Bootstrap inference for the finite population mean under complex sampling designs

Zhonglei Wang\(^1\) | Liuhua Peng\(^2\) | Jae Kwang Kim\(^3\) ©

Abstract
Bootstrap is a useful computational tool for statistical inference, but it may lead to erroneous analysis under complex survey sampling. In this paper, we propose a unified bootstrap method for stratified multi-stage cluster sampling, Poisson sampling, simple random sampling without replacement and probability proportional to size sampling with replacement. In the proposed bootstrap method, we first generate bootstrap finite populations, apply the same sampling design to each bootstrap population to get a bootstrap sample, and then apply studentization. The second-order accuracy of the proposed bootstrap method is established by the Edgeworth expansion. Simulation studies confirm that the proposed bootstrap method outperforms the commonly used Wald-type method in terms of coverage, especially when the sample size is not large.

Keywords
confidence interval, Edgeworth expansion, second-order accuracy

1 INTRODUCTION

The bootstrap is widely used to evaluate the uncertainty of an estimator and to construct confidence intervals for the parameter of interest. As one of the most widely used bootstrap methods,
Efron’s bootstrap (Efron, 1979) resamples the realized sample with replacement, and applies the same estimation formula to each bootstrap replicate. The validity of Efron’s bootstrap has been justified for many estimators under simple random sampling. Estimators include sample means, empirical and quantile processes, von Mises functionals (Bickel & Freedman, 1981), $U$-statistics (Athreya et al., 1984) and $M$-estimators (Lahiri, 1992). Furthermore, with proper studentization, Efron’s bootstrap enjoys second-order accuracy under regularity conditions; see Hall (1992), Helmers (1991) and Lahiri (1994) for sample means, $U$-statistics, and $M$-estimators respectively. However, Efron’s bootstrap may lead to erroneous statistical inference under complex survey sampling because it does not take the sampling design into account (Mashreghi et al., 2016; Rao & Wu, 1988).

There are two main approaches to generating bootstrap samples in survey sampling (Mashreghi et al., 2016). In the first approach, the same sampling design is applied to generated bootstrap finite populations. Gross (1980) proposed a without-replacement bootstrap method for variance estimation for stratified random sampling, using finite populations generated through duplication of the realized sample. Bickel and Freedman (1984) generalized the bootstrap method of Gross (1980) and established central limit theorems for two bootstrap methods under stratified random sampling with or without replacement. Booth et al. (1994) also generalized the bootstrap method of Gross (1980) and established the second-order accuracy of their proposed method. Sverchkov and Pfeffermann (2004) used a multinomial distribution to generate a fixed bootstrap finite population. Chen et al. (2019) proposed two pseudo-population bootstrap methods for handling missing data problems in survey sampling, and argued that their methods are widely applicable.

In the second approach, bootstrap samples are generated directly from the realized sample, and rescaling, or other techniques, are used to adjust the bootstrap variance under specific sampling designs. Rao and Wu (1988) proposed a rescaling bootstrap method for variance estimation and for constructing confidence intervals under stratified random sampling. They demonstrated that the proposed bootstrap-$t$ intervals are second-order accurate under the assumption of known stratum variances. Also see Rao and Wu (1984) and Chipperfield and Preston (2007). Rao et al. (1992) proposed a rescaling bootstrap method to cover non-smooth statistics, but did not discuss second-order accuracy. Sitter (1992b) proposed a mirror-match bootstrap method for complex sample designs, including stratified random sampling and two-stage cluster sampling. The second-order accuracy is established for stratified random sampling based on the assumption of known stratum variances. Through empirical studies, Sitter (1992a) concluded that the without-replacement bootstrap and the mirror-match bootstrap outperform the rescaling bootstrap; see McCarthy and Snowden (1985) and Kovar et al. (1988) for more numerical study results. Shao and Sitter (1996) generalized the rescaling bootstrap method (Rao & Wu, 1988) to deal with missing data in survey sampling. Also see Davison and Sardy (2007) and Mashreghi et al. (2014). Antal and Tillé (2011) proposed one-one resampling methods to estimate the variance without any rescaling or weighting for certain complex sampling designs. Beaumont and Patak (2012) proposed a general bootstrap method by rescaling the bootstrap weights, but they did not investigate the second-order accuracy of their method. Saegusa (2014) proposed a non-parametric bootstrap method using two different bootstrap weights for two-phase stratified sampling without replacement, and established the conditional weak convergence of the empirical process associated with the proposed bootstrap method. However, Saegusa (2014) did not discuss the second-order property of the proposed bootstrap method. For hypothesis testing in survey sampling, Kim et al. (2021) proposed a unified bootstrap approach by scaling the survey weights, but only the conditional weak convergence of the proposed bootstrap method is established.
Although the second-order accuracy of rescaling bootstrap methods has been investigated under stratified random sampling (Booth et al., 1994; Rao & Wu, 1988; Sitter, 1992b), it is still an open question whether such a desirable property also holds for bootstrap methods under other complex sampling designs, especially under stratified multi-stage cluster sampling (Sitter, 1992b, section 2.1). In this paper, we propose a unified bootstrap method to achieve second-order accurate statistical inference based on the studentized estimator. The result holds for stratified multi-stage cluster sampling as well as for Poisson sampling, simple random sampling without replacement (SRS), and probability proportional to size (PPS) sampling with replacement. The main idea of the proposed bootstrap method is to mimic the sampling mechanism to generate bootstrap samples from bootstrap finite populations.

The proposed bootstrap method differs from the existing methods in several aspects. Although the proposed bootstrap method can be regarded as a 'pseudo-population bootstrap method' (Mashreghi et al., 2016), most existing procedures generate bootstrap finite populations by duplicating the realized sample a fixed number of times (Bickel & Freedman, 1984; Booth et al., 1994; Gross, 1980). However, the proposed bootstrap method generates bootstrap finite populations using a stochastic mechanism based on the multinomial distribution. Sverchkov and Pfeffermann (2004) also proposed using the multinomial distribution to generate a bootstrap finite population, but they focused on estimating the mean squared error under single-stage sampling. The proposed bootstrap method also differs from Sverchkov and Pfeffermann (2004) in that a studentized estimator is used to make statistical inference. Unlike Booth et al. (1994), the second-order accuracy of the proposed bootstrap method is investigated under stratified multi-stage cluster sampling as well as Poisson sampling and PPS sampling with replacement. To our best knowledge, we are the first to establish second-order accuracy for the bootstrap method for complex sampling designs. Different from Rao and Wu (1988) and Sitter (1992b), we establish the second-order accuracy of the proposed bootstrap method based on the studentized estimator. Specifically, for each sampling design, we establish that the estimation error for the distribution function of the studentized estimator is \( o_p(n^{-1/2}) \) (DiCiccio & Romano, 1995) under regularity conditions. Thus, the proposed bootstrap method is more widely applicable than existing methods.

The Wald-type method is widely used in survey sampling, but we show that it seldom achieves second-order accuracy. Moreover, the Wald-type method may lead to inflated type-I errors for hypothesis testing when the sample size is not large; see Section 4 for details. Compared to the Wald-type method, the proposed bootstrap method yields a superior test statistic.

The remaining part of the paper is organized as follows. In Section 2, we propose bootstrap methods for Poisson sampling, SRS and PPS sampling with replacement. The second-order accuracy of the proposed bootstrap method is established for each sampling design. The proposed bootstrap method is extended to stratified multi-stage cluster sampling and theoretically investigated in Section 3. Simulations are conducted to compare the proposed bootstrap method with the Wald-type method in Section 4. Some concluding remarks are made in Section 5.

2 | Bootstrap Method for Single-Stage Sampling

In this section, we propose bootstrap methods for Poisson sampling, SRS and PPS sampling with replacement. Let \( F_N = \{y_{N,1}, \ldots, y_{N,N}\} \subset \mathbb{R} \) be a finite population with a known population size \( N \). A sequence of finite populations is considered when we investigate the theoretical
properties of the proposed bootstrap method. In this paper, the finite population is assumed to be non-stochastic, and the parameter of interest is the finite population mean $\hat{Y}_N = N^{-1} \sum_{i=1}^{N} Y_{N,i}$.

2.1 Poisson sampling

Poisson sampling generates a sample by $N$ independent Bernoulli experiments, one for each element in the finite population. Specifically, for $i = 1, \ldots, N$, let $I_{N,i} \sim \text{Ber}(\pi_{N,i})$ with $\pi_{N,i} \in (0, 1)$, where $\text{Ber}(\pi)$ is a Bernoulli distribution with success probability $\pi$. The sample is $\{Y_{N,i} : I_{N,i} = 1, i = 1, \ldots, N\}$, where $I_{N,i} = 1$ if the $i$-th element is sampled and 0 otherwise. Let $\Pi_N = \{\pi_{N,i}, \ldots, \pi_{N,N}\}$ be the set of inclusion probabilities with $\pi_{N,i} = E(I_{N,i})$ for $i = 1, \ldots, N$. The sample size, $n = \sum_{i=1}^{N} I_{N,i}$, is random, and let $n_0 = E(n) = \sum_{i=1}^{N} \pi_{N,i}$ be the expected sample size. Without loss of generality, we assume that the first $n$ elements, $\{y_{N,1}, \ldots, y_{N,n}\}$, are sampled. We assume that the corresponding inclusion probabilities $\{\pi_{N,1}, \ldots, \pi_{N,n}\}$ are known.

Given a sample generated by Poisson sampling, let $\hat{Y}_{N,Poi} = N^{-1} \sum_{i=1}^{n} y_{N,i}\pi_{N,i}^{-1}$ be the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) of $Y_N$. A variance estimator of $\hat{Y}_{N,Poi}$ is $\hat{V}_{N,Poi} = N^{-2} \sum_{i=1}^{n} y_{N,i}^2 (1 - \pi_{N,i})\pi_{N,i}^{-2}$. It can be shown that both $\hat{Y}_{N,Poi}$ and $\hat{V}_{N,Poi}$ are design-unbiased, where, by ‘an estimator $\hat{\theta}$ is design-unbiased for a population parameter $\theta$', we refer to $E(\hat{\theta}) = \theta$ with the expectation taken with respect to Poisson sampling; see Fuller (2009) section 1.2 for details. That is, all probability statements are with respect to the distribution defined by Poisson sampling. Let $T_{N,Poi} = \hat{V}_{N,Poi}^{-1/2}(\hat{Y}_{N,Poi} - \hat{Y}_N)$ be a studentized estimator. The Edgeworth expansion of the distribution function of $T_{N,Poi}$ can be established based on the following conditions.

(A1) There exist constants $0 < C_1 \leq C_2$ with respect to $n_0$ and $N$, such that $C_1 \leq n_0^{-1} N \pi_{N,1} \leq C_2$ for $i = 1, \ldots, N$. Besides, the expected sample size satisfies $n_0 \to \infty$ as $N \to \infty$.

(A2) The sequence of finite populations and inclusion probabilities satisfies $\lim_{N \to \infty} (n_0 V_{N,Poi}) = \sigma_1^2$, where $V_{N,Poi} = N^{-2} \sum_{i=1}^{n} y_{N,i}^2 (1 - \pi_{N,i})\pi_{N,i}^{-1}$ is the variance of $\hat{Y}_{N,Poi}$, and $\sigma_1^2$ is a positive constant.

(A3) $\limsup_{N \to \infty} N^{-1} \sum_{i=1}^{N} \phi_{N,i} < \infty$.

