CHAPTER

5 STATISTICAL INFERENCE OF RANKED SET SAMPLING VIA RESAMPLING METHODS

Saeid Amiri[1] **and Reza Modarres**[2]

[1]*Department of Natural and Applied Sciences, University of Wisconsin-Green Bay, Green Bay, WI, United States*

[2]*Department of Statistics, The George Washington University, Washington, DC, United States*

5.1 INTRODUCTION

Rank-based sampling (RSS) provides powerful inference alternatives to simple random sampling (SRS) and often leads to large improvements in the precision of estimators. Several variants of RSS are designed to further improve the performance of RSS. Theoretical underpinnings must be developed for such RSS designs. However, these results are often nontrivial due to many factors, including unknown parent distributions, small sample sizes, and nonidentical order statistics that form the cornerstones of any RSS design. These difficulties make bootstrap methods more attractive for RSS. Bootstrap is a well-known resampling method that provides accurate inference for SRS. The bootstrap method has also been explored in the different RSS contexts: confidence interval estimation (Hui et al., 2005), resampling algorithms (Modarres et al., 2006), one sample test (Amiri et al., 2014), confidence bands for the CDF (Frey, 2014), empirical likelihood (Amiri et al., 2016), censored RSS (Mahdizadeh and Strzalkowska-Kominiak, 2017), and tests of perfect ranking (Amiri et al., 2017).

In this work, we consider the parametric statistical inference of one sample and two samples. RSS is concerned with small sample sizes and distribution-free methods such as sign, sign ranked, and Mann−Whitney tests have been investigated by Bohn and Wolfe (1992) and Ozturk and Wolfe (2000a, 2000b). However, proposed nonparametric RSS tests might be sensitive to the ranking procedure. Fligner and MacEachern (2006) consider the center of the observations to eliminate the impact of ranking. The sensitivity of these tests to the ranking procedure occurs due to the use of the distribution function of the rth order statistic to test the mean. To overcome the sensitivity, we explore $H_0 : \mu_x = \mu_0$ and its two-sample variant using the t test statistics and show that the bootstrap methods provide more accurate inference. We will also consider RSS with different ranks sizes.

The reminder of the chapter is organized as follows. Section 5.2 provides an overview of the data structure of an RSS, and then formally defines the test statistics for one and two samples. Section 5.3 is devoted to the bootstrap methods. It gives a number of theoretical results that allow us to use the bootstrap methods. In Section 5.4, we compare the proposed methods using Monte Carlo simulation. The simulations show that a hybrid method, based on the average of the p-values of pivotal and nonpivotal bootstrap tests, outperforms the competing tests.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00005-8

5.2 STATISTICAL INFERENCE FOR RSS

Suppose a total number of n units are to be measured from the underlying population on the variable of interest. Let n sets of units, each of size k, be randomly chosen from the population using a simple random sampling (SRS) technique. The units of each set are ranked by any means other than actual quantification of the variable. Finally, one unit in each ordered set with a prespecified rank is measured on the variable. Let m_r be the number of measurements on units with rank r, $r = 1, \ldots, k$, such that $n = \sum_{r=1}^{k} m_r$. Let $X_{(r)j}$ denote the measurement on the jth measured unit with rank r. This results in a URSS of size n from the underlying population as $X_{(r)j}$; $r = 1, \ldots, k, j = 1, \ldots, m_r$. When $m_r = m$, $r = 1, \ldots, k$, URSS reduces to the balanced RSS. It is worth mentioning that, in ranked set sampling designs, $X_{(1)j}, \ldots, X_{(k)j}$ are independent order statistics (as they are obtained from independent sets) and each $X_{(r)j}$ provides information about a different stratum of the population. One can represent the structure of a URSS as follows:

$$\mathcal{X}_r = \left\{X_{(r)1}, X_{(r)2}, \ldots, X_{(r)m_r}\right\} \overset{i.i.d.}{\sim} F_{(r)}, \quad r = 1, \ldots, k_1,$$

where $F_{(r)}$ is the distribution function (df) of the rth order statistic. The second sample can be generated using the same procedure. We assume the second sample is generated using k_2 which can be different from $k = k_1$

$$\mathcal{Y}_r = \{Y_{(r)1}, Y_{(r)2}, \ldots, Y_{(r)m_r}\} \overset{i.i.d.}{\sim} G_{(r)}, \quad r = 1, \ldots, k_2.$$

It is of interest to test $H_0 : F(x) \overset{d}{=} G(x - \Delta)$. Specifically, we are concerned with the null hypothesis $H_0 : \mu_x = \mu_y + \Delta$ versus $H_0 : \mu_x \neq \mu_y + \Delta$. Two sample tests are commonly used to determine whether the samples come from the same unknown distribution. In our setting, we assume X and Y are collected with different ranks sizes. Therefore, even under the same parent distributions, the variance of the estimator would not be the same.

