J. Chosun Natural Sci.
Vol. 15, No. 1 (2022) pp.9-18
https://doi.org/10.13160/ricns.2022.15.1.9

Comparison of the Power of Bootstrap Two-Sample Test and Wilcoxon Rank Sum Test for Positively Skewed Population

Sunyeong Heo†

Department of statistics, Changwon National University, Changwon, Korea

Abstract

This research examines the power of bootstrap two-sample test, and compares it with the powers of two-sample $t-$test and Wilcoxon rank sum test, through simulation. For simulation work, a positively skewed and heavy tailed distribution was selected as a population distribution, the chi-square distributions with three degrees of freedom, χ^2_3. For two independent samples, the fist sample was selected from χ^2_3. The second sample was selected independently from the same χ^2_3 as the first sample, and calculated $d+ax$ for each sampled value x, a randomly selected value from χ^2_3. The d in $d+ax$ has from 0 to 5 by 0.5 interval, and the a has from 1.0 to 1.5 by 0.1 interval. The powers of three methods were evaluated for the sample sizes 10,20,30,40,50. The null hypothesis was the two population medians being equal for Bootstrap two-sample test and Wilcoxon rank sum test, and the two population means being equal for the two-sample $t-$test. The powers were obtained using r program language; wilcox.test() in r base package for Wilcoxon rank sum test, t.test() in r base package for the two-sample $t-$ test, boot.two.bca() in r wBoot pacakge for the bootstrap two-sample test.

Simulation results show that the power of Wilcoxon rank sum test is the best for all 330 (n, a, d) combinations and the power of two-sample $t-$ test comes next, and the power of bootstrap two-sample comes last. As the results, it can be recommended to use the classic inference methods if there are widely accepted and used methods, in terms of time, costs, sometimes power.

Keyword: Bootstrap inference, chi-square distribution, power of test, two-sample $t-$test, Wilcoxon rank sum test.

(**Recceived** February 21, 2022; **Revised** March 8, 2022; **Accepted** March 18, 2022)

1. Introduction

The comparison of two independent samples is a fundamental inference procedure in statistics. Data scientists often use the terms of A-B test instead of the two-sample test. The two-sample test has been widely applied in many research fields: education, psychology, chemistry, marketing, clinical trials, and so on. In the two-sample problem. the most general application is the testing of equality between two location measures.

Traditional way for testing the equality of two location parameters is the two-sample $t-$

† Corresponding author: syheo@changwon.ac.kr

test. Two-sample t-test is a parametric procedures to test the equality of two population means, and assumes the sampled populations having normal distributions. When the sampled populations are unknown or non-normally distributed, it requires large sample sizes for applying central limit theory.

However, there are many situations in which the distribution of sampled population is unknown or non-normal but the sample sizes are not large enough to apply central limit theorem.

Nonparametric inference methods do not depend on the specific distribution of the sampled population, and so is often called distribution-free method[1].

A nonparametric alternative to two-sample t-test is Wilcoxon rank sum test. Wilcoxon rank sum test assumes only the sampled populations having continuous distribution, no matter what shape, and is based on the sum of ranks of the sample, having the smaller or equal sample size between two samples, when two samples are pooled into a single ordered array. When a population is non-normally distributed, the median is generally much more appropriate location parameter than mean. The null hypothesis of Wilcoxon rank sum test is that the two sampled populations have the same distribution except for possible difference between two location measures, medians. More details about Wilcoxon rank sum test refer to Gibbons (1993), Conover (1980), Sprent and Smeeton (2001), and so on.[1-3].

Another nonparametric alternative to two-sample t-test is bootstrap hypothesis test. Bootstrap inference does not assume a specific population distribution nor require the sampling distribution of the statistic to be used for testing.

Bootstrap is one of resampling methods. Bootstrap procedure resamples many subsamples with replacement from the original sample, and calculates the test statistic of interest from each resample. This process allows to estimate the distribution of the statistic of interest, and to estimate the test statistics's standard error, and to produce confidence interval about the parameter being concerned. and to perform hypothesis test. The best thing about bootstrap inference is that the inference is possible for numerous statistics when their exact forms of sampling distributions are unknown; for instance, a difference between two sample medians, a ratio of two correlation coefficients, and more complicate statistics. One of weaknesses of bootstrap procedure is that it requires a lot of computation, and such a shortcoming makes bootstrap method to be rarely used for a long time even though it has been developed long ago. However, the rapid advances in computing technology and the lowered computer prices recently make many researchers use this method instead of classic statistical methods. It is easy to see that data analysts, who have interested in big data analysis, use bootstrap method for data analyses.

