Electronic Journal of Statistics
Vol. 9 (2015) 535–566
ISSN: 1935-7524
DOI: 10.1214/15-EJS1011

Inference from small and big data sets with error rates*

Miklós Csörgő and Masoud M. Nasari†

School of Mathematics and Statistics of Carleton University
Ottawa, ON, Canada
e-mail: mcsorgo@math.carleton.ca; mmnasari@math.carleton.ca

Abstract: In this paper we introduce randomized t-type statistics that will be referred to as *randomized pivots*. We show that these randomized pivots yield central limit theorems with a significantly smaller error as compared to that of their classical counterparts under the same conditions. This constitutes a desirable result when a relatively small number of data is available. When a data set is too big to be processed, or when it constitutes a random sample from a super-population, we use our randomized pivots to infer about the mean based on significantly smaller sub-samples. The approach taken is shown to relate naturally to estimating distributions of both small and big data sets.

MSC 2010 subject classifications: Primary 62E20; secondary 62G09.
Keywords and phrases: Randomized t-pivots, Berry-Esséen bounds, improved CLT's, small and moderate samples.

Received July 2014.

Contents

*Research supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant of M. Csörgő.
†Corresponding author.

1. Introduction

In this paper we address the problem of making inference about the population mean when the available sample is either small or big. In case of having a small sample, we develop a randomization technique that yields central limit theorems (CLT's) with a significantly smaller magnitude of error that would compensate for the lack of sufficient information as a result of not having a large number of observations. In case of having a relatively small sample, our randomization technique provides an efficient alterative to the computationally intensive bootstrap (cf. Efron [12] and Efron and Tibshirani [13]). Assigning appropriate random weights to the data, the randomization approach used in this paper yields the same asymptotic accuracy as that of the bootstrap without a need for repeated re-sampling from the original data (cf. Hall [17] for details on the asymptotic accuracy of the bootstrap).

For more on the bootstrap, we refer to Shao and Tu [26] and Lahiri [20] for bootstrapping time series data. We note also that the randomly weighted pivotal quantities that are to be introduced in (3) and (4) are normalized randomly weighted partial sums of the original data. This, in turn, suggests a closeness in nature between our approach to creating more accurate pivots to that of the so-called weighted bootstrap (cf., for example, Arenal-Gutiérrez and Matrán [1], Barbe and Bertail [2], Csörgő et al. [7], Mason and Newton [21] and Mason and Shao [22]). Despite the mentioned similarity, our viewpoint in this article is fundamentally different from the bootstrap. Unlike the bootstrap, we use randomization to create direct more accurate pivots for the parameter of interest in hand, i.e., the mean. To illustrate the idea, we mention that the weighted t-statistic as in (3) is used as a direct pivot for the sample mean $\bar{X}_n$ rather than being used to estimate the cut-off points of the sampling distribution of the classical t-pivot for the mean, as it is the case in the bootstrap. In contrast we introduce a randomized and more accurate direct pivot, as in (4), for the population mean.

The randomization framework in this paper also accommodates the super-population perspective in which a finite population of numbers is viewed as a random sample drawn from an imaginary super-population (cf. Hartley and Silken [18]). Adopting this view, via randomization, we also study important characteristics, such as the mean and distribution, of a finite population in this context.

Unless stated otherwise, $X, X_1, \ldots$ throughout are assumed to be independent random variables with a common distribution function F (i.i.d. random variables), mean $\mu := E_X X$ and variance $0 < \sigma^2 := E_X(X-\mu)^2 < +\infty$. Based on $X_1, \ldots, X_n$, a random sample on X, for each integer $n \geq 1$, define

$$\bar{X}_n := \sum_{i=1}^{n} X_i/n \text{ and } S_n^2 := \sum_{i=1}^{n} (X_i - \bar{X}_n)^2/n,$$

the sample mean and sample variance, respectively, and consider the classical Student t-statistic

$$T_n(X) := \frac{\bar{X}_n}{S_n/\sqrt{n}} = \frac{\sum_{i=1}^n X_i}{S_n\sqrt{n}} \tag{1}$$

that, in turn, on replacing X_i by $X_i - \mu$, $1 \le i \le n$, yields

$$T_n(X-\mu) := \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\sum_{i=1}^n (X_i - \mu)}{S_n\sqrt{n}}, \tag{2}$$

the classical Student t-pivot for the population mean μ.

Define now $T^{(1)}_{m_n,n}$ and $G^{(1)}_{m_n,n}$, randomized versions of $T_n(X)$ and $T_n(X-\mu)$ respectively, as follows:

$$T^{(1)}_{m_n,n} := \frac{\bar{X}_{m_n,n} - \bar{X}_n}{S_n\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}} = \frac{\sum_{i=1}^n \big(\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big)X_i}{S_n\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}}, \tag{3}$$

$$G^{(1)}_{m_n,n} := \frac{\sum_{i=1}^n \big|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big|\big(X_i - \mu\big)}{S_n\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}}, \tag{4}$$

where,

$$\bar{X}_{m_n,n} := \sum_{i=1}^n w_i^{(n)} X_i/m_n, \tag{5}$$

is the randomized sample mean and the weights $(w_1^{(n)}, \ldots, w_n^{(n)})$ have a multinomial distribution of size $m_n := \sum_{i=1}^n w_i^{(n)}$ with respective probabilities $1/n$, i.e.,

$$(w_1^{(n)}, \ldots, w_n^{(n)}) \stackrel{d}{=} \mathit{multinomial}(m_n; \frac{1}{n}, \ldots, \frac{1}{n}).$$

The just introduced respective randomized $T^{(1)}_{m_n,n}$ and $G^{(1)}_{m_n,n}$ versions of $T_n(X)$ and $T_n(X-\mu)$ can be computed via re-sampling from the set of indices $\{1, \ldots, n\}$ of $X_1, \ldots, X_n$ with replacement m_n times so that, for each $1 \le i \le n$, $w_i^{(n)}$ is the count of the number of times the index i of X_i is chosen in this re-sampling process.

Remark 1.1. In view of the preceding definition of $w_i^{(n)}$, $1 \le i \le n$, they form a row-wise independent triangular array of random variables such that $\sum_{i=1}^n w_i^{(n)} = m_n$ and, for each $n \ge 1$,

$$(w_1^{(n)}, \ldots, w_n^{(n)}) \stackrel{d}{=} \mathit{multinomial}(m_n; \frac{1}{n}, \ldots, \frac{1}{n}),$$

i.e., the weights have a multinomial distribution of size m_n with respective probabilities $1/n$. Clearly, for each n, $w_i^{(n)}$ are independent from the random sample X_i, $1 \le i \le n$. Weights denoted by $w_i^{(n)}$ will stand for triangular multinomial random variables in this context throughout.

Thus, $T^{(1)}_{m_n,n}$ and $G^{(1)}_{m_n,n}$ can simply be computed by generating, independently from the data, a realization of the random multinomial weights $(w^{(n)}_1, \ldots, w^{(n)}_n)$ as in Remark 1.1.

Define the similarly computable further randomized versions $T^{(2)}_{m_n,n}$ and $G^{(2)}_{m_n,n}$ of $T_n(X)$ and $T_n(X-\mu)$ respectively, as follows:

$$T^{(2)}_{m_n,n} := \frac{\bar{X}_{m_n,n} - \bar{X}_n}{S_{m_n,n}\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}} = \frac{\sum_{j=1}^n \big(\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big) X_i}{S_{m_n,n}\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}} \tag{6}$$

$$G^{(2)}_{m_n,n} := \frac{\sum_{i=1}^n \big|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big|\big(X_i - \mu\big)}{S_{m_n,n}\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}}, \tag{7}$$

where $S^2_{m_n,n}$ is the randomized sample variance, defined as

$$S^2_{m_n,n} := \sum_{i=1}^n w_i^{(n)} \big(X_i - \bar{X}_{m_n,n}\big)^2 / m_n. \tag{8}$$

Unlike $T_n(X)$ that can be transformed into $T_n(X-\mu)$, the Student pivot for μ as in (2) (cf. Giné *et al.* [16] for the asymptotic equivalence of the two), its randomized versions $T^{(1)}_{m_n,n}$ and $T^{(2)}_{m_n,n}$ do not have this straightforward property, i.e., they do not yield a pivotal quantity for the population mean $\mu = E_X X$ by simply replacing each X_i by $X_i - \mu$ in their definitions. We introduced $G^{(1)}_{m_n,n}$ and $G^{(2)}_{m_n,n}$ in this paper to serve as direct randomized pivots for the population mean μ, while $T^{(1)}_{m_n,n}$ and $T^{(2)}_{m_n,n}$ will now be viewed on their own as randomized pivots for the sample mean $\bar{X}_n$ in case of a big data set.

Our Theorem 2.1 and its corollaries will explain the higher order accuracy these randomized pivots provide for inference about the mean μ, as compared to that provided by $T_n(X-\mu)$.

Among the many outstanding contributions in the literature studying the asymptotic behavior of $T_n(X)$ and $T_n(X-\mu)$, our main tool in this paper, Theorem 2.1 below, relates mostly to Bentkus *et al.* [4], Bentkus and Götze [5], Pinelis [24] and Shao [27].

A short outline of the contributions of this paper reads as follows.

In Section 2 we derive the rates of convergence for $G^{(i)}_{m_n,n}$ and $T^{(i)}_{m_n,n}$, $i = 1, 2$, via establishing Berry-Esséen type results in Theorem 2.1 and its Corollaries 2.1–2.3. In Corollary 2.3 we show that, on taking $m_n = n$, $G^{(i)}_{m_n,n}$ and $T^{(i)}_{m_n,n}$, $i = 1, 2$, converge, in distribution, to the standard normal at the rate of $O(1/n)$. This rate is significantly better than the best possible $O(1/\sqrt{n})$ rate of convergence under similar moment conditions for the classical t-statistic $T_n(X)$ and its Student pivot $T_n(X-\mu)$, based on a random sample of size n. The latter $O(1/\sqrt{n})$ rate is best possible in the sense that it cannot be improved without restricting the class of distribution functions of the data, for example, to normal or symmetrical

distributions. In section 2 we also present numerical studies that well support our conclusion that, on taking $m_n = n$, $G^{(i)}_{m_n,n}$ and $T^{(i)}_{m_n,n}$, $i = 1, 2$, converge to standard normal at a significantly faster rate than that of the classical CLT. In Sections 4 and 5, the respective rates of convergence of the CLT's in Section 2 will be put to significant use. In Section 4, $G^{(i)}_{m_n,n}$, $i = 1, 2$, are studied as natural asymptotic pivots for the population mean $\mu = E_X X$. In section 5, $T^{(i)}_{m_n,n}$, $i = 1, 2$, are studied as natural asymptotic pivots for the sample mean $\bar{X}_n$ that closely shadows μ, when dealing with big data sets of univariate observations of n labeled units $\{X_1, \ldots, X_n\}$. In this case, instead of trying to process the entire data set that may even be impossible to do, sampling it via generating random weights independently from the data as in Remark 1.1 (cf. Section 5) makes it possible to use $T^{(2)}_{m_n,n}$ to construct an interval estimation for the sample mean $\bar{X}_n$ based on significantly smaller sub-samples. The latter confidence set for $\bar{X}_n$ in turn will be seen to contain the population mean μ as well, and with same rates of convergence, in terms m_n and n, as those established for having $\bar{X}_n$ in there. In Section 6 the sample and population distribution functions are studied along the lines of Sections 2–5. The proofs are given in Sections 7 and Appendices A and B.

For throughout use, we let $(\Omega_X, \mathfrak{F}_X, P_X)$ denote the probability space of the random variables $X, X_1, \ldots$, and $(\Omega_w, \mathfrak{F}_w, P_w)$ be the probability space on which the weights

$$\big(w_1^{(1)}, (w_1^{(2)}, w_2^{(2)}), \ldots, (w_1^{(n)}, \ldots, w_n^{(n)}), \ldots\big)$$

are defined. In view of the independence of these two sets of random variables, jointly they live on the direct product probability space $(\Omega_X \times \Omega_w, \mathfrak{F}_X \otimes \mathfrak{F}_w, P_{X,w} = P_X . P_w)$. For each $n \geq 1$, we also let $P_{.|w}(.)$ stand for the conditional probabilities given $\mathfrak{F}_w^{(n)} := \sigma(w_1^{(n)}, \ldots, w_n^{(n)})$ with corresponding conditional expected value $E_{.|w}(.)$.

2. The rate of convergence of the CLT's for $G^{(i)}_{m_n}$ and $T^{(i)}_{m_n}$, $i = 1, 2$

One of the efficient tools to control the error when approximating the distribution function of a statistic with that of a standard normal random variable is provided by Berry-Esséen type inequalities (cf., e.g., Serfling [25]), which provide upper bounds for the error of approximation for any finite number of observations in hand. It is well known that, on assuming $E_X|X - \mu|^3 < +\infty$, as the sample size n increases to infinity, the rate at which the Berry-Esséen upper bound for $\sup_{-\infty<t<+\infty} |P_X(T_n(X-\mu) \leq t) - \Phi(t)|$ vanishes is $O(n^{-1/2})$, where, and also throughout, Φ stands for the standard normal distribution function.