The following proposition can be used to establish the asymptotic normality of statistic under the null hypothesis.

Proposition 1: *Let F denote the cdf of a member of the family with $\int x^2 dF(x) < \infty$ and $\hat{F}_{(r)}$ is the empirical distribution function (edf) of the rth row. If $\vartheta_i = (\overline{X}_{(i)} - \mu_{(i)})$, then $(\vartheta_1, \ldots, \vartheta_k)$ converges in distribution to a multivariate normal distribution with mean vector zero and covariance matrix $diag(\sigma^2_{(1)}/m_1, \ldots, \sigma^2_{(k)}/m_k)$ where $\sigma^2_{(i)} = \int (x - \mu_{(i)})^2 dF_{(i)}(x)$ and $\mu_{(i)} = \int x dF_{(i)}(x)$.*

Proposition 1 suggests the following statistic for testing H_0: $\mu = \mu_0$,

$$Z = \frac{1}{k}\sum_{r=1}^{k} \overline{X}_{(r)} - \mu_0 \hat{\sigma} \overset{d}{\to} N(0, 1),$$

where $\hat{\sigma}^2$ is the plug-in estimator for the $V\left(\frac{1}{k}\sum_{r=1}^{k} \overline{X}_{(r)}\right)$,

$$\hat{\sigma}^2 = \frac{1}{k^2}\sum_{r=1}^{k} \frac{\hat{\sigma}^2_{(r)}}{m_r},$$

and $\sigma^2_{(r)}$ is the estimate of $V(\overline{X}_{(r)})$. Using the central limit theorem, one obtains a confidence interval where

$$P\left(\mu \in \left(\overline{X} + t_{\alpha/2, n-1}\frac{\sigma}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2, n-1}\frac{\sigma}{\sqrt{n}}\right)\right) \approx 1 - \alpha.$$

One needs $\sigma^2_{(r)}$ to estimate the variance of the mean. Hence it is necessary to have $m_r \geq 2$. The estimate of the variance for small sample sizes would be very inaccurate, suggesting that a pivotal statistic might be unreliable. We show in Section 5.4 that parametric statistics are very conservative. Bootstrap provides a nonparametric alternative to estimate the variance. The bootstrap method can be used to obtain the sampling distribution of the statistic of interest and allows for estimation of the standard error of any well-defined functional. Hence, bootstrap enables us to draw inferences when the exact or the asymptotic distribution of the statistic of interest is unavailable. A procedure of generating resamples to calculate the variance is discussed in Section 5.3.

Proposition 1 can be used to obtain a test statistic for two samples $\mathscr{X}_1, \ldots, \mathscr{X}_{k_1}$ and $\mathscr{Y}_1, \ldots, \mathscr{Y}_{k_2}$. One can show that

$$T(\mathscr{X}, \mathscr{Y}) = \left(\frac{1}{k_1}\sum\nolimits_{r=1}^{k_1} \overline{X}_{(r)} - \frac{1}{k_2}\sum\nolimits_{r=1}^{k_2} \overline{Y}_{(r)}\right) - (\mu_1 - \mu_2)\hat{\sigma} \xrightarrow{d} N(0,1),$$

where

$$\hat{\sigma}^2 = \frac{1}{k_1^2}\sum_{r_1=1}^{k_1} \frac{\hat{\sigma}^2_{(r_1)}}{m_{r_1}} + \frac{1}{k_2^2}\sum_{r_2=1}^{k_2} \frac{\hat{\sigma}^2_{(r_2)}}{m_{r_2}}.$$

We can consider the parametric statistical inference for the skewed distribution: let $X_1, \ldots, X_n$ be i.i.d. random variable with the mean μ and finite variance σ^2. Since the characteristic function of S_n converges to $e^{-t^2/2}$, the characteristic function of the standard normal, $\sqrt{n}S_n = \sqrt{n}(\mu - \mu)/\sigma$, is asymptotically normally distributed with zero mean and unit variance. To take the sample skewness into account, the following proposition obtains the Edgeworth expansion of $\sqrt{n}S_n$.