This research is for comparing the power of the bootstrap inference with the classic statistical inference's through simulation. The two-sample t-test is the most powerful test when the sampled population is normally distributed. So, an asymmetric and positively skewed distribution is selected as a population distribution. From the population distribution, two independent samples are defined and selected. From the selected samples, the powers of two-sample t-test, Wilcoxon rank sum test, and bootstrap test were evaluated

and compared.

2. Bootstrap Two-Sample Test

The justification of bootstrap inference rests on three aspects; the similarity of the sample distribution with the population distribution, the original sample size n, and the number of resamples, R[4].

Bootstrapping considers the original sample as the population, and selects replicate samples from the original sample with replacement. So, the original sample must well represent the sampled population.

The sample size, n, and the number of resamples, R, for the justification of bootstrap inference depend on the hypothesis being tested and the level of significance[1,4]. Mooney and Duval (1993) suggest the original sample size of 30~50. Efron and Tibshirani (1986) suggests the number of resamples R of at least 1,000 for the confidence interval estimation at the level of significance $\alpha = 0.05$ [6]. For More about n and R refer to Mooney and Duval (1993) and references therein.

Hypothesis test is directly connected with confidence interval. If a test statistic value calculated from a sample is located within $1-\alpha$ level acceptance region, then we accept the null hypothesis, and reject it if not[7].

There are several methods for bootstrap confidence intervals[4,8]. Mooney and Duval (1993) presents four methods; the normal approximation method, the percentile method, the bias-corrected method (BC method), and the percentile-t method.

The normal approximation method is useful when the test statistic has normal distribution but its variance is unknown. The percentile method constructs an $(1-\alpha)100\%$ confidence interval that includes all values of $\hat{\theta}_i$, the test statistic value from the ith resample, between the $100(\alpha/2)^{\text{th}}$ and $100(1-\alpha/2)^{\text{th}}$ percentiles of the bootstrapped sampling distribution of $\hat{\theta}$, the test statistic. The BC method was suggested by Efron(1982) to overcome drawbacks of the percentile method, and to adjust the bootstrapped sampling distribution to center on the point estimator $\hat{\theta}$, a test statistic value calculated from the original sample. Finally, the percentil-t has proposed to overcome problems the BC method has. For more details about bootstrap confidence interval methods refer to Efron (1982), Efron and Tibshirani (1993), Mooney and Duval (1993), and Johnson (2001)[1,8-10].

This research used the BC method to calculate the powers of bootstrap two-sample test.

We can refer to many statistic textbooks for two-sample $t-$ test, and references herein for Wilcoxon rank sum test.

3. Power of Bootstrap Two-sample Test

3.1 Simulation design

Let x_1 and x_2 be independent random variables, and for constant (d, a),

$$\begin{aligned} x_1 &\sim \chi_3^2 \\ x_2 &\sim d + a\,\chi_3^2 \end{aligned} \tag{1}$$

where χ_3^2 is a chi-square distribution with three degrees of freedom.

Also assume that $x_{11}, x_{12}, \cdots, x_{1n}$ is a random sample from the distribution of x_1, and $x_{21}, x_{22}, \cdots, x_{2n}$ is a random sample from

the distribution of x_2 and independent of the first sample x_{1i}'s.

The skewness of chi-square distribution with three degrees of freedom, $p = 3$,

$$\alpha_3 = \sqrt{\frac{8}{p}} = \approx 1.63$$

and its kurtosis is

$$\alpha_4 = 3 + \frac{12}{p} = 7.$$

So the distribution of χ^2_3 is asymmetric and positively skewed, and has heavier tail than normal distribution.

The mean and variance of chi-square distribution with p degrees of freedom are p and $2p$. Therefore, the mean and variance of random variable x_1 are 3 and 6, and the mean and variance of x_2 are $d+3a$ and is $6a^2$.

The hypotheses we are here interested in are

$$H_0 : \theta_1 = \theta_2 \text{ vs. } H_1 : \theta_1 < \theta_2,$$

where θ_i is the central measure of the distribution of x_i, $i = 1, 2$. If two independent samples are selected from normal distribution, then θ_i will be μ_i, population mean. However, chi-square distribution with three degrees of freedom has asymmetric and heavy tail, median will be a more proper central measure than mean, and θ_i will be Me_i, the ith population median.

