0.4111	-0.1

** t : Student's t -test applied to untransformed data

w : Wilcoxon test

$t(l)$: transformed t -test by a maximum likelihood method

$t(n)$: new transformed t -test by a normal goodness-of-fit method

Based on the results of various tests in Table 1, it follows that all four tests provide identical conclusion of the acceptance of the null hypothesis at 5% level of significance, with transformed two tests, $t(l)$ and $t(n)$, outperforming the other two tests with p -values 0.4229 and 0.4111.

It is to be noted that the conclusion of t -test, whatever it is, may be misleading because data do not provide any evidence of normality, a violation of applicability of t -test. Wilcoxon test assumes that the two distributions are identical, and is a popular alternative to Student's t -test for comparing two populations with respect to locations (medians). On the other hand, the conclusion of both transformed t -tests appears to be valid because transformations were intended to achieve normality.

Example 2

For this example, we simulate sample X from a $G(2,1)$ distribution and the sample Y from $0.8 + G(2,1)$ distribution. Thus, in the population distributions of X and Y , an absolute mean difference is $|\Delta x| = |\mu_x - \mu_y| = 0.8$.

In other words, we simulate two samples X and Y under alternative hypothesis $H_1: \Delta x \neq 0$. For the convenience of the presentation, we round up the values of the simulated data to two decimal places and are presented as follows:

<i>X</i> :	1.06	1.88	3.68	1.13	2.08	4.84	1.42	1.29	0.37	2.43
<i>Y</i> :	1.73	2.74	1.85	3.74	1.95	2.73	4.25	2.35	1.94	2.08
	2.45	2.49	3.08	1.71	3.48					

We wish to test $H_0: \Delta x = 0$ against the two sided alternative $H_1: \Delta x \neq 0$ using various tests discussed in this article.

The summary statistics of two simulated samples are as follows: for sample *X*, mean=2.02 and skewness=1.17; for sample *Y* mean=2.57 and skewness=0.89. The histograms and boxplots of *X* and *Y* are presented in Figure 2 to understand the shape of the simulated data.

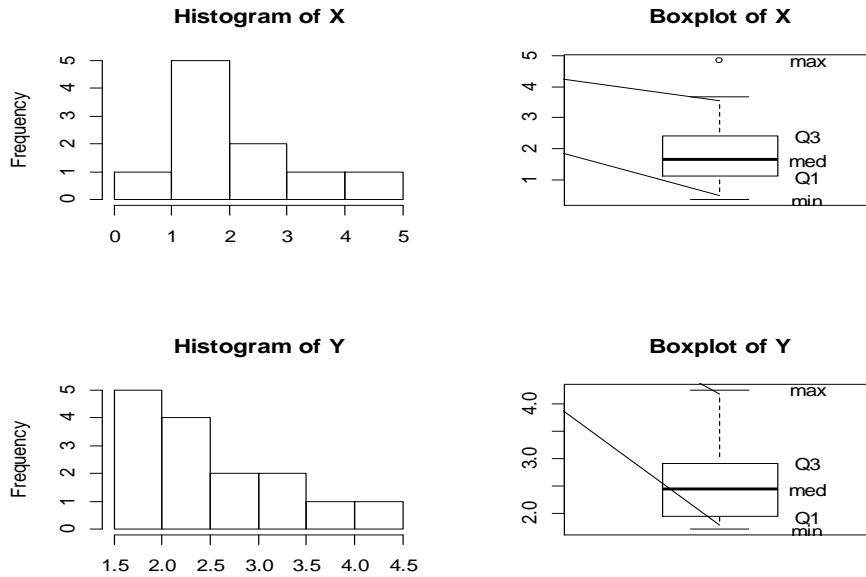


Figure 2. Histograms and boxplots of samples *X* and *Y* in Example 2 with their shapes

Since the samples *X* and *Y* come from the populations with identical variance but different means, we expect that various test statistics would be able to assess the inequality of the two means with stronger evidence. The results of various tests with corresponding *p*-values are reported in Table 2.

Table 2. Test statistics and *p*-values for various tests for simulated data in Example 2

Tests	Test statistic	<i>p</i> -value	λ
<i>t</i>	-1.3101	0.2031	-
<i>w</i>	43.500	0.8055	-
<i>t(l)</i>	-1.696	0.1034	0.5
<i>t(n)</i>	-2.0487	0.0487	-0.2

The results of various tests in Table 2, suggest that among all four tests only the new proposed test *t(n)* test provides the correct decision by rejecting the null hypothesis of equality of two means with *p*-value 0.0487 at 5% level of significance. Other tests provide evidence of accepting the null hypothesis with *p* values 0.2031 (untransformed test *t*), 0.8055 (Wilcoxon test *w*) and 0.1034 (transformed test based on the ML method), respectively.