Proposition 2: *If $E\left(Y_i^6\right) < \infty$ and Cramer's condition holds, the asymptotic distribution function of $\sqrt{n}S_n$ is*

$$P(\sqrt{n}S_n \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}\gamma(ax^2 + b)\phi(x) + O(n^{-1}),$$

where a and b are known constants, γ is an estimable constant, and Φ and ϕ denote the standard normal distribution and density functions, respectively.

Hall (1992) suggested two functions,

$$S_1(t) = t + a\hat{\gamma}t^2 + \frac{1}{3}a^2\hat{\gamma}^2t^3 + n^{-1}b\hat{\gamma},$$
$$S_2(t) = (2an^{-\frac{1}{2}}\hat{\gamma})^{-1}\left\{\exp\left(2an^{-\frac{1}{2}}\hat{\gamma}t\right) - 1\right\} + n^{-1}b\hat{\gamma},$$

where $a = 1/3$ and $b = 1/6$. Zhou and Dinh (2005) suggested

$$S_3(t) = t + t^2 + \frac{1}{3}t^3 + n^{-1}b\hat{\gamma}.$$

Using $S_i(t)$, for $i = 1, 2, 3$, one can construct new confidence intervals for μ as

$$(\hat{\mu} - S_i(n^{-1/2}t_{1-\alpha/2,n-1})\hat{\sigma}, \hat{\mu} - S_i(n^{-1/2}t_{\alpha/2,n-1})\hat{\sigma}),$$

where $t_{1-\alpha/2,n-1}$ is the $1 - \alpha/2$ quartile of the t distribution. However, use of the sample skewness in the asymptotic distribution makes the inference less reliable, especially for the parametric methods. For example, the asymptotic distribution of test for the coefficient of variation depends on the skewness. This parameter makes the inference for coefficient of variation inaccurate, see Amiri (2016). It is of interest to study this problem using a fully nonparametric approach via the bootstrap.

5.3 BOOTSTRAP METHOD

Bootstrap resampling is a well-known statistical method to conduct statistical inference. Bootstrap mimics the underlying distribution of the observations by resampling from the URSS sample. Several papers have explored the application of bootstrap in RSS. URSS bootstrap was considered in Amiri et al. (2014). The idea of URSS bootstrap is to obtain a sample of size n_0 from each stratum in order to transform the URSS to an RSS dataset. The RSS dataset is then resampled to provide inference. Amiri et al. (2017) consider more general resampling techniques that obtain resamples from the entire dataset instead the resampling each stratum. The procedure is described below.

Algorithm:

1. Select a row randomly and select an observation, continue until k observations have been collected (obviously any row can appear more than once).

$$\text{Order them as } X^{\diamond}_{(1)} \le \ldots \le X^{\diamond}_{(k)} \text{ and retain } X^{*}_{(r)1} = X^{\diamond}_{(r)}.$$

2. Perform steps 1–2 m_r times and collect $X^*_{(r)1}, \ldots, X^*_{(r)m_r}$.
3. Perform step 3 for $r = 1, \ldots, k$.
4. Repeat steps 1–4, B times to obtain the bootstrap samples.

Using step 1 of the algorithm,

$$\{X^{\diamond}_1, \ldots, X^{\diamond}_k\} \sim \hat{F}_n(t) = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{m_r}\sum_{j=1}^{m_r} I(X_{(r)j} \le t),$$

and using steps 2 and 3,

$$\mathbf{X}^*_r = \{X^*_{(r)1}, X^*_{(r)2}, \ldots, X^*_{(r)m_r}\} \sim \hat{F}_{(r)}(.), \tag{5.1}$$

where $\hat{F}_{(r)}(t) = \frac{1}{m_r}\sum_{j=1}^{m_r} I(X_{(r)j} \le t)$. Let

$$\hat{F}^*_{(r)}(t) = \frac{1}{m_r}\sum_{j=1}^{m_r} I\left(X^*_{(r)j} \le t\right), \tag{5.2}$$

$$\hat{F}^*_n(t) = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{m_r}\sum_{j=1}^{m_r} I\left(X^*_{(r)j} \le t\right). \tag{5.3}$$

Amiri et al. (2017) proved the following propositions for the proposed bootstrap algorithm. These properties are essential to draw inference using the resamples.

