Journal of Statistical Theory and Practice (2024) 18:16
https://doi.org/10.1007/s42519-024-00370-x

ORIGINAL ARTICLE

Smoothed Bootstrap Methods for Hypothesis Testing

Asamh S. M. Al Luhayb[1] · Tahani Coolen-Maturi[2] · Frank P. A. Coolen[2]

Accepted: 7 February 2024 / Published online: 4 March 2024

Abstract
This paper demonstrates the application of smoothed bootstrap methods and Efron's methods for hypothesis testing on real-valued data, right-censored data and bivariate data. The tests include quartile hypothesis tests, two sample medians and Pearson and Kendall correlation tests. Simulation studies indicate that the smoothed bootstrap methods outperform Efron's methods in most scenarios, particularly for small datasets. The smoothed bootstrap methods provide smaller discrepancies between the actual and nominal error rates, which makes them more reliable for testing hypotheses.

Keywords Achieved significance level · Banks' bootstrap · Bootstrap confidence interval · Efron's bootstrap · Smoothed bootstrap

1 Introduction

The bootstrap method, as introduced by Efron [13], is a nonparametric statistical method proposed to specify the variability of sample estimates. The method has been widely used in the literature for a variety of statistical problems [17] as it is easy to apply and overall provides good results. When the distribution is unknown, the bootstrap method could be of great practical use [10].

For univariate real-valued data, Efron [13] introduced the bootstrap method, which is used in many real-world applications; see Efron and Tibshirani [17], Davison and Hinkley [10] and Berrar [5] for more details. For an original data set of size n, boot-

✉ Tahani Coolen-Maturi
tahani.maturi@durham.ac.uk

Asamh S. M. Al Luhayb
a.alluhayb@qu.edu.sa

Frank P. A. Coolen
frank.coolen@durham.ac.uk

[1] Department of Mathematics, College of Science, Qassim University, P.O. Box 6644, Buraydah 51452, Saudi Arabia

[2] Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK

strap samples of size n are created by random sampling with replacement and then computing the function of interest based on each bootstrap sample. The empirical distribution of the results can be used as a proxy for the distribution of the function of interest. In the case of finite support, Banks [4] presented a smoothed bootstrap method by linear interpolation between consecutive observations. Banks' bootstrap method starts with ordering the n observations of the original sample, where it is assumed that there are no ties, and taking the $n+1$ intervals of the partition of the support created by the n ordered observations. Each interval is assigned probability $\frac{1}{n+1}$. To generate one Banks' bootstrap sample, n intervals are resampled, and then one observation is drawn uniformly from each chosen interval. With Banks' bootstrap method, it is allowed to sample from the whole support, and ties occur with probability 0 in the bootstrap samples. This is contrary to Efron's method, where the process is restricted to resampling from the original data set [13]. In the case of underlying distributions with infinite support, Coolen and BinHimd [8] generalised Banks' bootstrap method by assuming distribution tail(s) for the first and last interval.

Efron [14] presented the bootstrap method for right-censored data, which is widely used in survival analysis; see Efron and Tibshirani [4, 16]. This bootstrap version is very similar to the method presented for univariate real-valued data, where multiple bootstrap samples of size n are created by resampling from the original sample, and the function of interest is computed based on each bootstrap sample. The empirical distribution of those resulting values can be used as a good proxy for the distribution of the function of interest. Al Luhayb et al. [2] generalized Banks' bootstrap method based on the right-censoring $A_{(n)}$ assumption [9]. The generalised bootstrap method produced better results; see Al Luhayb [1] and Al Luhayb et al. [2] for more details.

Efron and Tibshirani [16] introduced the bootstrap method for bivariate data, where again, multiple bootstrap samples are generated by resampling from the original data set, and the function of interest is computed based on each bootstrap sample. The empirical distribution of the resulting values can be a good proxy for the distribution of the function of interest. However, Efron's bootstrap method often produces poor results when working with small data sets. To address this issue, Al Luhayb et al. [3] proposed three new smoothed bootstrap methods. These methods rely on applying Nonparametric Predictive Inference on the marginals and modelling the dependence using parametric and nonparametric copulas. The new bootstrap methods have been shown to produce more accurate results. For further details, we refer the reader to Al Luhayb [1] and Al Luhayb et al. [3].

Classical statistical methods are widely used for testing statistical hypotheses, although their underlying assumptions are not always met, especially with complex data sets. To avoid these issues, Efron's bootstrap method has been used to test statistical hypotheses [16, 23, 24], which is easy to implement, and it provides good approximation results. However, it may not be suitable for small data sets and may include ties in the bootstrap samples. To overcome these limitations, various smoothed bootstrap methods have been proposed by Banks [4], Al Luhayb et al. [2] and Al Luhayb et al. [3] for real-valued data, right-censored data, and bivariate data, respectively. This paper investigates the use of these bootstrap methods for hypothesis testing and compares their results with those of Efron's methods.

This paper is organised as follows: Sect. 2 provides an overview of several bootstrap methods for real-valued univariate data, right-censored univariate data, and real-valued bivariate data. To illustrate their application, an example with data from the literature is presented in Sect. 3 using Efron's and Banks' bootstrap methods for hypothesis testing. Section 4 compares the smoothed bootstrap methods and Efron's bootstrap methods through simulations in various hypothesis tests, such as quartile hypothesis tests, two-sample medians, Pearson and Kendall correlation tests. Firstly, the smoothed bootstrap methods and Efron's bootstrap methods for real-valued univariate data and right-censored univariate data are used to compute the Type I error rates for quartile tests. Secondly, the achieved significance level is used to compute the Type I error rate for two-sample median tests. Lastly, for real-valued bivariate data, the smoothed bootstrap methods and Efron's bootstrap method are compared in computing the Type I error rates for Pearson and Kendall correlation tests. The final section provides some concluding remarks.

2 Bootstrap Methods for Different Data Types

When it comes to real-world applications, using traditional statistical methods can be challenging due to the mathematical assumptions involved. However, the use of bootstrap methods can provide a computer-based way of conducting statistical inference that doesn't require complex formulas. This paper demonstrates the use of different bootstrap methods for hypothesis testing. This section will provide an overview of multiple bootstrap methods that can be applied to real-valued data, right-censored data, and bivariate data.

2.1 Bootstrap Methods for Real-Valued Univariate Data

In this section, we will discuss two bootstrap methods for data that include only real-valued observations, namely Efron's bootstrap method and Banks' bootstrap method [4, 13]. These methods are used to measure the variability of sample estimates for a given function of interest $\theta(F)$, where F is a continuous distribution defined on the interval $[a, b]$. Suppose we have n independent and identically distributed random quantities $X_1, X_2, \ldots, X_n$ from the distribution F and the corresponding observations are $x_1, x_2, \ldots, x_n$.