(A4) For any $m \leq N$ with $m \to \infty$ as $N \to \infty$, there exists a subset $\{\ell_1, \ldots, \ell_m\} \subset \{1, \ldots, N\}$ satisfying the following two conditions:

(a) $\limsup_{m \to \infty} m^{-1} \sum_{i=1}^{m} \phi_{N,\ell_i} < \infty$.

(b) $G_{m,Poi}(x)$ converges to a non-lattice distribution $G_{Poi}(x)$ as $N \to \infty$, where $G_{m,Poi}(x)$ is the distribution function that assigns mass $1/m$ to $x_{N,\ell_i}$ for $i = 1, \ldots, m$, and $x_{N,i} = n_0 N^{-1} \pi_{N,i}^{-1} y_{N,i}$ for $i = 1, \ldots, N$.

Condition (A1) is commonly used in survey sampling; the first part of (A1) restricts the inclusion probabilities as used in Theorem 1.3.5 of Fuller (2009), and the second part is a mild condition on the expected sample size. Condition (A2) specifies the order of the variance associated with $\hat{Y}_{N,Poi}$ under Poisson sampling. Condition (A3) is a moment condition required for the Edgeworth expansion of the distribution function associated with $T_{N,Poi}$ and for the bootstrap counterpart, and it is common in the literature of the Edgeworth expansion. When $m = N$, (A4a) is essentially (A3). However, we isolate (A4a) from (A3) since (A4) is mainly about the existence of sub-sequences satisfying (a) and (b) simultaneously. Condition (A4b) is a non-lattice condition and is used to derive the Edgeworth expansion for the distribution functions.
associated with \( T_{N,Poi} \) and its bootstrap counterpart; see Feller (1971) section 16.6 and Bhattacharya and Rao (2010 section 19) for details. A distribution \( G(x) \) is non-latticed if \( \int \exp(\{tx\})dG(x) \neq 1 \) for all \( t \neq 0 \); see Babu and Singh (1984) for details. Specifically, if the finite population \( F_N \) is a random sample from a super-population model with sixth finite moment, then (A3) holds almost surely using the Etemadi’s strong law of large numbers (Athreya & Lahiri, 2006, theorem 8.2.7). If, in addition, \( \{n_0N^{-1/2}y_{N,i} : i = 1, \ldots, N \} \) is also a random sample with respect to a non-lattice super-population model, then (A4) also holds when \( \{\ell_1, \ldots, \ell_m \} \) is randomly selected from \( \{1, \ldots, N\} \) for any \( m \leq N \); see Theorem 1.3.1 of Fuller (2009) for details. For some specific surveys, however, the non-lattice condition (A4b) may not hold. For example, it is common to have binary responses \( y_{N,i} \in \{0, 1\} \) for \( i = 1, \ldots, N \). If we assume that \( \pi_{N,i} = n_0N^{-1} \), then \( G_{m,Poi}(x) \) cannot converge to a non-lattice distribution. It is beyond our scope to investigate the case where \( G_{Poi}(x) \) is a lattice distribution function.

**Remark 1** (Singh 1981) showed that, for the sample mean of independent and identically distributed random variables, Efron’s bootstrap is second-order accurate if and only if the underlying distribution is non-latticed. However, the non-lattice or Cramér-type condition may not be necessary for some other models. For example, Karabulut and Lahiri (1997) established the Edgeworth expansion for \( M \)-estimators of a linear regression parameter without Cramér-type (or non-lattice) conditions, and they showed that Efron’s bootstrap with certain modification is second-order accurate even when the error is generated from a lattice distribution. In this paper, the theoretical technique for Poisson sampling is adapted from the Edgeworth expansion for \( U \)-statistics (Helmers, 1991; Jing & Wang, 2003), where non-lattices conditions are commonly assumed. In addition, we do not assume randomness for the population elements.

**Theorem 1** Suppose that (A1)–(A4) hold. Then,

\[
F_{N,Poi}(z) = \Phi(z) + \left[ \frac{\mu^{(3)}_{N,Poi}}{6V_{3/2,N,Poi}}(1 - z^2) + \frac{\tau^{(3)}_{N,Poi}}{2V_{3/2,N,Poi}}z^2 \right] \phi(z) + o\left(n_0^{-1/2}\right) \tag{1}
\]

uniformly in \( z \in \mathbb{R} \), where \( F_{N,Poi}(z) \) is the distribution function of \( T_{N,Poi} \) under Poisson sampling, \( \Phi(z) \) is the distribution function of the standard normal distribution with the density function \( \phi(z) \), \( \mu^{(3)}_{N,Poi} = N^{-3}\sum_{i=1}^{N}y_{N,i}^3(1 - \pi_{N,i})\{(1 - \pi_{N,i})^2\pi_{N,i}^2 - 1\} \), and \( \tau^{(3)}_{N,Poi} = N^{-3}\sum_{i=1}^{N}y_{N,i}^3(1 - \pi_{N,i})^2\pi_{N,i}^{-2} \). In addition, \( \mu^{(3)}_{N,Poi}V_{3/2,N,Poi}^{-1} = O(n_0^{-1}) \) and \( \tau^{(3)}_{N,Poi}V_{3/2,N,Poi}^{-1} = O(n_0^{-1/2}) \).

Theorem 1 establishes the Edgeworth expansion of the distribution function associated with the studentized estimator \( T_{N,Poi} \). Different from the standard Edgeworth expansion for the standardized estimator (Hall, 1992), we have an additional term \( z^2\tau^{(3)}_{N,Poi}(2V_{3/2,N,Poi})^{-1}\phi(z) \) due to the variance estimator under Poisson sampling. Theorem 1 is proved by the technique of deriving the Edgeworth expansion for studentized \( U \)-statistics (Helmers, 1991; Jing & Wang, 2003); see Section S2 of the Supplementary Material for details. It is worth mentioning that condition (A4) ensures that there exists a subset \( \{\ell_1, \ldots, \ell_m\} \subset \{1, \ldots, N\} \) such that \( \{y_{N,\ell_1}, \ldots, y_{N,\ell_m}\} \) has a finite sixth moment and the characteristic function of \( \sum_{i=1}^{m}n_0^{-1/2}y_{N,\ell_i}(I_{N,i} - \pi_{N,i}) \) can be well controlled. This is essential in the proof of Theorem 1; see Lemma S5 in Section S12 of the Supplementary Material for details. By Theorem 1, if a standard normal distribution is used to approximate the distribution of \( T_{N,Poi} \) and \( \mu^{(3)}_{N,Poi}V_{3/2,N,Poi}^{-1}(1 - z^2)/6 + \tau^{(3)}_{N,Poi}V_{3/2,N,Poi}^{-1/2}z^2 \neq 0 \), then the corresponding
approximation error is of order $O(n_0^{-1/2})$. Therefore, the Wald-type method, which approximates $F_{N,Poi}(z)$ by a standard normal distribution, is generally not second-order accurate in terms of distributional approximation, and a similar conclusion applies to other sampling designs in this paper.

In order to approximate the distribution function of $T_{N,Poi}$, we propose a bootstrap method that mimics the sampling mechanism by two steps. First, a bootstrap finite population $F^*_N = \{y^*_{N,1}, \ldots, y^*_{N,N}\}$ and the corresponding bootstrap inclusion probabilities $\Pi^*_N = \{\pi^*_N, 1, \ldots, \pi^*_N\}$ are generated by a multinomial distribution based on the realized sample and the associated inclusion probabilities, and this step can be viewed as the reverse of the sampling procedure. Specifically, generate $(N^*_1, \ldots, N^*_n)$ by a multinomial distribution $MN(N; \rho_N)$ with $N$ trials and a probability vector $\rho_N = (\rho_{N,1}, \ldots, \rho_{N,n})$, where $\rho_{N,i} = \pi_N^{-1}(\sum_{j=1}^n \pi_N^{-1})^{-1}$ for $i = 1, \ldots, n$. Then, $F^*_N$ and $\Pi^*_N$ are constructed by repeating $y_{N,i}$ and $\pi_N$ $N^*_i$ times, respectively, for $i = 1, \ldots, n$. The bootstrap population mean is $\hat{Y}_N^* = N^{-1}\sum_{i=1}^N y_{N,i}^*$ and the second step is to generate a bootstrap sample $\{y_{N,i} : I^*_{N,i} = 1, i = 1, \ldots, N\}$ by Poisson sampling, where $I^*_{N,i} \sim \text{Ber}(\pi^*_N)$. Then, the bootstrap counterparts of $\hat{Y}_{N,Poi}$ and $\hat{V}_{N,Poi}$ are $\hat{Y}_{N,Poi}^* = N^{-1}\sum_{i=1}^N I^*_{N,Poi,i} y^*_{N,i}(\pi^*_N)^{-1}$ and $\hat{V}_{N,Poi}^* = N^{-2}\sum_{i=1}^N I^*_{N,Poi,i} (y^*_{N,i})^2(1 - \pi^*_N)(\pi^*_N)^{-2}$ respectively. Finally, a bootstrap studentized estimator is $T_{N,Poi}^* = (\hat{V}_{N,Poi}^*)^{-1/2}(\hat{Y}_{N,Poi}^* - \hat{Y}_N^*)$. The proposed bootstrap method is summarized in Algorithm 1 under Poisson sampling.

**Algorithm 1: Bootstrap method under Poisson sampling**

**Data:** realized sample $\{y_{N,1}, \ldots, y_{N,n}\}$, inclusion probabilities $\{\pi_{N,1}, \ldots, \pi_{N,n}\}$, population size $N$, bootstrap repetition number $M$.

**Result:** Bootstrap studentized estimators $\{T_{N,Poi}^*(m) : m = 1, \ldots, M\}$.

```
1 for i = 1, \ldots, n do
  2 \rho_{N,i} \leftarrow \pi N^{-1}(\sum_{j=1}^n \pi N^{-1})^{-1}
2 end
4 for m = 1, \ldots, M do
  5 Generate $(N^*_1, \ldots, N^*_n)$ by a multinomial distribution $MN(N; \rho_N)$
  6 Obtain $F^*_N$ and $\Pi^*_N$ by repeating $y_{N,i}$ and $\pi_{N,i} N^*_i$ times for $i = 1, \ldots, n$
  7 $\hat{Y}_N^* \leftarrow N^{-1}\sum_{i=1}^N y_{N,i}^*$
  8 for i = 1, \ldots, N do
    9 $I^*_{N,i} \sim \text{Ber}(\pi^*_N)$
10 end
11 $\hat{Y}_{N,Poi}^* \leftarrow N^{-1}\sum_{i=1}^N I^*_{N,Poi,i} y_{N,i}^*(\pi^*_N)^{-1}$
12 $\hat{V}_{N,Poi}^* \leftarrow N^{-2}\sum_{i=1}^N I^*_{N,Poi,i} (y_{N,i}^*)^2(1 - \pi^*_N)(\pi^*_N)^{-2}$
13 $T_{N,Poi}^*(m) \leftarrow (\hat{V}_{N,Poi}^*)^{-1/2}(\hat{Y}_{N,Poi}^* - \hat{Y}_N^*)$
14 end
```

In Algorithm 1, the bootstrap studentized estimator $T_{N,Poi}^*$ is centered by the bootstrap population mean $\hat{Y}_N^*$, not by $\hat{Y}_{N,Poi}$. Otherwise, if $\hat{Y}_{N,Poi}$ were used instead, additional variability would be induced due to the generation of bootstrap finite populations, so it would lead to erroneous statistical inference. The same argument applies to other sampling designs discussed in the paper. The computational complexity is $O(N)$ to generate bootstrap sampling indicators $\{I^*_{N,1}, \ldots, I^*_{N,N}\}$.
in Algorithm 1, and an equivalent but more computationally efficient algorithm is proposed in Section S1.1 of the Supplementary Material.