Proposition 3: *Let $F_{(r)}(t)$ denote the cdf of the rth row of a member of the family with the continuous density function, and $\hat{F}^*_{(r)}$ denote the edf of the rth row given in (Eq. (5.2)), it follows that*

$$\hat{F}^*_{(r)}(t) \xrightarrow{a.s.} F_{(r)}(t).$$

Proposition 4: *Let $F(t)$ denote the cdf of a member of the family with the continuous density function, and suppose $\mathscr{X}_1^*, \ldots, \mathscr{X}_k^*$ are samples obtained using the proposed bootstrap algorithm, it follows that*

$$\sup_{t \in \mathbb{R}} |\hat{F}_n^*(t) - F(t)| = 0.$$

Proposition 4 shows a desirable property for the bootstrap method that can be used to draw statistical inference. The direct application of bootstrap is in the estimation of variance. Suppose $\mathscr{X}_1^*, \ldots, \mathscr{X}_k^*$ and we are interested in $V(\theta(F_{(1)}, \ldots, F_{(K)})) = \frac{1}{k^2} \sum_{r=1}^{k} \frac{\sigma_{(r)}^2}{m_r}$. The plug-in estimation is $V(\theta(\hat{F}_{(1)}, \ldots, \hat{F}_{(K)})) = \hat{\sigma}^2 = \frac{1}{k^2} \sum_{r=1}^{k} \frac{\hat{\sigma}_{(r)}^2}{m_r}$ where $\hat{F}_{(r)}$ is the edf on the $r-$th stratum. Clearly, the plug-in estimate does not work for $m_r = 1$. However, one can use the proposed bootstrap to estimate the variance. Generate the resamples using the proposed algorithm and compute $\theta(\hat{F}_{(1)}^*, \ldots, \hat{F}_{(k)}^*) = \overline{X}^* = \frac{1}{k} \sum_{r=1}^{k} \overline{X}_{(r)}^*$, and repeat the procedure B times to obtain $\overline{X}_b^*, b = 1, \ldots, B$. The most important property of the bootstrap lies in the conditional independence, given the original sample. Hence, we view bootstrap resample as iid random samples and compute the sample mean and the sample variance with,

$$\overline{\overline{X}}^* = \frac{1}{B} \sum_{b=1}^{B} \overline{X}_b^*,$$

$$\hat{V}^* \left(\theta\left(\hat{F}_{(1)}, \ldots, \hat{F}_{(K)}\right)\right) = \frac{1}{B} \sum_{b=1}^{B} (\overline{X}_b^* - \overline{\overline{X}}^*)^2.$$

The confidence interval can be found using the bootstrap estimate of variance as,

$$\frac{1}{k} \sum_{r=1}^{k} \overline{X}_{(r)} \pm t_{\alpha/2, n-1} \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\overline{X}_b^* - -\overline{\overline{X}}^*\right)^2}.$$

The nonparametric confidence interval can be obtained using the percentile confidence interval

$$(\overline{X}_{\alpha/2}^*, \overline{X}_{1-\alpha/2}^*),$$

where $\overline{X}_{\alpha/2}^*$ is the $\alpha/2$ percentile of bootstrap resample mean.

5.4 NUMERICAL STUDY

This section is devoted to assessing the accuracy and comparisons of the proposed test statistics for finite sample sizes. We study the type I error rate and the statistical power. The proposed tests are based on the same simulated data in order to provide a meaningful comparison. The resampling is carried out using $B = 800$ resamples. In order to make a comparative evaluation of the testing procedures, we seek certain desirable features, such as robustness, power, and small sample test validity in terms of observed type I error rates. In the following, the significance and the power of the proposed tests are studied for different sample sizes.

To compare two group means: $H_0 : \mu_x = \mu_y + \delta$ vs. $H_0 : \mu_x \neq \mu_y + \delta$, the appropriate test statistic is

$$\boldsymbol{T}_0(\mathcal{X},\mathcal{Y}) = \left(\frac{1}{\boldsymbol{k}_1}\sum\nolimits_{r=1}^{k_1}\overline{\boldsymbol{X}}_{(r)} - \frac{1}{\boldsymbol{k}_2}\sum\nolimits_{r=1}^{k_2}\overline{\boldsymbol{Y}}_{(r)}\right) - \boldsymbol{\delta}\hat{\boldsymbol{\sigma}}, \tag{5.4}$$

where, $\hat{\sigma}^2 = \frac{1}{k_1^2}\sum\limits_{r_1=1}^{k_1}\frac{\hat{\sigma}^2_{(r_1)}}{m_{r_1}} + \frac{1}{k_2^2}\sum\limits_{r_2=1}^{k_2}\frac{\hat{\sigma}^2_{(r_2)}}{m_{r_2}}$. Under the null hypothesis, $T_0(\mathcal{X},\mathcal{Y}) \sim t_{n_1+n_2-2}$. We refer to this test as the parametric test and denote it with PT.