Efron's bootstrap method [13] is a nonparametric method proposed to measure the variability of sample estimates. It uses the empirical distribution function of the original sample, where each observation has the same probability of being selected. To create B resamples of size n, we randomly select observations with replacement from the original sample. We then calculate the function of interest $\hat{\theta}$ for each bootstrap sample to obtain $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$. The empirical distribution of these results approximates the sampling distribution of $\theta(F)$. Efron's bootstrap method is commonly used for hypothesis testing and has been shown to provide reliable results [17].

Banks' bootstrap method [4] is a smoothed bootstrap method for real-valued univariate data. The original data points are ordered as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, and the sample

space $[a, b]$ is divided into $n+1$ intervals by the observations, where the end points $x_{(0)}$ and $x_{(n+1)}$ are equal to a and b, respectively. Each interval $(x_{(i)}, x_{(i+1)})$ for $i = 0, 1, 2, \ldots, n$ is assigned a probability of $\frac{1}{n+1}$. To create a bootstrap sample, we randomly select n intervals with replacement, and then sample one observation uniformly from each selected interval. Based on the bootstrap sample, we calculate the function of interest and repeat this process B times to obtain $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$. The empirical distribution of these values approximates the sampling distribution of $\theta(F)$. Banks' bootstrap method is used for hypothesis testing in this paper and will be compared to Efron's bootstrap method in Sect. 4.

2.2 Bootstrap Methods for Right-Censored Univariate Data

This section presents Efron's bootstrap method [14] and the smoothed bootstrap method for right-censored data [1, 2]. Let $T_1, T_2, \ldots, T_n$ be independent and identically distributed event random variables from a distribution F supported on $\mathbb{R}^+$ and let $C_1, C_2, \ldots, C_n$ be independent and identically distributed right-censored random variables from a distribution G supported on $\mathbb{R}^+$. Furthermore, let $(X_1, D_1), (X_2, D_2), \ldots, (X_n, D_n)$ be the right-censored random variables, where each pair can be derived by

$$X_i = \begin{cases} T_i & \text{if } T_i \leq C_i \text{ (uncensored)} \\ C_i & \text{if } T_i > C_i \text{ (censored)} \end{cases} \tag{1}$$

$$D_i = \begin{cases} 1 & \text{if } X_i = T_i \text{ (uncensored)} \\ 0 & \text{if } X_i = C_i \text{ (censored)} \end{cases} \tag{2}$$

where $i = 1, 2, \ldots, n$. Let $(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)$ be the observations of the corresponding random quantities $(X_1, D_1), (X_2, D_2), \ldots, (X_n, D_n)$ and $\theta(F)$ is the function of interest, where this function can be estimated by $\theta(\hat{F})$.

Efron [14] proposed a nonparametric bootstrap method for data with right-censored observations. This method is similar to the one he proposed for real-valued data. In this method, the empirical distribution function of the original sample is used, so that each observation has an equal probability of $\frac{1}{n}$, regardless of whether it is an event or a censored observation. To apply this method, B bootstrap samples of size n are generated by randomly selecting observations from the original dataset with replacement. The function of interest is then calculated based on each bootstrap sample. This process results in values $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$, where the empirical distribution of these values can be a good estimate for the sampling distribution of $\theta(F)$. This bootstrap method is useful for testing the equality of average lifetimes over two populations [25], and it has been shown to provide good results in multiple statistical inferences, see Efron [15], Efron and Tibshirani [16, 17] for more details.

Another method for right-censored data is the smoothed bootstrap method, introduced by Al Luhayb [1] and Al Luhayb et al. [2]. This method generalises Banks' bootstrap method for right-censored data, and is based on the generalisation of the $A_{(n)}$ assumption for data that contains right-censored observations, proposed by Coolen and Yan [9]. To implement this method, the data support is divided into $n+1$ intervals

by the original data, and the right-censored $A_{(n)}$ assumption is used to assign specific probabilities to these intervals. For each bootstrap sample, n intervals are resampled with the assignment probabilities, and one observation is sampled from each interval. Performing these steps B times creates B bootstrap samples. Then, the function of interest is computed for each bootstrap sample, resulting in the values $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$. The empirical distribution of these values is used to estimate the sampling distribution of $\theta(F)$. In this paper, we use the smoothed bootstrap method for hypothesis testing and compare its performance to Efron's bootstrap method, with the comparison results presented in Sect. 4.

2.3 Bootstrap Methods for Bivariate Data

In this section, we will discuss Efron's bootstrap method [16] and three smoothed bootstrap methods for bivariate data [1, 3]. Let $(X_i, Y_i) \in \mathbb{R}^2$, for $i = 1, 2, \ldots, n$ denote independent and identically distributed random variables with a distribution of H. The observations corresponding to (X_i, Y_i) are (x_i, y_i). We are interested in $\theta(H)$, which is estimated by $\theta(\hat{H})$. To implement the bootstrap, Efron and Tibshirani [16] used the empirical distribution. The bootstrap method involves creating multiple bootstrap samples, say B, of size n by resampling with equal probability from the observed data. Based on each bootstrap sample, the function of interest is calculated, resulting in B values. The empirical distribution of these B values is used as a proxy for the distribution of the function of interest. This is the same approach as for univariate data. Several references use this bootstrap method for hypothesis testing. For further details, see e.g. Dolker et al. [11], MacKinnon [19] and Hesterberg [18].

In their recent work, Al Luhayb [1] and Al Luhayb et al. [3] proposed three different smoothed bootstrap methods for estimating the distribution of a function of interest. The first smoothed bootstrap method, referred to by SBSP, is based on the semi-parametric predictive method, which is proposed by Muhammad [20]. The second smoothed bootstrap method, referred to by SBNP, is based on the nonparametric predictive method introduced by Muhammad et al. [21]. These two methods divide the sample space into $(n + 1)^2$ squares (or blocks hereafter), each assigned with a certain probability. The third method, referred to by SEB, is based on uniform kernels, where each data point is surrounded by a block of size $b_X \times b_Y$, and the observation is located at the centre of its corresponding block, with b_X and b_Y being the chosen bandwidths for the kernel. To create a bootstrap sample, n blocks are resampled with the assignment probabilities, and one observation is sampled from each chosen block. This process is repeated multiple times, typically $B = 1000$ times, and based on each bootstrap sample, the function of interest is calculated. This results in B values, and the empirical distribution of these values is used to estimate the distribution of the function of interest.