The next theorem establishes the Edgeworth expansion for the distribution function of $T_{N,\text{Poi}}^*$.

**Theorem 2** Suppose that (A1)–(A4) hold. Then,

$$
F_{N,\text{Poi}}^*(z) = \Phi(z) + \left\{ \frac{\hat{\gamma}_{N,\text{Poi}}^3}{6V_{N,\text{Poi}}^{3/2}} (1 - z^2) + \frac{\hat{\gamma}_{N,\text{Poi}}^3}{2V_{N,\text{Poi}}^{3/2}} z^2 \right\} \phi(z) + o_p\left(n_0^{-1/2}\right) \quad (2)
$$

uniformly in $z \in \mathbb{R}$, where $F_{N,\text{Poi}}^*(z)$ is the distribution function of $T_{N,\text{Poi}}^*$ conditional on the bootstrap finite population $F_{N,\text{Poi}}^*$ and the bootstrap inclusion probabilities $\Pi_{N,\text{Poi}}^*$, and the stochastic order in (2) is with respect to the proposed bootstrap method as well as Poisson sampling to generate the realized sample.

The proof of Theorem 2, which is presented in Section S3 of the Supplementary Material, utilizes a similar technique for proving Theorem 1. This is due to the fact that $T_{N,\text{Poi}}^*$ shares the same form as $T_{N,\text{Poi}}$. However, additional efforts have been allocated on checking the validity of conditions (A1)–(A3) and Lemma S5 under the bootstrap framework; see Lemmas S6–S7 in Section S12 of the Supplementary Material for details.

According to Theorems 1–2, we conclude that

$$
sup_{z \in \mathbb{R}} \left| F_{N,\text{Poi}}^*(z) - F_{N,\text{Poi}}(z) \right| = o_p\left(n_0^{-1/2}\right), \quad (3)
$$

and it indicates that the proposed bootstrap method is second-order accurate with estimation error of order $o_p(n_0^{-1/2})$ in terms of distributional approximation. If the second term on the right-hand side of (1) does not vanish, the proposed bootstrap method outperforms the Wald-type method. Generally, the analytical form of $F_{N,\text{Poi}}^*(z)$ is inaccessible, so the empirical distribution of $\{T_{N,\text{Poi}}^{(m)} : m = 1, \ldots, M\}$ is used instead.

Next, we consider interval estimators for the parameter of interest $\hat{Y}_N$. For the proposed bootstrap method, the one-sided lower, one-sided upper and two-sided confidence intervals of confidence level $(1 - \alpha) \times 100\%$ are

$$
\left( \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},1-\alpha} \hat{V}_{N,\text{Poi}}^{1/2}, \infty \right), \left(-\infty, \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},a} \hat{V}_{N,\text{Poi}}^{1/2} \right) \quad \text{and} \quad \left( \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},1-\alpha/2} \hat{V}_{N,\text{Poi}}^{1/2}, \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},a/2} \hat{V}_{N,\text{Poi}}^{1/2} \right),
$$

(4)

where $\hat{q}_{N,\text{Poi},\gamma}$ is the $\gamma$-th empirical quantile of $\{T_{N,\text{Poi}}^{(m)} : m = 1, \ldots, M\}$. For the Wald-type method, the corresponding confidence intervals are

$$
\left( \hat{Y}_{N,\text{Poi}} - q_{1-\alpha} \hat{V}_{N,\text{Poi}}^{1/2}, \infty \right), \left(-\infty, \hat{Y}_{N,\text{Poi}} - q_{a} \hat{V}_{N,\text{Poi}}^{1/2} \right) \quad \text{and} \quad \left( \hat{Y}_{N,\text{Poi}} - q_{1-\alpha/2} \hat{V}_{N,\text{Poi}}^{1/2}, \hat{Y}_{N,\text{Poi}} - q_{a/2} \hat{V}_{N,\text{Poi}}^{1/2} \right),
$$

(5)

where $q_\gamma$ is the $\gamma$-th quantile of the standard normal distribution. The difference between those confidence intervals in (4)–(5) is that empirical quantiles of the bootstrap studentized estimators are used for the proposed bootstrap method, while those of the standard normal distribution are used for the Wald-type method.
To compare the coverage probabilities of the confidence intervals in (4)–(5), we consider the following Cornish-Fisher expansions for \( q_{N,\text{Poi},\gamma} \) and \( \hat{q}_{N,\text{Poi},\gamma}^* \), where \( q_{N,\text{Poi},\gamma} \) is the \( \gamma \)-th quantile with respect to \( F_{N,\text{Poi}}(z) \).

**Corollary 1** Suppose that (A1)–(A4) hold. If \( M \sim n_0 \log n_0 \), then, for any fixed \( \gamma \in (0, 1) \), we have

\[
q_{N,\text{Poi},\gamma} = q_\gamma - \left\{ \frac{\mu_{N,\text{Poi}}^{(3)}}{6V_{N,\text{Poi}}^{3/2}} (1 - q_\gamma^2) + \frac{\tau_{N,\text{Poi}}^{(3)}}{2V_{N,\text{Poi}}^{3/2}} q_\gamma^2 \right\} + o \left( n_0^{-1/2} \right),
\]

(6)

\[
\hat{q}_{N,\text{Poi},\gamma}^* = q_\gamma - \left\{ \frac{\mu_{N,\text{Poi}}^{(3)}}{6V_{N,\text{Poi}}^{3/2}} (1 - q_\gamma^2) + \frac{\tau_{N,\text{Poi}}^{(3)}}{2V_{N,\text{Poi}}^{3/2}} q_\gamma^2 \right\} + o_p \left( n_0^{-1/2} \right),
\]

(7)

where \( a_n \sim b_n \) is equivalent to \( a_n = O(b_n) \) and \( b_n = O(a_n) \), and the stochastic order in (7) is with respect to the proposed bootstrap method as well as the generation of the realized sample \( \{y_{N,1}, \ldots, y_{N,n}\} \) under Poisson sampling.

Since the empirical quantile \( \hat{q}_{N,\text{Poi},\gamma}^* \) is used to approximate \( q_{N,\text{Poi},\gamma} \), an additional condition, \( M \sim n_0 \log n_0 \), is assumed to guarantee the stochastic order \( o_p(n_0^{-1/2}) \) in (7); see Babu and Singh (1983) for details. By Theorem 1, we can use Taylor expansion of \( \Phi(q_{N,\text{Poi},\gamma}) \) around \( q = q_\gamma \) to prove (6), and (7) can be established in a similar manner. Thus, the proof of Corollary 1 is omitted.

By Theorem 1 and Corollary 1, we have

\[
\left| q_\gamma - q_{N,\text{Poi},\gamma} \right| = O \left( n_0^{-1/2} \right) \quad \text{and} \quad \left| \hat{q}_{N,\text{Poi},\gamma}^* - q_{N,\text{Poi},\gamma} \right| = o_p \left( n_0^{-1/2} \right),
\]

(8)

which indicates that the empirical quantile obtained by the proposed bootstrap method is closer to the truth compared to the normal quantile. Thus, the proposed bootstrap algorithm has an advantage over the Wald-type method in approximating critical points for hypothesis testing; see also section 3.4 of Hall (1987). However, it is not the main focus of this paper to discuss hypothesis testing problems.

According to Theorem 1 and Corollary 1, the coverage probabilities of one-sided lower and upper confidence intervals by the proposed bootstrap method are

\[
P \left( \hat{Y}_N > \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},1-a}^* \hat{V}_{N,\text{Poi}}^{1/2} \right) = 1 - \alpha + o_p \left( n_0^{-1/2} \right),
P \left( \hat{Y}_N < \hat{Y}_{N,\text{Poi}} - \hat{q}_{N,\text{Poi},a}^* \hat{V}_{N,\text{Poi}}^{1/2} \right) = 1 - \alpha + o_p \left( n_0^{-1/2} \right),
\]

while those for the Wald-type method are

\[
P \left( \hat{Y}_N > \hat{Y}_{N,\text{Poi}} - q_{1-a} \hat{V}_{N,\text{Poi}}^{1/2} \right) = 1 - \alpha + \left\{ \frac{\mu_{N,\text{Poi}}^{(3)}}{6V_{N,\text{Poi}}^{3/2}} (1 - q_{1-a}^2) + \frac{\tau_{N,\text{Poi}}^{(3)}}{2V_{N,\text{Poi}}^{3/2}} q_{1-a}^2 \right\} \phi(q_{1-a}) + o \left( n_0^{-1/2} \right)
\]

\[
= 1 - \alpha + O \left( n_0^{-1/2} \right),
\]

(9)
Thus, the coverage errors of the one-sided lower and upper confidence intervals are of order $o_p(n_0^{-1/2})$ for the proposed bootstrap method, but those are of order $O(n_0^{-1/2})$ for the Wald-type method. This conclusion justifies the second-order accuracy of the proposed bootstrap method and represents an advantage over the Wald-type method in terms of converge probabilities of one-sided lower and upper confidence intervals.

By Corollary 1, the coverage error of the two-sided bootstrap confidence interval is of order $O_p(n_0^{-1/2})$ for the proposed bootstrap method. According to (9)–(10), we have

\[
\mathbb{P}\left( \hat{Y}_N - q_{1-a/2} \hat{V}_{N,Poi}^{1/2} < \hat{Y}_N < \hat{Y}_N - q_{a/2} \hat{V}_{N,Poi}^{1/2} \right) = 1 - \alpha + o\left(n_0^{-1/2}\right),
\]

since $\mu_{N,Poi}^{(3)} / (6V_{N,Poi}^{3/2})(1 - z^2) + \tau_{N,Poi}^{(3)} (2V_{N,Poi}^{3/2}) z^2$ is an even function of $z$. Thus, the coverage error of the two-sided confidence interval is of order $o(n_0^{-1/2})$ for the Wald-type method as well. Although no theoretical advantage has been shown for the bootstrap two-sided confidence interval in terms of the coverage probability, the bootstrap two-sided confidence interval in (4) is still favored since its upper and lower limits have error rates $o_p(n_0^{-1/2})$, which outperform the Wald-type two-sided confidence interval. In addition, as discussed in section 3.6 of Hall (1992), an asymmetric equal-tailed confidence interval may convey important information. Similar discussion about the confidence intervals applies to other sampling designs in this paper.

**Remark 2** The analysis of second-order accuracy of the confidence intervals by the proposed bootstrap method above relies on the Cornish-Fisher expansion in Corollary 1. However, it is not the only way to obtain the desired results. For example, we can obtain similar results from (3) directly as in the proof of Corollary 5.1 of Das et al. (2019).

Even though not revealed theoretically, simulation studies in Section 4 indicate that the two-sided confidence interval by the proposed bootstrap method outperforms the Wald-type one in terms of the coverage. A high-order Edgeworth expansion is needed to further investigate the advantage of the bootstrap two-sided confidence interval over the Wald-type one (Hall, 1992). However, it is beyond the scope of this paper.