The bootstrap test, referred to as BT, is constructed as follows. Calculate the statistic given in (Eq. (5.4)), take the resamples according to the algorithm described in Section 5.3, and calculate the following statistic,

$$\boldsymbol{T}^*(\mathcal{X}^*,\mathcal{Y}^*,\mathcal{X},\mathcal{Y}) = \frac{\left(\frac{1}{k_1}\sum\nolimits_{r=1}^{k_1}\overline{\boldsymbol{X}}^*_{(r)} - \frac{1}{k_2}\sum\nolimits_{r=1}^{k_2}\overline{\boldsymbol{Y}}^*_{(r)}\right) - \left(\frac{1}{k_1}\sum\nolimits_{r=1}^{k_1}\overline{\boldsymbol{X}}_{(r)} - \frac{1}{k_2}\sum\nolimits_{r=1}^{k_2}\overline{\boldsymbol{Y}}_{(r)}\right)}{\hat{\boldsymbol{\sigma}}^*}, \tag{5.5}$$

where σ^* is the estimate of variance using the bootstrap samples. Generate B resamples and calculate the test statistics,

$$T_1^*(\mathcal{X}^*,\mathcal{Y}^*,\mathcal{X},\mathcal{Y}),\ldots,T_B^*(\mathcal{X}^*,\mathcal{Y}^*,\mathcal{X},\mathcal{Y}).$$

The approximate p-value can be estimated with

$$p^* = \frac{\#T_b^*(\mathcal{X}^*,\mathcal{Y}^*,\mathcal{X},\mathcal{Y}) \le T_0(\mathcal{X},\mathcal{Y})}{B},$$

$$p - \text{value} = \min\{p^*, 1-p^*\}.$$

Since the RSS often uses small sample sizes and the plug-in estimate of variance is not very accurate, one may consider a third approach and use nonpivotal test statistics to calculate the p-value. That is,

$$\begin{aligned} T_0(\mathcal{X},\mathcal{Y}) &= \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}_{(r)}\right) - \delta, \\ T^*(\mathcal{X}^*,\mathcal{Y}^*,\mathcal{X},\mathcal{Y}) &= \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}^*_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}^*_{(r)}\right) - \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}_{(r)}\right). \end{aligned} \tag{5.6}$$

This bootstrap test using the nonpivotal statistic is denoted as BNT.

We compare the following test statistics:

1. PT: Parametric two-sample t-test (Eq. (5.4));
2. BT: Bootstrap test (Eq. (5.5));
3. BNT: Nonpivotal bootstrap test (Eq. (5.6));
4. BHT : Hybrid test of BT and BNT.

Table 5.1 includes the simulation of the 10th percentile of p-value for the proposed methods with different sample sizes $(n_X, n_Y) = (k_1m, k_2m)$, and the following underlying distributions:

1. $X \stackrel{d}{=} Y \sim N(0,1)$,
2. $X \stackrel{d}{=} Y \sim exp(1.5)$,
3. $X \stackrel{d}{=} Y \sim logistic(1,1)$,
4. $X \stackrel{d}{=} Y \sim Gamma(1,2)$.

