DOI 10.1007/s10986-015-9283-y
Lithuanian Mathematical Journal, Vol. 55, No. 2, April, 2015, pp. 312–330

Gini's mean difference and variance as measures of finite populations scales

Andrius Čiginas [a,1] **and Dalius Pumputis** [b]

[a] Institute of Mathematics and Informatics, Vilnius University, Akademijos str. 4, LT-08663 Vilnius, Lithuania
[b] Lithuanian University of Educational Sciences, Studentų str. 39, LT-08106 Vilnius, Lithuania
(e-mail: andrius.ciginas@mii.vu.lt; dalius.pumputis@leu.lt)

Received September 1, 2014; revised March 30, 2015

Abstract. We consider Gini's mean difference statistic as an alternative to the empirical variance in the settings of finite populations, where simple random samples are drawn without replacement. In particular, we discuss specific (in the finite-population context) estimation strategies for a scale of the population, related to the alternative statistic in a possible presence of outliers in the data.

The paper also presents a wide comparative survey of properties of Gini's mean difference statistic and the empirical variance. It includes asymptotic properties of both statistics: the asymptotic normality, one-term Edgeworth expansions, and bootstrap approximations for Studentized versions of the statistics. Estimation of the variances and other parameters of the statistics is also studied, where we use auxiliary information available on the population elements. Theoretical results are illustrated by a simulation study.

MSC: 62E20

Keywords: sampling without replacement, sample variance, Gini's mean difference, robustness, asymptotic normality, second-order approximations

1 Introduction

Together with a location parameter, a spread (or scale) of survey population usually is a parameter of interest. If a statistician assumes the classical model of independent and identically distributed (i.i.d.) observations, then he has at his disposal a number of various parametric distribution families, for example, Gaussian, Cauchy, etc. Assume that he chooses a particular family (population model) for a further analysis of data. This family comes with its own location and scale measures, for instance, the normal distribution parameters "suggest" to measure the mean and variance of the survey population, and the Cauchy distribution is specified by the population median and interquartile range. The traditional statistics theory gives the answers of how to get efficient estimates of locations and scales in the case of commonly used population models. However, parametric statistical models, being comparatively convenient, are also known as nonrobust, that is, deviations from their

[1] The research of A. Čiginas is supported by European Union Structural Funds project "Postdoctoral Fellowship Implementation in Lithuania."

assumptions may lead to misleading conclusions. As it is often the instance, appearance of some or more outlying observations can strongly affect the quality of typical estimators of the population location and scale. Then, if we believe that these outliers are, for example, measurement errors, robust estimation methods can be a treatment of the problem. The pioneering book of Huber [23] on the formalization of robust estimation starts from the normal distribution scale estimation example showing the inefficiency of the empirical variance compared to the mean absolute deviation in the presence of outliers in the sample data. Gini's mean difference (GMD) statistic is, in a sense, similar to the latter statistic. This estimator, its properties, connections, and comparisons with the sample variance is the aim of our paper.

We consider the GMD statistic as an alternative to the empirical variance in the setting of finite population $\{1,\dots,N\}$ of elements with the corresponding set of real values $\mathcal{X}=\{x_1,\dots,x_N\}$ of the variable x under investigation and for the simple random sample $\{1,\dots,n\}$ of size $n<N$ *drawn without replacement* from the population with measurements $\mathbb{X}=\{X_1,\dots,X_n\}$ of the variable x. In particular, the population parameters

$$G=\binom{N}{2}^{-1}\sum_{1\leqslant i<j\leqslant N}|x_i-x_j| \tag{1.1}$$

and

$$V=\binom{N}{2}^{-1}\sum_{1\leqslant i<j\leqslant N}\frac{(x_i-x_j)^2}{2} \tag{1.2}$$

are two candidates to measure the scale of $\mathcal{X}$; the latter seems more natural since $\operatorname{Var}X_1=(N-1)V/N$. The corresponding unbiased estimators of these parameters are the GMD statistic

$$U_G=\binom{n}{2}^{-1}\sum_{1\leqslant i<j\leqslant n}|X_i-X_j| \tag{1.3}$$

and the empirical variance

$$U_V=\binom{n}{2}^{-1}\sum_{1\leqslant i<j\leqslant n}\frac{(X_i-X_j)^2}{2}. \tag{1.4}$$

As an alternative to (1.4), the GMD statistic, known better since Gini [18], is widely used in economics. Now it is an ordinary measure of dispersion of income distribution and in cases of variables similar to income; see the monograph of Yitzhaki and Schechtman [31], where, following the words of the authors, the commonly used variance-based analysis is "translated" into the Gini-based one. The use of GMD is not restricted to the measurements of economic inequality. Just like in the problems of economists, where data deviate from the normality, the parameter G and its estimator U_G can be used as dispersion measures for many kinds of statistical data. Our choice of the finite-population setting also has an economic motivation because, in economic surveys, the number N of surveyed objects or subjects is not necessarily that large (compared to the sample size) as to ignore the dependence between the observations in the set $\mathbb{X}$.

In Section 2, we consider three strategies of estimation of the finite-population scale related to the alternative U_G. We exploit two assumptions that are usually possible in the finite-population context: the so-called superpopulation assumption and availability of an auxiliary information on the population elements. We also carry out simulation experiments, where we analyze advantages and disadvantages of the strategies and compare them under populations without and with outliers.

The GMD statistic is one of several well-known universal estimators that are less sensitive to outliers than the empirical variance. The other ones are, for instance, the median absolute deviation (MAD) and the interquartile range (IQR), known as highly robust statistics, and the mean absolute deviation, which is of a similar sensitivity to outliers as GMD. The statistics MAD and IQR are efficient where data contain large outliers or/and the proportion of outliers is large, but they are relatively inefficient in the opposite case. Recently, it has been demonstrated by Gerstenberger and Vogel [17] that the GMD statistic combines advantages of the sample

variance and the absolute mean deviation in the i.i.d. case. In particular, at heavy-tailed distributions, GMD is, along with the absolute mean deviation, substantially more efficient than the standard deviation. However, looking from the point of the robust estimation theory, if we can link data to a parametric population model, then, in particular situations, there are more effective robust estimators of scale than the universal ones; see Huber [23]. On the other hand, we focus here on a unified improvement of the empirical variance.