### 2.2 Simple random sampling without replacement

Under SRS, a sample of size $n$ is selected with equal probability. One commonly used sampling procedure is to randomly select one element from the remaining population with equal probability after removing the selected ones, and the procedure is continued until we get a sample of size $n$. Different from the Poisson sampling, the sample size $n$ is non-stochastic under SRS, and the inclusion probability is $\pi_{N,i} = nN^{-1}$ for $i = 1, \ldots, N$. Without loss of generality, we assume that the first $n$ elements are sampled under SRS. The corresponding Horvitz-Thompson estimator
of \( \hat{Y}_N \) is \( \hat{Y}_{N,SRS} = n^{-1}\sum_{i=1}^n y_{N,i} \), and its variance estimator is \( \hat{V}_{N,SRS} = n^{-1}(1 - nN^{-1})s_{N,SRS}^2 \) with \( s_{N,SRS}^2 = n^{-1}\sum_{i=1}^n (y_{N,i} - \hat{Y}_{N,SRS})^2 \).

To investigate the Edgeworth expansion of the distribution function with respect to the studentized estimator \( T_{N,SRS} = \frac{\hat{Y}_{N,SRS}}{\hat{V}_{N,SRS}^{1/2}}(\hat{Y}_{N,SRS} - \hat{Y}_N) \), the following conditions are required.

1. \( \exists \kappa \in (0, 1) \) such that \( nN^{-1} \leq 1 - \kappa \) as \( N \to \infty \). Besides, \( n \to \infty \) as \( N \to \infty \).
2. The sequence of finite populations satisfies \( \lim_{N \to \infty} \sigma_{N,SRS}^2 = \sigma_2^2 \), where \( \sigma_{N,SRS}^2 = N^{-1}\sum_{i=1}^N (y_{N,i} - \hat{Y}_N)^2 \) and \( \sigma_2^2 \) is a positive constant.
3. There exists a positive constant \( \delta < 1 \) such that \( \lim_{N \to \infty} N^{-1}\sum_{i=1}^N y_{N,i} \delta + \delta < \infty \).
4. The distribution function \( G_{N,SRS}(x) \) converges to a non-lattice distribution \( G_{SRS}(x) \), where \( G_{N,SRS}(x) \) assigns probability \( 1/N \) to \( y_{N,1}, \ldots, y_{N,N} \).

Condition (B1) is a counterpart of (A1) and is used to rule out the trivial case when the sample size equals to the population size. Condition (B2) is similar to (A2), and it regulates the variance of \( F_N \) with respect to the distribution \( G_{N,SRS}(x) \). Different from (A2), \( \sigma_{N,SRS}^2 \) is not scaled by the sample size \( n \), since it is not the variance of \( \hat{Y}_{N,SRS} \), and it measures the dispersion of the finite population. The moment condition in (B3) is similar to (A3), and it guarantees the convergence of the variance estimator and the Edgeworth expansion for the distribution function of the studentized estimator. The non-lattice condition in (B4) is a counterpart of (A4b). By (B1)–(B4), Babu and Singh (1985 Theorem 2) established

\[
F_{N,SRS}(z) = \Phi(z) + \frac{(1 - n/N)^{1/2}\mu_{N,SRS}^{(3)}}{6n^{1/2}\sigma_{N,SRS}^3} \left\{ 3z^2 - \frac{1 - 2n/N}{1 - n/N}(z^2 - 1) \right\} \phi(z) + o\left( n^{-1/2} \right)
\]

uniformly in \( z \in \mathbb{R} \), where \( F_{N,SRS}(z) \) is the distribution function of \( T_{N,SRS} \), and \( \mu_{N,SRS}^{(3)} = -N^{-1}\sum_{i=1}^N (y_{N,i} - \hat{Y}_N)^3 \).

We propose a similar bootstrap method as for Poisson sampling to approximate the distribution of \( T_{N,SRS} \). Since the inclusion probabilities are the same for elements in the finite population \( P_x \) under SRS, we first generate \( (N_1^*, \ldots, N_n^*) \) from \( M(N; \rho_N) \) with \( \rho_N = (n^{-1}, \ldots, n^{-1}) \), and the corresponding bootstrap finite population \( P_{N,SRS}^* \) is obtained by repeating \( y_{N,i} \) \( N_i^* \) times for \( i = 1, \ldots, n \). Let \( \bar{Y}_N = N^{-1}\sum_{i=1}^n y_{N,i} = N^{-1}\sum_{i=1}^n N_i^* y_{N,i} \) be the bootstrap population mean. A bootstrap sample \( \{y_{N,i}^* : i \in S^*\} \) of size \( n \) is generated from \( P_{N,SRS}^* \) using SRS, where \( S^* \) is the index set of selected elements in the bootstrap sample. Let the bootstrap counterparts of \( \hat{Y}_{N,SRS} \) and \( \hat{V}_{N,SRS} \) be \( \hat{Y}_{N,SRS}^* = n^{-1}\sum_{i \in S^*} y_{N,i}^* \) and \( \hat{V}_{SRS}^* = n^{-1}(1 - nN^{-1})s_{N,SRS}^2 \), respectively, where \( s_{N,SRS}^2 = n^{-1}\sum_{i \in S^*} (y_{N,i}^* - \hat{Y}_{N,SRS}^*)^2 \). Then, the bootstrap studentized estimator is \( T_{N,SRS}^* = (\hat{V}_{SRS}^*)^{-1/2}(\hat{Y}_{N,SRS}^* - \bar{Y}_N) \). The detailed algorithm of the proposed bootstrap method is summarized by Algorithm 3 in the Appendix. To obtain a bootstrap sample by Algorithm 3, the computational complexity of selecting each element is \( O(N) \) if \( n = o(N) \). In Section S1.2 of the Supplementary Material, we propose an equivalent but more computationally efficient algorithm.

The next theorem establishes the Edgeworth expansion for the distribution function of \( T_{N,SRS}^* \).

**Theorem 3** Suppose that (B1)–(B4) hold. Then,

\[
F_{N,SRS}^*(z) = \Phi(z) + \frac{(1 - n/N)^{1/2}\mu_{N,SRS}^{(3)}}{6n^{1/2}\sigma_{N,SRS}^3} \left\{ 3z^2 - \frac{1 - 2n/N}{1 - n/N}(z^2 - 1) \right\} \phi(z) + o_p\left( n^{-1/2} \right)
\]
uniformly in \( z \in \mathbb{R} \), where \( F_{N,SRS}^*(z) \) is the distribution function of \( T_{N,SRS}^* \) conditional on the bootstrap finite population \( F_N^* \), and the stochastic order in (12) is with respect to the proposed bootstrap method as well as the generation of the realized sample \{y_{N,1}, \ldots, y_{N,N}\} under SRS.

The proof of Theorem 3 is in Section S4 of the Supplementary Material, and it is a combination of checking the validity of the bootstrap counterparts of conditions (B1)–(B4) and then applying estimator (Hansen & Hurwitz, 1943), \(\hat{\mu}_{N,SRS} \neq \Phi(z) + \frac{\mu_{N,PSS}^{(3)}}{6\sqrt{n}\sigma_{N,PSS}^3}(2z^2 + 1)\phi(z) + o(n^{-1/2})\) (13)

2.3 Probability proportional to size sampling with replacement

Different from Poisson sampling and SRS, PPS sampling with replacement generates a sample of size \( n \) by independently selecting \( n \) elements from the finite population \( F_N \) with selection probabilities \( P_N = \{p_{N,i} : i = 1, \ldots, N\} \), where \( p_{N,i} \in (0, 1) \) is the selection probability of \( y_{N,i} \) and \( \sum_{i=1}^N p_{N,i} = 1 \). In practice, the selection probability \( p_{N,i} \) is often proportional to a specific size measure. Since replication is possible under PPS sampling with replacement, instead of the Horvitz-Thompson estimator, the parameter of interest \( \hat{Y}_N \) is estimated by the Hansen–Hurwitz estimator (Hansen & Hurwitz, 1943), \(\hat{Y}_{N,PSS} = N^{-1}n^{-1}\sum_{i=1}^n Z_{N,i}, \text{ where } Z_{N,i} = p_{N,a(i)}y_{N,a(i)}, p_{N,a(i)} = p_{N,k} \text{ and } y_{N,a(i)} = y_{N,k} \text{ if } a(i) = k, \text{ and } a(i) \text{ is the index of the selected element for the } i\text{-th draw. A variance estimator of } \hat{Y}_{N,PSS} \text{ is } \hat{V}_{N,PSS} = \{n(n - 1)\}^{-1}\sum_{i=1}^n (N^{-1}Z_{N,i} - \hat{Y}_{N,PSS})^2.

Under PPS sampling with replacement, we need the following conditions to establish the Edgeworth expansion of the distribution function with respect to the studentized estimator \( T_{N,PSS} = \hat{V}_{N,PSS}^{1/2}(\hat{Y}_{N,PSS} - \hat{Y}_N) \).

(C1) There exist constants \( 0 < C_4 \leq C_5 \) with respect to \( N \), such that \( C_4 \leq Np_{N,i} \leq C_5 \) for \( i = 1, \ldots, N \). Besides, \( n \to \infty \) as \( N \to \infty \).

(C2) The sequence of finite populations and selection probabilities satisfies \( \lim_{N \to \infty} \sigma_{N,PSS}^2 = \sigma_3^2 \), where \( \sigma_{N,PSS}^2 = \sum_{i=1}^N p_{N,i}(N^{-1}p_{N,i}^2 - 1)\hat{Y}_N - \hat{Y}_N^2 \), and \( \sigma_3^2 \) is a positive constant.

(C3) The distribution \( G_{N,PSS}(x) \) is non-latticed, where \( G_{N,PSS}(x) \) assigns probability \( p_{N,i} \) to \( N^{-1}p_{N,i}y_{N,i} \) for \( i = 1, \ldots, N \).

Condition (C1) restricts the selection probabilities and the sample size, and (C2) rules out the degenerate case with \( \sigma_3 = 0 \) under PPS sampling with replacement. As (B2), \( \sigma_{N,PSS}^2 \) measures the dispersion of the finite population under PPS sampling with replacement, so it is not scaled by the sample size. Unlike Poisson sampling or SRS, we do not require \( G_{N,PSS}(x) \) to converge to a non-lattice distribution function in (C3). Instead, we assume \( G_{N,PSS}(x) \) is non-latticed, and it is used to establish the Edgeworth expansion under PPS sampling with replacement.

The next theorem establishes the Edgeworth expansion for the distribution function of \( T_{N,PSS} \).

**Theorem 4** Suppose that (A3) and (C1)–(C3) hold. Then,

\[
F_{N,PSS}(z) = \Phi(z) + \frac{\mu_{N,PSS}^{(3)}}{6\sqrt{n}\sigma_{N,PSS}^3}(2z^2 + 1)\phi(z) + o(n^{-1/2})
\]
uniformly in $z \in \mathbb{R}$, where $F_{N,\text{pps}}$ is the distribution function of $T_{N,\text{pps}}$ under PPS sampling with replacement, and $\mu_{N,\text{pps}}^{(3)} = \sum_{i=1}^{N} p_{N,i}(N^{-1} p_{N,i}^{-1} y_{N,i} - \bar{Y})^3 = O(1)$.