Furthermore, the latter rate is best possible in the sense that it cannot be improved without narrowing the class of distribution functions considered.

Our Berry-Esséen type inequalities for the respective conditional, given the weights $w_i^{(n)}$'s, distributions of $G^{(1)}_{m_n,n}$ and $T^{(1)}_{m_n,n}$, as in (4) and (3) respectively, and $G^{(2)}_{m_n,n}$ and $T^{(2)}_{m_n,n}$, as in (7) and (6) respectively, read as follows.

Theorem 2.1. *Assume that $E_X|X|^3 < +\infty$ and let $\Phi(.)$ be the standard normal distribution function. Also, for arbitrary positive numbers δ, ε, let $\varepsilon_1, \varepsilon_2 > 0$ be so that $\delta > (\varepsilon_1/\varepsilon)^2 + P_X(|S_n^2 - \sigma^2| > \varepsilon_1^2) + \varepsilon_2 > 0$, where, for $t \in \mathbb{R}$, $\Phi(t-\varepsilon) - \Phi(t) > -\varepsilon_2$ and $\Phi(t+\varepsilon) - \Phi(t) < \varepsilon_2$. Then, for all n, m_n we have*

$$
\begin{aligned}
(A)\quad & P_w\Big\{ \sup_{-\infty<t<+\infty} \Big| P_{X|w}(G^{(1)}_{m_n,n} \le t) - \Phi(t) \Big| > \delta \Big\} \\
& \le \delta_n^{-2}(1-\varepsilon)^{-3}(1-\frac{1}{n})^{-3}(\frac{n}{m_n^3} + \frac{n^2}{m_n^3})\{\frac{15m_n^3}{n^3} + \frac{25m_n^2}{n^2} + \frac{m_n}{n}\} \\
& + \varepsilon^{-2}\frac{m_n^2}{(1-\frac{1}{n})}\Big\{\frac{1-\frac{1}{n}}{n^3m_n^3} + \frac{(1-\frac{1}{n})^4}{m_n^3} + \frac{(m_n-1)(1-\frac{1}{n})^2}{nm_n^3} + \frac{4(n-1)}{n^3m_n} + \frac{1}{m_n^2} \\
& - \frac{1}{nm_n^2} + \frac{n-1}{n^3m_n^3} + \frac{4(n-1)}{n^2m_n^3} - \frac{(1-\frac{1}{n})^2}{m_n^2}\Big\},
\end{aligned}
$$

and also

$$
\begin{aligned}
(B)\quad & P_w\Big\{ \sup_{-\infty<t<+\infty} \Big| P_{X|w}(T^{(1)}_{m_n,n} \le t) - \Phi(t) \Big| > \varepsilon \Big\} \\
& \le \delta_n^{-2}(1-\varepsilon)^{-3}(1-\frac{1}{n})^{-3}(\frac{n}{m_n^3} + \frac{n^2}{m_n^3})\{\frac{15m_n^3}{n^3} + \frac{25m_n^2}{n^2} + \frac{m_n}{n}\} \\
& + \varepsilon^{-2}\frac{m_n^2}{(1-\frac{1}{n})}\Big\{\frac{1-\frac{1}{n}}{n^3m_n^3} + \frac{(1-\frac{1}{n})^4}{m_n^3} + \frac{(m_n-1)(1-\frac{1}{n})^2}{nm_n^3} + \frac{4(n-1)}{n^3m_n} + \frac{1}{m_n^2} \\
& - \frac{1}{nm_n^2} + \frac{n-1}{n^3m_n^3} + \frac{4(n-1)}{n^2m_n^3} - \frac{(1-\frac{1}{n})^2}{m_n^2}\Big\},
\end{aligned}
$$

where

$$
\delta_n := \frac{\delta - (\varepsilon_1/\varepsilon)^2 - P_X(|S_n^2 - \sigma^2| > \varepsilon_1^2) + \varepsilon_2}{CE_X|X-\mu|^3/\sigma^{3/2}},
$$

with C being a universal constant as in the Berry-Esséen upper bound for independent and not necessarily identically distributed summands (cf. page 33 of Serfling [25]).

The following result, a corollary to Theorem 2.1, gives the rate of convergence of the respective conditional CLT's for $G^{(1)}_{m_n,n}$ and $T^{(1)}_{m_n,n}$, as well as for $G^{(2)}_{m_n,n}$ and $T^{(2)}_{m_n,n}$.

Corollary 2.1. *Assume that $E_X|X|^3 < +\infty$. If $n,\ m_n \to +\infty$ in such a way that $m_n = o(n^2)$, then, for arbitrary $\delta > 0$, we have*

$$
(A)\quad P_w\Big\{ \sup_{-\infty<t<+\infty} \Big| P_{X|w}(G^{(1)}_{m_n,n} \le t) - \Phi(t) \Big| > \delta \Big\} = O\Big(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}\}\Big),
$$

$$
(B)\quad P_w\Big\{ \sup_{-\infty<t<+\infty} \Big| P_{X|w}(T^{(1)}_{m_n,n} \le t) - \Phi(t) \Big| > \delta \Big\} = O\Big(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}\}\Big).
$$

Moreover, if $E_X X^4 < +\infty$, if $n, m_n \to +\infty$ in such a way that $m_n = o(n^2)$ and $n = o(m_n^2)$ then, for $\delta > 0$, we also have

$$
(C)\quad P_w\Big\{ \sup_{-\infty<t<+\infty} \Big| P_{X|w}(G^{(2)}_{m_n,n} \le t) - \Phi(t) \Big| > \delta \Big\} = O\Big(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}, \frac{n}{m_n^2}\}\Big),
$$

$$(D)\quad P_w\Big\{\sup_{-\infty<t<+\infty}\Big|P_{X|w}(T^{(2)}_{m_n,n}\le t)-\Phi(t)\Big|>\delta\Big\}=O\Big(\max\{\frac{m_n}{n^2},\frac{1}{m_n},\frac{n}{m_n^2}\}\Big).$$

When $0<E_XX^2<+\infty$, the conditional $P_{X|w}$ CLT's for $G^{(i)}_{m_n,n}$ and $T^{(i)}_{m_n,n}$, $i=1,2$, whose respective rates of convergence are established in Corollary 2.1, can be concluded as direct consequences of a realization of the Lindeberg-Feller CLT (cf. Theorems 27.3 and 27.4 of Billingsley [6]) as formulated in Lemma 5.1 of Csörgő *et al.* [7] (cf. also Appendix B) that is also known as the Hájeck -Sidák Theorem (cf., e.g., Theorem 5.3 in Das Gupta [10]).

Remark 2.1. On taking $m_n=n$, when $E_X|X|^3<+\infty$, the rates of convergence of Corollary 2.1 for both $G^{(1)}_{m_n,n}$ and $T^{(1)}_{m_n,n}$ are of order $O(n^{-1})$. The same is true for $G^{(2)}_{m_n,n}$ and $T^{(2)}_{m_n}$ for $m_n=n$ when $E_XX^4<+\infty$.

Remark 2.2. When $E_XX^4<+\infty$, the extra term n/m_n^2 which appears in the rate of convergence of $G^{(2)}_{m_n,n}$ and $T^{(2)}_{m_n,n}$ in (C) and (D) of Corollary 2.1, is the rate at which $P_w\{P_{X|w}(|S^2_{m_n,n}-S^2_n|>\varepsilon_1)>\varepsilon_2\}$ approaches zero as $n,m_n\to+\infty$, where ε_1 and ε_2 are arbitrary positive numbers.

The conditional CLT's resulting from (A), (B), (C) and (D) of Corollary 2.1 imply respective unconditional CLT's in terms of the joint distribution $P_{X,w}$ as in the following Corollaries 2.2 and 2.3.

Corollary 2.2. *Assume that $E_X|X|^3<+\infty$. If $n,\ m_n\to+\infty$ in such a way that $m_n=o(n^2)$, then*

$$\sup_{-\infty<t<+\infty}\big|P_{X,w}(G^{(1)}_{m_n,n}\le t)-\Phi(t)\big| \le O\Big(\max\{\frac{m_n}{n^2},\frac{1}{m_n}\}\Big), \tag{9}$$

$$\sup_{-\infty<t<+\infty}\Big|P_{X,w}(T^{(1)}_{m_n,n}\le t)-\Phi(t)\Big| \le O\Big(\max\{\frac{m_n}{n^2},\frac{1}{m_n}\}\Big). \tag{10}$$

Moreover, if $E_XX^4<+\infty$, if $n,m_n\to+\infty$ in such a way that $m_n=o(n^2)$ and $n=o(m_n^2)$, then

$$\sup_{-\infty<t<+\infty}\big|P_{X,w}(G^{(2)}_{m_n,n}\le t)-\Phi(t)\big| \le O\Big(\max\{\frac{m_n}{n^2},\frac{1}{m_n},\frac{n}{m_n^2}\}\Big) \tag{11}$$

$$\sup_{-\infty<t<+\infty}\big|P_{X,w}(T^{(2)}_{m_n,n}\le t)-\Phi(t)\big| \le O\Big(\max\{\frac{m_n}{n^2},\frac{1}{m_n},\frac{n}{m_n^2}\}\Big) \tag{12}$$

The following Corollary 2.3, a trivial consequence of Corollary 2.2 on taking $m_n=n$, is of particular interest as it asserts that the rate at which each of the error terms of the CLT's therein vanishes happens at the optimal $O(1/n)$ rate. This is a significant improvement over the classical Berry-Esséen $O(1/\sqrt{n})$ rate of error for $T_n(X)$ and $T_n(X-\mu)$ on only assuming the same $E|X|^3<+\infty$ moment condition for $G^{(1)}_{n,n}$ and $T^{(1)}_{n,n}$, and $E_XX^4<+\infty$ as well in the case of $G^{(2)}_{n,n}$ and $T^{(2)}_{n,n}$. Further moment conditions would not improve the $O(1/n)$ rates of convergence in hand, as below.

Corollary 2.3. *When $m_n = n$, as $n \to +\infty$, we have*

$$\sup_{-\infty<t<+\infty} \left|P_{X,w}(G^{(1)}_{n,n} \le t) - \Phi(t)\right| \le O(1/n), \tag{13}$$

$$\sup_{-\infty<t<+\infty} \left|P_{X,w}(T^{(1)}_{n,n} \le t) - \Phi(t)\right| \le O(1/n), \tag{14}$$

$$\sup_{-\infty<t<+\infty} \left|P_{X,w}(G^{(2)}_{n,n} \le t) - \Phi(t)\right| \le O(1/n), \tag{15}$$

$$\sup_{-\infty<t<+\infty} \left|P_{X,w}(T^{(2)}_{n,n} \le t) - \Phi(t)\right| \le O(1/n), \tag{16}$$

where (13) and (14) hold true when $E_X|X|^3 < +\infty$, and (15) and (16) hold true when $E_X X^4 < +\infty$.

Remark 2.3. The bootstrap is a computationally extensive method that is widely practiced to infer about parameters of the population. A nonparametric bootstrap method of constricting a confidence interval (C.I.) for μ is called the bootstrap t-percentile (cf. Efron and Tibshirani [13]). In this method the Student t-statistic $T_n(X-\mu)$ is used as a pivot for μ and its percentiles are estimated by repeatedly drawing bootstrap sub-samples from the original data set. It is generally known that the bootstrap t-percentile is of second order efficient, i.e., approximating the sampling distribution of the pivot $T_n(X-\mu)$ results in an error of order $1/n$. In view of Remark 2.1 and (13) and (15) of Corollary 2.3, the randomized pivots $G^{(1)}_{n,n}$ and $G^{(2)}_{n,n}$ possess the same asymptotic property as the bootstrap. However, the use of $G^{(i)}_{n,n}$, $i = 1, 2$, has a number of advantages over the nonparametric bootstrap.

Firstly, making asymptotic inference about μ based on $G^{(i)}_{n,n}$, $i = 1, 2$, by no means is a computationally demanding process. In fact, they can be computed directly by generating only one realization of the multinomial weights $w^{(n)}_i$, $1 \le i \le n$. A *much* shorter computation running time is a significant benefit of the use of these randomized pivots. Moreover, the bootstrap inference can be effected by the number of bootstrap sub-samples (cf. Yatracos [29]). Clearly, $G^{(i)}_{n,n}$, $i = 1, 2$, are immune to this type of error that is introduced in a bootstrap based inference when an insufficiently large number of bootstrap sub-samples are drawn.

Secondly, in addition to the assumption of the existence and finiteness of a sufficiently large number of moments for the data, the efficiency of the bootstrap in the literature is usually limited to non-lattice, mainly continuous, distributions (cf., for example, Singh [28] and Hall [17]). On the other hand $G^{(1)}_{n,n}$, for instance, is of correct second order for both continuous and discrete data on assuming only $E_X|X|^3 < +\infty$. Moreover, in both discrete and continuous cases, it does not result in excessively wide C.I.'s for μ, i.e., it delivers smooth inference even when the sample is of discrete nature. The same is true for $G^{(2)}_{n,n}$ when $E_X X^4 < +\infty$. Discrete data of course are very common in applications. The number of insurance claims in a period of time, the number of car accidents

or earthquakes during a certain period of time in a city are just a few examples of discrete data appearing in practice for which $G_{n,n}^{(i)}$, $i = 1, 2$, provide a more accurate inference about their means (cf. Section 4) and related percentiles (cf. Section 6) in a nonparametric framework.