Table 1 Yearly maximum flow rates (gallons per second) at a gauging station in North Carolina

5550	4380	2370	3220	8050	4560	2100
6840	5640	3500	1940	7060	7500	5370
13,100	4920	6500	4790	6050	4560	3210
6450	5870	2900	5490	3490	9030	3100
4600	3410	3690	6420	10,300	7240	9130

Table 2 The 90% confidence intervals for the median based on Efron's bootstrap method and Banks' bootstrap method

	Efron's method	Banks' method
90% confidence interval	(4560, 6050)	(4532, 6167)

3 Example

In this section, we will explore an example using data from the literature on the maximum flow rates over a 100 year period at gauging stations on rivers in North Carolina [6]. The data is presented in Table 1, and it shows the maximum flow rates in gallons per second. Our goal is to investigate whether the median of the data is equal to 5400 gallons per second using a 90% confidence interval, using Efron's bootstrap method and Banks' bootstrap method.

To conduct the test, we first generate 1000 bootstrap data sets from the original data using each of the two bootstrap methods, resulting in 1000 bootstrap samples for each method. Then, we calculate the median for each bootstrap sample, and from the resulting values, we can define the 90% bootstrap confidence interval for the median by taking the 50th and 950th ordered values.

If the value 5400 is included in the confidence interval, we fail to reject the null hypothesis. Otherwise, we reject the null hypothesis. Table 2 presents the 90% confidence intervals for the median based on both Efron's and Banks' bootstrap methods. As the value 5400 falls within both confidence intervals, therefore we fail to reject the null hypothesis.

4 Comparison of the Bootstrap Methods

Hypothesis tests based on the bootstrap method are a type of computer-based statistical technique. Thanks to recent advancements in computational power, these tests have become practical for real-world applications. The basic idea behind the bootstrap method is simple to understand and doesn't rely on complex mathematical assumptions. In this section, we will conduct various tests for different types of data using the bootstrap method explained in Sect. 2.

4.1 Hypothesis Tests for Quartiles

In this section, we calculate the Type I error rates of quartile hypothesis tests based on bootstrap methods presented in Sect. 2.2. These methods are used when the data contains right-censored observations. To determine how well the bootstrap methods perform, we simulate datasets that include right-censored observations from two different scenarios. For the first scenario, we use the Beta distribution with parameters $\alpha = 1.2$ and $\beta = 3.2$, where α and β are the shape parameters, and the Uniform distribution with parameters $a = 0$ and $b = 1.82$ for event time observations and right-censored observations, respectively. The second scenario is defined as $T \sim \text{Log-Normal}(\mu = 0, \sigma = 1)$ and $C \sim \text{Weibull}(\alpha = 3, \beta = 3.7)$, where α is the shape parameter and β is the scale parameter (see Appendix). In both scenarios, the censoring proportion p in the generated datasets is 15%, and this is determined by setting the two parameters of the uniform distribution. For more information on how to fix the censoring proportion, we refer the reader to Wan [26] and Al Luhayb [1].

To compare Efron's bootstrap method with the smoothed bootstrap method, we generate $N = 1000$ datasets from each scenario. For each dataset, we apply each method $B = 1000$ times, resulting in 1000 bootstrap samples based on each method. We then compute the quartile of interest at each bootstrap sample and use the resulting values to define the $100(1 - 2\alpha)\%$ bootstrap confidence interval for the quartile. We count one if the value of the quartile specified in the null hypothesis is not included in the confidence interval; otherwise, we count zero. We repeat this procedure for all $N = 1000$ generated datasets, then count the number of times the null hypothesis was rejected over the 1000 trials. This ratio will be the Type I error rate of the quartile's hypothesis test with significance level 2α.

It's important to note that Efron's bootstrap samples often include some censored observations, so we use the Kaplan–Meier (KM) estimator to find their corresponding quartiles. Suppose we are interested in the median; we should find a time t such that $\hat{S}(t) = 0.50$ in each bootstrap sample. Unfortunately, in some samples, we cannot find that time t because there is no time such that $\hat{S}^{-1}(0.50) = t$. In this case, we have considered three options or solutions. The first option is to neglect all not applicable medians, so the $100(1 - 2\alpha)\%$ bootstrap confidence interval for the median is based on fewer than 1000 bootstrap samples. This option is referred to as $\text{E}_{(1)}$. The second option is to assume the median to be the maximum event time of that bootstrap sample. This is Efron's suggestion, which is used for each bootstrap sample whose median is not found by the KM estimator [12]. This option is referred to as $\text{E}_{(2)}$. Finally, we fit an Exponential distribution to the interval with a rate parameter of $\hat{\lambda}^* = -\ln(\hat{S}(t_{max}))/t_{max}$, where t_{max} is the maximum event time of the bootstrap sample and $\hat{S}(.)$ is the KM estimator. This allows us to find the corresponding median, X_{med}, with $X_{med} = -\ln(0.50)/\hat{\lambda}^*$. This suggestion is presented in Brown et al. [7], and we refer to it as $\text{E}_{(3)}$. In the last two cases, we can ensure that the confidence interval is based on 1000 bootstrap samples' medians.

In the tables, the NA represents the number of Efron's bootstrap samples where quartiles cannot be found, while ABS represents the number of cases where a bootstrap

sample containing only right-censored observations is replaced by another sample that includes at least one event time. These two numbers are out of 1,000,000.

We consider three different strategies for the smoothed bootstrap method when sampling observations from the $n+1$ intervals partitioning the sample space. The first strategy is to sample uniformly from all intervals, denoted by SB. The second strategy is to assume an exponential tail for each interval and sample from the tails to create the bootstrap samples, denoted by SB_{exp}. The third strategy is to sample uniformly from all intervals except the last intervals, for which we sample from the exponential tails. We refer to this strategy as SB_{Lexp}. By investigating how the sampling strategies affect the results, we can gain insight into the impact of different sampling methods on the smoothed bootstrap method.

Tables 3 and 4 show the results of the Type I error rates for the quartiles' hypothesis tests with significance levels 0.10 and 0.05 for simulated data sets in the first scenario. When the sample size is 10, the smoothed bootstrap with its three assumptions, SB, SB_{exp} and SB_{Lexp}, provides lower discrepancies between actual and nominal error rates for all quartiles' tests compared to Efron's bootstrap with its three assumptions, $\text{E}_{(1)}$, $\text{E}_{(2)}$ and $\text{E}_{(3)}$. The superiority of the smoothed bootstrap methods is due not only to the event observations obtained for the smoothed bootstrap samples, but also to the fact that the KM estimator used in Efron's bootstrap samples is often not able to find the quartiles, particularly the second and third ones. In 1,000,000 bootstrap samples, we cannot find the first, second and third quartiles in 228, 3736 and 32,821 bootstrap samples, respectively. As the sample size increases to 50, 100 and 500, both methods provide good results, but Efron's method is better, and the number of NA and ABS decreases toward zero. These decreases lead to equal results when $\text{E}_{(1)}$, $\text{E}_{(2)}$ and $\text{E}_{(3)}$ are used. Also, at these large sample sizes, SB, SB_{exp} and SB_{Lexp} provide approximately equal outcomes.