An additional premium to be paid is a relatively complex access to the properties of the GMD statistic. On the other hand, in these problems, U_G is more attractive than the other mentioned examples of universal estimators because of its smoothness (in a certain sense) or that it uses the complete sample information. In Section 3, an asymptotic analysis of distributions of the statistics U_G and U_V shows that their properties are similarly simple. To explain that, we apply an available theory of U- and L-statistics in the case of samples without replacement. In particular, statistics (1.3) and (1.4) are likely the most popular U-statistics of degree two, and (1.3) is also the L-statistic; see Serfling [30]. As the L-statistic, U_G is smooth in the sense that its weight function is smooth; see the same reference.

To be consistent with already known results, we first notice that expressions of the variance of U_G and its approximations are known since Nair [27] and Lomnicki [26] in the case of i.i.d. observations and since Glasser [19] for the simple random samples without replacement. Second, an efficient method to study variances and the asymptotic normality of the statistics U_G and U_V is Hoeffding's [22] decomposition for U-statistics. Much later, Zhao and Chen [32] used an analogous decomposition in the case of finite population. Third, similarly, the second-order approximation theory for samples without replacement has been realized following the case of i.i.d. observations: Kokic and Weber [25] and Bloznelis and Götze [7] follow Bickel et al. [3] on one-term Edgeworth approximations to the distributions of standardized U-statistics; the papers of Bloznelis [5, 6] on a one-term Edgeworth expansion for Studentized U-statistics and bootstrap approximations appeared after Helmers [21].

Since the true values of variances of the statistics are almost always unknown, we prefer considering the asymptotic normality, one-term Edgeworth expansions, and bootstrap approximations for Studentized versions of the statistics U_G and U_V. A basis for such a study is the general theory of Bloznelis and Götze [8] and Bloznelis [5, 6] (with the without-replacement bootstrap, of Booth et al. [9]), where, to ensure the validity of approximations, quite general smoothness conditions are imposed on parts of the Hoeffding decomposition of U-statistics. Theorems of Section 3 allow us to compare distributional properties of U_G and U_V much easier.

A successful application of the one-term Edgeworth expansion requires having good estimators (in the sense of an asymptotic consistency or a small mean square error) of the expansion parameters. In the case of symmetric statistics (symmetric functions of observations), including U-statistics, jackknife techniques are used to estimate these parameters; see Putter and van Zwet [29] and Bloznelis [4]. In particular cases of statistics, for example, U_G and U_V, there are more ways to construct estimators of the Edgeworth expansion parameters. For example, for L-statistics, including U_G, the bootstrap was used by Čiginas [14], and, assuming that the auxiliary information is available, calibration methods were applied by Pumputis and Čiginas [28]. In Section 4, we propose simple and efficient estimators of the parameters without any auxiliary information as well as using it. Similar estimators of the variances of U_G and U_V are considered as well. In Section 5, we discuss empirical Edgeworth expansions, based on the estimators of parameters, and bootstrap approximations. In Section 6, we compare the obtained estimation results for both statistics of interest in the simulation study. Here we are also interested in the role of outliers in populations. Conclusions of the paper are given in Section 7. Proofs of the theorems are given in the appendices.

2 Estimation of a scale

2.1 Outliers and estimation strategies

In the i.i.d. setup, for many common parametric population models, the sample variance is an efficient estimator under ideal or close to ideal conditions. However, if some of the sample data differ substantially from the other, then the GMD statistic can be a better choice because it puts smaller weights on extreme observations and thus lowers their impact on the estimation.

In the finite-population case, outliers are also less influential in case U_G is applied. To see this, let us write parameters (1.1) and (1.2) in a different form. Here and further, without loss of generality, we assume that $x_1 \leqslant \cdots \leqslant x_N$ and denote $\Delta_i = x_{i+1} - x_i$, $i = 1, \ldots, N-1$. Then, taking $x_j - x_i = \sum_{k=i}^{j-1} \Delta_k$, we obtain

$$G^2 = \frac{4}{N^2(N-1)^2}\left[\sum_{i=1}^{N-1} i^2(N-i)^2\Delta_i^2 + 2\sum_{1\leqslant i<j\leqslant N-1} ij(N-i)(N-j)\Delta_i\Delta_j\right]$$

and

$$V = \frac{1}{N(N-1)}\left[\sum_{i=1}^{N-1} i(N-i)\Delta_i^2 + 2\sum_{1\leqslant i<j\leqslant N-1} i(N-j)\Delta_i\Delta_j\right].$$

These expressions are connected via the system of equations

$$\Delta'_i\Delta'_j = \frac{4j(N-i)}{N(N-1)}\Delta_i\Delta_j, \qquad 1 \leqslant i \leqslant j \leqslant N-1, \tag{2.1}$$

where $\Delta'_i = x'_{i+1} - x'_i$, $i = 1, \ldots, N-1$, with the numbers $x'_1, \ldots, x'_N$ as its solution given $\mathcal{X}$. However, there is no solution, except in cases of very simple $\mathcal{X}$. Therefore, (2.1) can be viewed as a formal transformation only, which explains the assertion.

Model of outliers. For simple random samples without replacement, we assume the existence of so-called representative outliers. This notion was introduced by Chambers [10]. It means the assumptions that: outlying observations are not measurement errors; the unsampled population part should contain outliers too. Other types of outliers, together with their various modeling mechanisms, are also considered in the literature; see, for example, Chambers et al. [11], Béguin and Hulliger [2], and Alqallaf et al. [1], but their treatment is usually related to editing and imputation of survey data. Therefore, due to our focus on estimation, we consider the representative outliers only.