3. Numerical studies

In this section we use the statistical software R to conduct our numerical studies for comparing the performance of $G_{n,n}^{(1)}$ as in (13) of Corollary 2.3 to that of its classical counterpart $T_n(X - \mu)$.

In order to provide initial motivation for the more in-depth numerical studies as in Table 2 below, that indicate a significantly better performance of the pivot $G_{n,n}^{(1)}$ for μ over its classical counterpart $T_n(X - \mu)$, we first compare the empirical probabilities of coverage of these pivots for μ in Table 1. The nominal probability coverage for the one sided C.I.'s in Table 1 is 95% in terms of the standard normal cutoff point 1.644854. The C.I.'s in Table 1 are based on 1000 replications of the data $(X_1, \ldots, X_n)$ for both pivots $G_{n,n}^{(1)}$ and $T_n(X - \mu)$, and 1000 replications of $(w_1^{(n)}, \ldots, w_n^{(n)})$, with $\sum_{i=1}^n w_i^{(n)} = n$, for computing $G_{n,n}^{(1)}$. The intervals are obtained by setting:

$$G_{n,n}^{(1)} \leq 1.644854 \qquad \text{and} \qquad T_n(X - \mu) \leq 1.644854.$$

The empirical probabilities of coverage for each one of these pivots are presented in Table 1 for the distributions therein.

Table 1 below shows that the sampling distribution of $G_{n,n}^{(1)}$ in each case, even for small sample sizes, is close enough to the standard normal distribution. Namely, using standard normal percentiles, $G_{n,n}^{(1)}$, as a pivot for the population mean μ, tends to yield probability coverages that are near to the nominal 95% even for sample sizes for which the classical CLT for $T_n(X - \mu)$ provides less valid C.I.'s for μ.

In order to study in-depth the refinement provided by $G_{n,n}^{(1)}$ over the classical $T_n(X - \mu)$ in view of Corollary 2.3, in the following Table 2 we present some numerical illustrations of the rates of convergence of *one sided* C.I.'s for the

TABLE 1
Comparing the empirical probability coverage of pivot $G_{n,n}^{(1)}$ to $T_n(X - \mu)$

Distribution of Sample	n	coverage of $G_{n,n}^{(1)}$	coverage of $T_n(X - \mu)$
Binomial$(10, 0.1)$	20	0.956	0.964
	30	0.953	0.960
Exponential(1)	20	0.959	0.975
	30	0.956	0.968
Normal$(0, 1)$	20	0.945	0.931
	30	0.951	0.946
Beta$(5, 1)$	20	0.914	0.903
	30	0.949	0.909
Binomial$(10, .9)$	20	0.922	0.904
	30	0.956	0.936

TABLE 2
Comparing the pivot $G^{(1)}_{n,n}$ to $T_n(X-\mu)$

Distribution of Sample	n	prop $G^{(1)}$	prop $T_n(X-\mu)$
Binomial(10, 0.1)	20	0.745	0.486
	30	0.764	0.546
	40	0.768	0.511
Poisson(1)	20	0.552	0.322
	30	0.554	0.376
	40	0.560	0.364
Lognormal(0, 1)	20	0.142	0.000
	30	0.168	0.000
	40	0.196	0.000
Exponential(1)	20	0.308	0.016
	30	0.338	0.020
	40	0.432	0.044
Normal(0, 1)	20	0.566	0.486
	30	0.600	0.568
	40	0.634	0.612
Beta(5, 1)	20	0.074	0.000
	30	0.136	0.016
	40	0.234	0.058

population mean μ based on the pivot $G^{(1)}_{n,n}$ whose validity and the rate at which they approach to their nominal probability coverage are concluded in (13) of our Corollary 2.3 for $G^{(1)}_{n,n}$.

To construct our *asymptotic* 95% C.I.'s based on $G^{(1)}_{n,n}$, in Table 2 we use the standard normal 95% cutoff point 1.644854. All the one sided C.I.'s in Table 2 are asymptotic, with both pivots in hand having standard normal limiting distribution as $n \to +\infty$.

Table 2 displays the proportion of 500 generated one sided C.I.'s with empirical coverage probability value falling in the interval $[0.94, 0.96]$ right around the standard normal 95% coverage that we use. Each one of these 500 C.I.'s is constructed by generating 500 sets of i.i.d. observations $(X_1, \ldots, X_n)$, with n as displayed, from the indicated respective underlying distributions. For simulating each value of $G^{(1)}_{n,n}$, we also generate 500 sets of the multinoimal weights $(w^{(n)}_1, \ldots, w^{(n)}_n)$, with $\sum_{1\le i\le n} w^{(n)}_i = n$ and associated probability vector $(1/n, \ldots, 1/n)$.

Both Tables 1 and 2 indicate a highly satisfactory performance of the pivot $G^{(1)}_{n,n}$ as compared to Student t-confidence intervals.

To exhibit the performance of the pivot $G^{(1)}_{n,n}$ in Table 2, in addition to normal, we also consider data from skewed distributions. It is known that the Student t-distribution converges to standard normal at a rate of order $O(1/n)$. The numerical results in Table 2 show that, based on normal data, $G^{(1)}_{n,n}$ performs as well as the t-statistic $T_n(X-\mu)$. The latter is an empirical indication that $G^{(1)}_{n,n}$ does indeed converge to standard normal at the rate of $O(1/n)$.

In Table 2, we denote the proportions of the C.I.'s with empirical probability coverage values between 94% and 96% associated with the pivots $G^{(1)}_{n,n}$ and $T_n(X-\mu)$, respectively by *prop* $G^{(1)}$ and *prop* $T_n(X-\mu)$.

In Table 2 the standard normal 95% cutoff point 1.644854 was used for both pivots $G^{(1)}_{n,n}$ and $T_n(X-\mu)$. Furthermore, in Table 2, Lognormal(0,1) stands for the Lognormal distribution with mean zero and variance one.

4. Randomized asymptotic pivots for the population mean μ

We are now to present $G^{(1)}_{m_n,n}$ of (4) and $G^{(2)}_{m_n,n}$ of (7) as *direct* asymptotic randomized pivots for the population mean $\mu = E_X X$, first when only $0 < \sigma^2 := E_X(X-\mu)^2 < +\infty$ is assumed, followed by assuming $E_X|X|^3 < +\infty$ as in Remark 4.1, and $E_X X^4 < +\infty$ as in Remark 4.2.

We note that for the coinciding numerator terms of $G^{(1)}_{m_n,n}$ and $G^{(2)}_{m_n,n}$ we have

$$E_{X|w}\big(\sum_{i=1}^{n}|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}|(X_i-\mu)\big)=0. \tag{17}$$

Furthermore, given $w_i^{(n)}$'s, for the randomized weighted average

$$\sum_{i=1}^{n}|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}|(X_i-\mu) =: \bar{X}_{m_n,n}(\mu), \tag{18}$$

mutatis mutandis in verifying (86) in Appendix A, we conclude that when the original sample size n *is fixed* and $m := m_n$, then, *as* $m \to +\infty$, we have

$$\bar{X}_{m_n,n}(\mu) = \bar{X}_{m,n}(\mu) \to 0 \textit{ in probability} - P_{X,w}, \tag{19}$$

and the same holds true if $n \to +\infty$ as well.

In view of (17)

$$\frac{\sum_{i=1}^{n}|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}|X_i}{\sum_{j=1}^{n}|\frac{w_j^{(n)}}{m_n}-\frac{1}{n}|} =: \hat{X}_{m_n,n}. \tag{20}$$

is an unbiased estimator for μ with respect to $P_{X|w}$.

It can be shown that when $E_X X^2 < +\infty$, as $n, m_n \to +\infty$ such that $m_n = o(n^2)$, $\hat{X}_{m_n,n}$ is a consistent estimator for the population mean μ in terms of $P_{X,w}$, i.e.,

$$\hat{X}_{m_n,n} \to \mu \textit{ in probability} - P_{X,w}. \tag{21}$$

In Appendix A we give a direct proof for (21) for the important case when $m_n = n$, for which the CLT's in Corollary 2.1 hold true at the $O(1/n)$ rate.

As to $G^{(1)}_{m_n,n}$ of (4), on replacing $(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})$ by $|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}|$ in the proof of (a) of Corollary 2.1 of Csörgő *et al.* [7] (cf. Appendix B), *as* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$, when $0 < \sigma^2 := E_X(X-\mu)^2 < 0$, we arrive at

$$P_{X|w}(G^{(1)}_{m_n,n} \le t) \to \Phi(t) \textit{ in probability} - P_w \textit{ for all } t \in \mathbb{R}, \tag{22}$$

and, via Lemma 1.2 in S. Csörgő and Rosalsky [9], we conclude also the unconditional CLT

$$P_{X,w}(G^{(1)}_{m_n,n} \le t) \to \Phi(t) \textit{ for all } t \in \mathbb{R}. \tag{23}$$

Remark 4.1. When $E_X|X|^3 < +\infty$ and $n, m_n \to +\infty$ so that $m_n = o(n^2)$, then, in addition to (22), we have (A) of Corollary 2.1 as well, and, in addition to (23), we also have (9) and (13) as in Corollaries 2.2 and 2.3 respectively.

When $E_X X^2 < +\infty$, in Appendix A we show that when n *is fixed* and $m := m_n \to +\infty$, the randomized sample variance $S^2_{m_n,n}$, as defined in (8), converges in probability-$P_{X,w}$ to the sample variance S^2_n, i.e., we have (cf. (87) in Appendix A or Remark 2.1 of Csörgő *et al.* [7])

$$S^2_{m,n} \to S^2_n \text{ in probability} - P_{X,w}. \tag{24}$$

For related results along these lines in terms of u- and v-statistics, we refer to Csörgő and Nasari [8], where, in a more general setup, we establish in probability and almost sure consistencies of randomized u- and v-statistics.

In Appendix A we also show that, when $E_X X^2 < +\infty$, *if* $n,\ m_n \to +\infty$ *so that* $n = o(m_n)$, then we have (cf. (87) in Appendix A)

$$\left(S^2_{m_n,n} - S^2_n\right) \to 0 \text{ in probability} - P_{X,w}. \tag{25}$$

When $E_X X^4 < +\infty$, the preceding convergence also holds true if, instead of $n = o(m_n)$, we assume $n = o(m_n^2)$ (cf. the proof of (C) and (D) of Corollary 2.1).

On combining (25) with the CLT in (23), when $E_X X^2 < +\infty$, *as* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$ *and* $n = o(m_n)$, the following unconditional CLT holds true as well in terms of $P_{X,w}$

$$G^{(2)}_{m_n,n} \xrightarrow{d} Z, \tag{26}$$

where, and also throughout, $\xrightarrow{d}$ stands for convergence in distribution, $G^{(2)}_{m_n,n}$ is as defined in (7), and Z stands for a standard normal random variable.

Remark 4.2. Assuming that $E_X X^4 < +\infty$ and $n, m_n \to +\infty$ so that $m_n = o(n^2)$ and $n = o(m_n^2)$, then we have (11) and (15) as in Corollaries 2.2 and 2.3 respectively, i.e., then the unconditional CLT

$$G^{(2)}_{m_n,n} \xrightarrow{d} Z \tag{27}$$

holds true in terms of $P_{X,w}$ at the therein indicated respective rates of convergence, and we have (C) of Corollary 2.1 as well, i.e.,

$$P_{X|w}(G^{(2)}_{m_n,n} \le t) \to \Phi(t) \text{ in probability} - P_w \text{ for all } t \in \mathbb{R} \tag{28}$$

at the therein indicated rate of convergence.

With $G^{(1)}_{m_n,n}$ and $G^{(2)}_{m_n,n}$ in mind as direct asymptotic pivots for μ, the CLT's as in (22) and (23), as well as their respective versions as spelled out in Remark 4.1, together with the CLT's as in (26), (27) and (28), can be used to construct exact size asymptotic C.I.'s for the population mean $\mu = E_X X$. Thus, in terms of $G^{(1)}_{m_n,n}$, *as* $n, m_n \to +\infty$ *and* $m_n = o(n^2)$, we conclude as follows, a $1-\alpha$ size asymptotic C.I. for the population mean $\mu = E_X X$, which is valid both in terms of the conditional $P_{X|w}$ and in unconditional $P_{X,w}$ distributions as in

(22) and (23) respectively, as well as with rates of convergence as in Remark 4.1:

$$\hat{X}_{m_n,n} - z_{\alpha/2}\frac{S_n D_n}{\sum_{j=1}^{n}|\frac{w_j^{(n)}}{m_n} - \frac{1}{n}|} \leq \mu \leq \hat{X}_{m_n,n} + z_{\alpha/2}\frac{S_n D_n}{\sum_{j=1}^{n}|\frac{w_j^{(n)}}{m_n} - \frac{1}{n}|} \tag{29}$$

where $z_{\alpha/2}$ satisfies $P(Z \geq z_{\alpha/2}) = \alpha/2$, $\hat{X}_{m_n,n}$ is as in (20) and $D_n := \sqrt{\sum_{j=1}^{n}(\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}$.