Since $T_{N,\text{pps}}$ can be rewritten as a studentized sample mean of a sequence of i.i.d. random variables, Theorem 4 follows from the classical result on Edgeworth expansion of studentized $t$-statistics (Hall, 1987); see Section S6 of the Supplementary Material for details.

We propose a bootstrap method to mimic the sampling mechanism of PPS sampling with replacement. Similar as Poisson sampling, let $\rho_{N,i} = p_{N,i}^{-1} (\sum_{i=1}^{N} p_{N,i}^{-1})^{-1}$ for $i = 1, \ldots, n$. Then, a bootstrap finite population $F^\ast_{N}$ and $\{p^*_1, \ldots, p^*_N\}$ are generated by drawing a random vector ($N^*_1, \ldots, N^*_n$) from a multinomial distribution $\text{MN}(N; \rho_N)$ with $\rho_N = (\rho_{N,1}, \ldots, \rho_{N,n})$, where $N^*_i$ is the number of replications of $y_{N,i}$ and $p_{N,i}$ respectively. Since $\sum_{i=1}^{N} p_{N,i}^* \neq 1$ generally, the bootstrap selection probabilities are $P^* = \{(C^*_N)^{-1} p^*_1, \ldots, (C^*_N)^{-1} p^*_N\}$, where $C_N = \sum_{i=1}^{N} p_{N,i}^*$. Denote $\bar{Y}^*_N = N^{-1} \sum_{i=1}^{N} y^*_N, y^*_N = N^{-1} \sum_{i=1}^{N} N^*_i y_{N,i}$. Let $\{y^*_{N,b}(i) : i = 1, \ldots, n\}$ be a bootstrap sample generated by PPS sampling with replacement based on $F^*_N$ and $P^*_N$, where $b(i)$ is the index selected at the $i$th draw, and $y^*_{N,b}(i) = y^*_{N,k}$ and $p^*_{N,b}(i) = p^*_k$ if $b(i) = k$ for $i = 1, \ldots, n$. Let $\hat{Y}^*_{N,\text{pps}} = n^{-1} \sum_{i=1}^{n} N^{-1} (C^*_N)^{-1} y^*_N - \bar{Y}^*_N$, $\hat{V}^*_{N,\text{pps}} = n^{-2} \sum_{i=1}^{n} (N^{-1} C^*_N)^{-1} y^*_N - \bar{Y}^*_N$, and $T^*_{N,\text{pps}} = (\hat{V}^*_{N,\text{pps}})^{-1/2}(\hat{Y}^*_{N,\text{pps}} - \bar{Y}^*_N)$. The detailed algorithm of the proposed bootstrap method is summarized by Algorithm 4 in the Appendix. The computational complexity of Algorithm 4 is $O(N)$ to generate an element for the bootstrap sample, and an equivalent but more computationally efficient algorithm is proposed in Section S1.3 of the Supplementary Material.

**Theorem 5** Suppose that (A3) and (C1)–(C3) hold. Then,

$$F^*_{N,\text{pps}}(z) = \Phi(z) \frac{\mu_{N,\text{pps}}^{(3)}}{6 \sqrt{n} \sigma_{N,\text{pps}}^3}(2z^2 + 1) \phi(z) + o_p \left(n^{-1/2}\right) \quad (14)$$

uniformly in $z \in \mathbb{R}$, where $F^*_{N,\text{pps}}(z)$ is the distribution function of $T^*_{N,\text{pps}}$ conditional on $F^*_{N}$ and $P^*_N$, and the stochastic order in (14) is with respect to the proposed bootstrap method as well as the generation of the realized sample under PPS sampling with replacement.

Theorem 5 establishes the Edgeworth expansion for the distribution function of $T^*_{N,\text{pps}}$ conditional on the bootstrap finite population $F^*_N$ and selection probabilities $P^*_N$. The proof of Theorem 3 is in Section S7 of the Supplementary Material. Same as $T_{N,\text{pps}}$, $T^*_{N,\text{pps}}$ is in the form of a studentized $t$-statistic, so the proof of Theorem 5 follows by checking the bootstrap counterparts of the conditions in Theorem 4. By comparing (13) with (14), we conclude that the proposed bootstrap method is second-order accurate under PPS sampling with replacement.

By Theorems 4–5, we can obtain the Cornish-Fisher expansions for PPS sampling with replacement in Section S8 of the Supplementary Material and similar conclusions for the three types of confidence intervals in Section 2.1.

### 3 STRATIFIED MULTI-STAGE CLUSTER SAMPLING

Stratified multi-stage cluster sampling is widely used in socio-economic surveys, such as the face-to-face sampling design for the American National Election Studies (DeBell et al., 2018).
Under stratified multi-stage cluster sampling, the finite population is partitioned into $H$ mutually exclusive strata, and there are $N_h$ clusters (or primary sampling units, PSU’s) in the $h$-th stratum for $h = 1, \ldots, H$. Let $Y = \sum_{h=1}^{H} Y_h$ be the population total and $N = \sum_{h=1}^{H} M_h$ be the population size, where $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ is the total of the $h$-th stratum, and $Y_{hi}$ is the total of the ($hi$)-th PSU, $M_h = \sum_{i=1}^{N_h} M_{hi}$ is the size of the $h$-th stratum, and $M_{hi}$ is the size of the ($hi$)-th PSU. For the ease of notation, we implicitly assume that the population elements are indexed by ‘$\cdot$’. Assume that the stratum sizes $\{N_1, \ldots, N_H\}$ as well as the cluster sizes $\{M_{hi} : h = 1, \ldots, H; i = 1, \ldots, N_h\}$ are known. The parameter of interest is the population mean $\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$, where $W_h = M_h N^{-1}$ and $\bar{Y}_h = M_{h}^{-1} Y_h$.

For $h = 1, \ldots, H$, suppose that $n_h \geq 2$ PSUs are selected using PPS sampling with replacement based on the selection probabilities $\{p_{hi} : i = 1, \ldots, N_h\}$ satisfying $\sum_{i=1}^{N_h} p_{hi} = 1$, and subsamples are taken independently from the selected PSU’s; see Krewski and Rao (1981) for details. Denote $\hat{Y}_{a(h)}$ to be a design-unbiased estimator of $\hat{Y}_{a(h)}$ with respect to the sampling on the second and subsequent stages, where $a(h)$ is the index of the $i$-th selected PSU in the $h$-th stratum. Then, a design-unbiased estimator of $\bar{Y}$ is $\hat{Y} = \sum_{h=1}^{H} W_h \hat{Y}_h$, where $\hat{Y}_h = n_h^{-1} \sum_{i=1}^{N_h} \hat{Y}_{a(h)}$, and $\hat{Y}_{h} = M_{h}^{-1} \hat{Y}_{h} M_{hi} p_{hi}^{-1}$. It can be shown that a design-unbiased variance estimator for $\hat{Y}$ is $\hat{\bar{V}} = \sum_{h=1}^{H} W_h \hat{V}_h$, where $\hat{V}_h = \{n_h(n_h - 1)\}^{-1} \sum_{i=1}^{N_h} (\hat{Y}_{a(h)} - \hat{Y}_h)^2$.

In this section, we mainly adopt the framework of Krewski and Rao (1981), and the following conditions are assumed to establish the Edgeworth expansion for the distribution function with respect to the studentized estimator $T_{MS} = \hat{\bar{V}}^{-1/2} (\bar{Y} - \bar{Y})$.

(D1) For $h = 1, \ldots, H$, $n_h \to \infty$. In addition, there exist constants $0 < d_1 \leq d_2$, such that $d_1 \leq \inf_{1 \leq h_1, h_2 \leq H} (n_{h_1}/n_{h_2}) \leq \sup_{1 \leq h_1, h_2 \leq H} (n_{h_1}/n_{h_2}) \leq d_2$.

(D2) There exists a positive constant $d_3$ such that $\max_{1 \leq h \leq H} (HW_h) \leq d_3$.

(D3) As $H \to \infty$, there exists a positive constant $V_0$ such that $\lim_{n \to \infty} n \sum_{h=1}^{H} W_h^2 n_h^{-1} V_h = V_0$, where $n = \sum_{h=1}^{H} n_h$, and $V_h = \var{\hat{Y}_{a(h)}}$ for $h = 1, \ldots, H$.

(D4) There exist constants $0 < d_4 \leq d_5$, such that $d_4 \leq \min_{1 \leq h \leq H} (M_h M_{hi}^{-1} p_{hi}) \leq \max_{1 \leq i \leq N_h, 1 \leq h \leq H} (M_h M_{hi}^{-1} p_{hi}) \leq d_5$.

(D5) $\max_{1 \leq i \leq N_h, 1 \leq h \leq H} |\hat{Y}_{hi}| = O(1)$.

(D6) There exists a positive constant $\delta < 1$, such that $\max_{1 \leq i \leq N_h, 1 \leq h \leq H} E_{II} |\hat{Y}_{hi}|^{1+\delta} \leq \epsilon$, where $E_{II}$ is the expectation with respect to the second and subsequent sampling stages, and $\hat{Y}_{hi}$ is a design-unbiased estimator of the cluster mean $\bar{Y}_{hi}$ for $h = 1, \ldots, H$ and $i = 1, \ldots, N_h$.

(D7) There exists a positive constant $d_0 < 1$ such that $\max_{1 \leq i \leq N_h, 1 \leq h \leq H} \sup_{|t| \geq \epsilon} \left| E_{II} \{ \exp(i t \hat{Y}_{hi}) \} \right| \leq d_0$ for all $\epsilon > 0$.

By (D1), the number of selected PSU’s are of the same order for different strata. Different from Krewski and Rao (1981), $n_h$ diverges to guarantee the validity of the proposed bootstrap method. Condition (D2) rules out the existence of strata with extremely large stratum sizes. Condition (D3) shows the asymptotic property of the variance of $\hat{Y}$, which is similar to condition C4 of Krewski and Rao (1981). Condition (D4) is similar to (C1), and it is used to restrict the selection probabilities within each stratum. Condition (D5) assumes that the cluster means are uniformly bounded. Condition (D6) is similar to condition C1 of Krewski and Rao (1981), and it restricts the second and subsequent sampling stages. Condition (D7) is a non-lattice assumption, which is used to derive the Edgeworth expansion of the distribution functions associated with the studentized estimator and its bootstrap counterpart.
Let \( V = \text{var}(\hat{Y} \mid F_{MS}) \), \( X_{hi} = V^{-1/2}N^{-1}n_h^{-1}M_{a(h)}p_{a(h)}^{-1}\hat{Y}_{a(h)} \), \( \mu_h = E(X_{hi} \mid F_{MS}) = V^{-1/2}N^{-1}n_h^{-1}\sum_{j=1}^{N_{h}}M_{hj}Y_{hj} \) and \( \mu_h^{(3)} = E \{ (X_{hi} - \mu_h)^3 \mid F_{MS} \} \) for \( h = 1, \ldots, H \) and \( i = 1, \ldots, n_h \), where \( F_{MS} \) is the finite population under stratified multi-stage cluster sampling.

**Theorem 6** Suppose that (D1)–(D7) hold. If \( H = o(n(\log n)^{-2}) \), then

\[
F_{MS}(z) = \Phi(z) + 6^{-1}\sum_{h=1}^{H}n_h\mu_h^{(3)}(2z^2 + 1)\phi(z) + o\left(n^{-1/2}\right)
\]

uniformly in \( z \in \mathbb{R} \), where \( F_{MS}(z) \) is the distribution function of \( T_{MS} \) conditional on \( F_{MS} \). In addition, \( \mu_h^{(3)} = O(n^{-3/2}) \) for \( h = 1, \ldots, H \).