Based on the performances of two examples presented, it seems reasonable to recommend the new transformed test for skewed data.

5. Simulation Study

In this section, we carry out a simulation study to compare the finite sample performance of the various tests described in this article, along with the proposed *t*-test. All simulations are performed by using the statistical software R, with values of $\lambda \in I = \{-1:0.1:1\}$. Under the null model, the samples *X* and *Y* are simulated from $G(a, b)$ population where *a* is the shape parameter and *b* is the scale parameter. Under alternative model, the samples *X* and *Y* are simulated from $G(a, b) + \Delta x$ and $G(a, b)$ populations, respectively, with the mean difference $\Delta x > 0$. The mean difference Δx is arbitrarily chosen from the set {0.15, 0.25, 0.50, 0.65, 1.25} to ensure a testing power away from 0 and 1 for the purpose of the comparisons. Note that the skewness of $G(a, b)$ distribution is $\gamma_1 = 2/\sqrt{a}$. In simulations, we choose different values of the parameter *a* to allow varying levels of skewness of the simulated samples. We fix the value of the parameter *b* at 1 since it does not affect the skewness of the simulated data. In all simulations, the Monte Carlo size is considered 5,000. The power of various tests is estimated from the proportion of rejection of null hypothesis under alternative over a Monte Carlo simulation of size 5,000 at 5% level of significance. In a similar manner, the level of significance is estimated from the proportion of the rejection of the null hypothesis over a Monte

Carlo simulation of size 5,000 at 5% level of significance when the null hypothesis is true. Table 3 provides the values of the parameter a used in the simulation of samples X and Y to allow varying values of the skewness.

Table 3. Values of a and γ_1 used in simulations of X and Y

Shape parameter a	Skewness γ_1
0.25	4.0
0.5	2.8
1	2.0
2	1.4
10	0.6

Table 4 provides estimated power of the simulation study for varying values of shape parameter a , sample sizes (m, n) and the mean difference $\Delta x = \mu_x - \mu_y$.

Table 5 provides estimated rejection rates under the null distributions at 5% level of significance, along with mean and standard deviation of the estimated transformation parameter λ by maximum likelihood ($\hat{\lambda}_l$) and univariate goodness of fit technique ($\hat{\lambda}_n$) over a Monte Carlo simulation of size 5,000.

6. Results and Discussion

The results of Example 1 in section 4 suggest that all the four tests applied to compare means of checkout times of two grocery checkers lead to the identical conclusion of acceptance of equality of two locations with p -values 0.3716 (Student's t), 0.2394 (Wilcoxon test) and 0.4229 (transformed t -test by a maximum likelihood method) and 0.4111. However, given the fact that the X and Y seem to have a positive skewed distributions, as are evident by histograms and boxplots in Figure 1, and the skewness (1.71 for X and 0.86 for Y), one may be doubtful about the conclusion of the Student's t -test. In reference to the conclusions of four tests applied to Example 2, only the proposed new test $t(n)$ could make a correct conclusion of the rejection of the null hypothesis given the fact that the data were generated under the alternative. Thus, the proposed test outperforms other tests in the right decision-making.

Table 4. Simulated power of various tests at 5% significance level over 5,000 samples