When $E_X X^4 < +\infty$, then we can replace S_n by $S_{m_n,n}$ in (29), and then the thus obtained $1-\alpha$ size asymptotic C.I. for the population mean μ holds true in terms of $G^{(2)}_{m_n,n}$ via both of the respective CLT's as in (27) and (28) with respective rates of convergence as indicated in Remark 4.2.

In view of Remark 4.1, on taking $m_n = n$, when $E_X|X|^3 < +\infty$, then both CLT's as in (22) and (23) hold true with a $O(1/n)$ rate of convergence (cf. Remark 2.1 and (9) of Corollary 2.3). Hence, the $1-\alpha$ size asymptotic C.I. for μ as in (29) is also achieved at that rate in both cases. The same conclusion remains true on replacing S_n by $S_{m_n,n}$ in (29) and taking $m_n = n$ when $E_X X^4 < +\infty$ (cf. Remarks 4.2 and 2.1, and (15) of Corollary 2.3).

5. Randomized asymptotic pivots for the sample and population means of big data sets

Big data sets problems represent a new era of having too many data in one sample that in some cases need to be stored on several machines and, on occasions, even on thousands of machines. In some cases dealing with samples of this volume directly is virtually impossible.

We use a method of sub-sampling from the original data of size n in such a way that only the picked elements of the original sample are to be used to infer about a parameter of interest of the population. This can be done by generating a realization of multinomial random variables $(w_1^{(n)}, \ldots, w_n^{(n)})$ with size $m_n = \sum_{i=1}^{n} w_i^{(n)}$, where $m_n << n$, independently from the data. The generated weights are to be put in a one-to-one correspondence with the indices of the members of the original sample. Then, *only* those data in the sample are to be observed whose corresponding weights are not zero.

We note that in the case when the sub-sample size m_n and the sample size n are so that $m_n/n \leq 0.05$, then, for each i, $1 \leq i \leq n$, $w_i^{(n)}$ is either zero or one almost all the time. In this case, practically, we are then sampling *without* replacement from the original massive data set of size n in the context of Remark 1.1.

A closer look at our approach of sub-sampling from a big data set reveals a close connection of this approach to the *super-population viewpoint* of *finite populations* as in Hartley and Sielken [18]. According to this viewpoint, a finite population of size n can be seen as a random sample drawn from an imaginary super-population. This is the sampling step that is viewed as an imaginary step

by Hartley and Sielken. To study the latter finite population as a sample, a smaller sub-sample of size m_n without replacement is then drawn from it. This viewpoint agrees with that of ours with the proviso that in our approach the super-population and the process of sampling from it can be either *imaginary* or *real*. While Hartley and Sielken [18] aimed at studying only the finite population, our approach permits studying not only the finite population itself but also its parent super-population via drawing a sub-sample as explained above. We only impose moment assumptions on the parent super-population.

The super-population viewpoint works perfectly for big data sets as well, for they can be viewed as finite populations and their uncomputable characteristics, such as their means, are to be estimated on their own. In this case the sample is so big that, to a large extent, it portrays the population, viewed now as a real super-population, from which it was drawn.

The numerical characteristics of a big data set, viewed as a finite population, should be fairly close to their super-population counterparts. For instance, the sample mean of a give data set $\{X_1, \ldots, X_n\}$ of large size n will be seen to deviate from the population mean only by a negligible error in the context of this section. The same will be seen to be true for the sample percentiles and their population counterparts in Section 6.

In this section, in view of the finite population viewpoint, we construct confidence sets for the sample mean, $\bar{X}_n$, of a large i.i.d. sample. In case of big data sets when the mean of the parent super-population is also to be captured, these confidence sets can in turn be used to serve as C.I.'s for the population mean μ, due to closeness of the two parameters $\bar{X}_n$ and μ (cf. (37) and (38)).

To begin with, we consider the associated numerator term of $T^{(i)}_{m_n,n}$, $i = 1, 2$, and write

$$\begin{aligned} \sum_{i=1}^{n} \Big(\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\Big) X_i &= \frac{1}{m_n} \sum_{i=1}^{n} w_i^{(n)} X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \\ &= \bar{X}_{m_n,n} - \bar{X}_n. \end{aligned} \tag{30}$$

We note that when the original sample size n *is* assumed to be *fixed*, then on taking *only one* large sub-sample of size $m := m_n$, via re-sampling the set of indices of the observations with replacement as in Remark 1.1, when $E_X X^2 < +\infty$, as $m \to +\infty$, we have

$$\bar{X}_{m,n} \to \bar{X}_n \; \textit{in probability} - P_{X,w} \tag{31}$$

(cf. (86) of Appendix A).

Further to (31), *as* $n,\ m_n \to +\infty$, then (cf. (86) in Appendix A)

$$\big(\bar{X}_{m_n,n} - \bar{X}_n\big) \to 0 \; \textit{in probability} - P_{X,w}. \tag{32}$$

As to $T^{(1)}_{m_n,n}$, and further to (32), we have that $E_{X|w}(\sum_{i=1}^{n}(\frac{w_i^{(n)}}{m_n} - \frac{1}{n})X_i) = 0$ and, *if* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$, then (cf. part (a) of Corollary 2.1 of Csörgő *et al.* [7] and Appendix B)

$$P_{X|w}(T^{(1)}_{m_n,n} \le t) \to P(Z \le t) \; \textit{in probability} - P_w \; \textit{for all } t \in \mathbb{R}. \tag{33}$$

Consequently, when $E_X X^2 < +\infty$, as $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$, we also arrive at

$$P_{X,w}(T^{(1)}_{m_n,n} \le t) \to P(Z \le t) \text{ for all } t \in \mathbb{R}, \tag{34}$$

an unconditional CLT.

Remark 5.1. When $E_X|X|^3 < +\infty$ and $n, m_n \to +\infty$ so that $m_n = o(n^2)$, then, in addition to (33), we have (B) of Corollary 2.1 as well, and in addition to (34), we also have (10) and (14) as in Corollaries 2.2 and 2.3 respectively.

Furthermore, in view of the CLT as in (34) and conclusion (25), *as* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$ *and* $n = o(m_n)$, *in terms of probability-*$P_{X,w}$ we conclude the unconditional CLT

$$T^{(2)}_{m_n,n} \xrightarrow{d} Z, \tag{35}$$

where $T^{(2)}_{m_n,n}$ is as defined in (6).

Remark 5.2. Assuming that $E_X X^4 < +\infty$ and $n, m_n \to +\infty$ so that $m_n = o(n^2)$ and $n = o(m_n^2)$, we then have (12) and (16) as in Corollaries 2.2 and 2.3 respectively, i.e., then the unconditional CLT as in (35), in terms of $P_{X,w}$, holds true at the therein indicated respective rates of convergence. Naturally, under the same conditions, as $n, m_n \to +\infty$, we have (D) of Corollary 2.1 as well, i.e.,

$$P_{X|w}(T^{(2)}_{m_n,n} \le t) \longrightarrow \Phi(t) \text{ in probability} - P_w \text{ for all } t \in \mathbb{R} \tag{36}$$

at the therein indicated rate of convergence.

Remark 5.3. Considering that our approach to randomizing the original sample in this section coincides with drawing a smaller sub-sample of size m_n with replacement from the original big data set $\{X_1, \ldots, X_n\}$ via re-sampling its index set $\{1, \ldots, n\}$ as in Remark 1.1, it is important to note that in order to compute both $\bar{X}_{m_n,n}$ and $S^2_{m_n,n}$, as in (5) and (8), respectively, only those X_i's are needed whose $w_i^{(n)} \neq 0$. This means that both $\bar{X}_{m_n,n}$ and $S^2_{m_n,n}$ are computable based only on the smaller sub-sample rather than the entire original big data set.

Under their respective conditions the CLT's as in (35) and (36) can be used to construct confidence sets for the sample mean $\bar{X}_n$ that is an unknown parameter in our present context.

We spell out the one based on $T^{(2)}_{m_n,n}$ as in (35) that is also valid in terms of (36), i.e., both in the context of Remark 5.2. Accordingly, *when* $E_X X^4 < +\infty$ *and* $m_n, n \to +\infty$ *so that* $m_n = o(n^2)$ and $n = o(m_n^2)$, then for any $\alpha \in (0,1)$, we conclude a $1-\alpha$ size asymptotic confidence set for $\bar{X}_n$, at the indicated rates of convergence, as follows

$$\bar{X}_{m_n,n} - z_{\alpha/2} S_{m_n,n} D_n \le \bar{X}_n \le \bar{X}_{m_n,n} + z_{\alpha/2} S_{m_n,n} D_n, \tag{37}$$

where $z_{\alpha/2}$ is as in (29), and $D_n := \sqrt{\sum_{j=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}$.

When $E_X|X| < +\infty$, as $n \to +\infty$, we have that $\bar{X}_n - \mu =: \varepsilon_n = o(1)$, almost surely in P_X-probability, as $n \to +\infty$. Since, the original sample size n of a big data set is already very large to begin with, ε_n is already negligible with high P_X-probability. Consequently, the confidence set (37) for $\bar{X}_n$ can actually be viewed as a $(1-\alpha)$ size asymptotic C.I. as well for the population mean μ, by simply rewriting it as follows

$$\bar{X}_{m_n,n} - z_{\alpha/2} S_{m_n,n} D_n \le \mu + \varepsilon_n \le \bar{X}_{m_n,n} + z_{\alpha/2} S_{m_n,n} D_n, \tag{38}$$

where $z_{\alpha/2}$ and D_n are as in (37).

We emphasize that (37) and (38) are identical statements under the conditions as spelled out right above (37). The asymptotic negligibility of the error sequence ε_n in (38) can, however, be studied on its own as $n \to +\infty$, freely from the identical conditions for (37) and (38) that $m_n = o(n^2)$ and $n = o(m_n^2)$, as $n, m_n \to +\infty$.

To further elaborate on the fact that (38) should work well as an asymptotic $(1-\alpha)$ size C.I. for the population mean μ in the case of a big data set, we make use of some well known classical results on the complete convergence of $\bar{X}_n$ to μ under two or more moment conditions for X.

We first mention the Erdős-Hsu-Robbins theorem (cf. [19, 14] and [15]) that concludes

$$\sum_{n=1}^{+\infty} P_X\left(|\bar{X}_n - \mu| > \epsilon\right) < +\infty, \ \textit{for every } \epsilon > 0,$$

if and only if $E_X X^2 < +\infty$. Thus, in addition to concluding that $\bar{X}_n - \mu = \varepsilon_n = o(1)$ almost surely-P_X, we also infer that, for any $\epsilon > 0$, $\{P_X(|\varepsilon_n| > \epsilon)\}_{n=1}^{+\infty}$ approaches zero at a rate faster than $O(1/n)$. In other words, as $n \to +\infty$, ε_n approaches zero in probability-P_X at a rate faster than the best possible rate of convergence for $T^{(2)}_{m_n,n}$ (cf. Corollary 2.3). Therefore, even when assuming only a two moment condition, (38) captures $\bar{X}_n$ and μ simultaneously with a high P_X-probability, that is, typically, $1 - 1/(n \log^2 n)$.

Further along these lines, we also mention the Baum and Katz theorem in [3] that asserts

$$\sum_{n=1}^{+\infty} n^{r/p-2} P_X\left(|\bar{X}_n - \mu| > \epsilon\, n^{1/p-1}\right) < +\infty,$$

for every $\epsilon > 0$ and some $p \in (0,2)$, if and only if $E_X|X|^r < +\infty$. Thus, when $E_X X^4 < +\infty$, then for a big sample of size $n = 10^6$, for example, with $p = 1$

$$P_X(|\varepsilon_n| \le \epsilon) \gtrapprox 1 - \frac{1}{10^{18}(\log 10^6)^2} \ \textit{for any } \epsilon > 0.$$

This shows that $\varepsilon_n = \bar{X}_n - \mu$ in (38) becomes arbitrarily small at a very fast rate in probability-P_X in terms of the original big sample size n, without paying attention to how n and m_n relate to each other when arriving at the asymptotic $(1-\alpha)$ size confidence set for covering $\bar{X}_n$ as in (37). Hence, the confidence set

(37) for the unknown sample mean $\bar{X}_n$ of a big data set of size n, viewed as in (38), is also seen to be an asymptotic $(1-\alpha)$ size C.I. for the unknown population mean μ under the same conditions that are used to arrive at having (37).

We now also illustrate how one goes about constructing the coinciding random boundaries in (37) and (38) in general, and then in case of having a big sample of size $n = 10^6$, as a convenient example.