More formally, denote by $0 \leqslant p \leqslant N$ the number of outliers in the population. Assume that the population elements $\{i_1, \ldots, i_p\} \subseteq \{1, \ldots, N\}$ belong to a different population, but this phenomenon is not known until the sample $\mathbb{X}$ has not been obtained. Then the corresponding values from the sequence $x_1, \ldots, x_N$ are treated as outliers. In the random sample $\mathbb{X}$, the number of outliers is random and equals the number of elements in the set $\{i_1, \ldots, i_p\} \cap \{1, \ldots, n\}$.

The proportion p/N of outliers can be restricted without a significant loss of generality. In particular, as it is pointed by Huber [23], a part of gross errors (outliers) in samples usually is not larger than 10%. An interesting note on this issue is given by Chhikara and Feiveson [12]: "*... it is reasonable to consider three potential outliers in a data set of* 10 *observations, but it is unrealistic to expect* 30 *outliers out of a data set of* 100 *observations. In the latter case, the outlier detection problem becomes one of discrimination between two or more classes of data.*" Similarly, for finite populations, if a large portion of outliers is expected in the population, they are typically neutralized (with the help of auxiliary information) by applying stratified sampling designs, that is, collecting potential outliers into a separate stratum. Another but similar solution in this case is postratification.

Estimation strategies. For a finite population, we consider the following specific ways of applying the GMD statistic as the alternative to the sample variance.

- (S_1) Assume that the fixed numbers $x_1, \ldots, x_N$ are realizations of the i.i.d. random variables $X_1^*, \ldots, X_N^*$ (superpopulation model) from a parametric family of distributions with the scale parameter that is a one-argument function of $\sqrt{\operatorname{Var} X_1^*}$. Then the scale of $\mathcal{X}$ is treated as the same function of $\sqrt{V}$, and the estimator of the argument $\sqrt{V}$ is taken to be of the form aU_G, where $a > 0$ is a constant compensating a bias.

(S_2) In the presence of the well-correlated and completely known auxiliary variable z with values $\mathcal{Z} = \{z_1, \dots, z_N\}$ in the population, the scale measure is $\sqrt{V}$, and its estimator is aU_G with the correction $a > 0$ evaluated from the set $\mathcal{Z}$.

(S_3) The parameter G itself is treated as the scale of $\mathcal{X}$, and the GMD statistic U_G is its estimator.

Case (S_1) is close to the parametric statistics. In the i.i.d. setting, the multipliers a, which ensure that aU_G is the unbiased estimator of $\sqrt{V}$, are known for commonly used parametric families, e.g., $a = \sqrt{\pi}/2$ for the normal distributions, and $a = 1$ for the exponential distributions. See also Tables 2–3 in [17], for some other distributions. Therefore, assuming the existence of a superpopulation, we use the same constants for the estimation in the finite population. If a good auxiliary information $\mathcal{Z}$ is available, then these theoretical a should not differ so much from the corresponding values obtained in case (S_2), where $a > 0$ is evaluated from $aG = \sqrt{V}$, using $\mathcal{Z}$ instead of $\mathcal{X}$. If the scatter of $\mathcal{X}$ cannot be linked to a family of distributions, for example, if it is a mixture of two unknown distributions and there is no other additional information, then we propose strategy (S_3).

2.2 Numerical analysis

We compare efficiencies of strategies (S_1) and (S_2) with respect to the common estimation by $\sqrt{U_V}$ in the presence of outliers. We consider two populations with values of the variable x generated respectively from two different parametric families: the normal distributions $\mathcal{N}(\mu, \sigma^2)$ and the gamma distributions $\mathcal{G}(k, \theta)$ with shape k and scale θ, where the variance is equal to $k\theta^2$. In the case of a gamma distribution, the correction $a = k^{-1/2}(2 - 4I_{0.5}(k+1, k))^{-1}$ depends on k, where $I_t(u, v)$ is a regularized incomplete Beta function. For each of these populations, we consecutively increase the part of outliers in the population as follows. First, we select some particular population elements randomly without replacement. Second, we replace their values by the new ones generated from the same family of distributions, but with different parameters, and we fix these values. In the next steps, the set of outlying elements is enlarged by selecting from those that still do not belong to the outliers.

In particular, the distributions are: $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 9)$, the latter of which is aimed to generate outliers; $\mathcal{G}(3, 1/\sqrt{3})$ (then $a = 8\sqrt{3}/15$), and $\mathcal{G}(3, \sqrt{3})$, which is also meant for outliers. We take $N = 1000$, $n = 200$, and consecutively construct populations with $p = 0, 20, 40, 60, 80, 100$ outliers.

The fixed values of the auxiliary information $\mathcal{Z}$ are generated by the linear regression $z_i = 3 + 2x_i + \varepsilon_i$, where ε_i, $i = 1, \dots, N$, are i.i.d. random variables from $\mathcal{N}(0, \vartheta^2)$. Since the set $\mathcal{X}$ is different for different p, the collections $\mathcal{Z}$ are different too.

To better understand the role of the auxiliary information in strategy (S_2), we simulate different correlations ρ_{zx} between $\mathcal{Z}$ and $\mathcal{X}$. The correlation is controlled by the variance ϑ^2 in the linear model. Thus, we choose the variance in order to have $\rho_{zx} = 0.9, 0.7, 0.5$, and 0.1 approximately. Tables 1–2 present the comparison of estimation methods by means of mean square errors and biases.

It is seen in Table 1 that strategy (S_1) improves the estimator $\sqrt{U_V}$, where the proportion p/N is smaller. For p/N larger than 0.04, (S_1) becomes inefficient (by $\mathrm{MSE}(\cdot)$) because its bias is large since the fixed correction a is a too coarse approximation to the mix of the normal distributions. Strategy (S_2) is the best under a strong correlation between x and z because the estimation bias is well corrected. The efficiency of (S_2) decreases with a decrease in the correlation ρ_{zx}, but, with the mechanism used to generate $\mathcal{Z}$, the estimation is not relatively inaccurate even for correlation 0.1. This means a certain robustness of this strategy.