Theorem 6 establishes the Edgeworth expansion of the distribution function associated with the studentized estimator \( T_{MS} \) under stratified multi-stage cluster sampling. The condition that \( H = o(n(\log n)^{-2}) \) is satisfied when \( n_h/(\log H) \rightarrow \infty \). The proof of Theorem 6, which utilizes the technique of deriving the Edgeworth expansion of multivariate sample means (Babu & Singh, 1983; Liu, 1988), is in Section S9 of the Supplementary Material.

Under stratified multi-stage cluster sampling, since sampling is conducted independently for each stratum and we only consider the variability due to the first-stage cluster sampling as in Krewski and Rao (1981), the bootstrap method is proposed to generate \( F^*_h = \{ \hat{Y}_{h1}, \ldots, \hat{Y}_{hN_h} \} \) by \( \{ \hat{Y}_{a(hi)} : i = 1, \ldots, n_h \} \) for \( h = 1, \ldots, H \), and the sampling designs on the second and subsequent stages are ignored, where \( \hat{Y}_h \) is the cluster mean of the bootstrapped (hi)-th cluster associated with size \( M_{hi}^* \). The generation of \( F^*_h \) is the same as the first step of the bootstrap method for PPS sampling with replacement. Specifically, for \( h = 1, \ldots, H, F^*_h, \{ p_{h1}, \ldots, p_{hN_h} \} \) and the bootstrap cluster sizes \( \{ M_{h1}^*, \ldots, M_{hN_h}^* \} \) contain \( N_{hi}^* \) replicates of \( \hat{Y}_{a(hi)}, p_{a(hi)} \) and \( M_{a(hi)} \), respectively, where \( (N_{h1}^*, \ldots, N_{hN_h}^*) \) is generated by a multinomial distribution \( MN(N_h; \rho_h) \) with \( N_h \) trials and a probability vector \( \rho_h = (\rho_{h1}, \ldots, \rho_{hN_h}) \), \( \rho_{hi} = p_{a(hi)}^{-1}\sum_{j=1}^{N_{h}}p_{a(hj)}^{-1} \) for \( i = 1, \ldots, n_h \). To guarantee that the summation of bootstrap selection probabilities equals to 1 within each stratum, the bootstrap selection probabilities are \( P^*_h = \{ (C_h^*)^{-1}p_{h1}^*, \ldots, (C_h^*)^{-1}p_{hN_h}^* \} \) for the \( h \)th stratum, where

\[
C_h^* = \sum_{i=1}^{N_h}p_{hi}^*.
\]

That is, only the bootstrap cluster means are generated by the proposed bootstrap method, and the bootstrap stratum size \( M_{h}^* = \sum_{i=1}^{N_h}M_{hi}^* \) and the bootstrap population size \( N^* = \sum_{h=1}^{H}M_{h}^* \) are generally different from original stratum size \( M_h \) and population size \( N \). Based on the generated bootstrap finite population and selection probabilities, PPS sampling with replacement is conducted independently within each stratum to obtain the index set of the bootstrap sample \( b(hi) : h = 1, \ldots, H; i = 1, \ldots, n_h \), where \( b(hi) \) is the index selected at the \( i \)th draw from \( F^*_h \) with selection probabilities \( P^*_h \) for \( h = 1, \ldots, H \). Then, the bootstrap studentized estimator is \( T_{MS}^* = (\hat{V}^{*})^{-1/2}(\hat{Y}^{*} - \hat{\mu}^{*}) \), where \( \hat{V}_{hi} = (N_{hi}^*)^{-1}M_{hi}^* \) for \( h = 1, \ldots, H, \hat{V}_{b(hi)} = (M_{hi}^*)^{-1}(\hat{Y}_{b(hi)} - \hat{\mu}^{*})^{-1}(p_{b(hi)}^{*})^{-1}C_{hi}^* \), \( \hat{Y}_{hi} = n_h^{-1}\sum_{i=1}^{N_{hi}} \hat{Y}_{b(hi)} \) is the bootstrap estimator of the stratum mean, \( \hat{V} = \sum_{h=1}^{H}W_{h}^*\hat{V}_{hi} \) is the bootstrap estimator of the population mean \( \hat{V} = \sum_{h=1}^{H}W_{hi}\hat{Y}_{hi} = (M_{hi}^*)^{-1}(\sum_{i=1}^{N_{hi}}N_{a(hi)}M_{a(hi)}\hat{Y}_{a(hi)}) \) is the bootstrap stratum mean, \( \hat{V}_{hi} = (N_h(n_h - 1))^{-1}(\sum_{i=1}^{N_{hi}}(\hat{Y}_{b(hi)} - \hat{Y}_{hi})^2) \) is the bootstrap variance estimator of \( \hat{Y}_{hi} \), and \( \hat{V} = \sum_{h=1}^{H}(W_{hi})^2\hat{V}_{hi} \) is the bootstrap variance estimator of \( \hat{Y} \). The detailed algorithm of the proposed bootstrap method is summarized in Algorithm 2. Since PPS sampling with replacement is conducted to select PSU’s, we can improve the computational efficiency of Algorithm 2 by a similar algorithm in Section S1.3 of the Supplementary Material.
Suppose that (D1)–(D7) hold. Moreover, if $H = o(n (\log n)^{-6})$, then

$$F_{*MS}^*(z) = \Phi(z) + 6^{-1} \sum_{h=1}^{H} n_h \mu_{h}^{(3)} (2z^2 + 1) \phi(z) + o_p\left(n^{-1/2}\right)$$  \hspace{1cm} (15)$$

uniformly in $z \in \mathbb{R}$, where $F_{*MS}^*(z)$ is the distribution function of $T_{MS}^*$ conditional on $F_{*MS}^*$, and $T_{MS}^* = \bigcup_{h=1}^{H} T_{h}^*.$
4 | SIMULATION STUDY

4.1 | Single-stage sampling

In this section, we test the performance of the proposed bootstrap method under single-stage sampling designs. A finite population $F_N = \{y_{N,1}, \ldots, y_{N,N}\}$ is generated by $y_{N,i} \sim \text{Gamma}(1/30, 30)$ for $i = 1, \ldots, N$, where $\text{Gamma}(\alpha, \beta)$ is a gamma distribution with shape parameter $\alpha$ and scale parameter $\beta$. We are interested in constructing $(1-\alpha) \times 100\%$ confidence intervals for the finite population mean $\bar{Y}_N$ under the following sampling designs:

(a) Poisson sampling. The inclusion probability is $\pi_{N,i} = n_0 z_{N,i} \left(\sum_{j=1}^{N} z_{N,j}\right)^{-1}$ for $i = 1, \ldots, N$, where $n_0$ is the expected sample size, $z_{N,i} = \log(s_{N,i} + 3)$, and $s_{N,i} | y_{N,i} \sim \text{Gamma}(1/30, y_{N,i})$.

(b) SRS with sample size $n_0$.

(c) PPS sampling with replacement. The selection probability is $p_{N,i} = n_0^{-1} \pi_{N,i}$ for $i = 1, \ldots, N$, and the sample size is $n_0$, where $\pi_{N,i}$ is defined in (a).

In this simulation study, we consider $N \in \{10,000, 60,000\}$, $n_0 \in \{150,300\}$, and $\alpha \in \{0.1, 0.05, 0.01\}$.

For a specific sampling design, denote $\hat{Y}_N$ and $\hat{\nu}_N$ to be the point and variance estimators obtained from the realized sample, $F_N(z)$ to be the distribution function of $\hat{\nu}_N^{1/2} (\hat{Y}_N - \bar{Y}_N)$, and $\hat{Y}_N^{(m)}$ and $\hat{\nu}_N^{(m)}$ to be the corresponding point and variance estimators obtained from the $m$-th bootstrap sample; see Section 2 for details. For the ease of notation, we have omitted the index with respect to each Monte Carlo simulation. We compare the proposed bootstrap method with $M = 5,000$ bootstrap repetitions and the Wald-type method based on the following $(1-\alpha) \times 100\%$ confidence intervals for $\bar{Y}_N$.

(a) One-sided lower confidence interval. For the proposed bootstrap method, we use $(\hat{Y}_N - q_{1-\alpha}^{*} \hat{\nu}_N^{1/2}, \infty)$, where $q_{\gamma}^{*}$ is the $\gamma$-th empirical quantile of $\{\tau^{(m)} : m = 1, \ldots, M\}$, $\hat{\tau}^{(m)} = (\hat{\nu}_N^{(m)})^{-1/2} (\hat{Y}_N^{(m)} - \bar{Y}_N^{(m)})$, and $\bar{Y}_N^{(m)}$ is the bootstrap population mean with respect to the $m$-th bootstrap sample. We use $(\hat{Y}_N - q_{1-\alpha} \hat{\nu}_N^{1/2}, \infty)$ for the Wald-type method.

(b) One-sided upper confidence interval. We consider $(-\infty, \hat{Y}_N - q_{\alpha}^{*} \hat{\nu}_N^{1/2})$ and $(-\infty, \hat{Y}_N - q_{\alpha/2} \hat{\nu}_N^{1/2})$ for the proposed bootstrap method and the Wald-type method respectively.

(c) Two-sided confidence interval. We consider $(\hat{Y}_N - q_{1-\alpha/2}^{*} \hat{\nu}_N^{1/2}, \hat{Y}_N - q_{\alpha/2}^{*} \hat{\nu}_N^{1/2})$ and $(\hat{Y}_N - q_{1-\alpha/2} \hat{\nu}_N^{1/2}, \hat{Y}_N - q_{\alpha/2} \hat{\nu}_N^{1/2})$ for the proposed bootstrap method and the Wald-type method respectively.

For each setup, 5,000 Monte Carlo simulations are conducted, and the two methods are compared in terms of the coverage rates and average lengths of the confidence intervals. In this section, we only report the results for Poisson sampling, and those for SRS and PPS are very similar; see Section S13.1 of the Supplementary Material for details.

Figure 1 summarizes the coverage rates under different setups for Poisson sampling. Generally, the confidence intervals generated by the proposed bootstrap method have better coverage rates compared with those obtained from the Wald-type method, regardless of population sizes, sample sizes, confidence levels and types of confidence intervals. As the sample size increases, the coverage rates improve for both methods, but the performance of the Wald-type method
Figure 1 Coverage rates of confidence intervals with different confidence levels for the proposed bootstrap method (Boot) and the Wald-type method (Wald) under Poisson sampling. For the types of confidence interval, ‘OL’, ‘OU’ and ‘TS’ stand for one-sided lower, one-sided upper and two-sided confidence intervals respectively. The two rows correspond to $N = 10,000$ and $N = 60,000$, and the two columns correspond to $n_0 = 150$ and $n_0 = 300$. [Colour figure can be viewed at wileyonlinelibrary.com]

is questionable even when the sample size is large. Compared with the Wald-type method, the proposed bootstrap method provides more accurate one-sided confidence intervals due to its second-order accuracy; see Theorems 1–2 for details. For example, the coverage rate of the one-sided upper confidence interval with 90% confidence level is less than 0.8 for the Wald-type method when $n_0 = 300$, but it is much closer to 0.9 for the one by the proposed bootstrap method. Although, by Theorem 1, a symmetric two-sided confidence interval by the Wald-type method also guarantees the second-order accuracy theoretically, its numerical coverage rate is questionable as shown in Figure 1. To investigate the advantage of the bootstrap two-sided confidence interval, a higher-order Edgeworth expansion is needed; see Section 2.1 for details. Besides, we can also conclude that the population size has little impact on the coverage rates of the confidence intervals as the sample size is negligible compared with the population size in our setup.