First of all we emphasize that in the asymptotic confidence set (37) for $\bar{X}_n$ of a big data set, the bounds in hand, are computed by generating, independently from the entire data set, a realization of the random multinomial weights $(w_1^{(n)}, \ldots, w_n^{(n)})$ as in Remark 1.1. Thus, instead of trying to process the entire big data set $\{X_1, \ldots, X_n\}$ in order to compute $\bar{X}_n$, sampling it only via its index set $\{1, \ldots, n\}$ as above, we end up estimating $\bar{X}_n$ in terms of a confidence set as in (37) that can be based on significantly smaller sub-samples of size m_n of the entire big data set of size n, without having to deal with the latter directly, whenever $E_X X^4 < +\infty$ and $m_n = o(n^2)$ and $n = o(m_n^2)$ (cf. Remark 5.2). In this case the rate of convergence of the conditional CLT as in (36), as well as its unconditional CLT as in (35), is

$$O\big(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}, \frac{n}{m_n^2}\}\big) \tag{39}$$

in view of (D) of Corollary 2.1 and (12) of Corollary 2.2 respectively.

We note that, on account of having $n = o(m_n^2)$, as $m_n, n \to +\infty$ we cannot consider taking $m_n = n^{1/2}$ in the context of (39). We may however consider taking

$$m_n = n^{1/2} n^{\delta},\ 0 < \delta < 1/2, \tag{40}$$

and then the rate of convergence in (39) reduces to

$$O(n^{-2\delta}),\ 0 < \delta < 1/2. \tag{41}$$

For example, on taking $\delta = 1/4$, then $m_n = n^{3/4}$, and the rate of convergence for covering $\bar{X}_n$ as in (37) converts becomes $O(n^{-1/2})$, that coincides with that of the classical CLT for the Student t-statistic and pivot (cf. (1) and (2)). For instance, in this case, for a big sample of size $n = 10^6$, the CLT of (36) and its unconditional version for $T^{(2)}_{m_n,n}$ are both applied with a sub-sample of size $m_{10^6} = \sum_{i=1}^{10^6} w_i^{(10^6)} = (10^6)^{3/4} \approx 31{,}623$, where the random multinomially distributed weights $(w_1^{(10^6)}, \ldots, w_n^{(10^6)})$ are generated independently from the data $\{X_1, \ldots, X_{10^6}\}$ with respective probabilities $1/10^6$, i.e.,

$$(w_1^{(10^6)}, \ldots, w_{10^6}^{(10^6)}) \overset{d}{=} multinomial(31{,}623;\ \frac{1}{10^6}, \ldots, \frac{1}{10^6}). \tag{42}$$

These multinomial weights, in turn, are used to construct a $(1-\alpha)$ size confidence set *à la* (37), covering the unobserved mean $\bar{X}_{10^6}$, as well as the unknown population mean μ, with an error proportional to 0.001 (cf. (41) with $\delta = 1/4$).

More reduction of the sub-sample size m_n can, for example, be achieved by taking

$$m_n = n^{1/2} \log\log n \tag{43}$$

instead of that in (40) and, via (39), arriving at the rate of convergence

$$O(1/(\log\log n)^2) \tag{44}$$

for the CLT's in hand, instead of that in (41). For instance, if we again consider having a big sample of size $n = 10^6$, then (43) yields a sub-sample of size $m_n = 10^3 \log\log 10^6 \approx 2{,}626$, and constructing a $(1-\alpha)$ size confidence set *à la* (37), will cover the unobserved $\bar{X}_{10^6}$, as well as the unknown population mean μ, with an error proportional to $1/(\log\log 10^6)^2 \approx 1/7$. The latter increased error, as compared to the previous example with respective sub-sample size $m_{10^6} = 31{,}623$, is due to the much reduced sub-sample of size $m_{10^6} = 2{,}626$ in this context. This scenario can also be viewed in terms of using normal $z_{\alpha/2}$ percentiles for the Student t-pivot $T_n(X-\mu)$ when estimating the population mean μ on the basis of $n = 49$ i.i.d. observations with an error proportional to $1/\sqrt{49} = 1/7$.

6. Randomized CLT's and C.I.'s for the empirical and theoretical distributions with application to big data sets

In this section we put our randomization technique into use for estimating the percentiles of a population based on a given sample. When the sample size is relatively small or moderate, then our randomization technique as in Section 4 provides increased accuracy in making inference about the percentiles of the original population. When, dealing with super-populations with unobservable samples, such as big data sets or big finite populations, the results in this section rhyme with those discussed in Section 5.

Let $X, X_1, X_2, \ldots$ be independent real valued random variables with a common distribution function F as before, but now without assuming the existence of any finite moments for X. Let $\{X_1, \ldots, X_n\}$ be a random sample of size $n \geq 1$ on X and, for each n, define the empirical distribution function

$$F_n(x) := \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)/n, \ x \in \mathbb{R}, \tag{45}$$

and the sample variance of the indicator variables $\mathbb{1}(X_i \leq x)$

$$S_n^2(x) := \frac{1}{n}\sum_{i=1}^{n} \big(\mathbb{1}(X_i \leq x) - F_n(x)\big)^2 = F_n(x)(1-F_n(x)), \ x \in \mathbb{R}. \tag{46}$$

With $m_n = \sum_{i=1}^{n} w_i^{(n)}$ and the multinomial weights as in Remark 1.1,

$$\big(w_1^{(n)}, \ldots, w_n^{(n)}\big) \stackrel{d}{=} multinomial\big(m_n; \frac{1}{n}, \ldots, \frac{1}{n}\big),$$

that are independent from the random sample of n labeled units $\{X_1, \ldots, X_n\}$, define the randomized standardized empirical process

$$\begin{aligned}
\alpha^{(1)}_{m_n,n}(x) &:= \frac{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n}) \mathbb{1}(X_i \le x)}{\sqrt{F(x)(1-F(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}} \\
&= \frac{\sum_{i=1}^n \frac{w_i^{(n)}}{m_n} \mathbb{1}(X_i \le x) - F_n(x)}{\sqrt{F(x)(1-F(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2} } \\
&= \frac{F_{m_n,n}(x) - F_n(x)}{\sqrt{F(x)(1-F(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}}, \ x \in \mathbb{R}
\end{aligned} \tag{47}$$

where

$$F_{m_n,n}(x) := \sum_{i=1}^n \frac{w_i^{(n)}}{m_n} \mathbb{1}(X_i \le x), \ x \in \mathbb{R}, \tag{48}$$

is the randomized empirical distribution function.

We note that, point-wise in $x \in \mathbb{R}$,

$$E_{X|w}(F_{m_n,n}(x)) = F(x) = E_{X,w}(F_{m_n,n}(x)). \tag{49}$$

Define also the randomized sub-sample variance of the indicator random variables $\mathbb{1}(X_i \le x)$ by putting

$$\begin{aligned}
S^2_{m_n,n}(x) &:= \sum_{i=1}^n w_i^{(n)} \big(\mathbb{1}(X_i \le x) - F_{m_n,n}(x)\big)^2 / m_n \\
&= F_{m_n,n}(x)(1 - F_{m_n,n}(x)), \ x \in \mathbb{R}.
\end{aligned} \tag{50}$$

With n *fixed and* $m = m_n \to +\infty$, along the lines of (31) we arrive at

$$F_{m_n,n}(x) \longrightarrow F_n(x) \ \textit{in probability} - P_{X,w}, \ \textit{point} - \textit{wise in} \ x \in \mathbb{R}, \tag{51}$$

and, consequently, point-wise in $x \in \mathbb{R}$, *as* $m = m_n \to +\infty$,

$$S^2_{m_n,n}(x) \longrightarrow F_n(x)(1 - F_n(x)) = S^2_n(x) \ \textit{in probability} - P_{X,w}. \tag{52}$$

Furthermore, *à la* (32), *as* $n, m_n \to +\infty$, *point-wise in* $x \in \mathbb{R}$, we conclude

$$\big(F_{m_n,n}(x) - F_n(x)\big) \longrightarrow 0 \ \textit{in probability} - P_{X,w}, \tag{53}$$

that, in turn, *point-wise in* $x \in \mathbb{R}$, *as* $n, m_n \to +\infty$, implies

$$\big(S^2_{m_n,n} - S^2_n(x)\big) \longrightarrow 0 \ \textit{in probability} - P_{X,w}, \tag{54}$$

with $S^2_{m_n,n}$ and $S^2_n(x)$ respectively as in (50) and (46).

We wish to note and emphasize that, unlike in (25), for concluding (54), *we do not have to assume that* $n = o(m_n)$ as $n, m_n \to +\infty$.

Further to the randomized standardized empirical process $\alpha^{(1)}_{n,m_n}(x)$, we now define the following Studentized/self-normalized versions with $x \in \mathbb{R}$, as follows:

$$\hat{\alpha}^{(1)}_{m_n,n}(x) := \frac{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})\mathbb{1}(X_i \le x)}{\sqrt{F_n(x)(1-F_n(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}} \tag{55}$$

$$\hat{\hat{\alpha}}^{(1)}_{m_n,n}(x) := \frac{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})\mathbb{1}(X_i \le x)}{\sqrt{F_{m_n,n}(x)(1-F_{m_n,n}(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}} \tag{56}$$

$$\hat{\alpha}^{(2)}_{m_n,n}(x) := \frac{\sum_{i=1}^n \big|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big|\big(\mathbb{1}(X_i \le x) - F(x)\big)}{\sqrt{F_n(x)(1-F_n(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}} \tag{57}$$

$$\hat{\hat{\alpha}}^{(2)}_{m_n,n}(x) := \frac{\sum_{i=1}^n \big|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\big|\big(\mathbb{1}(X_i \le x) - F(x)\big)}{\sqrt{F_{m_n,n}(x)(1-F_{m_n,n}(x))}\sqrt{\sum_{j=1}^n (\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}}. \tag{58}$$

Clearly, on replacing X_i by $\mathbb{1}(X_i \le x)$ and μ by $F(x)$, $x \in \mathbb{R}$, in the formula in (18), we arrive at the respective statements of (17) and (19) in this context. Also, replacing X_i by $\mathbb{1}(X_i \le x)$ in the formula as in (20), we conclude the statement of (21) with μ replaced by $F(x)$, $x \in \mathbb{R}$.

As to the latter statement, on letting

$$\hat{F}_{m_n,n}(x) := \frac{\sum_{i=1}^n |\frac{w_i^{(n)}}{m_n} - \frac{1}{n}|\mathbb{1}(X_i \le x)}{\sum_{j=1}^n |\frac{w_j^{(n)}}{m_n} - \frac{1}{n}|}, \tag{59}$$

as $n, m_n \to +\infty$, such that $m_n = o(n^2)$, point-wise in $x \in \mathbb{R}$, by virtue of (21),

$$\hat{F}_{m_n,n}(x) \longrightarrow F(x) \textit{ in probability} - P_{X,w}. \tag{60}$$

In Lemma 5.2 of Csörgő *et al.* [7] it is shown that, *if* $m_n, n \to +\infty$ *so that* $m_n = o(n^2)$, *then*

$$M_n := \frac{\max_{1\le i\le n}\big(\frac{w_i^{(n)}}{n} - \frac{1}{n}\big)^2}{\sum_{j=1}^n \big(\frac{w_j^{(n)}}{n} - \frac{1}{n}\big)^2} \to 0 \textit{ in probability} - P_w. \tag{61}$$

This, mutatis mutandis, combined with (a) of Corollary 2.1 of Csörgő *et al.* [7], *as* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$, yields

$$P_{X|w}\big(\hat{\alpha}^{(s)}_{m_n,n}(x) \le t\big) \to P(Z \le t) \textit{ in probability} - P_w, \textit{ for all } x, t \in \mathbb{R}, \tag{62}$$

with $s = 1$ and also for $s = 2$, and via Lemma 1.2 in S. Csörgő and Rosalsky [9], this results in having also the unconditional CLT

$$P_{X,w}\big(\hat{\alpha}^{(s)}_{m_n,n}(x) \le t\big) \to P(Z \le t) \text{ for all } x, t \in \mathbb{R}, \tag{63}$$

with $s = 1$ and also for $s = 2$.

On combining (63) and (54), *as* $n, m_n \to +\infty$ *so that* $m_n = o(n^2)$, when $s = 1$ in (63), we conclude

$$\hat{\hat{\alpha}}^{(1)}_{m_n,n}(x) \xrightarrow{d} Z \tag{64}$$

and, when $s = 2$ in (63), we arrive at

$$\hat{\hat{\alpha}}^{(2)}_{m_n,n}(x) \xrightarrow{d} Z \tag{65}$$

for all $x \in \mathbb{R}$.

Remark 6.1. The Berry-Esséen type inequality (A) of our Theorem 2.1 continues to hold true for $\hat{\alpha}^{(2)}_{m_n,n}(x)$, and so does also (B) of Theorem 2.1 for $\hat{\alpha}^{(1)}_{m_n,n}(x)$, without the assumption $E_X|X|^3 < +\infty$, for the indicator random variable $\mathbb{1}(X \le x)$ requires no moments assumptions.