Table 2 shows similar results for the asymmetric gamma distributions. Here outliers affect the estimators more because the distribution of outliers has a larger mean (location) in addition. Therefore, strategies (S_1) and (S_2) are efficient for smaller proportions p/N than in Table 1.

We conclude that strategies (S_1) and (S_2), and the GMD statistic thereby, are efficient with respect to $\sqrt{U_V}$ if there is a small percent of outliers in the population. Moreover, there is no loss in the efficiency of the strategies if there are no outliers in the population.

Table 1. $\mathcal{N}(0,1)$ with outliers $\mathcal{N}(0,9)$. Accuracy according to $10 \times (\mathrm{BIAS}(\cdot), \sqrt{\mathrm{MSE}(\cdot)})$

p/N	$\sqrt{U_V}$	(S_1)	$\rho_{zx} = 0.9$; (S_2)	$\rho_{zx} = 0.7$; (S_2)	$\rho_{zx} = 0.5$; (S_2)	$\rho_{zx} = 0.1$; (S_2)
0.00	$(-0.01, 0.48)$	$(-0.06, 0.47)$	$(-0.02, 0.47)$	$(\ 0.00, 0.47)$	$(-0.05, 0.47)$	$(\ 0.03, 0.47)$
0.02	$(-0.03, 0.86)$	$(-0.41, 0.73)$	$(-0.13, 0.63)$	$(-0.32, 0.68)$	$(-0.40, 0.72)$	$(-0.31, 0.65)$
0.04	$(-0.03, 0.99)$	$(-0.67, 0.97)$	$(-0.16, 0.75)$	$(-0.44, 0.84)$	$(-0.63, 0.95)$	$(-0.47, 0.80)$
0.06	$(-0.03, 1.10)$	$(-0.92, 1.22)$	$(-0.24, 0.87)$	$(-0.71, 1.07)$	$(-0.86, 1.17)$	$(-0.85, 1.14)$
0.08	$(-0.03, 1.10)$	$(-0.95, 1.26)$	$(-0.28, 0.91)$	$(-0.58, 1.02)$	$(-0.92, 1.23)$	$(-0.93, 1.24)$
0.10	$(-0.03, 1.22)$	$(-1.18, 1.48)$	$(-0.21, 0.98)$	$(-0.91, 1.29)$	$(-1.06, 1.39)$	$(-1.01, 1.32)$

Table 2. $\mathcal{G}(3, 1/\sqrt{3})$ with outliers $\mathcal{G}(3, \sqrt{3})$. Accuracy according to $10 \times (\mathrm{BIAS}(\cdot), \sqrt{\mathrm{MSE}(\cdot)})$

p/N	$\sqrt{U_V}$	(S_1)	$\rho_{zx} = 0.9$; (S_2)	$\rho_{zx} = 0.7$; (S_2)	$\rho_{zx} = 0.5$; (S_2)	$\rho_{zx} = 0.1$; (S_2)
0.00	$(-0.03, 0.71)$	$(-0.18, 0.64)$	$(-0.21, 0.65)$	$(-0.42, 0.73)$	$(-0.60, 0.84)$	$(-0.62, 0.85)$
0.02	$(-0.09, 1.40)$	$(-1.09, 1.40)$	$(-0.48, 1.04)$	$(-1.05, 1.36)$	$(-1.48, 1.70)$	$(-1.54, 1.75)$
0.04	$(-0.08, 1.42)$	$(-1.24, 1.55)$	$(-0.66, 1.18)$	$(-1.21, 1.53)$	$(-1.69, 1.91)$	$(-1.73, 1.94)$
0.06	$(-0.07, 1.47)$	$(-1.56, 1.87)$	$(-0.76, 1.35)$	$(-1.71, 1.99)$	$(-1.98, 2.23)$	$(-2.14, 2.36)$
0.08	$(-0.12, 1.93)$	$(-2.24, 2.56)$	$(-1.08, 1.71)$	$(-2.06, 2.41)$	$(-2.70, 2.94)$	$(-2.86, 3.10)$
0.10	$(-0.13, 1.99)$	$(-2.56, 2.89)$	$(-1.31, 1.97)$	$(-2.54, 2.88)$	$(-2.85, 3.14)$	$(-3.03, 3.30)$

3 Theoretical properties of the statistics

3.1 Hoeffding's decompositions and variances

The statistic $U = U_n(\mathbb{X}) = \sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j)$, where the function $h\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies $h(x,y) = h(y,x)$, is called the U-statistic of degree two. For the cases of the GMD statistic U_G and the sample variance U_V, we have

$$h(X_1, X_2) = \binom{n}{2}^{-1} |X_1 - X_2|$$

and

$$h(X_1, X_2) = \binom{n}{2}^{-1} \frac{(X_1 - X_2)^2}{2},$$

respectively. Following [5], the Hoeffding decomposition of the U-statistic is

$$U = \mathbf{E}\, U + U_1 + U_2, \tag{3.1}$$

where $U_1 = \sum_{i=1}^{n} g_1(X_i)$ and $U_2 = \sum_{1 \leqslant i < j \leqslant n} g_2(X_i, X_j)$ are centered and uncorrelated linear and quadratic parts, respectively. Here, for $1 \leqslant k \leqslant N$,

$$g_1(x_k) = (n-1)\frac{N-1}{N-2}\, \mathbf{E}\left(h(X_1, X_2) - \mathbf{E}\, h(X_1, X_2) \;\middle|\; X_1 = x_k\right)$$

and, for $1 \leqslant k \neq l \leqslant N$,

$$g_2(x_k, x_l) = h(x_k, x_l) - \mathbf{E}\, h(X_1, X_2) - (n-1)^{-1}\big(g_1(x_k) + g_1(x_l)\big).$$

The so-called first- and second-order influence functions g_1 and g_2 have usually a different impact on the variance of U-statistic. As in cases of any other linearization techniques, it is expected that the linear part

in (3.1) dominates the remainder in the sense of variance size. In particular, we consider the structures of variances of the statistics U_G and U_V by formula (2.6) in [8]:

$$\operatorname{Var} U = \frac{n(N-n)}{N-1}\sigma_1^2 + \binom{n}{2}\binom{N-n}{2}\binom{N-2}{2}^{-1}\sigma_2^2, \tag{3.2}$$

where $\sigma_1^2 = \mathbf{E}\, g_1^2(X_1)$ and $\sigma_2^2 = \mathbf{E}\, g_2^2(X_1, X_2)$. Let us elaborate the statistics of interest.