a	(m, n)	$\Delta x = 0.15$				$\Delta x = 0.25$			
		t	w	$t(l)$	$t(n)$	t	w	$t(l)$	$t(n)$
0.25	(10,10)	0.173	0.478	0.584	0.657	0.341	0.656	0.766	0.817
	(15,15)	0.199	0.668	0.761	0.872	0.399	0.838	0.911	0.965
	(20,20)	0.219	0.799	0.873	0.958	0.463	0.936	0.968	0.993
	(25,25)	0.255	0.879	0.934	0.985	0.506	0.971	0.990	0.999
	(15,10)	0.169	0.577	0.702	0.809	0.362	0.740	0.852	0.888
	(20,15)	0.212	0.722	0.822	0.937	0.413	0.874	0.941	0.980
	(25,20)	0.241	0.825	0.908	0.977	0.481	0.944	0.978	0.996
0.50	(10,10)	0.078	0.180	0.230	0.270	0.159	0.324	0.396	0.462
	(15,15)	0.097	0.264	0.332	0.426	0.200	0.469	0.551	0.678
	(20,20)	0.120	0.354	0.422	0.550	0.241	0.599	0.668	0.807
	(25,25)	0.131	0.420	0.505	0.646	0.263	0.705	0.769	0.890
	(15,10)	0.080	0.241	0.311	0.431	0.173	0.392	0.490	0.645
	(20,15)	0.102	0.299	0.387	0.538	0.224	0.528	0.629	0.793
	(25,20)	0.124	0.388	0.490	0.648	0.257	0.626	0.726	0.868
1	(10,10)	0.222	0.301	0.362	0.403	0.346	0.452	0.516	0.569
	(15,15)	0.303	0.458	0.525	0.609	0.453	0.620	0.700	0.784
	(20,20)	0.379	0.591	0.647	0.740	0.553	0.772	0.811	0.886
	(25,25)	0.424	0.682	0.740	0.825	0.638	0.849	0.892	0.941
	(15,10)	0.259	0.390	0.467	0.561	0.399	0.538	0.627	0.729
	(20,15)	0.326	0.517	0.600	0.703	0.503	0.689	0.763	0.859
	(25,20)	0.397	0.615	0.707	0.806	0.600	0.793	0.852	0.923
2	(10,10)	0.181	0.196	0.236	0.255	0.274	0.301	0.348	0.376
	(15,15)	0.250	0.304	0.347	0.372	0.381	0.450	0.502	0.541
	(20,20)	0.313	0.399	0.445	0.475	0.484	0.587	0.641	0.681
	(25,25)	0.365	0.472	0.530	0.570	0.565	0.678	0.730	0.774
	(15,10)	0.212	0.254	0.312	0.344	0.316	0.371	0.439	0.485
	(20,15)	0.269	0.330	0.395	0.435	0.430	0.509	0.582	0.632
	(25,20)	0.334	0.421	0.485	0.526	0.522	0.627	0.689	0.737
10	(10,10)	0.087	0.073	0.093	0.100	0.129	0.121	0.138	0.145
	(15,15)	0.107	0.104	0.121	0.125	0.190	0.183	0.207	0.216
	(20,20)	0.131	0.132	0.141	0.149	0.239	0.236	0.256	0.265
	(25,25)	0.157	0.159	0.172	0.178	0.283	0.289	0.308	0.317
	(15,10)	0.097	0.091	0.105	0.115	0.146	0.148	0.164	0.174
	(20,15)	0.120	0.116	0.133	0.141	0.210	0.204	0.232	0.240
	(25,20)	0.149	0.146	0.163	0.169	0.246	0.249	0.273	0.283

Table 5. Estimated rejection rates at 5% level, along with average and standard deviation (S.D) of $\hat{\lambda}$