Table 5.1 Observed α-levels of the Proposed Tests at $\alpha = 0.1$

			m					**m**			
dist.	k_1, k_2	**test**	**3**	**4**	**5**	**6**	**dist.**	**3**	**4**	**5**	**6**
$X \overset{d}{=} Y \sim N(0,1)$	(3,3)	PT	0.315	0.226	0.188	0.169	$X \overset{d}{=} Y \sim exp(1.5)$	0.317	0.231	0.191	0.165
		BT	0.133	0.119	0.115	0.111		0.144	0.127	0.116	0.108
		BNT	0.080	0.093	0.099	0.101		0.076	0.099	0.101	0.102
		BHT	0.095	0.100	0.104	0.104		0.098	0.105	0.104	0.103
	(3,4)	PT	0.310	0.226	0.194	0.174		0.316	0.237	0.186	0.178
		BT	0.117	0.117	0.115	0.11		0.127	0.116	0.108	0.117
		BNT	0.077	0.095	0.097	0.103		0.088	0.100	0.100	0.111
		BHT	0.087	0.101	0.104	0.104		0.095	0.104	0.101	0.111
	(3,5)	PT	0.316	0.236	0.196	0.184		0.333	0.243	0.196	0.177
		BT	0.132	0.119	0.110	0.116		0.132	0.117	0.113	0.112
		BNT	0.090	0.104	0.094	0.103		0.088	0.102	0.110	0.107
		BHT	0.100	0.106	0.098	0.108		0.094	0.105	0.111	0.108
	(4,4)	PT	0.312	0.225	0.189	0.166		0.316	0.217	0.196	0.182
		BT	0.114	0.110	0.106	0.106		0.125	0.105	0.107	0.115
		BNT	0.092	0.095	0.100	0.101		0.089	0.097	0.103	0.111
		BHT	0.094	0.098	0.103	0.100		0.098	0.096	0.105	0.113
	(4,5)	PT	0.316	0.218	0.197	0.176		0.307	0.230	0.198	0.161
		BT	0.112	0.107	0.109	0.105		0.118	0.112	0.110	0.097
		BNT	0.092	0.092	0.098	0.102		0.091	0.101	0.103	0.094
		BHT	0.095	0.094	0.102	0.104		0.097	0.102	0.104	0.093
$X \overset{d}{=} Y \sim logistic(1,1)$	(3,3)	PT	0.331	0.244	0.204	0.178	$X \overset{d}{=} Y \sim Gamma(1,2)$	0.338	0.231	0.202	0.185
		BT	0.136	0.124	0.127	0.109		0.133	0.118	0.121	0.114
		BNT	0.072	0.090	0.111	0.112		0.082	0.090	0.109	0.112
		BHT	0.077	0.099	0.116	0.106		0.086	0.098	0.111	0.109
	(3,4)	PT	0.338	0.240	0.204	0.185		0.337	0.251	0.196	0.179
		BT	0.136	0.122	0.123	0.116		0.138	0.126	0.111	0.116
		BNT	0.091	0.103	0.112	0.114		0.093	0.105	0.105	0.113
		BHT	0.093	0.105	0.116	0.112		0.101	0.111	0.106	0.110

(*Continued*)

Table 5.1 Observed α-levels of the Proposed Tests at $\alpha = 0.1$ *Continued*

dist.	k_1, k_2	test	m: 3	4	5	6	dist.	m: 3	4	5	6
	(3,5)	PT	0.353	0.244	0.209	0.191		0.333	0.256	0.203	0.175
		BT	0.145	0.124	0.119	0.121		0.131	0.132	0.120	0.113
		BNT	0.101	0.100	0.107	0.117		0.093	0.110	0.106	0.106
		BHT	0.105	0.103	0.109	0.114		0.095	0.117	0.110	0.105
	(4,4)	PT	0.323	0.245	0.184	0.180		0.331	0.233	0.197	0.182
		BT	0.121	0.121	0.102	0.113		0.121	0.112	0.105	0.111
		BNT	0.088	0.104	0.099	0.118		0.095	0.101	0.105	0.114
		BHT	0.090	0.107	0.098	0.114		0.092	0.102	0.103	0.108
	(4,5)	PT	0.337	0.238	0.202	0.189		0.339	0.242	0.197	0.178
		BT	0.126	0.115	0.115	0.111		0.124	0.117	0.111	0.102
		BNT	0.099	0.100	0.113	0.115		0.093	0.102	0.109	0.108
		BHT	0.096	0.101	0.110	0.111		0.091	0.102	0.107	0.104

Since X and Y are generated from the same distributions, it is expected that an accurate test maintains the nominal level. In the frequentist approach, the appealing property of the p-value is its (asymptotic) uniformity on $Unif(0,1)$ under the null hypothesis. When a test statistic is conservative (or liberal), the actual type I error of the test will be small (large) compared with the nominal level. For a conservative (or liberal) test, the power values can be misleading. It is easy to see that a conservative p-value hardly rejects an incorrect null hypothesis and a liberal test easily rejects a correct null hypothesis too often and both lead to incorrect inferences.