GMD statistic. To find the influence functions, we rewrite (1.3) in the alternative form

$$U_G = \binom{n}{2}^{-1} \sum_{j=1}^{n} (2j - n - 1) X_{j:n},$$

where $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ are order statistics of the observations $\mathbb{X}$, and apply the Hoeffding decomposition results to the L-statistics from Čiginas [13]. Denote $a_i = (2i - N)/N$, $1 \leqslant i \leqslant N-1$. Then, for $1 \leqslant k \leqslant N$,

$$g_1(x_k) = -\frac{2}{n}\frac{N}{N-2}\sum_{i=1}^{N-1}\left(\mathbf{I}_{\{i \geqslant k\}} - \frac{i}{N}\right) a_i \Delta_i,$$

where $\mathbf{I}_{\{\cdot\}}$ is the indicator function, and, for $1 \leqslant k < l \leqslant N$,

$$g_2(x_k, x_l) = -\frac{4}{n(n-1)}\sum_{i=1}^{N-1} \phi_{k,l}(i)\Delta_i,$$

where

$$\phi_{k,l}(i) = \begin{cases} i(i-1)/A & \text{if } 1 \leqslant i < k, \\ -(i-1)(N-i-1)/A & \text{if } k \leqslant i < l, \\ (N-i-1)(N-i)/A & \text{if } l \leqslant i < N, \end{cases}$$

with $A = (N-1)(N-2)$. Next, direct calculations yield the following expressions of variance decomposition (3.2) components:

$$\sigma_1^2 = \frac{4}{n^2}\frac{1}{(N-2)^2}\left[\sum_{i=1}^{N-1} i(N-i)a_i^2\Delta_i^2 + 2\sum_{1 \leqslant i < j \leqslant N-1} i(N-j)a_i a_j \Delta_i \Delta_j\right] \tag{3.3}$$

and

$$\sigma_2^2 = \frac{16}{n^2(n-1)^2}\frac{1}{N(N-1)^2(N-2)}\left[\sum_{i=1}^{N-1} i(i-1)(N-i-1)(N-i)\Delta_i^2 + 2\sum_{1 \leqslant i < j \leqslant N-1} i(i-1)(N-j-1)(N-j)\Delta_i\Delta_j\right]. \tag{3.4}$$

Sample variance. Denote the population moments as $b_1 = \mathbf{E}\, X_1$ and $\mu_k = \mathbf{E}(X_1 - b_1)^k$ for $k = 2, \ldots, 6$. Then, for $1 \leqslant k \leqslant N$,

$$g_1(x_k) = \frac{1}{n}\frac{N}{N-2}\big[(x_k - b_1)^2 - \mu_2\big] \tag{3.5}$$

and, for $1 \leqslant k < l \leqslant N$,

$$g_2(x_k, x_l) = \frac{1}{n(n-1)}\left\{(x_k - x_l)^2 + \frac{2N}{(N-1)(N-2)}\mu_2 - \frac{N}{N-2}\big[(x_k - b_1)^2 + (x_l - b_1)^2\big]\right\}. \tag{3.6}$$

After straightforward calculations, we obtain the following formulas:

$$\sigma_1^2 = \frac{1}{n^2}\left(\frac{N}{N-2}\right)^2(\mu_4 - \mu_2^2) \tag{3.7}$$

and

$$\sigma_2^2 = \frac{4}{n^2(n-1)^2}\frac{N}{(N-1)(N-2)}\left(\frac{N^2 - 3N + 3}{N-1}\mu_2^2 - \mu_4\right). \tag{3.8}$$

In fact, various expressions of $\operatorname{Var} U_V$ are known in the literature. For comparison, we mention only the expression that appeared in [24].

3.2 Asymptotic normality

Common inferences about statistics are based on knowledge of their distributions. If exact distributions cannot be found, then, for samples of quite a large size, the normal approximation to distributions is usually appropriate. Here, for the statistics under investigation, we present sufficient and simple conditions under which the distribution function

$$F_{nS}(y) = \mathbf{P}\{U - \mathbf{E}\,U \leqslant yS\} \tag{3.9}$$

of the Studentized U-statistic is asymptotically normal as the sample size increases. Here

$$S^2 = S^2(\mathbb{X}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{i=1}^{n}\big(U_{n-1}(\mathbb{X}\backslash X_i) - \bar{U}\big)^2, \quad \text{where } \bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_{n-1}(\mathbb{X}\backslash X_i), \tag{3.10}$$

is the jackknife estimator of variance for any U-statistic.

In the finite population asymptotics, the population size increases together with the sample size. We denote $n_* = \min\{n, N-n\}$, which tends to infinity as n does in the i.i.d. setup. Next, to be correct in the formulation of asymptotic results, a sequence of values $\mathcal{X}_r = \{x_{r,1}, \dots, x_{r,N_r}\}$ in the populations with $N_r \to \infty$ as $r \to \infty$ and a sequence of statistics $U_{n_r}(\mathbb{X}_r)$, where $\mathbb{X}_r = \{X_{r,1}, \dots, X_{r,n_r}\}$ is a sample drawn without replacement from $\mathcal{X}_r$, should be considered. Further, we omit the subscript r in these and other quantities for notational simplicity.