Table 1 summarizes the average lengths of two-sided confidence intervals based on the 5,000 Monte Carlo simulations under Poisson sampling. In general, the two-sided confidence intervals of the proposed bootstrap method are approximately twice as wide as those from the Wald-type method. As the confidence level $1 - \alpha \times 100\%$ increases, the proposed bootstrap method generates a wider two-sided confidence interval than the Wald-type method. The narrower two-sided confidence intervals for the Wald-type method results in lower coverage rates as shown in Figure 1. Thus, the Wald-type method may lead to inflated type-I error for hypothesis testing, even if the sample size is large. As the population size increases, the average length of the two-sided confidence intervals remains almost the same for both methods due to the fact that the sample size is negligible compared with the population size.
TABLE 1  Average lengths of two-sided confidence intervals with different confidence levels, including ‘90%’, ‘95%’ and ‘99%’, under Poisson sampling

| Methods | \( n_0 = 150 \) | \( n_0 = 300 \) |
|---------|----------------|----------------|
|         | 90%  | 95%  | 99%  | 90%  | 95%  | 99%  |
| \( N = 10,000 \) | Boot | 2.17 | 2.90 | 4.90 | 1.25 | 1.58 | 2.35 |
|          | Wald | 1.34 | 1.60 | 2.10 | 0.98 | 1.17 | 1.54 |
| \( N = 60,000 \) | Boot | 2.22 | 2.98 | 5.08 | 1.28 | 1.62 | 2.42 |
|          | Wald | 1.34 | 1.60 | 2.10 | 0.99 | 1.18 | 1.56 |

Boot, the proposed bootstrap method; Wald, the Wald-type method.

FIGURE 2  Approximation of \( F_N(z) \) by the proposed bootstrap method (Boot) and the Wald-type method (Wald) under Poisson sampling. ‘Empi’ corresponds to the empirical distribution of the studentized estimators based on 500,000 Monte Carlo simulations. The two rows correspond to \( N = 10,000 \) and \( N = 60,000 \), and the two columns correspond to \( n_0 = 150 \) and \( n_0 = 300 \) [Colour figure can be viewed at wileyonlinelibrary.com]

As the sample size increases, the two-sided confidence intervals get narrower for both methods.

In addition, we also compare the two methods in terms of accuracy in approximating the distribution function \( F_N(z) \) for \( z \in \{-3, -2.5, \ldots, 2.5, 3\} \). Since the analytical form of \( F_N(z) \) is hard to derive, it is approximated by an empirical distribution based on 500,000 Monte Carlo simulations. For the proposed bootstrap method, the distribution function is estimated by

\[
\hat{F}_N^*(z) = M^{-1} \sum_{m=1}^{M} 1(T^{*(m)} \leq z)
\]  

(16)
for each Monte Carlo simulation, where \(1(x \leq y) = 1\) if \(x \leq y\) and 0 otherwise. A standard normal distribution is used for the Wald-type method. Figure 2 summarizes the simulation results under Poisson sampling, and the values for the proposed bootstrap method are the average of \(\hat{F}_N(z)\) over the 5,000 Monte Carlo simulations. Generally, due to its second-order accuracy, the proposed bootstrap method outperforms the Wald-type method in terms of approximation accuracy of the distribution function \(F_N(z)\) regardless of population sizes and sample sizes. As the sample size increases, the approximation gets better for both methods, but we can still see the advantage of the proposed bootstrap method over the Wald-type method, especially when \(z\) is small.

4.2 Stratified two-stage cluster sampling

In this section, we test the performance of the proposed bootstrap method under stratified two-stage cluster sampling. A finite population \(P_{MS} = \{y_{hij} : h = 1, \ldots , H; i = 1, \ldots , M_h; j = 1, \ldots , M_{hi}\}\) is generated using the following setup. For \(h = 1, \ldots , H\), \(M_h\) is independently generated from \(\{50, 51, \ldots , 80\}\) with equal probability. For \(h = 1, \ldots , H\), \(M_{hi}\) is independently generated from \(\{40, 41, \ldots , 70\}\) with equal probability. For \(h = 1, \ldots , H\), \(i = 1, \ldots , M_h\) and \(j = 1, \ldots , M_{hi}\), \(y_{hij} = a_h + 2b_{hi} + \epsilon_{hij}\), where \(a_h \sim \text{Ex}(10)\) is the stratum effect, \(b_{hi} \sim \text{Ex}(6)\) is the cluster effect, \(\epsilon_{hij} \sim \text{Ex}(4)\), and \(\text{Ex}(\lambda)\) is an exponential distribution with scale parameter \(\lambda\). Within each stratum, we use PPS sampling with replacement to generate \(n_1\) clusters, and the selection probability is proportional to the size of the corresponding PSU. Within each selected PSU, SRS is conducted to generate a sample of size \(n_2\). Based on a realized sample, we are interested in constructing \((1 - \alpha) \times 100\%\) confidence intervals for the population mean \(Y = \sum_{h=1}^{H} \sum_{i=1}^{M_h} y_{hij}\), where \(N = \sum_{h=1}^{H} \sum_{i=1}^{M_h} M_{hi}\) is the population size. In this simulation, we consider \(H \in \{5,10\}\), \(n_1 \in \{5,10\}\), \(n_2 \in \{5,10\}\) and \(\alpha \in \{0.1, 0.05, 0.01\}\). However, due to the page limit, we only show the results for the cases with \(n_2 = 5\) and relegate those with \(n_2 = 10\) to Section S13.2 of the Supplementary Material.

We still consider the three types of confidence intervals in Section 4.1, and 5,000 Monte Carlo simulations are conducted for the proposed bootstrap method with \(M = 5,000\) bootstrap iterations and the Wald-type method. Figure 3 summarizes the coverage rates under stratified two-stage cluster sampling with \(n_2 = 5\). Generally, the confidence intervals from the proposed bootstrap method have better coverage rates than those from the Wald-type methods, regardless of stratum sizes, the number of selected clusters within each stratum and confidence levels. As \(H\) or \(n_1\) increases, the coverage rates improve for both methods since a large sample is available. However, even when \(H = 10\) and \(n_1 = 10\), the propose bootstrap method still outperforms the Wald-type method. For example, when the confidence level is 90%, the coverage rate of the one-sided upper interval by the Wald-type method is much smaller 0.9, but the coverage rate of the proposed bootstrap method is close to 0.9.

Table 2 summarizes the average lengths of two-sided confidence intervals based on the 5,000 Monte Carlo simulations under stratified two-stage cluster sampling with \(n_2 = 5\). The confidence intervals obtained from the proposed bootstrap method are generally wider than those from the Wald-type method under different setups, especially when \(H\) or \(n_1\) is small. As the confidence level \((1 - \alpha) \times 100\%\) increases, the proposed bootstrap method generates a wider confidence interval than the Wald-type method. As \(H\) or \(n_1\) increases, the confidence intervals get narrower for both methods. When both \(H\) and \(n_1\) are large, the two-sided confidence intervals have almost the
FIGURE 3 Coverage rates of confidence intervals with different confidence levels for the proposed bootstrap method (Boot) and the Wald-type method (Wald) under stratified two-stage cluster sampling with $n_2 = 5$. For the types of confidence interval, ‘OL’, ‘OU’ and ‘TS’ stand for one-sided lower, one-sided upper and two-sided confidence intervals respectively. The two rows correspond to $H = 5$ and $H = 10$, and the two columns correspond to $n_1 = 5$ and $n_1 = 10$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Average lengths of two-sided confidence intervals with different confidence levels, including ‘90%’, ‘95%’ and ‘99%’, under stratified two-stage cluster sampling with $n_2 = 5$

|          | $n_1 = 5$ |         |         |          |         |         |          |         |         |
|----------|-----------|---------|---------|----------|---------|---------|----------|---------|---------|
|          | $n_1 = 5$ |         |         |          |         |         |          |         |         |
|          | $n_1 = 5$ | $n_1 = 10$ |
| Methods  | 90%      | 95%     | 99%     | 90%      | 95%     | 99%     | 90%      | 95%     | 99%     |
| $H = 5$  | Boot     | 10.46   | 13.19   | 19.84    | 6.62    | 8.15    | 11.38    |
|          | Wald     | 8.16    | 9.72    | 12.77    | 5.89    | 7.01    | 9.22     |
| $H = 10$ | Boot     | 6.00    | 7.37    | 10.38    | 4.07    | 4.90    | 6.62     |
|          | Wald     | 5.54    | 6.60    | 8.67     | 3.93    | 4.69    | 6.16     |

Boot, the proposed bootstrap method; Wald, the Wald-type method.

same length for both methods, and the coverage rates are also approximately the same as shown in Figure 3.

As in Section 4.1, we also compare the proposed bootstrap method with the Wald-type method in terms of approximation accuracy of the distribution function of the studentized estimator, and the true distribution function is approximated by an empirical one based on 500,000 Monte Carlo simulations. Figure 4 summarizes the simulation results under stratified two-stage cluster sampling with $n_2 = 5$, and the values for the proposed bootstrap method are the average of similar empirical distributions in (16) over the 5,000 Monte Carlo simulations. Generally, due to its second-order accuracy, the proposed bootstrap method outperforms the Wald-type method in approximating the distribution function $F_N(z)$ accurately regardless of stratum sizes.
and the number of clusters selected with each stratum. For example, when $H = 5$ and $n_1 = 5$, the proposed bootstrap method approximates the empirical distribution function much better than the Wald-type method, especially for a small $z$. As $H$ and $n_1$ increase, the approximation of both methods improves, but we can still observe the advantage of the proposed bootstrap method over the Wald-type method.

5 | CONCLUSION

In this paper, we propose bootstrap methods under stratified multi-stage cluster sampling as well as some commonly used single-stage sampling designs, and rigorously show that the proposed bootstrap methods are second-order accurate. Unlike the popular rescaling bootstrap methods in survey sampling, we generate bootstrap finite populations, from which the same sampling design is conducted to generate bootstrap samples. Thus, the proposed bootstrap method mimics the sampling mechanism directly. Since the proposed bootstrap method is based on an asymptotically pivotal statistic, it is necessary to estimate the variance of the design-unbiased estimator, and such a variance estimator is usually available under commonly used sampling designs. Numerical results show that the proposed bootstrap method outperforms the Wald-type method in terms of coverage rates of confidence intervals and approximation accuracy of the distribution function of the studentized estimator.

The main focus of the proposed bootstrap method is to estimate the population mean, and it may be interesting to extend this method to make inference on the parameters defined through estimating equations. The proposed bootstrap method can be used to handle analytic inference
with informative sampling (Pfeffermann & Sverchkov, 1999). Application to M-estimation or Z-estimation using the empirical process theory (Han & Wellner, 2021) can be an interesting research topic as well.

**ACKNOWLEDGEMENT**
The authors thank the anonymous referee, the Associate Editor, professor Simon Wood, and professor Wayne Fuller for their constructive comments which have greatly improved the paper. Open access funding provided by the Iowa State University Library. [Correction added on 24 May 2022, after first online publication: ISUL funding statement has been added.]