a	(m, n)	Estimated level of significance				Average		S.D	
		t	w	$t(l)$	$t(n)$	$\hat{\lambda}_l$	$\hat{\lambda}_n$	$\hat{\lambda}_l$	$\hat{\lambda}_n$
0.25	(10,10)	0.031	0.041	0.051	0.054	0.136	0.170	0.062	0.081
	(15,15)	0.036	0.048	0.056	0.056	0.137	0.167	0.053	0.075
	(20,20)	0.037	0.050	0.053	0.054	0.137	0.163	0.050	0.061
	(25,25)	0.042	0.051	0.055	0.056	0.136	0.163	0.048	0.055
	(15,10)	0.035	0.052	0.055	0.056	0.137	0.168	0.056	0.087
	(20,15)	0.040	0.047	0.050	0.054	0.138	0.166	0.050	0.065
	(25,20)	0.039	0.051	0.053	0.055	0.136	0.164	0.049	0.058
0.50	(10,10)	0.042	0.045	0.054	0.055	0.194	0.226	0.102	0.171
	(15,15)	0.038	0.036	0.045	0.051	0.200	0.228	0.078	0.115
	(20,20)	0.040	0.044	0.047	0.050	0.204	0.227	0.065	0.088
	(25,25)	0.041	0.045	0.047	0.049	0.204	0.226	0.058	0.077
	(15,10)	0.042	0.044	0.050	0.054	0.198	0.227	0.087	0.132
	(20,15)	0.043	0.047	0.052	0.053	0.203	0.228	0.070	0.100
	(25,20)	0.040	0.043	0.047	0.051	0.203	0.227	0.061	0.081
1	(10,10)	0.040	0.040	0.046	0.051	0.241	0.271	0.170	0.254
	(15,15)	0.047	0.045	0.052	0.055	0.249	0.273	0.130	0.177
	(20,20)	0.045	0.048	0.049	0.053	0.253	0.271	0.109	0.139
	(25,25)	0.043	0.047	0.047	0.049	0.257	0.273	0.096	0.118
	(15,10)	0.049	0.051	0.057	0.058	0.246	0.274	0.149	0.208
	(20,15)	0.050	0.049	0.052	0.054	0.250	0.272	0.117	0.153
	(25,20)	0.044	0.047	0.050	0.053	0.255	0.273	0.102	0.128
2	(10,10)	0.040	0.039	0.047	0.053	0.269	0.293	0.273	0.377
	(15,15)	0.052	0.046	0.054	0.054	0.271	0.289	0.211	0.270
	(20,20)	0.046	0.045	0.051	0.054	0.282	0.297	0.175	0.209
	(25,25)	0.048	0.048	0.050	0.051	0.289	0.302	0.155	0.181
	(15,10)	0.044	0.044	0.048	0.051	0.269	0.287	0.236	0.307
	(20,15)	0.049	0.045	0.052	0.055	0.280	0.297	0.194	0.239
	(25,20)	0.052	0.049	0.053	0.054	0.279	0.291	0.164	0.193
10	(10,10)	0.047	0.042	0.050	0.053	0.260	0.238	0.565	0.653
	(15,15)	0.043	0.038	0.046	0.048	0.276	0.276	0.475	0.542
	(20,20)	0.045	0.044	0.048	0.051	0.304	0.302	0.414	0.462
	(25,25)	0.051	0.051	0.052	0.053	0.299	0.307	0.374	0.414
	(15,10)	0.044	0.043	0.048	0.052	0.276	0.264	0.508	0.586
	(20,15)	0.052	0.048	0.055	0.056	0.282	0.280	0.443	0.503
	(25,20)	0.049	0.046	0.051	0.054	0.292	0.295	0.391	0.437

As we look at the simulation results presented in Table 4, it is evident that the new transformed test $t(n)$ provides the maximum power for all sample sizes, equal ($m = n$) and unequal ($m \neq n$), among all four tests considered. We consider equal sample sizes ($m = n$) at 10, 15, 20 and 25. Note that the lower value of the shape parameter a corresponds to the higher value of the skewness. To evaluate the performance for varying values of skewness, we consider values of a from 0.25 to 10 with arbitrary increases to its values to cause skewness to decrease from 4 to 0.6 as appeared in Table 3. It is also evident that all tests demonstrate higher power as mean difference Δx and sample size increase. The new test $t(n)$ has always performed the best in terms of estimated testing power; the second best has been the $t(l)$ test. However, as expected, the nonparametric test w has demonstrated higher power than the Student's t -test. Also, the differences in power among four tests have decreased as the skewness of the distribution has decreased. It makes sense because Wilcoxon and transformed tests are expected to perform better for skewed distribution; the higher the skewness, the better is their performance with respect to the testing power. As we see, overall, the proposed new test $t(n)$ outperforms all other three tests in terms of the power, for all sample size and skewness considered in the simulation.

From the simulated results presented in Table 5, it appears that the estimated level of significance for Student's t ranges from 0.031 to 0.052, for a 5% nominal level of significance, throughout the simulation, under null hypothesis. Indeed, the estimated levels of significance seem to be underestimated for all sample sizes for highly skewed distributions (e.g., $a = 0.25, 0.50$) and approach the nominal level as the skewness decreases ($a = 1, 2, 10$). The estimated rejection rates for Wilcoxon test is close to the nominal level of 5%, with estimated values ranging from 0.036 to 0.052, under null hypothesis. On the other hand, the estimated rejection rates for both versions of transformed tests are comparable at 5% level of significance, with estimated values ranging from 0.045 to 0.057 for $t(l)$ test, and 0.048

to 0.058 for $t(n)$ test, under null hypothesis.