In the second scenario, we should note that the data space is $(0, \infty)$, which is different from the first scenario where the support is (0, 1), so the last intervals for the smoothed method are not bounded. In this case, we can only use smoothed bootstrap assumptions SB_{exp} and SB_{Lexp}, not SB. Tables 5 and 6 present the results of Type I error rates for the quartiles' hypothesis tests with significance levels of 0.10 and 0.05, respectively. The SB_{exp} and SB_{Lexp} methods again outperform Efron's method in defining the Type I error rates when the sample size is small. As the sample size gets large, both methods perform well, as observed in Tables 3 and 4.

In a special case where data includes only failures, with no censored observations, we will use Banks' bootstrap method and Efron's bootstrap method, which are presented in Sect. 2.1, to compute the Type I error rates for the quartiles' hypothesis tests. In the simulations, we use Beta($\alpha = 1.2, \beta = 3.2$) to create data sets and repeat the same comparison procedure as in the previous simulations. Tables 7 and 8 present the Type I error rates for the quartiles' hypothesis tests based on Banks' and Efron's methods with significance levels of 0.10 and 0.05, respectively. Banks' bootstrap method performs better, particularly when $n = 10$ and $2\alpha = 0.05$. As the sample size gets large, both methods perform well.

Table 3 Type I error rates with significance level $2\alpha = 0.10$, $T \sim \text{Beta}(\alpha = 1.2, \beta = 3.2)$, $C \sim \text{Unif}(a = 0, b = 1.82)$ and $p = 0.15$

H_0 :		$Q_1 = 0.117$						$Q_2 = 0.236$						$Q_3 = 0.396$					
n	Measures	SB	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$
10	Type I	0.103	0.103	0.105	0.107	0.107	0.107	0.090	0.096	0.097	0.151	0.151	0.149	0.068	0.110	0.111	0.200	0.202	0.172
	NA	–	–	–	228	0	0	–	–	–	3736	0	0	–	–	–	32,821	0	0
	ABS	–	–	–	12	12	12	–	–	–	12	12	12	–	–	–	12	12	12
50	Type I	0.098	0.098	0.101	0.108	0.108	0.108	0.126	0.126	0.114	0.117	0.117	0.117	0.121	0.121	0.126	0.108	0.107	0.107
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	56	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
100	Type I	0.100	0.100	0.098	0.100	0.100	0.100	0.120	0.120	0.117	0.104	0.104	0.104	0.133	0.133	0.134	0.114	0.114	0.114
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
500	Type I	0.104	0.104	0.104	0.100	0.100	0.100	0.126	0.126	0.126	0.110	0.110	0.110	0.121	0.121	0.121	0.094	0.094	0.094
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0

Table 4 Type I error rates with significance level $2\alpha = 0.05$, $T \sim \text{Beta}(\alpha = 1.2, \beta = 3.2)$, $C \sim \text{Unif}(a = 0, b = 1.82)$ and $p = 0.15$

H_0 :		$Q_1 = 0.117$						$Q_2 = 0.236$						$Q_3 = 0.396$					
n	Measures	SB	SB_{exp}	SB_{Lexp}	$\text{E}_{(1)}$	$\text{E}_{(2)}$	$\text{E}_{(3)}$	SB	SB_{exp}	SB_{Lexp}	$\text{E}_{(1)}$	$\text{E}_{(2)}$	$\text{E}_{(3)}$	SB	SB_{exp}	SB_{Lexp}	$\text{E}_{(1)}$	$\text{E}_{(2)}$	$\text{E}_{(3)}$
10	Type I	0.051	0.049	0.050	0.088	0.088	0.088	0.046	0.048	0.050	0.070	0.070	0.068	0.020	0.072	0.065	0.183	0.181	0.146
	NA	–	–	–	228	0	0	–	–	–	3736	0	0	–	–	–	32,821	0	0
	ABS	–	–	–	12	12	12	–	–	–	12	12	12	–	–	–	12	12	12
50	Type I	0.054	0.054	0.045	0.059	0.059	0.059	0.066	0.066	0.070	0.069	0.069	0.069	0.067	0.067	0.067	0.059	0.059	0.059
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	56	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
100	Type I	0.047	0.047	0.052	0.045	0.045	0.045	0.057	0.057	0.061	0.057	0.057	0.057	0.078	0.078	0.083	0.061	0.061	0.061
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
500	Type I	0.058	0.058	0.058	0.054	0.054	0.054	0.062	0.062	0.062	0.054	0.054	0.054	0.072	0.072	0.072	0.049	0.049	0.049
	NA	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0
	ABS	–	–	–	0	0	0	–	–	–	0	0	0	–	–	–	0	0	0

Table 5 Type I error rates with significance level $2\alpha = 0.10$, $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$) and $p = 0.15$

H_0 :		$Q_1 = 0.509$					$Q_2 = 1$					$Q_3 = 1.963$				
n	Measures	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$
10	Type I	0.092	0.096	0.103	0.103	0.103	0.098	0.093	0.128	0.119	0.126	0.104	0.108	0.287	0.304	0.172
	NA	–	–	1813	0	0	–	–	23,589	0	0	–	–	167,582	0	0
	ABS	–	–	61	61	61	–	–	61	61	61	–	–	61	61	61
50	Type I	0.089	0.092	0.121	0.121	0.121	0.095	0.092	0.106	0.106	0.106	0.118	0.115	0.118	0.119	0.119
	NA	–	–	0	0	0	–	–	2	0	0	–	–	18,178	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
100	Type I	0.084	0.090	0.100	0.100	0.100	0.097	0.102	0.101	0.101	0.101	0.119	0.117	0.116	0.117	0.117
	NA	–	–	0	0	0	–	–	0	0	0	–	–	1421	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
500	Type I	0.103	0.106	0.098	0.098	0.098	0.106	0.103	0.104	0.104	0.104	0.120	0.120	0.112	0.112	0.112
	NA	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0

Table 6 Type I error rates with significance level $2\alpha = 0.05$, $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$) and $p = 0.15$