For simulation, d was chosen from 0 to 5 by 0.5 interval, and a from 1.0 to 1.5 by 0.1 interval, and n from 10 to 50 by 10 interval, and so the power was calculated for 330 (d, a, n) combinations for each method. When $d = 0$ and $a = 1$, two samples are selected from a equal distribution, and so two central measures are equal, too. On the other hand, when $d > 0$ and $a = 1$, the two sampled population have equal variance, but the population distribution of x_2 has bigger mean and median as much as d than the x_1's. When $d > 0$ and $a > 1$, two population distributions have different central measures and variances, and the x_2's distribution has bigger mean, median, and variance than x_1's.

The power of two-sample t-test was obtained using the function t.test() in r base r package such as

t.test(x_1, x_2, mu=0, var.equal=T, alt="less")

when $a = 1$. The option var.equal=T was changed to var.equal=F when $a > 1.0$.

For the power of Wilcoxon rank sum test, the function wilcox.test() in r base package was used such as

wilcox.test(x_1, x_2, mu=0, alt="less", exact=T)

when $n = 10$, and for $n > 10$ the option exact=F was used instead of exact=T. So, when $n = 10$, the power of Wilcoxon rank sum is the power of exact test, and when $n > 10$ the power of asymptotic test.

The power of bootstrap test was calculated using boot.two.bca() using the r package wBoot such as

boot.two.bca(x_1, x_2,median, stacked=F,
null.hyp=0, alt="less", R=1000)

where R=1000 means the number of bootstrap replicates.

3.2 Simulation results

Table 1. shows the power of two sample t-test for population mean differences. Table 2. shows the power of Wilcoxon rank sum test for two sample median differences, and Table 3. the powers of bootstrap test using the bias-corrected method (BC method). The powers for all three tests were calculated using the same data. The level of significance is $\alpha = 0.05$.

Table 4. shows the powers of three tests when $a = 1$, two populations having equal variance and different d. Fig 1. shows the power of three test when $a = 1$ and $n = 10$, and Fig 2. when $a = 1$ and $n = 50$. When the populations have equal variance, all three tests achieve the nominal level of significance, $\alpha = 0.05$, which is the power when $d = 0$, even though Wilcoxon rank sum test has the smallest significance level and a little less than $\alpha = 0.05$, and two sample t-test comes next but nearly equal to $\alpha = 0.05$ at all n, and the bootstrap two-sample test applied BC method has the biggest significance levels, which are slightly larger than the nominal level $\alpha = 0.05$ except $n = 30$. Fig 1, and Fig 2. shows that as d increase, the powers of three tests also increase, but the power of Wilcoxon rank sum test is the best, and then two sample t-test, and bootstrap test using BC method is the last. Two-sample t-test is developed for normal populations but it shows good performances for a relatively large positively skewed distribution like this χ^2_3, comparing to bootstrap test.

Table 1. through Table 3. shows that for all selected value (a, d), Wilcoxon rank sum test has the best power, and two sample t-test next, and the bootstrap test comes last.

Fig 1. Powers of two-sample test when variances are equal and sample size $n = 10$ for alternative differences of location parameters.

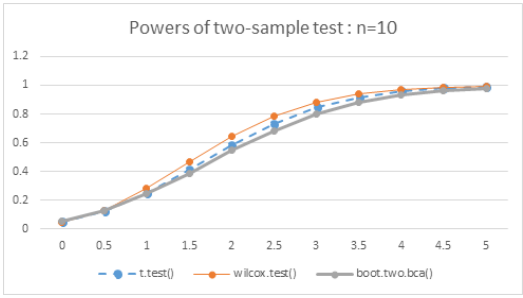

Fig 2. Powers of two-sample test when variances are equal and sample size $n = 50$ for alternative differences of location parameters.