Denote $\tau^2 = n(1 - n/N)$ for short. The Erdős, Rényi [16], and Hájek [20] Lindeberg-type condition

$$\sigma_1^{-2}\,\mathbf{E}\,g_1^2(X_1)\mathbf{I}_{\{|g_1(X_1)| > \varepsilon\tau\sigma_1\}} = o(1) \quad \text{as } n_* \to \infty \text{ for every } \varepsilon > 0, \tag{3.11}$$

imposed on the linear part of U-statistic, is necessary for the normality of asymptotically linear statistics with the growth of size n_*. This condition, together with the moment conditions ensuring the asymptotic linearity, is sufficient for the statistics U_G and U_V by the following limit theorem.

Theorem 1. *Assume that $n_* \to \infty$. Let* (3.11) *be satisfied. Assume that for all n_**: (i) $\mathbf{E}\,X_1^2 \leqslant C_1 < \infty$ *for U_G and* (ii) $\mathbf{E}\,X_1^4 \leqslant C_2 < \infty$ *for U_V. Then, for U_G and U_V,* (3.9) *tends to the standard normal distribution function $\Phi(y)$ for every $y \in \mathbb{R}$.*

3.3 True one-term Edgeworth expansions

When the sample size is not large, the normal approximation to (3.9) can be inaccurate. Then the one-term Edgeworth expansion

$$H_{nS}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)\alpha + 3(y^2+1)\kappa}{6\tau}\,\varphi(y) \tag{3.12}$$

for the Studentized U-statistics, constructed in [5], can serve as an improvement. Here $\varphi(y)$ is the standard normal density function, and

$$\alpha = \sigma_1^{-3}\,\mathbf{E}\,g_1^3(X_1) \quad \text{and} \quad \kappa = \sigma_1^{-3}\tau^2\,\mathbf{E}\,g_2(X_1,X_2)g_1(X_1)g_1(X_2)$$

are the population characteristics. Next, we present detailed expressions of these parameters for both statistics of interest.

GMD statistic. Routine but tedious combinatorial calculations lead to

$$\begin{aligned}\alpha = -\sigma_1^{-3}\frac{8}{n^3}\frac{1}{(N-2)^3}\Bigg[&\sum_{i=1}^{N-1} i(N-2i)(N-i)a_i^3\Delta_i^3 + 3\sum_{1\leqslant i<j\leqslant N-1} i(N-2i)(N-j)a_i^2a_j\Delta_i^2\Delta_j \\ &+ 3\sum_{1\leqslant i<j\leqslant N-1} i(N-2j)(N-j)a_ia_j^2\Delta_i\Delta_j^2 \\ &+ 6\sum_{1\leqslant i<j<m\leqslant N-1} i(N-2j)(N-m)a_ia_ja_m\Delta_i\Delta_j\Delta_m\Bigg]\end{aligned} \tag{3.13}$$

and

$$\kappa = -\sigma_1^{-3}\tau^2\frac{16}{n^3(n-1)}\frac{N}{(N-1)^2(N-2)^3}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\sum_{m=1}^{N-1} c_{ijm}a_ja_m\Delta_i\Delta_j\Delta_m, \tag{3.14}$$

where

$$c_{ijm} = \begin{cases} i(i-1)(N-m)[N-j-1+N^{-1}j(m-j)] & \text{if } i\leqslant j\leqslant m,\\ i(i-1)(N-j)[N-m-1+N^{-1}m(m-j)] & \text{if } i\leqslant m<j,\\ j(N-m)[(i-1)(N-i-1) & \\ \quad +N^{-1}\{(N-i)(N-i-1)(i-j)+i(i-1)(m-i)\}] & \text{if } j<i<m,\\ m(N-j)[(i-1)(N-i-1) & \\ \quad +N^{-1}\{i(i-1)(i-j)+(N-i-1)(N-i)(m-i)\}] & \text{if } m<i<j,\\ j(N-i-1)(N-i)[m-1+N^{-1}(N-m)(m-j)] & \text{if } j<m\leqslant i,\\ m(N-i-1)(N-i)[j-1+N^{-1}(N-j)(m-j)] & \text{if } m\leqslant j\leqslant i. \end{cases}$$

These formulas are new in the literature.

Sample variance. With straightforward calculations we can arrive at the following results:

$$\alpha = \sigma_1^{-3}\frac{1}{n^3}\left(\frac{N}{N-2}\right)^3\left(2\mu_2^3 - 3\mu_4\mu_2 + \mu_6\right) \tag{3.15}$$

and

$$\kappa = \sigma_1^{-3}\tau^2 \frac{2}{n^3(n-1)}\left(\frac{N}{N-2}\right)^3 \frac{1}{N-1}\left(-(N-2)\mu_3^2 - \frac{2N-1}{N-1}\mu_4\mu_2 + \frac{N}{N-1}\mu_2^3 + \mu_6\right). \tag{3.16}$$

Note that (3.16) can be simplified (approximated) by leaving the term with μ_3^2 in the brackets only. For comparison, expressions similar to these can be identified in the Edgeworth approximation given by Kokic and Weber [25] for the standardized sample variance.

Whereas the error of normal approximation is typically of order $O(n_*^{-1/2})$ (see, e.g., [32] for the case of standardized U-statistics), the error of the true (with known parameters α and κ) one-term Edgeworth approximation (3.12) is of order $o(n_*^{-1/2})$ under certain conditions. The first condition from those is the asymptotical nonlatticeness of the linear part of the U-statistic: for every $\varepsilon > 0$ and every $B > 0$,

$$\liminf_{n_*\to\infty} \sup_{\varepsilon<|t|<B} \left|\mathbf{E}\exp\{\mathrm{i}t\sigma_1^{-1}g_1(X_1)\}\right| < 1; \tag{3.17}$$

see [8]. This and other specific conditions sufficient for the statistics U_G and U_V are summarized in the following theorem.

Theorem 2. *Assume that $n_* \to \infty$ and $(1-n/N)\tau \to \infty$. Let* (3.17) *be satisfied. Assume that, for some $\delta > 0$ and for all n_*:* (i) $\mathbf{E}\,|X_1|^{6+\delta} \leqslant C_1 < \infty$ *for U_G and* (ii) $\mathbf{E}\,|X_1|^{12+\delta} \leqslant C_2 < \infty$ *for U_V. Then, we have*

$$\sup_{y\in\mathbb{R}}\left|F_{nS}(y) - H_{nS}(y)\right| = o\big(n_*^{-1/2}\big) \quad \text{as } n_* \to \infty$$

for U_G and U_V.