H_0 :		$Q_1 = 0.509$					$Q_2 = 1$					$Q_3 = 1.963$				
n	Measures	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	SB_{exp}	SB_{Lexp}	$E_{(1)}$	$E_{(2)}$	$E_{(3)}$
10	Type I	0.040	0.045	0.084	0.084	0.084	0.047	0.050	0.070	0.069	0.070	0.065	0.069	0.250	0.268	0.138
	NA	–	–	1813	0	0	–	–	23,589	0	0	–	–	167,582	0	0
	ABS	–	–	61	61	61	–	–	61	61	61	–	–	61	61	61
50	Type I	0.047	0.049	0.066	0.066	0.066	0.054	0.054	0.056	0.056	0.056	0.066	0.059	0.062	0.062	0.062
	NA	–	–	0	0	0	–	–	2	0	0	–	–	18,178	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
100	Type I	0.042	0.046	0.047	0.047	0.047	0.050	0.047	0.049	0.049	0.049	0.061	0.065	0.065	0.065	0.065
	NA	–	–	0	0	0	–	–	0	0	0	–	–	1421	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
500	Type I	0.054	0.057	0.054	0.054	0.054	0.047	0.043	0.050	0.050	0.050	0.070	0.066	0.062	0.062	0.062
	NA	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0
	ABS	–	–	0	0	0	–	–	0	0	0	–	–	0	0	0

Table 7 Type I error rates with significance level $2\alpha = 0.10$, Beta($\alpha = 1.2, \beta = 3.2$) and $p = 0$

H_0 :	$Q_1 = 0.117$		$Q_2 = 0.236$		$Q_3 = 0.396$	
n	Banks	Efron	Banks	Efron	Banks	Efron
10	0.102	0.099	0.080	0.136	0.081	0.096
50	0.089	0.113	0.099	0.112	0.099	0.111
100	0.099	0.103	0.113	0.109	0.095	0.103
500	0.097	0.103	0.101	0.102	0.087	0.091

Table 8 Type I error rates with significance level $2\alpha = 0.05$, Beta($\alpha = 1.2, \beta = 3.2$) and $p = 0$

H_0 :	$Q_1 = 0.117$		$Q_2 = 0.236$		$Q_3 = 0.396$	
n	Banks	Efron	Banks	Efron	Banks	Efron
10	0.052	0.089	0.046	0.064	0.014	0.086
50	0.046	0.059	0.058	0.060	0.055	0.069
100	0.043	0.042	0.054	0.060	0.054	0.058
500	0.052	0.058	0.057	0.056	0.040	0.042

4.2 The Two-Sample Problem

When conducting a hypothesis test $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 \neq \theta_2$, where θ_1 and θ_2 represent the function of interest in the first and second populations respectively, the achieved significance level (ASL) is used to draw a conclusion. ASL is defined as the probability of observing at least the same value as $\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$, when the null hypothesis is true,

$$ASL = \text{Prob}_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\} \tag{3}$$

The smaller the value of ASL, the stronger the evidence against H_0. The value $\hat{\theta}$ is fixed at its observed value, and the quantity $\hat{\theta}^*$ has the null hypothesis distribution, which is the distribution of $\hat{\theta}$ if H_0 is true [17].

Efron and Tibshirani [17] used the achieved significance level to test whether the two populations have equal mean or not. Suppose we have two samples $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ from possibly different probability distributions, and we wish to test the null hypothesis $H_0 : \theta_1 = \theta_2$. Efron's bootstrap method is used to approximate the ASL value, then H_0 is rejected when $\widehat{ASL} < 2\alpha$. The algorithm to test the null hypothesis based on the bootstrap methods is as follows

(i) Combine $\mathbf{z}$ and $\mathbf{y}$ samples together, so we get a sample $\mathbf{x}$ of size $n + m$. Thus, $\mathbf{x} = \{z_1, z_2, \ldots, z_n, y_1, y_2, \ldots, y_m\}$
(ii) Draw B bootstrap samples of size $n+m$ with replacement from $\mathbf{x}$, and call the first n observations $\mathbf{z}^{*b}$ and the remaining m observations $\mathbf{y}^{*b}$ for $b = 1, 2, \ldots, B$.
(iii) For each bootstrap sample, we compute the means of $\mathbf{z}^{*b}$ and $\mathbf{y}^{*b}$, then find $A^{*b} = \overline{\mathbf{z}}^{*b} - \overline{\mathbf{y}}^{*b}$, $b = 1, 2, \ldots, B$.

Table 9 Type I error rates with significance level $2\alpha = 0.10$, and all samples created by $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$), where $p = 0.15$

n	SB_{exp}	SB_{Lexp}	$E_{(2)}$	$E_{(3)}$
10	0.078	0.075	0.091	0.089
50	0.079	0.079	0.090	0.090
100	0.100	0.101	0.107	0.107
500	0.105	0.101	0.104	0.104

(iv) The achieved significance level ASL can be approximated by

$$\widehat{ASL} = \frac{\sum_{b=1}^{B} \{A^{*b} \geq A_{obs}\}}{B} \tag{4}$$

where $A_{obs} = \bar{\mathbf{z}} - \bar{\mathbf{y}}$, and $\bar{\mathbf{z}}$ and $\bar{\mathbf{y}}$ are the sample means of the two original samples.

We will employ the proposed strategy in this section to examine whether the two samples have the same median ($Q_2^1 = Q_2^2$) or not. To conduct these tests, we will use the bootstrap methods presented in Sect. 2.2 and make comparisons through simulations. Specifically, we will calculate the Type I error rate for the following hypothesis test:

$$H_0 : Q_2^1 = Q_2^2 \quad \text{versus} \quad H_1 : Q_2^1 \neq Q_2^2 \tag{5}$$

In order to compare different bootstrap methods through simulation, we first generate two datasets of size n using the second scenario proposed in Sect. 4.1. We compute the medians of these datasets, $\hat{Q}_2^1$ and $\hat{Q}_2^2$, and calculate $A_{obs} = \hat{Q}_2^1 - \hat{Q}_2^2$. Next, we combine the two datasets so that they form a new dataset of size $2n$. Then, for each bootstrap method, we draw 1000 samples of size $2n$, and call the first n observations $\mathbf{z}^{*b}$ and the remaining n observations $\mathbf{y}^{*b}$ for $b = 1, 2, \ldots, B$. We compute $A^{*b} = \hat{Q}_2(\mathbf{z}^{*b}) - \hat{Q}_2(\mathbf{y}^{*b})$ for each bootstrap sample, resulting in 1000 A^* values. Finally, we calculate the ASL value and reject H_0 if $\widehat{ASL} < 2\alpha$. We repeat this process $B = 1000$ times and count the number of times we reject the null hypothesis. We take the ratio of rejected hypotheses out of 1000 trials and consider the method with the ratio closest to 2α as the best method. The final results of the simulations are presented in Tables 9 and 10 for two different significance levels.

As the sample space of the underlying distribution is $[0, \infty)$, we only consider SB_{exp} and SB_{Lexp} for the smoothed bootstrap method. For Efron's method, we consider $E_{(2)}$ and $E_{(3)}$ as they are guaranteed to find the median of each set in each bootstrap sample. Tables 9 and 10 present the Type I error rates of the hypothesis test in Equation (5) with significance levels of 0.10 and 0.05, respectively. The SB_{exp} and SB_{Lexp} methods generally provide lower actual Type I error rates compared to $E_{(2)}$ and $E_{(3)}$ at different sample sizes. However, $E_{(2)}$ and $E_{(3)}$ provide smaller discrepancies between the actual and nominal Type I error levels, especially when the sample size is small. When $n = 500$, all methods provide almost identical results.