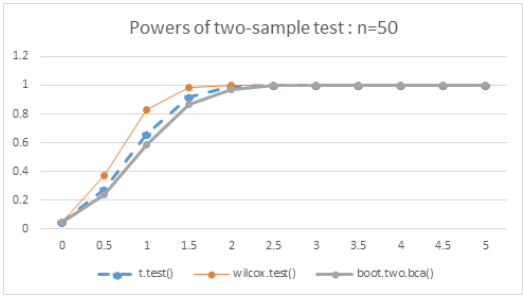

Table 1. Power of two sample t-test for testing $H_0 : \mu_1 = \mu_2$ vs. $H_0 : \mu_1 < \mu_2$ when two independent samples are selected from chi-square distribution with three degrees of freedom $(\alpha = 0.05)$.

a	sample size (n)	d										
		0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1.0	10	0.0491	0.1243	0.2497	0.4144	0.5805	0.7323	0.8449	0.9154	0.9560	0.9788	0.9907
	20	0.0480	0.1585	0.3726	0.6197	0.8193	0.9309	0.9752	0.9931	0.9986	0.9998	1.0000
	30	0.0461	0.1982	0.471	0.7624	0.9254	0.983	0.9968	0.9996	1.0000	1.0000	1.0000
	40	0.0493	0.2345	0.5749	0.8563	0.972	0.9965	0.9994	1.0000	1.0000	1.0000	1.0000
	50	0.0501	0.2752	0.6622	0.9176	0.9905	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000
1.1	10	0.0747	0.1684	0.3074	0.4770	0.6379	0.7761	0.8702	0.9317	0.9651	0.9831	0.9916
	20	0.0950	0.2518	0.4840	0.7214	0.8770	0.9537	0.9838	0.9957	0.9994	1.0000	1.0000
	30	0.1099	0.3224	0.6269	0.8549	0.9596	0.9921	0.9989	0.9998	1.0000	1.0000	1.0000
	40	0.1280	0.3983	0.7317	0.9275	0.9890	0.9986	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.1391	0.4722	0.8146	0.9662	0.9965	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
1.2	10	0.1096	0.2204	0.3746	0.5405	0.6931	0.8175	0.8988	0.9452	0.9736	0.9882	0.9939
	20	0.1591	0.3573	0.5950	0.7955	0.9165	0.9705	0.9907	0.9980	0.9998	1.0000	1.0000
	30	0.2055	0.4631	0.7414	0.9158	0.9794	0.9961	0.9993	0.9999	1.0000	1.0000	1.0000
	40	0.2489	0.5733	0.8455	0.9670	0.9953	0.9993	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.2956	0.6608	0.9105	0.9888	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.3	10	0.1531	0.2787	0.4364	0.5990	0.7411	0.8480	0.9190	0.9582	0.9797	0.9906	0.9955
	20	0.2443	0.4633	0.6939	0.8561	0.9452	0.9813	0.9946	0.9990	0.9998	1.0000	1.0000
	30	0.3199	0.6049	0.8383	0.9512	0.9894	0.9979	0.9997	1.0000	1.0000	1.0000	1.0000
	40	0.4008	0.7188	0.9163	0.9853	0.9981	0.9997	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.4802	0.8071	0.9609	0.9952	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.4	10	0.2000	0.3361	0.5007	0.6530	0.7845	0.8779	0.9339	0.9661	0.9845	0.9927	0.9968
	20	0.3425	0.5661	0.7691	0.9017	0.9627	0.9874	0.9965	0.9997	1.0000	1.0000	1.0000
	30	0.4510	0.7197	0.8993	0.9733	0.9945	0.9990	0.9998	1.0000	1.0000	1.0000	1.0000
	40	0.5649	0.8303	0.9592	0.9936	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.6543	0.9004	0.9843	0.9984	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1,5	10	0.2529	0.3974	0.5591	0.7008	0.8184	0.9003	0.9463	0.9734	0.9887	0.9942	0.9974
	20	0.4433	0.6610	0.8314	0.9326	0.9759	0.9922	0.9984	0.9997	1.0000	1.0000	1.0000
	30	0.5842	0.8147	0.9400	0.9852	0.9972	0.9995	0.9999	1.0000	1.0000	1.0000	1.0000
	40	0.7024	0.9024	0.9812	0.9970	0.9997	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.7939	0.9519	0.9943	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2. Power of Wilcoxon rank sum test for testing $H_0 : Me_1 = Me_2$ vs. $H_0 : Me_1 < Me_2$ when two independent samples are selected from chi-square distribution with three degrees of freedom ($\alpha = 0.05$).