The estimated average and standard deviation of $\hat{\lambda}_l$ and $\hat{\lambda}_n$ over 5,000 simulations under null hypothesis are also reported in Table 5, where the search for $\hat{\lambda}_l$ and $\hat{\lambda}_n$ is made in the interval $[-1,1]$ with an increment of 0.1. It follows that the average and standard deviation of $\hat{\lambda}_l$ and $\hat{\lambda}_n$ depend on the levels of skewness of the distributions, with standard deviation of both decreasing with the increase of the sample sizes for a given value of skewness. In terms of average and standard deviation values of $\hat{\lambda}_l$ and $\hat{\lambda}_n$, similar conclusions apply under the alternative hypothesis where powers are calculated and therefore, are not reported in Table 4 to avoid redundancy.

7. Concluding Remarks

This article proposes a new transformed t -test where the Box-Cox transformation to normality is achieved via a univariate normal goodness-of-fit test. To this end, we i) apply Shapiro and Wilk test to the combined standardized transformed samples to fit into the $N(0,1)$ distribution, ii) estimate the best transformation to normality by observing the maximum p -value from the Shapiro and Wilk test for all possible values of $\lambda \in \{-1:0.1:1\}$ and iii) apply student's t -test to the best normal transformed samples to compare location parameters (means). The performance of the new test over Student's t -test, Wilcoxon test and an existing transformed t -test achieved via likelihood method has been justified by two examples, and simulations where data comes from skewed distributions (gamma distribution). It is evident that the new test is appropriate for estimating the level of significance and is more powerful than other three tests considered for skewed distributions. It is also clear that higher the skewness, the better are the transformed t -tests in terms of the testing power, with the new transformed $t(n)$ test performing the best. It makes sense because if the data is less skewed or almost no skewed at all, the power transformation will not be needed or appropriate. It follows that the power of all tests is sensitive to the mean difference and sample size; the power of all tests increases with the increase in the mean difference of two population means and the size of the samples. Given the performance of the proposed new t -test, in terms of estimated power under the alternative hypothesis, and estimated level of significance under the null hypothesis, researchers can practice the proposed test with confidence. Overall, the Wilcoxon test is better in power than the Student's t -test and transformed t -tests are better than the Wilcoxon test with the new proposed test $t(n)$ demonstrating the highest power. If researchers are too concern about the estimated level of significance, they might consider Wilcoxon test because of its robustness. However, if power is of the concern, the new test performs the best.

Acknowledgements

The authors would like to thank the editorial team and referees for useful comments and suggestions.

References

- Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford University Press, London.
- Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.*, 76, 296-311. <https://doi.org/10.1080/01621459.1981.10477649>
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformation (with discussion). *J. Roy. Statist. Soc. Ser. B.*, 26, 211-252.
- Chen, H. (1995). Tests following transformations. *Ann. Statist.*, 23, 1587-1593. <https://doi.org/10.1214/aos/1176324314>
- Chen, H., & Loh, W. Y. (1992). Bounds on ARE's of tests following Box-Cox transformations. *Ann. Statist.*, 20, 1485-1500. <https://doi.org/10.1214/aos/1176348780>
- Desu, M. M., & Raghavarao, D. (2004). *Nonparametric Statistical Methods for Complete and Censored Data*. Chapman & Hall/CRC Press. <https://doi.org/10.1201/9781482285895>
- Gibbons, D., & Chakraborti, S. (2014). *Nonparametric Statistical Inference*. (4th ed.). Revised and Expanded. Taylor & Francis.
- Hernandez, F., & Johnson, R. A. (1980). The large-sample behavior of transformations to normality. *J. Amer. Statist. Assoc.*, 75, 855-861. <https://doi.org/10.1080/01621459.1980.10477563>
- Hinkley, D. V. (1975). On Power transformations to symmetry. *Biometrika*, 62, 101-111. <https://doi.org/10.1093/biomet/62.1.101>
- Islam, M. K., & Chen, H. (2007). Transformed test for Homogeneity of variances. *Journal of Probability and Statistical Science*, 5, 151-170.
- Kvam, P. H., & Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering*, Wiley. <https://doi.org/10.1002/9780470168707>

- Lehmann, E. L. (1994). *Testing Statistical Hypotheses*. Chapman & Hall, New York and London.
- Mann, H. B., & Whitney, D. R. (1947). On a test whether one of the two random samples is stochastically larger than the other. *Ann. Math. Statist.*, *18*, 50-60. <https://doi.org/10.1214/aoms/1177730491>
- Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley: Reading, MA.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Verzani, J. (2005). *Using R for Introductory Statistics*. Chapman & Hall/CRC Press.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80-83. <https://doi.org/10.2307/3001968>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).