Table 10 Type I error rates with significance level $2\alpha = 0.05$, and all samples created by $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$), where $p = 0.15$

n	SB_{exp}	SB_{Lexp}	$E_{(2)}$	$E_{(3)}$
10	0.025	0.025	0.031	0.031
50	0.039	0.041	0.039	0.039
100	0.047	0.046	0.049	0.049
500	0.043	0.042	0.043	0.043

Table 11 Type I error rates with significance level $2\alpha = 0.10$, the first samples from $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$), where $p = 0.15$ and the second samples from $T \sim$ Weibull($\alpha = 1, \beta = 1.443$), $C \sim$ Exponential($\lambda = 0.12$), where $p = 0.15$

n	SB_{exp}	SB_{Lexp}	$E_{(2)}$	$E_{(3)}$
10	0.082	0.079	0.083	0.083
50	0.103	0.105	0.095	0.095
100	0.101	0.097	0.093	0.093
500	0.089	0.092	0.084	0.084

Table 12 Type I error rates with significance level $2\alpha = 0.05$, the first samples from $T \sim$ Log-Normal($\mu = 0, \sigma = 1$), $C \sim$ Weibull($\alpha = 3, \beta = 3.7$), where $p = 0.15$ and the second samples from $T \sim$ Weibull($\alpha = 1, \beta = 1.443$), $C \sim$ Exponential($\lambda = 0.12$), where $p = 0.15$

n	SB_{exp}	SB_{Lexp}	$E_{(2)}$	$E_{(3)}$
10	0.030	0.027	0.038	0.038
50	0.046	0.047	0.046	0.046
100	0.041	0.043	0.034	0.034
500	0.045	0.047	0.043	0.043

In previous simulations, we created both samples in each run from a single scenario, but now we want to create samples from two different scenarios. In each run, the first sample is created from $T \sim$ Log-Normal($\mu = 0, \sigma = 1$) and $C \sim$ Weibull($\alpha = 3, \beta = 3.7$), while the second sample is created from $T \sim$ Weibull($\alpha = 1, \beta = 1.443$) and $C \sim$ Exponential($\lambda = 0.12$), where $p = 0.15$ in both scenarios (see Appendix). We aim to investigate how the bootstrap methods perform when the two samples have different distributions but the same median (which is equal to 1). Tables 11 and 12 show the Type I error rates with significance levels of 0.10 and 0.05, respectively. All methods perform well at different sample sizes, and the results are close to the nominal size 2α, particularly when the sample size is large.

4.3 Pearson Correlation Test

In Sect. 2.3, we present smoothed bootstrap methods and compare them to Efron's method. We compute the Type I error rate to determine the superiority of each method, where a method is considered superior if its corresponding Type I error rate is closer to the significance level of 2α. In this section, we simulate data sets from two different distributions to compare the methods. For the first scenario, we generate data sets from Gumbel copula, where the marginals X and Y both follow the standard uniform distri-

bution. The second scenario is Clayton copula where X follows the normal distribution with mean 1 and standard deviation 1, and Y follows the normal distribution with mean 5 and standard deviation 3. For both scenarios, we consider three dependence levels of ρ and three sample sizes with two significance levels. We also include the dependence parameters of copulas and their concordance measure Kendall's τ. The cumulative distribution functions of Gumbel copula and Clayton copula are, respectively, given by [22]

$$C_g(u, v|\theta_g) = \exp\left(-\left[(-\ln(u))^{\theta_g} + (-\ln(v))^{\theta_g}\right]^{1/\theta_g}\right) \tag{6}$$

$$C_c(u, v|\theta_c) = \max\left[\left(u^{-\theta_c} + v^{-\theta_c} - 1\right)^{-1/\theta_c}, 0\right] \tag{7}$$

where all marginals are uniformly distributed on [0,1].

To compute the Type I error rate for the null hypothesis of $\rho = \rho^\star$ based on a bootstrap method, we create $N = 1000$ data sets with sample size n and dependence level $\rho = \rho^\star$ from one of the scenarios presented above. For each generated data set, we apply each bootstrap method $B = 1000$ times and compute the Pearson correlation of each bootstrap sample. We order the 1000 Pearson correlation bootstrapped values from lowest to highest and obtain the $100(1 - 2\alpha)\%$ bootstrap confidence interval. If the null hypothesis value is not included in the confidence interval, we reject H_0 and count 1; otherwise, we do not reject H_0 and count 0. The number of times that the null hypothesis was rejected over the 1000 trials will be the Type I error rate.

Table 13 presents the Type I error rates based on the bootstrap methods, where the significance level is 0.10. For a small sample size of $n = 10$, the SBSP and SBNP methods provide error rates closer to the nominal rate of 0.10 compared to Efron's and the smoothed Efron's methods. However, the SBNP method is the best when $\rho = 0.4$ and 0.8. When n increases to 50 and 100, all methods decrease the discrepancies between the actual and nominal error rates, but the SBNP method is the superior one in most cases.

With a significance level of 0.05, the actual Type I error rates based on the bootstrap methods are listed in Table 14. The SBSP and SBNP methods again provide lower discrepancies between the nominal and actual Type I error rates compared to Efron's and the smoothed Efron's methods, especially when $n = 10$. When the sample size increases to 50 and 100, all methods perform better, but the SBNP method is the best one in most settings.

In the second scenario, we simulate $N = 1000$ data sets with dependence level $\rho = \rho^\star$, and we compute Type I error rates using the bootstrap methods as shown in Tables 15 and 16. For $n = 10$, the SBSP method provides the closest results to the nominal error rates at most levels of ρ. As n increases to 50 and 100, its performance worsens for $H_0 : \rho = 0.8$ because the underlying distribution is not symmetric. At these large sample sizes, the SBNP, Efron and SEB methods perform better than the SBSP method, particularly the SBNP method. The SBNP method provides the lowest discrepancies between the nominal and actual error rates in most cases, in both significance levels of 0.10 and 0.05; however, when $n = 10$ and $\rho = 0, 0.4$, the SBNP method provides very small error rates.