a	sample size (n)	d										
		0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1.0	10*	0.0470	0.1313	0.2836	0.4710	0.6457	0.7873	0.8847	0.9414	0.9691	0.9843	0.9927
	20	0.0485	0.1986	0.4796	0.7567	0.9089	0.9742	0.9940	0.9989	1.0000	1.0000	1.0000
	30	0.0478	0.2608	0.6312	0.8944	0.9802	0.9973	0.9995	1.0000	1.0000	1.0000	1.0000
	40	0.0482	0.3190	0.7440	0.9555	0.9950	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.0487	0.3756	0.8323	0.9813	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.1	10*	0.0738	0.1792	0.3462	0.5319	0.7025	0.8272	0.9076	0.9541	0.9775	0.9895	0.9945
	20	0.0898	0.2918	0.5802	0.8216	0.9393	0.9851	0.9966	0.9995	1.0000	1.0000	1.0000
	30	0.1020	0.3898	0.7445	0.9372	0.9898	0.9988	0.9999	1.0000	1.0000	1.0000	1.0000
	40	0.1147	0.4815	0.8529	0.9788	0.9977	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.1273	0.5627	0.9149	0.9935	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.2	10*	0.1038	0.2295	0.4086	0.5910	0.7509	0.8589	0.9264	0.9641	0.9820	0.9915	0.9956
	20	0.1446	0.3916	0.6760	0.8728	0.9600	0.9905	0.9982	0.9999	1.0000	1.0000	1.0000
	30	0.1825	0.5212	0.8347	0.9663	0.9950	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000
	40	0.2224	0.6330	0.9167	0.9895	0.9993	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.2528	0.7237	0.9609	0.9976	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.3	10*	0.1394	0.2830	0.4632	0.6491	0.7884	0.8852	0.9427	0.9713	0.9869	0.9931	0.9964
	20	0.2159	0.4866	0.7561	0.9113	0.9754	0.9946	0.9989	1.0000	1.0000	1.0000	1.0000
	30	0.2862	0.6406	0.8938	0.9804	0.9973	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000
	40	0.3500	0.7561	0.9560	0.9949	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.4133	0.8397	0.9833	0.9993	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.4	10*	0.1797	0.3393	0.5224	0.6965	0.8214	0.9075	0.9551	0.9769	0.9899	0.9949	0.9974
	20	0.2998	0.5807	0.8158	0.9380	0.9831	0.9964	0.9996	1.0000	1.0000	1.0000	1.0000
	30	0.3996	0.7392	0.9338	0.9893	0.9984	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	40	0.4931	0.8478	0.9766	0.9981	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.5752	0.9096	0.9924	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1,5	10*	0.2278	0.3937	0.5745	0.7403	0.8535	0.9266	0.9634	0.9821	0.9915	0.9962	0.9983
	20	0.3884	0.6584	0.8621	0.9582	0.9902	0.9979	0.9999	1.0000	1.0000	1.0000	1.0000
	30	0.5161	0.8207	0.9613	0.9939	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	40	0.6257	0.9081	0.9880	0.9992	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.7163	0.9564	0.9971	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

* The power of Wilcoxon rank sum test when n=10 was obtained by the exact test.

Table 3. Power of bootstrap two-sample test for testing $H_0 : Me_1 = Me_2$ vs. $H_0 : Me_1 < Me_2$ when two independent samples are selected from chi-square distribution with three degrees of freedom ($\alpha = 0.05$).