Clearly the PT leads to an overly conservative test, i.e., fails to reject the null hypothesis when it should. However, this problem tends to diminish with an increase in sample size. Here, BT and BNP have better performances, and are closer to the actual p-value. It is noteworthy that for $m = 3$, $(k_1, k_2) = (3,3), (3,4)$ which have very small sample sizes $((n_X, n_Y) = (9,9), (9,12))$, BT and BNT are conservative and liberal, respectively. It is of interest to explore the average of the p-values. We refer to this hybrid test as the BHT method. Clearly BHT has better performance.

To compare the statistical power, we consider

1. $X \sim N(0,1), Y \sim N(0.5,1)$,
2. $X \sim exp(1), Y \sim exp(1.5)$,
3. $X \sim logistic(0,1), Y \sim logistic(1,1)$,
4. $X \sim Gamma(1,1), Y \sim Gamma(1,2)$.

Since X and Y are generated from different distributions with different parameters, a powerful test should reject the null hypothesis with high probability. The result is presented in Table 5.2. Since the PT is conservative for small sample sizes, we expect a large value for the power. However, this power value is overly optimistic and not accurate. Among the bootstrap tests, BHT has better power than BNT. BHT has less power than PT, keeping in mind that PT performs conservatively for small sample sizes.

5.5 CONCLUSIONS

A considerable amount of research has been conducted in the past few decades to advance the theoretical foundation of RSS and present its applications. RSS draws on additional information from inexpensive and easily obtained sources to collect a more representative sample. In this work, we review the statistical tests of means under one and two samples. RSS is often applied with small sample sizes. Presenting nonparametric methods and exploring the performance of test statistics are essential in obtaining a better understanding of their behavior. In our empirical study, we mainly considered small samples and compared the performance of proposed tests using Monte Carlo investigations under different distributions. We proposed a hybrid method, based on the average of the p-values of pivotal and nonpivotal bootstrap tests and demonstrate its better performance. The hybrid method provides a more accurate inference for small sample sizes and enables one to maintain the nominal level with comparable power.

Table 5.2 The Empirical Power of the Proposed Tests

dist.	k_1, k_2	test	m = 3	m = 4	m = 5	m = 6	dist.	m = 3	m = 4	m = 5	m = 6
$X \sim N(0,1), Y \sim N(0.5,1)$	(3,3)	PT	0.552	0.553	0.605	0.666	$X \sim exp(1), Y \sim exp(1.5)$	0.590	0.604	0.652	0.703
		BT	0.342	0.422	0.502	0.583		0.395	0.481	0.559	0.631
		BNT	0.224	0.359	0.457	0.553		0.267	0.403	0.510	0.602
		BHT	0.276	0.392	0.485	0.571		0.322	0.444	0.537	0.619
	(3,4)	PT	0.607	0.626	0.670	0.724		0.640	0.678	0.733	0.789
		BT	0.391	0.476	0.568	0.653		0.439	0.556	0.645	0.722
		BNT	0.269	0.423	0.522	0.617		0.309	0.488	0.596	0.683
		BHT	0.325	0.452	0.547	0.639		0.369	0.524	0.623	0.706
	(3,5)	PT	0.628	0.657	0.722	0.771		0.676	0.714	0.773	0.819
		BT	0.427	0.517	0.613	0.691		0.479	0.598	0.687	0.759
		BNT	0.314	0.455	0.576	0.664		0.353	0.521	0.635	0.714
		BHT	0.367	0.489	0.596	0.677		0.407	0.562	0.662	0.739
	(4,4)	PT	0.670	0.702	0.755	0.822		0.702	0.749	0.808	0.859
		BT	0.445	0.563	0.655	0.747		0.491	0.624	0.721	0.803
		BNT	0.356	0.514	0.630	0.729		0.389	0.567	0.679	0.773
		BHT	0.399	0.546	0.647	0.741		0.441	0.596	0.703	0.790
	(4,5)	PT	0.708	0.740	0.817	0.859		0.754	0.793	0.845	0.893
		BT	0.491	0.599	0.725	0.794		0.562	0.685	0.771	0.846
		BNT	0.407	0.569	0.702	0.776		0.459	0.628	0.735	0.823
		BHT	0.449	0.581	0.716	0.788		0.514	0.661	0.756	0.835
$X \sim logistic(0,1)$, $Y \sim logistic(1,1)$	(3,3)	PT	0.481	0.440	0.454	0.476	$X \sim Gamma(1,1)$, $Y \sim Gamma(1,2)$	0.640	0.695	0.752	0.808
		BT	0.257	0.302	0.347	0.380		0.437	0.563	0.654	0.736
		BNT	0.142	0.233	0.314	0.377		0.245	0.452	0.617	0.734
		BHT	0.175	0.263	0.327	0.381		0.320	0.510	0.635	0.737
	(3,4)	PT	0.529	0.500	0.535	0.562		0.668	0.720	0.791	0.844
		BT	0.323	0.365	0.438	0.488		0.450	0.581	0.684	0.773
		BNT	0.211	0.302	0.398	0.454		0.241	0.468	0.666	0.777
		BHT	0.252	0.337	0.421	0.470		0.321	0.531	0.676	0.781