Table 13 Type I error rates with significance level 0.10, Gumbel copula, $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	1	$\rho = 0$	0.114	0.120	0.139	0.142	0.105	0.113	0.106	0.105	0.106	0.102	0.107	0.106
0.275	1.3793	$\rho = 0.4$	0.137	0.129	0.147	0.149	0.136	0.122	0.128	0.127	0.129	0.105	0.109	0.106
0.610	2.5641	$\rho = 0.8$	0.133	0.075	0.189	0.184	0.129	0.123	0.121	0.126	0.126	0.103	0.111	0.107

Table 14 Type I error rates with significance level 0.05, Gumbel copula, $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	1	$\rho = 0$	0.064	0.072	0.085	0.081	0.046	0.051	0.053	0.058	0.056	0.052	0.055	0.057
0.275	1.3793	$\rho = 0.4$	0.075	0.070	0.100	0.098	0.067	0.079	0.080	0.075	0.066	0.061	0.061	0.058
0.610	2.5641	$\rho = 0.8$	0.079	0.034	0.131	0.127	0.074	0.070	0.078	0.076	0.080	0.066	0.071	0.071

Table 15 Type I error rates with significance level 0.10, Clayton copula, $X \sim \text{Normal}(\mu = 1, \sigma = 1)$ and $Y \sim \text{Normal}(\mu = 5, \sigma = 3)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	0	$\rho = 0$	0.119	0.026	0.144	0.147	0.119	0.097	0.117	0.115	0.116	0.097	0.102	0.102
0.259	0.6990	$\rho = 0.4$	0.142	0.039	0.167	0.165	0.150	0.102	0.122	0.125	0.135	0.114	0.116	0.119
0.630	3.4054	$\rho = 0.8$	0.144	0.175	0.189	0.196	0.218	0.110	0.141	0.132	0.277	0.104	0.111	0.118

Table 16 Type I error rates with significance level 0.05, Clayton copula, $X \sim \text{Normal}(\mu = 1, \sigma = 1)$ and $Y \sim \text{Normal}(\mu = 5, \sigma = 3)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	0	$\rho = 0$	0.065	0.009	0.086	0.086	0.053	0.048	0.060	0.064	0.063	0.056	0.054	0.058
0.259	0.6990	$\rho = 0.4$	0.088	0.012	0.108	0.105	0.076	0.048	0.066	0.068	0.083	0.066	0.063	0.070
0.630	3.4054	$\rho = 0.8$	0.080	0.039	0.130	0.132	0.147	0.051	0.079	0.079	0.200	0.054	0.062	0.063

4.4 Kendall Correlation Test

In Sect. 4.3, we computed the Type I error rate for the Pearson correlation test using different sample sizes and dependence levels. In this section, we aim to repeat the same comparisons, but this time, we will use the Kendall correlation test instead. We will use the same scenarios, generating datasets with $n = 10$, 50 and 100, and dependence levels of $\tau = 0$, 0.4 and 0.8, with significance levels of 0.10 and 0.05.

To generate data sets and apply the bootstrap methods, we will use the Gumbel copula, where both marginals follow Uniform(0,1). From Tables 17 and 18, we can see that the SBSP method performs well when $\tau = 0$ across all different sample sizes. However, it performs poorly as the sample size increases for $\tau = 0.4$ and 0.8. This is in contrast to the results based on SBNP, Efron's, and smoothed Efron's methods. These methods provide lower error rates than the nominal levels when the sample size is small at all different dependence levels. As n increases to 50 and 100, the error rates become closer to the nominal level 2α.

Tables 19 and 20 present the Type I error rates for the Kendal correlation test at different dependence levels with significance levels of 0.10 and 0.05, respectively. When $\tau = 0$ and $n = 10$, the error rate based on the SBNP method is significantly lower than the nominal level 2α, while the results of other methods are close to the nominal levels. As the sample size increases to 50 and 100, all methods provide good results. If there is a strong relation between the variables, it is recommended to use either Efron's bootstrap method or the SEB method. These methods are both able to produce good results because they have much less effect than the SBSP and SBNP methods on the observation's rank, which is the basis for computing the Kendall correlation.

5 Concluding Remarks

In this paper, we explored how the proposed smoothed bootstrap methods can be used to compute Type I error rates for different hypothesis tests and compare their results to Efron's bootstrap methods through simulations. The smoothed bootstrap methods are applied to real-valued data, right-censored data and bivariate data. For real-valued data and right-censored data, we test the null hypothesis that quartiles are equal to those of the underlying distributions. We also test whether two sample medians are equal, regardless of whether the two samples are from the same underlying distribution or not. For bivariate data, we compute the Type I error rates for Pearson and Kendall correlation tests.

We found that the smoothed bootstrap methods perform better when the sample size is small for real-valued and right-censored data, providing lower discrepancies between actual and nominal error rates. As the sample size gets larger, all bootstrap methods provide good results, but Efron's methods mostly perform better for the third quartile. For the two-sample median test, we use the achieved significance level to test whether the two samples have equal medians or not. All bootstrap methods performed well, and the Type I error rates are close to the nominal levels.

Table 17 Type I error rates of Kendall correlation test with significance level 0.10, Gumbel copula, $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	1	$\tau = 0$	0.103	0.076	0.073	0.076	0.103	0.094	0.087	0.092	0.105	0.096	0.102	0.099
0.4	1.667	$\tau = 0.4$	0.120	0.059	0.078	0.065	0.133	0.107	0.110	0.100	0.128	0.100	0.100	0.098
0.8	5	$\tau = 0.8$	0.047	0.062	0.094	0.046	0.132	0.076	0.076	0.077	0.130	0.077	0.081	0.070

Table 18 Type I error rates of Kendall correlation test with significance level 0.05, Gumbel copula, $X \sim \text{Unif}(0, 1)$ and $Y \sim \text{Unif}(0, 1)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	1	$\tau = 0$	0.057	0.035	0.037	0.040	0.048	0.041	0.041	0.045	0.055	0.047	0.049	0.052
0.4	1.667	$\tau = 0.4$	0.063	0.032	0.038	0.035	0.071	0.053	0.055	0.055	0.079	0.047	0.052	0.049
0.8	5	$\tau = 0.8$	0.021	0.025	0.021	0.025	0.072	0.039	0.032	0.037	0.068	0.043	0.038	0.042

Table 19 Type I error rates of Kendall correlation test with significance level 0.10, Clayton copula, $X \sim \text{Normal}(\mu = 1, \sigma = 1)$ and $Y \sim \text{Normal}(\mu = 5, \sigma = 3)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	0	0	0.103	0.020	0.073	0.076	0.103	0.087	0.087	0.092	0.105	0.095	0.102	0.099
0.4	1.333	0.4	0.125	0.037	0.089	0.074	0.140	0.094	0.101	0.099	0.121	0.090	0.098	0.089
0.8	8	0.8	0.049	0.918	0.110	0.046	0.165	0.456	0.078	0.080	0.160	0.169	0.088	0.094

Table 20 Type I error rates of Kendall correlation test with significance level 0.05, Clayton copula, $X \sim \text{Normal}(\mu = 1, \sigma = 1)$ and $Y \sim \text{Normal}(\mu = 5, \sigma = 3)$