a	sample size (n)	d										
		0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1.0	10	0.0560	0.1313	0.2433	0.3902	0.5458	0.6851	0.7973	0.8804	0.9324	0.9595	0.9786
	20	0.0512	0.1548	0.3346	0.5609	0.7539	0.8815	0.9537	0.9833	0.9956	0.9988	0.9999
	30	0.0500	0.1795	0.4169	0.6889	0.8744	0.9617	0.9903	0.9985	0.9999	1.0000	1.0000
	40	0.0521	0.2129	0.5120	0.7904	0.9413	0.9891	0.9982	0.9996	0.9999	1.0000	1.0000
	50	0.0512	0.2406	0.5865	0.8630	0.9716	0.9966	0.9998	1.0000	1.0000	1.0000	1.0000
1.1	10	0.0860	0.1707	0.2929	0.4492	0.5974	0.7285	0.8284	0.9005	0.9421	0.9674	0.9815
	20	0.0855	0.2226	0.4176	0.6378	0.8043	0.9097	0.9670	0.9879	0.9970	0.9994	0.9999
	30	0.0956	0.2661	0.5258	0.7691	0.9143	0.9763	0.9939	0.9989	0.9999	1.0000	1.0000
	40	0.1046	0.3227	0.6251	0.8586	0.9651	0.9940	0.9991	0.9998	0.9999	1.0000	1.0000
	50	0.1116	0.3721	0.7088	0.9159	0.9841	0.9984	1.0000	1.0000	1.0000	1.0000	1.0000
1.2	10	0.1159	0.2147	0.3456	0.5014	0.6444	0.7648	0.8517	0.9154	0.9500	0.9731	0.9850
	20	0.1325	0.2902	0.4982	0.7013	0.8440	0.9330	0.9759	0.9922	0.9982	0.9998	0.9999
	30	0.1531	0.3603	0.6208	0.8308	0.9441	0.9835	0.9964	0.9996	0.9999	1.0000	1.0000
	40	0.1800	0.4408	0.7262	0.9081	0.9799	0.9965	0.9996	0.9999	1.0000	1.0000	1.0000
	50	0.2003	0.5118	0.8068	0.9515	0.9921	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000
1.3	10	0.1511	0.2565	0.3981	0.5508	0.6863	0.7958	0.8736	0.9273	0.9577	0.9773	0.9875
	20	0.1919	0.2902	0.5706	0.7594	0.8771	0.9514	0.9825	0.9950	0.9984	0.9998	0.9999
	30	0.2281	0.4610	0.7071	0.8783	0.9614	0.9901	0.9983	0.9997	1.0000	1.0000	1.0000
	40	0.2752	0.5562	0.8037	0.9428	0.9883	0.9984	0.9996	0.9999	1.0000	1.0000	1.0000
	50	0.3120	0.6361	0.8750	0.9715	0.9955	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
1.4	10	0.1869	0.3038	0.4469	0.5992	0.7246	0.8220	0.8946	0.9376	0.9657	0.9805	0.9886
	20	0.2538	0.4402	0.6406	0.8041	0.9040	0.9648	0.9871	0.9968	0.9990	0.9999	1.0000
	30	0.3082	0.5546	0.7742	0.9128	0.9737	0.9931	0.9987	0.9999	1.0000	1.0000	1.0000
	40	0.3790	0.6545	0.8643	0.9643	0.9931	0.9989	0.9996	1.0000	1.0000	1.0000	1.0000
	50	0.4381	0.7422	0.9206	0.9835	0.9980	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1,5	10	0.2280	0.3516	0.4943	0.6378	0.7609	0.8467	0.9081	0.9473	0.9710	0.9838	0.9903
	20	0.3199	0.5080	0.6978	0.8419	0.9263	0.9740	0.9904	0.9973	0.9994	0.9999	1.0000
	30	0.3998	0.6376	0.8308	0.9397	0.9813	0.9952	0.9993	0.9999	1.0000	1.0000	1.0000
	40	0.4828	0.7431	0.9105	0.9776	0.9956	0.9995	0.9999	1.0000	1.0000	1.0000	1.0000
	50	0.5611	0.8218	0.9502	0.9906	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4. Powers of three tests, two sample t-test, Wilcoxon rank sum test, bootstrap test with BC method when two independent samples are selected from chi-square distribution with three degrees of freedom and the population variances are equal, $a = 1.0$ $(\alpha = 0.05)$.

test	sample size (n)	a										
		0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
t-test	10	0.0491	0.1243	0.2497	0.4144	0.5805	0.7323	0.8449	0.9154	0.9560	0.9788	0.9907
	20	0.0480	0.1585	0.3726	0.6197	0.8193	0.9309	0.9752	0.9931	0.9986	0.9998	1.0000
	30	0.0461	0.1982	0.471	0.7624	0.9254	0.983	0.9968	0.9996	1.0000	1.0000	1.0000
	40	0.0493	0.2345	0.5749	0.8563	0.972	0.9965	0.9994	1.0000	1.0000	1.0000	1.0000
	50	0.0501	0.2752	0.6622	0.9176	0.9905	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000
wilcoxon test	10*	0.0470	0.1313	0.2836	0.4710	0.6457	0.7873	0.8847	0.9414	0.9691	0.9843	0.9927
	20	0.0485	0.1986	0.4796	0.7567	0.9089	0.9742	0.9940	0.9989	1.0000	1.0000	1.0000
	30	0.0478	0.2608	0.6312	0.8944	0.9802	0.9973	0.9995	1.0000	1.0000	1.0000	1.0000
	40	0.0482	0.3190	0.7440	0.9555	0.9950	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000
	50	0.0487	0.3756	0.8323	0.9813	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
bootstrap test	10	0.0560	0.1313	0.2433	0.3902	0.5458	0.6851	0.7973	0.8804	0.9324	0.9595	0.9786
	20	0.0512	0.1548	0.3346	0.5609	0.7539	0.8815	0.9537	0.9833	0.9956	0.9988	0.9999
	30	0.0500	0.1795	0.4169	0.6889	0.8744	0.9617	0.9903	0.9985	0.9999	1.0000	1.0000
	40	0.0521	0.2129	0.5120	0.7904	0.9413	0.9891	0.9982	0.9996	0.9999	1.0000	1.0000
	50	0.0512	0.2406	0.5865	0.8630	0.9716	0.9966	0.9998	1.0000	1.0000	1.0000	1.0000