Remark 6.2. In view of Remark 6.1, in the context of this section, (A) and (B) of Corollary 2.1 read as follows: As $n, m_n \to +\infty$ in such a way that $m_n = o(n^2)$, then, mutatis mutandis, (A) and (B) hold true for $\hat{\alpha}^{(1)}_{m_n,n}(x)$ and $\hat{\alpha}^{(2)}_{m_n,n}(x)$, with $O(\max\{m_n/n^2, 1/m_n\})$ in both. Consequently, statements (9) and (10) of Corollary 2.2 also read similarly for $\hat{\alpha}^{(1)}_{m_n,n}$ and $\hat{\alpha}^{(2)}_{m_n,n}(x)$ in terms of the conditions and the rates of convergence. Thus, on taking $m_n = n$, we immediately obtain the optimal $O(n^{-1})$ rate conclusion of Remark 2.1 in this context as well, i.e., uniformly in $t \in \mathbb{R}$ and point-wise in $x \in \mathbb{R}$ for $\hat{\alpha}^{(1)}_{m_n,n}(x)$ and $\hat{\alpha}^{(2)}_{m_n,n}(x)$.

Remark 6.3. As to the rate of convergence of the respective CLT's in terms of $P_{X,w}$ as in (64) and (65), and also in terms of $P_{X|w}$, via (C) and (D) of Corollary 2.1, for $\hat{\hat{\alpha}}^{(1)}_{m_n,n}(x)$ and $\hat{\hat{\alpha}}^{(2)}_{m_n,n}(x)$, as $n, m_n \to +\infty$ in such away that $m_n = O(n^2)$, we obtain the rate $O(\max\{m_n/n^2, 1/m_n\})$. Thus, on taking $m_n = n$, we conclude the optimal rate of convergence $O(n^{-1})$ for $\hat{\hat{\alpha}}^{(1)}_{m_n,n}(x)$ and $\hat{\hat{\alpha}}^{(2)}_{m_n,n}(x)$, uniformly in $t \in \mathbb{R}$ and point-wise in $x \in \mathbb{R}$.

The CLT's for $\hat{\alpha}^{(1)}_{m_n,n}$ and $\hat{\hat{\alpha}}^{(1)}_{m_n,n}$ can be used to construct point-wise confidence sets for the empirical distribution function $F_n(.)$, while those for $\hat{\alpha}^{(2)}_{m_n,n}$ and $\hat{\hat{\alpha}}^{(2)}_{m_n,n}$ provide point-wise C.I.'s for the distribution function $F(.)$. We spell out the ones, respectively based on $\hat{\hat{\alpha}}^{(1)}_{m_n,n}$ and $\hat{\hat{\alpha}}^{(2)}_{m_n,n}$, that are valid both in terms of $P_{X|w}$ and $P_{X,w}$ with the rate of convergence $O(\max\{m_n/n^2, 1/m_n\})$ (cf. Remark 6.3). Thus, as $n, m_n \to +\infty$ so that $m_n = o(n^2)$, the CLT's in hand respectively result in the following asymptotically exact $(1-\alpha)$ size C.I.'s, for any $\alpha \in (0,1)$ and point-wise in $x \in \mathbb{R}$:

$$F_{m_n,n}(x) - z_{\alpha/2} S_{m_n,n}(x) D_n \le F_n(x) \le F_{m_n,n}(x) + z_{\alpha/2} S_{m_n,n}(x) D_n \tag{66}$$

$$\hat{F}_{m_n,n}(x) - z_{\alpha/2}\frac{S_{m_n,n}(x)D_n}{\sum_{j=1}^{n}|\frac{w_j^{(n)}}{m_n} - \frac{1}{n}|} \leq F(x) \leq \hat{F}_{m_n,n}(x) + z_{\alpha/2}\frac{S_{m_n,n}(x)D_n}{\sum_{j=1}^{n}|\frac{w_j^{(n)}}{m_n} - \frac{1}{n}|} \tag{67}$$

with $z_{\alpha/2}$ as in (29), $D_n := \sqrt{\sum_{j=1}^{n}(\frac{w_j^{(n)}}{m_n} - \frac{1}{n})^2}$, $S_{m_n,n}(x) = F_{m_n,n}(x)(1 - F_{m_n,n}(x))$ as in (50), $\hat{F}_{m_n,n}(x)$ as in (48), and $F_{m_n,n}(x)$ as in (59).

On taking $m_n = n$, then, for each $x \in \mathbb{R}$, both of the preceding C.I.'s achieve their nominal level at the optimal rate of $O(n^{-1})$. This is a significant achievement in capturing the population distribution by (67), for each $x \in \mathbb{R}$, when the available sample is of moderate size or small.

In case of having a big data set of size n, when processing the entire data set may not be possible, then both $F_n(.)$ and $F(.)$ are to be estimated. In this case the confidence set (66) can serve not only for covering $F_n(x)$, but $F(x)$ as well with any desirable accuracy for each $x \in \mathbb{R}$. Namely, on putting $\varepsilon_n(x) = F_n(x) - F(x)$, $x \in \mathbb{R}$, we simply re-write it as follows

$$F_{m_n,n}(x) - z_{\alpha/2}S_{m_n,n}(x)D_n \leq F(x) + \varepsilon_n(x) \leq F_{m_n,n}(x) + z_{\alpha/2}S_{m_n,n}(x)D_n \tag{68}$$

and argue via the Glivenko-Cantelli theorem that in case of big data sets ε_n is negligible with any desired accuracy for each $x \in \mathbb{R}$ at a fast enough rate of convergence as $n \to +\infty$, without paying attention to how m_n and n relate to each other when arriving at the asymptotic $(1 - \alpha)$ size confidence set that covers $F_n(x)$ for each $x \in \mathbb{R}$ as in (66). This, in turn, is guaranteed by the Dvoretzky-Kiefer-Wolfowitz [11] inequality that asserts for all $\epsilon > 0$

$$P_X(\sup_{-\infty<x<+\infty}|\varepsilon_n(x)| > \epsilon) \leq 2\exp(-2n\epsilon^2). \tag{69}$$

On summing in (69), one concludes the Glivenko-Cantelli theorem at the indicated exponentially fast rate of convergence to zero in P_X-probability that of course also holds true point-wise in $x \in \mathbb{R}$ for $\varepsilon_n(x)$ as in (68). Thus, the error induced when estimating $F(x)$, point-wise in $x \in \mathbb{R}$, as in (68) is practically zero for data sets of big size n.

For example, in view of inequality (69), where the best possible constant 2 in front of the exponential function is due to Massart [23], when a large sample of size $n = 10^6$ is at hand, then we have

$$P_X(\sup_{-\infty<x<+\infty}|\varepsilon_n(x)| > \epsilon) \leq 2\exp(-2\epsilon^2(10^6)) \tag{70}$$

for all $\epsilon > 0$. Thus, practically, the confidence set (66) for $F_n(x)$ is also a C.I. for $F(x)$ in the case of big data sets of size n.

Another spectacular illustration of the negligibility of $\varepsilon_n(x)$ in (68) is provided by taking $\epsilon = (\log n/n)^{1/2}$ in (69).

Recall now that as $n, m_n \to +\infty$ in such a way that $m_n = o(n^2)$, then the rate of convergence for having the $(1 - \alpha)$ size confidence set (66) for $F_n(x)$,

and also for $F(x)$, in view of (68), for $x \in \mathbb{R}$, is $O(m_n/n^2, 1/m_n)$. Consequently, when drawing a significantly smaller sub-sample of size $m_n = n^{1/2}$, for example, the rate of convergence becomes $O(n^{-1/2})$ that coincides with the rate of convergence of the classical CLT for the Student t-statistic and pivot, based on n observations as in (1) and (2) respectively. Needless to say that in case of a big data set, a sub-sample of size $m_n = n^{1/2}$ can be a huge reduction in the number of observations that we are to deal with instead of the original sample that, in our approach, results in the same magnitude of error as that of the classical CLT when the entire sample of size n is to be observed.

To illustrate the reduction provided by our confidence set (66) when it used to cover $F_n(x)$ or $F(x)$, point-wise in $x \in \mathbb{R}$, we consider a big data set of size $n = 10^6$. By generating the random weights $(w_1^{(10^6)}, \ldots, w_{10^6}^{(10^6)})$, with $m_{10^6} = \sum_{i=1}^{10^6} w_i^{(10^6)} = \sqrt{10^6} = 1000$, independently from the original sample (cf. Remark 1.1), our confidence set (66) to capture $F_n(x)$ is achieved with an error proportional to $1/1000$. Recalling also that in this case $\varepsilon_n = F_n(x) - F(x)$ is negligible already (cf. (70)), we also conclude that (66) captures $F(x)$ with an error proportional to $1/1000$.

7. Proofs

Proof of Theorem 2.1

Due to similarity of the two cases we only give the proof of part (A) of this theorem. The proof relies on the fact that, via conditioning on the weights $w_i^{(n)}$'s, $\sum_{i=1}^n |\frac{w_i^{(n)}}{m_n} - \frac{1}{n}|(X_i - \mu)$ is a sum of independent and non-identically distributed random variables. This in turn enables us to use a Berry-Esséen type inequality for self-normalized sums of independent and non-identically distributed random variables. Also, some of the ideas in the proof are similar to those of Slutsky's theorem.

We now write

$$\begin{aligned} G_{m_n,n}^{(1)} &= \frac{\sum_{i=1}^n \left|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\right|(X_i - \mu)}{\sigma\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}} + \frac{\sum_{i=1}^n \left|\frac{w_i^{(n)}}{m_n} - \frac{1}{n}\right|(X_i - \mu)}{\sigma\sqrt{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2}} \big(\frac{\sigma}{S_n} - 1\big) \\ &=: Z_{m_n} + Y_{m_n}. \end{aligned} \tag{71}$$

In view of the above setup, for $t \in \mathbb{R}$ and $\varepsilon_1 > 0$, we have

$$\begin{aligned} -P_{X|w}(|Y_{m_n}| > \varepsilon) \quad &+ \quad P_{X|w}(Z_{m_n} \le t - \varepsilon) \\ &\le \quad P_{X|w}(G_{m_n,n}^{(1)} \le t) \\ &\le \quad P_{X|w}(Z_{m_n} \le t + \varepsilon) + P_{X|w}(|Y_{m_n}| > \varepsilon). \end{aligned} \tag{72}$$

Observe now that for $\varepsilon_1 > 0$ we have

$$P_{X|w}(|Y_{m_n}| > \varepsilon) \le P_{X|w}\big(|Z_{m_n}| > \frac{\varepsilon}{\varepsilon_1}\big) + P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big). \tag{73}$$

One can readily see that

$$\begin{aligned} P_{X|w}\big(|Z_{m_n}| > \frac{\varepsilon}{\varepsilon_1}\big) &\leq (\frac{\varepsilon_2}{\varepsilon_1})^2 \frac{\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2 E_X(X_1-\mu)^2}{\sigma^2 \sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2} \\ &= (\frac{\varepsilon_1}{\varepsilon})^2. \end{aligned}$$

Combining now the preceding conclusion with (73), (72) can be replaced by

$$\begin{aligned} & -(\frac{\varepsilon_1}{\varepsilon})^2 - P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big) + P_{X|w}(Z_{m_n} \leq t - \varepsilon) \\ \leq\ & P_{X|w}(G^{(1)}_{m_n,n} \leq t) \\ \leq\ & (\frac{\varepsilon_1}{\varepsilon})^2 + P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big) + P_{X|w}(Z_{m_n} \leq t + \varepsilon). \end{aligned} \tag{74}$$

Now, the continuity of the normal distribution Φ allows us to choose $\varepsilon_2 > 0$ so that $\Phi(t+\varepsilon) - \Phi(t) < \varepsilon_2$ and $\Phi(t-\varepsilon) - \Phi(t) > -\varepsilon_2$. This combined with (74) yields

$$\begin{aligned} & -(\frac{\varepsilon_1}{\varepsilon})^2 - P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big) + P_{X|w}(Z_{m_n} \leq t - \varepsilon) - \Phi(t-\varepsilon) - \varepsilon_2 \\ \leq\ & P_{X|w}(G^{(1)}_{m_n,n} \leq t) - \Phi(t) \\ \leq\ & (\frac{\varepsilon_1}{\varepsilon})^2 + P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big) + P_{X|w}(Z_{m_n} \leq t + \varepsilon) - \Phi(t+\varepsilon) + \varepsilon_2. \end{aligned} \tag{75}$$

We now use the Berry-Esséen inequality for independent and not necessarily identically distributed random variables (cf., e.g., Serfling [25]) to write

$$P_{X|w}(Z_{m_n} \leq t + \varepsilon_1) - \Phi(t+\varepsilon_1) \leq (\frac{CE_X|X-\mu|^3}{\sigma^{3/2}}) \frac{\sum_{i=1}^n |\frac{w_i^{(n)}}{m_n} - \frac{1}{n}|^3}{\big(\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2\big)^{3/2}}$$

and

$$P_{X|w}(Z_{m_n} \leq t - \varepsilon_1) - \Phi(t-\varepsilon_1) \geq (\frac{-CE_X|X-\mu|^3}{\sigma^{3/2}}) \frac{\sum_{i=1}^n |\frac{w_i^{(n)}}{m_n} - \frac{1}{n}|^3}{\big(\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2\big)^{3/2}},$$

where C is a universal constant as in the Berry-Esséen inequality in this context (cf. page 33 of Serfling [25]).