4 Estimation of parameters

4.1 Estimators of variances

The jackknife variance estimator defined by (3.10) is universal for U- and other statistics, but it is not the best one for particular statistics. In [28], bootstrap and calibrated estimators, constructed for general L-statistics, are comparatively complex. Here, for both statistics of interest, we have explicit expressions of their variances. Therefore, more natural and simpler estimators of the variances are possible. We give here, in fact, plug-in estimators of variances, replacing the population moments by their empirical counterparts in the parameters σ_1^2 and σ_2^2 that define variance (3.2).

GMD statistic. Denote $\Delta_{i:n} = X_{i+1:n} - X_{i:n}$ and $A_i = (2i-n)/n$ for $1 \leqslant i \leqslant n-1$. Then the estimators of the variance components (3.3) and (3.4) are

$$\hat\sigma_{1G}^2 = \frac{4}{n^4}\left(\frac{N}{N-2}\right)^2\left[\sum_{i=1}^{n-1} i(n-i)A_i^2\Delta_{i:n}^2 + 2\sum_{1\leqslant i<j\leqslant n-1} i(n-j)A_iA_j\Delta_{i:n}\Delta_{j:n}\right] \tag{4.1}$$

and

$$\hat\sigma_{2G}^2 = \frac{16}{n^4(n-1)^4}\frac{N}{N-2}\left[\sum_{i=1}^{n-1} i(i-1)(n-i-1)(n-i)\Delta_{i:n}^2 + 2\sum_{1\leqslant i<j\leqslant n-1} i(i-1)(n-j-1)(n-j)\Delta_{i:n}\Delta_{j:n}\right]. \tag{4.2}$$

Denote by $\hat\sigma_G^2$ the estimator of the variance of U_G obtained by plugging (4.1) and (4.2) into (3.2).

Sample variance. Denote the sample moments by $m_k = n^{-1}\sum_{i=1}^{n}(X_i - n^{-1}\sum_{j=1}^{n} X_j)^k$ for $k = 2, \dots, 6$. Replacing the central moments in (3.7) and (3.8) by the corresponding empirical moments, we get

$$\hat{\sigma}_{1V}^2 = \frac{1}{n^2}\left(\frac{N}{N-2}\right)^2 (m_4 - m_2^2) \tag{4.3}$$

and

$$\hat{\sigma}_{2V}^2 = \frac{4}{n^2(n-1)^2}\frac{N}{(N-1)(N-2)}\left(\frac{N^2-3N+3}{N-1}m_2^2 - m_4\right). \tag{4.4}$$

Let $\hat{\sigma}_V^2$ denote the estimator of the variance of U_V obtained by plugging (4.3) and (4.4) into (3.2).

4.2 Estimators for parameters defining Edgeworth expansions

In order to apply the one-term Edgeworth approximation (3.12) to the distribution functions of statistics, the parameters α and κ must be evaluated. First, in Case A, analogously to the variance estimation case, we construct estimators of the parameters directly from the explicit expressions available. Second, in Case B, we assume that the auxiliary variable z is at our disposal with known values $\{z_1, \dots, z_N\}$ for all population elements. It is expected in this case that z is well correlated with the study variable x. Then the estimators below are immediately obtained from the true values of parameters.

GMD statistic. Case A. With the notation used for the variance estimator, according to formulas (3.13) and (3.14), the estimators are

$$\begin{aligned}\hat{\alpha}_G = -\hat{\sigma}_{1G}^{-3}\frac{8}{n^6}\left(\frac{N}{N-2}\right)^3 \Bigg[&\sum_{i=1}^{n-1} i(n-2i)(n-i)A_i^3\Delta_{i:n}^3 + 3\sum_{1\leqslant i<j\leqslant n-1} i(n-2i)(n-j)A_i^2A_j\Delta_{i:n}^2\Delta_{j:n} \\ &+ 3\sum_{1\leqslant i<j\leqslant n-1} i(n-2j)(n-j)A_iA_j^2\Delta_{i:n}\Delta_{j:n}^2 \\ &+ 6\sum_{1\leqslant i<j<m\leqslant n-1} i(n-2j)(n-m)A_iA_jA_m\Delta_{i:n}\Delta_{j:n}\Delta_{m:n}\Bigg]\end{aligned} \tag{4.5}$$

and

$$\hat{\kappa}_G = -\hat{\sigma}_{1G}^{-3}\tau^2\frac{16}{n^5(n-1)^3}\left(\frac{N}{N-2}\right)^3 \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\sum_{m=1}^{n-1} C_{ijm}A_jA_m\Delta_{i:n}\Delta_{j:n}\Delta_{m:n}, \tag{4.6}$$

with the case function

$$C_{ijm} = \begin{cases} i(i-1)(n-m)[n-j-1+n^{-1}j(m-j)] & \text{if } i \leqslant j \leqslant m, \\ i(i-1)(n-j)[n-m-1+n^{-1}m(m-j)] & \text{if } i \leqslant m < j, \\ j(n-m)[(i-1)(n-i-1) & \\ \quad +n^{-1}\{(n-i)(n-i-1)(i-j)+i(i-1)(m-i)\}] & \text{if } j < i < m, \\ m(n-j)[(i-1)(n-i-1) & \\ \quad +n^{-1}\{i(i-1)(i-j)+(n-i-1)(n-i)(m-i)\}] & \text{if } m < i < j, \\ j(n-i-1)(n-i)[m-1+n^{-1}(n-m)(m-j)] & \text{if } j < m \leqslant i, \\ m(n-i-1)(n-i)[j-1+n^{-1}(n-j)(m-j)] & \text{if } m \leqslant j \leqslant i. \end{cases}$$

Case B. Having an additional information, the ordered sequence of the values $z_1, \dots, z_N$ is used instead of $x_1 \leqslant \dots \leqslant x_N$ in expressions (3.13) and (3.14) of the true parameters α and κ. Denote the resulting estimates by ${}_z\hat{\alpha}_G$ and ${}_z\hat{\kappa}_G$.