* The power of Wilcoxon rank sum test when n=10 was calculated by the exact test.

4. Conclusion

As a computationally intensive method, bootstrap inference requires much more computation than classic statistical inference, like two-sample $t-$ test or Wilcoxon rank sum test.

This paper examined the power of bootstrap two-sample test, and compared it with the powers of two-sample $t-$ test and Wilcoxon rank sum test through simulation. For this, two independent samples were selected from chi-square distribution with three degrees of freedom, and powers were calculated for various location differences and scale differences between two samples.

Two-sample $t-$ test assumes that the samples are selected from normal population, or that the sample sizes are large. In this research, we examined the power for the sample sizes 10~50. Sample sizes greater than 30 are generally considered as large in statistics. The simulation results shows that the two sample $t-$ test has higher power than the bootstrap two-sample test at all selected (a, d, n) values, scale and location constants in equation (1), and sample size.

As a distribution-free method, Wilcoxon rank sum test is generally used for small samples. Wilcoxon rank sum test usually conducts the exact test for small samples, and the asymptotic test for large samples. The simulation results shows that Wilcoxon rank sum test for the equality of two population medians has the highest power among three tests; two-sample $t-$ test, Wilcoxon rank sum test, bootstrap two-sample test.

Through the simulation results, we can find

that the bootstrap two-sample test about the equality of population medians has the lowest power among three tests when the sampled population has positively skewed distribution with heavy tail, χ_3^2.

As the computation techniques are rapidly developed and the cost for large computation becomes low, some people seem to prefer using bootstrap methods to the classic statistical methods. Bootstrap inference will be good when there are no mathematically exact form of distributions of test statistics (e.g., Burce etc., 2020)[11]. However, if there are classic inference methods which are widely accepted and used like two-sample $t-$test or Wilcoxon rank sum test, then it will be efficient in terms of time and cost, sometimes in terms of power too, to use them.

References

[1] Gibbons J. D., "Nonparametric Statistics: An Introduction", Sage University Paper series on Quantitiative Application in the Social Science, 07-090, Newbury Park, CA: Sage, 1993.

[2] Conover, W. J., Practical Nonparametric Statistics (2nd ed), New York: Wiley, 1980.

[3] Sprent, P. and Smeeton, N. C., Applied Nonparametric Statistical Methods, 3rd ed., Chapman & Hall/CRC, London, 2001.

[4] Mooney, C. Z. and Duval, R. D., "Bootstrapping: A Nonparametric Approach to Statistical Inference", Sage University Paper series on Quantitiative Application in the Social Science, 07-095, Newbury Park, CA; Sage, 1993.

[5] Efron, B., "Bootstrap methods: Another look at the jackknife." Annals of Statistics, Vol. 7, pp. 1-26, 1979.

[6] Efron, B. and Tibshirani, R. "Bootstrap methods for Standard errors, confidence intervals, and other measures of statistical accuracy", Statistical Science 1: pp. 54-77, 1986.

[7] Good, P. I., Resampling Methods: A Practical Guide to Data Analysis, Birkhäuser, Boston, 1999.

[8] Efron, B., The jackknife, the bootstrap, and other resampling plans, Philadelphia: Society for Industrial and Applied Mathematics, 1982.

[9] Efron, B. and Tibshirani, R. An introduction to the Bootstrap, London, Chapman & Hill, 1993.

[10] Johnson, R. W., "An introduction to the bootstrap", Teaching Statistics, Vol. 23, No. 2, pp. 49-54, 2001.

[11] Bruce, P., Bruce, A., and Gedeck, P., Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python 2nd ed., O'Reilly Median, 2020.