**DATA AVAILABILITY STATEMENT**
* Functions.R: it contains the code for the proposed bootstrap method under different sampling designs.
* Demo.R: It contains the demo code based on our simulation setup.
* Readme.txt: It is a brief description about the two files above.

**ORCID**
Jae Kwang Kim [https://orcid.org/0000-0002-0246-6029](https://orcid.org/0000-0002-0246-6029)

**REFERENCES**
Antal, E. & Tillé, Y. (2011) A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106, 534–543.
Athreya, K.B. & Lahiri, S.N. (2006) Measure theory and probability theory. New York: Springer.
Athreya, K.B., Ghosh, M., Low, L.Y. & Sen, P.K. (1984) Laws of large numbers for bootstrapped U-statistics. *Journal of Statistical Planning and Inference*, 9, 185–194.
Babu, G.J. & Singh, K. (1983) Inference on means using the bootstrap. *The Annals of Statistics*, 11, 999–1003.
Babu, G.J. & Singh, K. (1984) On one term Edgeworth correction by Efron’s bootstrap. *Sankhyā A*, 46, 219–232.
Babu, G.J. & Singh, K. (1985) Edgeworth expansions for sampling without replacement from finite populations. *Journal of Multivariate Analysis*, 17, 261–278.
Beaumont, J.F. & Patak, Z. (2012) On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127–148.
Bhattacharya, R.N. & Rao, R.R. (2010) *Normal approximation and asymptotic expansions*. Philadelphia: SIAM.
Bickel, P.J. & Freedman, D.A. (1981) Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9, 1196–1217.
Bickel, P.J. & Freedman, D.A. (1984) Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470–482.
Booth, J.G., Butler, R.W. & Hall, P. (1994) Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282–1289.
Chen, S., Haziza, D., Léger, C. & Mashreghi, Z. (2019) Pseudo-population bootstrap methods for imputed survey data. *Biometrika*, 106, 369–384.
Chipperfield, J. & Preston, J. (2007) Efficient bootstrap for business surveys. *Survey Methodology*, 33, 167–172.
Das, D., Gregory, K. & Lahiri, S.N. (2019) Perturbation bootstrap in adaptive lasso. *The Annals of Statistics*, 47, 2080–2116.
Davison, A.C. & Sardy, S. (2007) Resampling variance estimation in surveys with missing data. *Journal of Statistics*, 23, 371–386.
DeBell, M., Amsbary, M., Meldener, V., Brock, S. & Maisel, N. (2018) Methodology report for the ANES 2016 time series study. Tech. rep., American National Election Studies, USA.

DiCiccio, T.J. & Romano, J.P. (1995) On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statistica Sinica*, 5, 141–160.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1–26.

Feller, W. (1971) *An introduction to probability theory and its applications*, vol. 2. New York: John Wiley.

Fuller, W.A. (2009) *Sampling statistics*. Hoboken: John Wiley.

Gross, S. (1980) Median estimation in sample surveys. In: *Proceedings of the Survey Research*, 181–184.

Hall, P. (1987) Edgeworth expansion for student’s t statistic under minimal moment conditions. *Annals of Probability*, 15, 920–931.

Hall, P. (1992) *The bootstrap and Edgeworth expansion*. New York: Springer.

Han, Q. & Wellner, J. A. (2021) Complex sampling designs: uniform limit theorems and applications. *The Annals of Statistics*, 49, 459.

Hansen, M.H. & Hurwitz, W.N. (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333–362.

Helmers, R. (1991) On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistic. *The Annals of Statistics*, 19, 470–484.

Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

Jing, B.-Y. & Wang, Q. (2003) Edgeworth expansion for U-statistics under minimal conditions. *The Annals of Statistics*, 31, 1376–1391.

Karabulut, I. & Lahiri, S.N. (1997) Two-term edgeworth expansion for M-estimators of a linear regression parameter without Cramér-type conditions and an application to the bootstrap. *Journal of the Australian Mathematical Society (Series A)*, 62, 361–370.

Kim, J., Rao, J.N.K. & Wang, Z. (2021) Hypotheses testing from complex survey data using bootstrap weights: A unified approach. Available from: https://arxiv.org/abs/1902.08944

Kovar, J.G., Rao, J.N.K. & Wu, C.F.J. (1988) Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25–45.

Krewski, D. & Rao, J.N.K. (1981) Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 1010–1019.

Lahiri, S.N. (1992) On bootstrapping M-estimators. *Sankhya A*, 54, 157–190.

Lahiri, S.N. (1994) On two-term Edgeworth expansion and bootstrap approximations for studentized multivariate M-estimators. *Sankhya A*, 56, 201–226.

Liu, Y.R. (1988) Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16, 1696–1708.

Mashreghi, Z., Léger, C. & Haziza, D. (2014) Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *Canadian Journal of Statistics*, 42, 142–167.

Mashreghi, Z., Haziza, D. & Léger, C. (2016) A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1–52.

McCarthy, P.J. & Snowden, C.B. (1985) The bootstrap and finite population sampling. In *Vital and Health Statistics*, no. in Series 2, DHHS Pub. No. (PHS) 85–1369. Washington. U.S. Government Printing Office: Public Health Service.

Pfeffermann, D. & Sverchkov, M. (1999) Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā Series B*, 61, 166–186.

Rao, J.N.K. & Wu, C.F.J. (1984) Bootstrap inference for sample surveys. In: *Proceedings of the Survey Research*, 106–112.

Rao, J.N.K. & Wu, C.F.J. (1988) Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231–241.

Rao, J.N.K., Wu, C.F.J. & Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209–217.

Saegusa, T. (2014) Bootstrapping two-phase sampling. Available from: https://arxiv.org/pdf/1406.5580.pdf

Shao, J. & Sitter, R.R. (1996) Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278–1288.
Singh, K. (1981) On the asymptotic accuracy of efron’s bootstrap. *The Annals of Statistics*, 6, 1187–1195.
Sitter, R.R. (1992a) Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135–154.
Sitter, R.R. (1992b) A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755–765.
Sverchkov, M. & Pfeffermann, D. (2004) Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79–92.

**SUPPORTING INFORMATION**
Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Wang, Z., Peng, L. & Kim, J.K. (2022) Bootstrap inference for the finite population mean under complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1150–1174. Available from: [https://doi.org/10.1111/rssb.12506](https://doi.org/10.1111/rssb.12506)

**APPENDIX A. ALGORITHMS**
This section contains the proposed algorithms under SRS and PPS sampling with replacement, where \( A \setminus B = \{x : x \in A, x \notin B \} \) for two sets \( A \) and \( B \).

*Algorithm 3: Bootstrap method under SRS*

- **Data:** realized sample \( \{y_{N,1}, \ldots, y_{N,n}\} \), population size \( N \), bootstrap repetition number \( M \).
- **Result:** Bootstrap studentized estimators \( \{\hat{T}^{(m)}_{N,SRS} : m = 1, \ldots, M \} \).

1. for \( m = 1, \ldots, M \) do
2. \hspace{1em} Generate \( (N^*_1, \ldots, N^*_n) \) by a multinomial distribution \( MN(N; \rho_N) \) with \( \rho_N = (n^{-1}, \ldots, n^{-1}) \)
3. \hspace{1em} Obtain \( \hat{x}^*_N \) by replicating \( y_{N,i} \) for \( N^*_i \) times for \( i = 1, \ldots, n \)
4. \hspace{1em} \( \hat{Y}^{*}_{N,SRS} \leftarrow n^{-1} \sum_{i=1}^{n} Y^*_{N,i} \)
5. \hspace{1em} \( U^{*} \leftarrow \{1, \ldots, N\} \quad S^* \leftarrow \emptyset \) for \( i = 1, \ldots, n \) do
6. \hspace{2em} Generate \( J^*_i \) at random from \( U^{*} \setminus \{J^*_i\} \) with equal selection probability for each element
7. \hspace{2em} \( U^* \leftarrow U^{*} \setminus \{J^*_i\} \)
8. \hspace{2em} \( S^* \leftarrow S^* \cup \{J^*_i\} \)
9. end
10. \( \hat{Y}^*_{N,SRS} \leftarrow n^{-1} \sum_{i \in S^*} Y^*_{N,i} \)
11. \( s^2_{N,SRS} \leftarrow n^{-1} \sum_{i \in S^*} (Y^*_{N,i} - \hat{Y}^*_{N,SRS})^2 \)
12. \( \hat{Y}^*_{SRS} \leftarrow n^{-1} (1 - nN^{-1}) s^2_{N,SRS} \)
13. \( T^{(m)}_{N,SRS} \leftarrow (\hat{Y}^*_{N,SRS} - \hat{Y}^*_{N})^{-1/2} (\hat{Y}^*_{N,SRS} - \hat{Y}^*_{N}) \)
14. end
Algorithm 4: Bootstrap method under PPS sampling with replacement

Data: realized sample \( (y_{N,a(i)} : i = 1, \ldots, n) \), selection probabilities \( (p_{N,a(i)} : i = 1, \ldots, n) \), population size \( N \), bootstrap repitition number \( M \).

Result: Bootstrap studentized estimators \( \{T_{N,PPS}^{(m)} : m = 1, \ldots, M\} \).

1. for \( i = 1, \ldots, n \) do
   2. \( \rho_{N,i} \leftarrow \rho_{N,a(i)}^{-1}(\sum_{j=1}^{n} \rho_{N,a(j)})^{-1} \)
   3. end

4. for \( m = 1, \ldots, M \) do
   5. Generate \( \{N_{a(1)}^{*}, \ldots, N_{a(n)}^{*}\} \) by a multinomial distribution \( \text{MN}(N; \rho_{N}) \)
   6. \( C_{N}^{*} \leftarrow \sum_{i=1}^{n} \hat{p}_{N,i}^{*} \)
   7. Obtain \( \hat{y}_{N}^{*} \) and \( \{p_{N,1}^{*}, \ldots, p_{N,N}^{*}\} \) by replicating \( y_{N,a(i)} \) and \( p_{N,a(i)} \) for \( N_{a(i)}^{*} \) times for \( i = 1, \ldots, n \)
   8. \( \hat{p}_{N}^{*} \leftarrow (C_{N}^{*})^{-1} p_{N,1}^{*}, \ldots, (C_{N}^{*})^{-1} p_{N,N}^{*} \)
   9. \( \hat{y}_{N,PPS}^{*} \leftarrow N^{-1} \sum_{i=1}^{n} y_{N,i}^{*} \)
   10. for \( i = 1, \ldots, n \) do
       11. Generate \( b(i) \) from \( \{1, \ldots, N\} \) with selection probabilities \( \hat{p}_{N,i}^{*} \)
   12. end
   13. \( \hat{y}_{N,PPS}^{*} \leftarrow n^{-1} \sum_{i=1}^{n} N^{-1} C_{N}^{*}(p_{N,b(i)}^{*})^{-1} y_{N,b(i)}^{*} \)
   14. \( \hat{y}_{N,PPS}^{*} \leftarrow n^{-2} \sum_{i=1}^{n} (N^{-1} C_{N}^{*}(p_{N,b(i)}^{*})^{-1} y_{N,b(i)}^{*} - \hat{y}_{N,PPS}^{*})^{2} \)
   15. \( T_{N,PPS}^{(m)} \leftarrow (\hat{y}_{N,PPS}^{*} - \hat{y}_{N}^{*})^{-1/2} \)
   16. end