	(3,5)	PT	0.580	0.567	0.598	0.631		0.706	0.740	0.813	0.861
		BT	0.382	0.452	0.508	0.560		0.476	0.597	0.714	0.799
		BNT	0.254	0.379	0.448	0.506		0.259	0.479	0.689	0.798
		BHT	0.305	0.412	0.475	0.531		0.337	0.540	0.708	0.800
	(4,4)	PT	0.539	0.548	0.581	0.624		0.780	0.819	0.880	0.927
		BT	0.321	0.394	0.456	0.525		0.595	0.721	0.817	0.887
		BNT	0.216	0.329	0.440	0.521		0.403	0.634	0.788	0.869
		BHT	0.251	0.362	0.449	0.524		0.488	0.686	0.805	0.879
	(4,5)	PT	0.587	0.598	0.654	0.700		0.798	0.855	0.909	0.946
		BT	0.372	0.456	0.553	0.615		0.611	0.752	0.851	0.911
		BNT	0.265	0.397	0.517	0.593		0.410	0.667	0.835	0.905
		BHT	0.306	0.426	0.539	0.606		0.509	0.716	0.847	0.912

REFERENCES

Amiri, S., 2016. Revisiting inference of coefficient of variation: nuisances parameters. Stat 5 (1), 234–241.
Amiri, S., Jozani, M.J., Modarres, R., 2014. Resampling unbalanced ranked set samples with applications in testing hypothesis about the population mean. J. Agric., Biol., Environ. Stat. 19 (1), 1–17.
Amiri, S., Jozani, M.J., Modarres, R., 2016. Exponentially tilted empirical distribution function for ranked set samples. J. Korean Stat. Soc. 45 (2), 176–187.
Amiri, S., Modarres, R., Zwanzig, S., 2017. Tests of perfect judgment ranking using pseudo samples. Comput. Stat. 32 (4), 1309–1322.
Bohn, L.L., Wolfe, D.A., 1992. Nonparametric two-sample procedures for ranked-set samples data. J. Am. Stat. Assoc. 87 (418), 552–561.
Fligner, M.A., MacEachern, S.N., 2006. Nonparametric two-sample methods for ranked set sample data. J. Am. Stat. Assoc. 101, 1107–1118.
Frey, J., 2014. Bootstrap confidence bands for the CDF using ranked-set sampling. J. the Korean Stat. Soc. 43 (3), 453–461.
Hall, P., 1992. On the removal of skewness by transformation.. J. R. Stat. Soc.. Ser. B 221–228.
Hui, T.P., Modarres, R., Zheng, G., 2005. Bootstrap confidence interval estimation of mean via ranked set sampling linear regression. J. Stat. Comput. Simul. 75 (7), 543–553.
Mahdizadeh, M., Strzalkowska-Kominiak, E., 2017. Resampling based inference for a distribution function using censored ranked set samples. Comput. Stat. 1–24.
Modarres, R., Hui, T.P., Zheng, G., 2006. Resampling methods for ranked set samples. Comput. Stat. Data Anal. 51 (2), 1039–1050.
Ozturk, O., Wolfe, D.A., 2000a. Alternative ranked set sampling protocols for the sign test. Stat. Probab. Lett. 47 (1), 15–23.
Ozturk, O., Wolfe, D.A., 2000b. An improved ranked set two-sample mann-whitney-wilcoxon test. Can. J. Stat. 28 (1), 123–135.
Zhou, X.H., Dinh, P., 2005. Nonparametric confidence intervals for the one-and two-sample problems. Biostatistics 6 (2), 187–200.

FURTHER READING

Hall, P., 2013. The Bootstrap and Edgeworth Expansion. Springer Science & Business Media, New York.