Incorporating these approximations into (75) we arrive at

$$\begin{aligned} & \sup_{-\infty<t<+\infty} \big|P_{X|w}(G^{(1)}_{m_n,n} \leq t) - \Phi(t)\big| \\ & \leq (\frac{\varepsilon_1}{\varepsilon})^2 + P_X\big(|S_n^2 - \sigma^2| > \varepsilon_1^2\big) + (\frac{CE_X|X-\mu|^3}{\sigma^{3/2}}) \frac{\sum_{i=1}^n |\frac{w_i^{(n)}}{m_n} - \frac{1}{n}|^3}{\big(\sum_{i=1}^n (\frac{w_i^{(n)}}{m_n} - \frac{1}{n})^2\big)^{3/2}} + \varepsilon_2. \end{aligned}$$

From the preceding relation we conclude that

$$P_w\Big(\sup_{-\infty<t<+\infty}\big|P_{X|w}(G^{(1)}_{m_n,n}\le t)-\Phi(t)\big|>\delta\Big)\le P_w\Big(\frac{\sum_{i=1}^n\big|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}\big|^3}{\big(\sum_{i=1}^n(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})^2\big)^{3/2}}>\delta_n\Big) \tag{76}$$

with δ_n as defined in the statement of Theorem 2.1.

For $\varepsilon>0$, the right hand side of (76) is bounded above by

$$\begin{aligned}&P_w\Big\{\sum_{i=1}^n\big|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}\big|^3>\frac{\delta_n(1-\varepsilon)^{\frac{3}{2}}(1-\frac{1}{n})^{\frac{3}{2}}}{m_n^{\frac{3}{2}}}\Big\}\\&+P_w\Big(\Big|\frac{m_n}{1-\frac{1}{n}}\sum_{i=1}^n(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})^2-1\Big|>\varepsilon\Big)\\&=:\Pi_1(n)+\Pi_2(n).\end{aligned}$$

We bound $\Pi_1(n)$ above by

$$\begin{aligned}&\delta_n^{-2}(1-\varepsilon)^{-3}(1-\frac{1}{n})^{-3}m_n^{-3}(n+n^2)E_w(w_1^{(n)}-\frac{m_n}{n})^6\\&=\delta_n^{-2}(1-\varepsilon)^{-3}(1-\frac{1}{n})^{-3}m_n^{-3}(n+n^2)\{\frac{15m_n^3}{n^3}+\frac{25m_n^2}{n^2}+\frac{m_n}{n}\}.\end{aligned} \tag{77}$$

As for $\Pi_2(n)$, recalling that $E_w(\sum_{i=1}^n(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})^2)=\frac{(1-\frac{1}{n})}{m_n}$, an application of Chebyshev's inequality yields

$$\begin{aligned}\Pi_2(n)&\le\frac{m_n^2}{\varepsilon^2(1-\frac{1}{n})^2}E_w\Big(\sum_{i=1}^n(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})^2-\frac{(1-\frac{1}{n})}{m_n}\Big)^2\\&=\frac{m_n^2}{\varepsilon^2(1-\frac{1}{n})^2}\,E_w\Big\{\Big(\sum_{i=1}^n(\frac{w_i^{(n)}}{m_n}-\frac{1}{n})^2\Big)^2-\frac{(1-\frac{1}{n})^2}{m_n^2}\Big\}^2\\&=\frac{m_n^2}{\varepsilon^2(1-\frac{1}{n})^2}\Big\{nE_w\Big(\frac{w_1^{(n)}}{m_n}-\frac{1}{n}\Big)^4\\&\quad+n(n-1)E_w\Big[\Big(\frac{w_1^{(n)}}{m_n}-\frac{1}{n}\Big)^2\Big(\frac{w_2^{(n)}}{m_n}-\frac{1}{n}\Big)^2\Big]-\frac{(1-\frac{1}{n})^2}{m_n^2}\Big\}.\end{aligned} \tag{78}$$

We now use the fact that $w^{(n)}$'s are multinomially distributed to compute the preceding relation. After some algebra it turns out that it can be bounded above by

$$\begin{aligned}&\frac{m_n^2}{\varepsilon^2(1-\frac{1}{n})^2}\Big\{\frac{1-\frac{1}{n}}{n^3m_n^3}+\frac{(1-\frac{1}{n})^4}{m_n^3}+\frac{(m_n-1)(1-\frac{1}{n})^2}{nm_n^3}+\frac{4(n-1)}{n^3m_n}+\frac{1}{m_n^2}\\&\qquad-\frac{1}{nm_n^2}+\frac{n-1}{n^3m_n^3}+\frac{4(n-1)}{n^2m_n^3}-\frac{(1-\frac{1}{n})^2}{m_n^2}\Big\}.\end{aligned} \tag{79}$$

Incorporating (77) and (79) into (76) completes the proof of part (A) of Theorem 2.1. □

Proof of Corollary 2.1

The proofs of parts (A) and (B) of this corollary are immediate consequences of Theorem 2.1.

To prove parts (C) and (D) of this corollary, in view of Theorem 2.1 it suffices to show that, for arbitrary $\varepsilon_1, \varepsilon_2 > 0$, as $n, m_n \to +\infty$,

$$P_w\big(P_{X|w}(|S_{m_n,n} - S_n^2| > \varepsilon_1) > \varepsilon_2\big) = O(\frac{n}{m_n^2}). \tag{80}$$

To prove the preceding result we first note that

$$\begin{aligned} S_{m_n,n}^2 - S_n^2 &= \sum_{1\le i\ne j\le n} \big(\frac{w_i^{(n)} w_j^{(n)}}{m_n(m_n-1)} - \frac{1}{n(n-1)}\big)\frac{(X_i - X_j)^2}{2} \\ &= \sum_{1\le i\ne j\le n} \big(\frac{w_i^{(n)} w_j^{(n)}}{m_n(m_n-1)} - \frac{1}{n(n-1)}\big)\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big). \end{aligned}$$

By virtue of the preceding observation, we proceed with the proof of (80) by first letting $d_{i,j}^{(n)} := \frac{w_i^{(n)} w_j^{(n)}}{m_n(m_n-1)} - \frac{1}{n(n-1)}$ and writing

$$\begin{aligned} &P_w\big\{P_{X|w}\big(\big|\sum_{1\le i\ne j\le n} d_{i,j}^{(n)}\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big| > \varepsilon_1\big)\varepsilon_2\big\} \\ &\le P_w\big\{E_{X|w}\big(\sum_{1\le i\ne j\le n} d_{i,j}^{(n)}\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big)^2 > \varepsilon_1^2\varepsilon_2\big\}. \end{aligned} \tag{81}$$

Observe now that

$$\begin{aligned} &E_{X|w}\big(\sum_{1\le i\ne j\le n} d_{i,j}^{(n)}\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big)^2 \\ &= E_X\big(\frac{(X_1 - X_2)^2}{2} - \sigma^2\big)^2 \sum_{1\le i\ne j\le n} (d_{i,j}^{(n)})^2 \\ &+ \sum_{\substack{1\le i,j,k\le n \\ i,j,k \text{ are distinct}}} d_{i,j}^{(n)} d_{i,k}^{(n)} E_X\big(\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big(\frac{(X_i - X_k)^2}{2} - \sigma^2\big)\big) \\ &+ \sum_{\substack{1\le i,j,k,l\le n \\ i,j,k,l \text{ are distinct}}} d_{i,j}^{(n)} d_{k,l}^{(n)} E_X\big(\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big(\frac{(X_k - X_l)^2}{2} - \sigma^2\big)\big). \end{aligned} \tag{82}$$

We note that in the preceding relation, since i, j, k are distinct, we have that

$$\begin{aligned} &E_X\big(\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big)\big(\frac{(X_i - X_k)^2}{2} - \sigma^2\big)\big) \\ = \;& E\big\{E\big(\frac{(X_i - X_j)^2}{2} - \sigma^2 | X_i\big)E\big(\frac{(X_i - X_k)^2}{2} - \sigma^2 | X_i\big)\big\} = \frac{E_X(X_1^2 - \sigma^2)}{4}. \end{aligned}$$

Also, since i, j, k, l are distinct, we have that

$$E_X\big((\frac{(X_i - X_j)^2}{2} - \sigma^2)(\frac{(X_k - X_l)^2}{2} - \sigma^2)\big) = E_X^2\big(\frac{(X_i - X_j)^2}{2} - \sigma^2\big) = 0.$$

Therefore, in view of (82) and (81), the proof of (80) follows if we show that

$$\sum_{1\le i\ne j\le n} (d_{i,j}^{(n)})^2 = O_{P_w}(\frac{1}{m_n^2}) \tag{83}$$

and

$$\sum_{\substack{1\le i,j,k\le n\\ i,j,k \text{ are distinct}}} d_{i,j}^{(n)} d_{i,k}^{(n)} = O_{P_w}(\frac{n}{m_n^2}). \tag{84}$$

Noting that, as $n, m_n \to +\infty$,

$$E_w\big\{\sum_{1\le i\ne j\le n} (d_{i,j}^{(n)})^2\big\} \sim \frac{1}{m_n^2}$$

and

$$E_w\big|\sum_{\substack{1\le i,j,k\le n\\ i,j,k \text{ are distinct}}} d_{i,j}^{(n)} d_{i,k}^{(n)}\big| \le n^3 E_w(d_{1,2}^{(n)})^2 \sim \frac{n}{m_n^2}.$$

The preceding two conclusions imply (83) and (84), respectively. Now the proof of Corollary 2.1 is complete. □

Proof of Corollary 2.2

The proof of this result is relatively easy. Due to their similarity, we only give the proof for (9) of Corollary 2.2 as follows. With arbitrary positive δ, as $n, m_n \to +\infty$ so that $m_n = o(n^2)$, in view of (A) of Corollary 2.1, we have

$$\begin{aligned}\sup_{-\infty<t<+\infty} \big|P_{X,w}\big(G_{m_n,n}^{(1)} \le t\big) - \Phi(t)\big| &\le \delta + 2P\big(\big|P_{X|w}(G_{m_n,n}^{(1)} \le t) - \Phi(t)\big| > \delta\big)\\ &= \delta + O\Big(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}\}\Big).\end{aligned} \tag{85}$$

Now for any given arbitrary small $\epsilon > 0$, take the arbitrary $\delta > 0$, in (85) so that $\epsilon - \delta > 0$. Thus, there exists an $n_0 = n(\epsilon, \delta)$ and $m_{n_0} = m_n(\epsilon, \delta)$ so that for all $n \ge n_0$ and $m_n \ge m_{n_0}$, with $m_n = o(n^2)$ we have $O(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}\}) < \epsilon$, i.e., the indicated upper bound in (85) becomes arbitrary small at the rate $O(\max\{\frac{m_n}{n^2}, \frac{1}{m_n}\})$ as $n, m_n \to +\infty$ so that $m_n = o(n^2)$. □

Appendix A

Consider the original sample $\{X_1, \ldots, X_n\}$ and assume that the sample size $n \ge 1$ is fixed. We are now to show that when n is fixed, as $m \to +\infty$, we have

$\bar{X}_{m,n} \to \bar{X}_n$ in probability $P_{X,w}$ as in (32). To do so, without loss of generality we assume that $\mu = 0$. Let $\varepsilon_1, \varepsilon_2 > 0$, and write

$$\begin{aligned} P_w\big\{P_{X|w}\big(\big|\bar{X}_{m,n} - \bar{X}_n\big| > \varepsilon_1\big) > \varepsilon_2\big\} &\le P_w\big\{E_{X|w}\big(\sum_{i=1}^{n}\big(\frac{w_i^{(n)}}{m} - \frac{1}{n}\big)X_i\big)^2 > \varepsilon_1^2\varepsilon_2\big\} \\ &= P_w\big\{\sum_{i=1}^{n}\big(\frac{w_i^{(n)}}{m} - \frac{1}{n}\big)^2 > \sigma^{-2}\varepsilon_1^2\varepsilon_2\big\} \\ &\le \sigma^2\varepsilon_1^{-2}\varepsilon_2^{-1}\ nE_w\big(\frac{w_1^{(n)}}{m} - \frac{1}{n}\big)^2 \\ &\le \sigma^{-2}\varepsilon_1^{-2}\varepsilon_2^{-1}\frac{(1-\frac{1}{n})}{m} \to 0,\ as\ m \to \infty. \end{aligned} \tag{86}$$

The preceding conclusion means that $P_{X|w}(|\bar{X}_{m,n} - \bar{X}_n| > \varepsilon_1) \to 0$ in probability-P_w. Hence, by the dominated convergence theorem, we conclude that $\hat{X}_m \to \bar{X}_n$ in probability $P_{X,w}$. □