$n =$			10				50				100			
τ	θ	H_0 :	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB	SBSP	SBNP	Efron	SEB
0	0	0	0.057	0.006	0.037	0.040	0.048	0.033	0.041	0.045	0.055	0.055	0.049	0.052
0.4	1.333	0.4	0.065	0.013	0.041	0.032	0.076	0.039	0.050	0.046	0.067	0.045	0.054	0.048
0.8	8	0.8	0.020	0.749	0.027	0.019	0.096	0.307	0.044	0.028	0.107	0.103	0.043	0.040

For the Pearson correlation test, the SBSP and SBNP methods lead to lower discrepancies between actual and nominal Type I error rates compared to Efron's and smoothed Efron's methods when the sample size is small. For large sample sizes, all methods provide good results. However, the SBNP method performs better in most dependence levels. In situations where the data distribution is asymmetric, the SBSP method does not perform well, particularly when τ is not close to zero, which results from the Normal copula assumption.

For the Kendall correlation test, it is recommended to use either Efron's bootstrap method or the SEB method, particularly when the underlying distribution is asymmetric and has a strong Kendall correlation. Their influences on the observations rank are much less than those of the SBSP and SBNP methods. When $\tau = 0$ and the sample size is small, all bootstrap methods perform well, and as n gets large, their performances improve and the Type I error rates become closer to the nominal level 2α.

In conclusion, we used the bootstrap methods for real-valued data, right-censored data and bivariate data to compute Type I error rates for different hypothesis tests. Future research could focus on applying these bootstrap methods to compute power or Type II error rates for some hypothesis tests.

Acknowledgements Asamh Al Luhayb was a PhD student at Durham University, supported by a scholarship from the Deanship of Scientific Research at Qassim University. During his studies, he worked under the supervision of Prof. Frank Coolen and Dr. Tahani Coolen-Maturi.

Declarations

Conflicts of interest The author states that there is no conflict of interest.

Appendix

The probability density functions for the distributions used in each scenario to generate right-censored data.

Scenario 1:

Beta distribution for event times:

$f(t) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\beta(\alpha,\beta)}$; $t \in [0, 1]$ where $\alpha = 1.2$ and $\beta = 3.2$.

Uniform distribution for censored times:

$g(c) = \frac{1}{b-a}$; $c \in [a, b]$ where $a = 0$ and $b = 1.82$.

Scenario 2:

Log-Normal distribution for event times:

$f(t) = \frac{1}{t\sqrt{2\pi}} \exp(-\frac{(\ln(t))^2}{2});\ t \in (0, \infty)$.
Weibull distribution for censored times:
$g(c) = \frac{\alpha}{\beta}(\frac{c}{\beta})^{\alpha-1} \exp(-(\frac{c}{\beta})^{\alpha});\ c \in [0, \infty)$ where $\alpha = 3$ and $\beta = 3.7$.
Scenario 3:
Weibull distribution for event times:
$f(t) = \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1} \exp(-(\frac{t}{\beta})^{\alpha});\ t \in [0, \infty)$ where $\alpha = 1$ and $\beta = 1.443$.
Exponential distribution for censored times:
$g(c) = \lambda \exp(-\lambda c);\ c \in [0, \infty)$ where $\lambda = 0.12$.

References

1. Al Luhayb ASM (2021) Smoothed bootstrap methods for right-censored data and bivariate data. PhD thesis, Durham University. http://etheses.dur.ac.uk/14096
2. Al Luhayb ASM, Coolen FPA, Coolen-Maturi T (2023) Smoothed bootstrap for right-censored data. Commun Stat Theory Methods. https://doi.org/10.1080/03610926.2023.2171708
3. Al Luhayb ASM, Coolen-Maturi T, Coolen FPA (2023) Smoothed bootstrap methods for bivariate data. J Stat Theory Pract 17(3):1–37. https://doi.org/10.1007/s42519-023-00334-7
4. Banks DL (1988) Histospline smoothing the Bayesian bootstrap. Biometrika 75:673–684
5. Berrar D (2019) Introduction to the non-parametric bootstrap. In: Encyclopedia of bioinformatics and computational biology. Academic Press, Oxford, pp 766–773
6. Boos DD (2003) Introduction to the bootstrap world. Stat Sci 18(2):168–174
7. Brown BW, Hollander M, Korwar RM (1974) Nonparametric tests of independence for censored data with applications to heart transplant studies. In: Proschan F, Serfling RJ (eds) Reliability and biometry. SIAM, Philadelphia, pp 327–354
8. Coolen FPA, BinHimd S (2020) Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov–Smirnov test. J Stat Theory Pract 14:1–13
9. Coolen FPA, Yan KJ (2004) Nonparametric predictive inference with right-censored data. J Stat Plan Inference 126:25–54
10. Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge
11. Dolker M, Halperin S, Divgi DR (1982) Problems with bootstrapping Pearson correlations in very small bivariate samples. Psychometrika 47(4):529–530
12. Efron B (1967) The two-sample problem with censored data. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 4. University of California Press, Berkeley, pp 831–853
13. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
14. Efron B (1981) Censored data and the bootstrap. J Am Stat Assoc 76:312–319
15. Efron B (1982) The jackknife, the bootstrap, and other resampling plans, vol 38. Society for Industrial and Applied Mathematics, Philadelphia
16. Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1:54–77
17. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, London
18. Hesterberg T (2011) Bootstrap. Wiley Interdiscip Rev Comput Stat 3(6):497–526
19. MacKinnon JG (2009) Bootstrap hypothesis testing. In: Handbook of computational econometrics, pp 183–213
20. Muhammad N (2016) Predictive inference with copulas for bivariate data. PhD thesis, Durham University, UK
21. Muhammad N, Coolen FPA, Coolen-Maturi T (2016) Predictive inference for bivariate data with nonparametric copula. Am Inst Phys AIP Conf Proc 1750(1):0600041–0600048. https://doi.org/10.1063/1.4954609
22. Muhammad N, Coolen-Maturi T, Coolen FPA (2018) Nonparametric predictive inference with parametric copulas for combining bivariate diagnostic tests. Stat Optim Inf Comput 6(3):398–408

23. Rasmussen JL (1987) Estimating correlation coefficients: bootstrap and parametric approaches. Psychol Bull 101(1):136–139
24. Strube MJ (1988) Bootstrap type I error rates for the correlation coefficient: an examination of alternate procedures. Psychol Bull 104(2):290–292
25. Vaman H, Tattar P (2022) Survival analysis. Chemical Rubber Company Press, Boca Raton
26. Wan F (2017) Simulating survival data with predefined censoring rates for proportional hazards models. Stat Med 36:721–880