We are now to show that the randomized sample variance $S^2_{m_n,n}$ is an in probability consistent estimator of the ordinary sample variance S_n^2 for each fixed n, when $m \to +\infty$. Employing now the u-statistic representation of the sample variance enables us to rewrite $S^2_{m_n,n}$, as in (8), as follows

$$S^2_{m_n,n} = \frac{\sum_{1\le i\le j\le n} w_i^{(n)} w_j^{(n)}(X_i - X_j)^2}{2m(m-1)}.$$

In view of the preceding formula, we have

$$S^2_{m_n,n} - S_n^2 = \sum_{1\le i\ne j\le n}\big(\frac{w_i^{(n)}w_j^{(n)}}{2m(m-1)} - \frac{1}{2n(n-1)}\big)(X_i - X_j)^2.$$

Now, for $\varepsilon_1, \varepsilon_2 > 0$, we write

$$\begin{aligned} & P_w\Big(P_{X|w}\big(\Big|S^2_{m_n,n} - S_n^2\Big| > \varepsilon_1\big) > \varepsilon_2\Big) \\ = \ & P_w\big(P_{X|w}\big(\big|\sum_{1\le i\ne j\le n}\big(\frac{w_i^{(n)}w_j^{(n)}}{m(m-1)} - \frac{1}{n(n-1)}\big)(X_i - X_j)^2\big| > 2\varepsilon_1\big) > \varepsilon_2\big) \\ \le \ & P_w\big(\sum_{1\le i\ne j\le n}\big|\frac{w_i^{(n)}w_j^{(n)}}{m_n(m_n-1)} - \frac{1}{n(n-1)}\big|E_X(X_i - X_j)^2 > 2\varepsilon_1\varepsilon_2\big) \\ \le \ & P_w\big(\sum_{1\le i\ne j\le n}\big|\frac{w_i^{(n)}w_j^{(n)}}{m_n(m_n-1)} - \frac{1}{n(n-1)}\big| > \varepsilon_1\varepsilon_2\sigma^{-2}\big). \end{aligned} \tag{87}$$

The preceding relation can be bounded above by:

$$
\begin{aligned}
&\varepsilon_1^{-2}\varepsilon_2^{-2}\sigma^4\Big\{n(n-1)E_w\big(\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big)^2\\
&+n(n-1)(n-2)E_w\Big(\big|\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big|\big|\frac{w_1^{(n)}w_3^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big|\Big)\\
&+n(n-1)(n-2)(n-3)E_w\Big(\big|\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big|\big|\frac{w_3^{(n)}w_4^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big|\Big)\Big\}\\
&\le\varepsilon_1^{-2}\varepsilon_2^{-2}\sigma^4\Big\{n(n-1)E_w\big(\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big)^2\\
&+n(n-1)(n-2)E_w\big(\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big)^2\\
&+n(n-1)(n-2)(n-3)E_w\big(\frac{w_1^{(n)}w_2^{(n)}}{m(m-1)}-\frac{1}{n(n-1)}\big)^2\Big\}\\
&=\varepsilon_1^{-2}\varepsilon_2^{-2}\sigma^4\big\{n(n-1)\\
&+n(n-1)(n-2)+n(n-1)(n-2)(n-3)\big\}\Big\{\frac{1}{n^4m^2}+\frac{n}{n^4m^2}+\frac{n^2}{n^4m^2}\Big\}.
\end{aligned}
$$

Clearly, the latter term approaches zero when $m\to+\infty$, for each fixed n. By this we have shown that $S^2_{m_n,n}\to S^2_n$ in probability-$P_{X,w}$, when n is fixed and only $m\to+\infty$. □

Proof of the consistency of $\hat{X}_{m_n,n}$ in (21)

We give the proof of (21) for $m_n=n$, noting that the proof below remains the same for $m_n\le n$ and it can be adjusted for the case $m_n=kn$, where k is a positive integer. In order to establish (21) when $m_n=n$, we first note that

$$
E_{X|w}\big(\sum_{i=1}^{n}\big|\frac{w_i^{(n)}}{n}-\frac{1}{n}\big|\big)=2\big(1-\frac{1}{n}\big)^n
$$

and, with $\varepsilon_1,\varepsilon_2,\varepsilon_3>0$, we proceed as follows.

$$
\begin{aligned}
&P_w\big\{P_{X|w}\big(\big|\hat{X}_{m_n,n}-\mu\big|>\varepsilon_1\big)>\varepsilon_2\big\}\\
&\le P_w\Big\{P_{X|w}\Big(\big|\hat{X}_{m_n,n}-\mu\big|>\varepsilon_1\Big)>\varepsilon_2,\Big|\sum_{=1}^{n}\big|\frac{w_j^{(n)}}{n}-\frac{1}{n}\big|-2\big(1-\frac{1}{n}\big)^n\Big|\le\varepsilon_3\Big\}\\
&+P_w\Big\{\Big|\sum_{j=1}^{n}\big|\frac{w_j^{(n)}}{n}-\frac{1}{n}\big|-2\big(1-\frac{1}{n}\big)^n\Big|>\varepsilon_3\Big\}\\
&\le P_w\big\{P_{X|w}\big(\big|\sum_{i=1}^{n}\big|\frac{w_j^{(n)}}{n}-\frac{1}{n}\big|(X_i-\mu)\big|>\varepsilon_1\big(2\big(1-\frac{1}{n}\big)^n-\varepsilon_3\big)\big)>\varepsilon_2\big\}
\end{aligned}
$$

$$+\varepsilon_3^{-2}E_w\big(\sum_{j=1}^{n}|\frac{w_j^{(n)}}{n}-\frac{1}{n}|-2(1-\frac{1}{n})^n\big)^2$$
$$\le P_w\big\{\sum_{i=1}^{n}(\frac{w_i^{(n)}}{n}-\frac{1}{n})^2>\sigma^{-2}\big(2(1-\frac{1}{n})^n-\varepsilon_3\big)^2\varepsilon_2\big\}$$
$$+\varepsilon_3^{-2}\big\{nE_w(\frac{w_1^{(n)}}{n}-\frac{1}{n})^2+n(n-1)E_w\big(|\frac{w_1^{(n)}}{n}-\frac{1}{n}||\frac{w_2^{(n)}}{n}-\frac{1}{n}|\big)-4(1-\frac{1}{n})^{2n}\big\}$$
$$=: K_1(n)+K_2(n).$$

A similar argument to that in (86) implies that, as $n\to+\infty$, and then $\varepsilon_3\to 0$, we have $K_1(n)\to 0$. As to $K_2(n)$, we note that

$$E_w(\frac{w_1^{(n)}}{n}-\frac{1}{n})^2 = n^{-2}(1-\frac{1}{n})$$
$$E_w\big(|\frac{w_1^{(n)}}{n}-\frac{1}{n}||\frac{w_2^{(n)}}{n}-\frac{1}{n}|\big) = -n^{-3}+4n^{-2}(1-\frac{1}{n})^n(1-\frac{1}{n-1})^n.$$

Observing now that, as $n\to+\infty$,

$$n(n-1)E_w\big(|\frac{w_1^{(n)}}{n}-\frac{1}{n}||\frac{w_2^{(n)}}{n}-\frac{1}{n}|\big)-4(1-\frac{1}{n})^{2n}\to 0,$$

we conclude that, as $n\to+\infty$, $K_2(n)\to 0$. By this we have concluded the consistency of $\hat{X}_{m_n,n}$ for the population mean μ, when $m_n=n$. □

Appendix B

The convergence in distribution of the partial sums of the form $\sum_{i=1}^{n}w_i^{(n)}X_i$ associated with $T_{m_n,n}^{(i)}$, $i=1,2$, were also studied in the context of the bootstrap by Csörgő *et al.* [7] via conditioning on the weights (cf. Theorem 2.1 and Corollary 2.2 therein). We note that the latter results include only randomly weighted statistics that are similar to $T_{m_n,n}^{(i)}$, $i=1,2$, which are natural pivots for the sample mean $\bar{X}_n$. In view of the fact that $G_{m_n,n}^{(i)}$, $i=1,2$, as defined by (4) and (7), are natural pivots for the population mean $\mu:=E_XX$, in a similar fashion to Theorem 2.1 and its Corollary 2.2 of Csörgő *et al.* [7], here we state conditional CLT's, given the weights $w_i^{(n)}$'s, where $(w_1^{(n)},\ldots,w_i^{(n)}) \stackrel{d}{=} multinomial(m_n;\,1/n,\ldots,1/n)$, for the partial sums $\sum_{i=1}^{n}|\frac{w_i^{(n)}}{m_n}-\frac{1}{n}|(X_i-\mu)$. The proofs of these results are essentially identical to that of Corollary 2.2 of Csörgő *et al.* [7] in view of the more general setup in terms of notations in the latter paper.

Theorem A.1. *Let $X,X_1,\ldots$ be real valued i.i.d. random variables with mean μ and variance σ^2, where $0<\sigma^2<+\infty$.*

(a) If $m_n,n\to\infty$, in such a way that $m_n=o(n^2)$, then

$$P_{X|w}(G_{m_n,n}^{(1)}\le t)\longrightarrow\Phi(t)\ \textit{in probability}-P_w\ \textit{for all}\ t\in\mathbb{R}.$$

(b) If $m_n, n \to \infty$ in such a way that $m_n = o(n^2)$ and $n = o(m_n)$, then

$$P_{X|w}(G^{(2)}_{m_n,n} \le t) \longrightarrow \Phi(t) \text{ in probability} - P_w, \text{for all } t \in \mathbb{R}.$$

Acknowledgements

We wish to thank two referees, an Associate Editor and the Editor for their careful reading and discussion of our manuscript, for their probing questions and suggestions that have led to an improved presentation of our results.

References

[1] Arenal-Gutiérrez, E. and Matrán, C. (1996). A zero-one law approach to the central limit theorem for the weighted bootstrap mean. *Annals of Probability* **24**, 532–540. MR1387650

[2] Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer-Verlag. MR2195545

[3] Baum, L. E. and Katz, M. (1965). Convergence rates in the law of large numbers. *The American Mathematical Society* **120**, 108–123. MR0198524

[4] Bentkus, V., Bloznelis, M. and Götze, F. (1996). A Berry-Esséen bound for Student's statistic in the non-i.i.d. case. *Journal of Theoretical Probability* **9**, 765–796. MR1400598

[5] Bentkus, V. and Götze, F. (1996). The Berry-Esséen bound for Student's statistic. *Annals of Probability* **24**, 491–503. MR1387647

[6] Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons, Second Edition. MR0830424

[7] Csörgő, M., Martsynyuk, Yu. V. and Nasari, M. M. (2014). Another look at bootstrapping the Student t-statistic. *Mathematical Methods of Statistics* **23**(4), 256–278. MR3295059

[8] Csörgő, M. and Nasari, M. M. (2013). Asymptotics of randomly weighted u- and v-statistics: Application to bootstrap. *Journal of Multivariate Analysis* **121**, 176–192. MR3090476

[9] Csörgő, S. and Rosalsky, A. (2003). A survey of limit laws for bootstrapped sums. *International Journal of Mathematics and Mathematical Sciences* **45**, 2835–2861.

[10] Das Gupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer. MR2664452

[11] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics* **27**, 642–669. MR0083864

[12] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26. MR0515681

[13] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, London. MR1270903

[14] Erdős, P. (1949). On a theorem of Hsu and Robbins. *Annals of Mathematical Statistics* **20**, 165–331. MR0030714
[15] Erdős, P. (1950). Remark on my paper "On a theorem of Hsu and Robbins". *Annals of Mathematical Statistics* **21**, 138. MR0032970
[16] Giné, E., Götze, F. and Mason, D. M. (1997). When is the Student t-statistic asymptotically standard normal? *Annals of Probability* **25**, 1514–1531. MR1457629
[17] Hall, P. (1995). *The Bootstrap and Edgeworth Expansion*. Springer. MR1145237
[18] Hartley, H. O. and Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics* **31**, 411–422. MR0386084
[19] Hsu, P. L. and Robbins, H. (1947). Complete convergence and the law of large numbers. *Proceeding of the National Academy of Science* **33**, 25–31. MR0019852
[20] Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer-Verlag, New York, Inc. MR2001447
[21] Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Annals of Statistics* **20**, 1611–1624. MR1186268
[22] Mason, D. M. and Shao, Q. M. (2001). Bootstrapping the Student t-statistic. *Annals of Probability* **29**, 1435–1450. MR1880227
[23] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability* **18**, 1269–1283. MR1062069
[24] Pinelis, I. (2012). On the Berry-Esséen bound for the Student statistic. arXiv:1101.3286 [math.ST].
[25] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165
[26] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer. MR1351010
[27] Shao, Q. M. (2005). An explicit Berry-Esséen bound for Student's t-statistic via Stein's method. In *Stein's Method and Applications, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* **5**, 143–155. Singapore University Press, Singapore. MR2205333
[28] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics* **9**, 1187–1195. MR0630102
[29] Yatracos, Y. (2002). Assessing the quality of bootstrap samples and of the bootstrap estimates obtained with finite resampling. *Statistics and Probability Letters* **59**, 281–292. MR1932871