Sample variance. Case A. From population parameters (3.15) and (3.16) we have the following plug-in estimators:

$$\hat{\alpha}_V = \hat{\sigma}_{1V}^{-3} \frac{1}{n^3} \left(\frac{N}{N-2} \right)^3 \left(2m_2^3 - 3m_4 m_2 + m_6 \right) \tag{4.7}$$

and

$$\hat{\kappa}_V = \hat{\sigma}_{1V}^{-3} \tau^2 \frac{2}{n^3(n-1)} \left(\frac{N}{N-2} \right)^3 \frac{1}{N-1} \left(-(N-2)m_3^2 - \frac{2N-1}{N-1} m_4 m_2 + \frac{N}{N-1} m_2^3 + m_6 \right). \tag{4.8}$$

Case B. In (3.15) and (3.16), the central population moments μ_k are evaluated using the values $z_1, \dots, z_N$. Then, denote the new estimates by ${}_z\hat{\alpha}_V$ and ${}_z\hat{\kappa}_V$.

5 Empirical Edgeworth and bootstrap approximations

Replacing the population parameters α and κ in Edgeworth expansion (3.12) with their estimators, we obtain the so-called empirical Edgeworth expansion. If particular estimators of the parameters are asymptotically consistent, then, under the conditions of Theorem 2, the empirical Edgeworth expansion approximates distribution function (3.9) with an error of the same order, but in probability. Bloznelis [4] constructed consistent jackknife estimators of the parameters. Bootstrap and calibrated estimators of the parameters were considered by Čiginas [14] and Pumputis and Čiginas [28], respectively. Here, for each of the statistics U_G and U_V, we have two new versions of the empirical Edgeworth expansion.

GMD statistic. By the results in Section 4.2 we have the empirical Edgeworth expansion

$$\widehat{H}_{nSG}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)\hat{\alpha}_G + 3(y^2+1)\hat{\kappa}_G}{6\tau} \varphi(y), \tag{5.1}$$

and, in the case where the auxiliary information is available, the approximation is

$${}_z\widehat{H}_{nSG}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)_z\hat{\alpha}_G + 3(y^2+1)_z\hat{\kappa}_G}{6\tau} \varphi(y), \tag{5.2}$$

which is not a random function because the values of the variable z are treated as fixed in the population.

Sample variance. The corresponding approximations to the distribution function of the Studentized sample variance are

$$\widehat{H}_{nSV}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)\hat{\alpha}_V + 3(y^2+1)\hat{\kappa}_V}{6\tau} \varphi(y) \tag{5.3}$$

and

$${}_z\widehat{H}_{nSV}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)_z\hat{\alpha}_V + 3(y^2+1)_z\hat{\kappa}_V}{6\tau} \varphi(y), \tag{5.4}$$

where the latter does not depend on the sample.

Estimators of the parameters α and κ in expansion (5.3) are asymptotically consistent under the conditions of Theorem 2. The efficiency of the other empirical Edgeworth expansions is examined in the simulation study in Section 6.

It is known that, in general, nonparametric bootstrap approximations to distributions of statistics are usually of a similar accuracy as one-term Edgeworth expansions. We consider here the finite-population bootstrap scheme introduced by Booth et al. [9]. We apply the results of Bloznelis [6], where the accuracy of this bootstrap method is considered for U-statistics.

The bootstrap approximation to distribution (3.9) is constructed as follows. Write $N = kn + l$, where $0 \leqslant l < n$. Then, given the sample $\mathbb{X}$, the empirical population $\widetilde{\mathcal{X}}$ of size N is formed by taking k copies of $\mathbb{X}$ and, if $l > 0$, adding the remaining l values, which are the simple random sample $\mathbb{Y} = \{Y_1, \dots, Y_l\}$ drawn without replacement from the set $\mathbb{X}$. With this particular bootstrap population $\widetilde{\mathcal{X}}$, we can turn already to the estimator of (3.9), despite that it is only one of $\binom{n}{l}$ empirical populations. Next, we draw the simple random sample $\widetilde{\mathbb{X}} = \{\widetilde{X}_1, \dots, \widetilde{X}_n\}$ without replacement from $\widetilde{\mathcal{X}}$. Denote by $\widetilde{U} = U_n(\widetilde{\mathbb{X}})$ the bootstrap estimator for the statistic of interest and introduce the corresponding jackknife estimator $\widetilde{S}^2 = S^2(\widetilde{\mathbb{X}})$ of the variance of $\widetilde{U}$ under the given population $\widetilde{\mathcal{X}}$. Then the bootstrap approximation to (3.9) is

$$\widetilde{F}_{nS}(y) = \mathbf{P}\left\{\widetilde{U} - \mathbf{E}(\widetilde{U} \mid \mathbb{X}, \mathbb{Y}) \leqslant y\widetilde{S} \;\middle|\; \mathbb{X}\right\}, \tag{5.5}$$

which averages over all possible empirical populations. The following theorem is on the validity of this approximation for the statistics U_G and U_V.

Theorem 3. *Assume that the conditions of Theorem* 2 *are satisfied. Then, we have*

$$\sup_{y\in\mathbb{R}} \left|F_{nS}(y) - \widetilde{F}_{nS}(y)\right| = o_P\big(n_*^{-1/2}\big) \quad \textit{as } n_* \to \infty$$

for U_G and U_V.

Denote by $\widetilde{F}_{nSG}(y)$ and $\widetilde{F}_{nSV}(y)$ the bootstrap approximations for the statistics U_G and U_V, respectively.

6 Numerical modeling

In this section, we illustrate the theoretical results on the second-order approximations to distribution functions of the Studentized GMD statistic and the Studentized sample variance by numerical examples according to the data framework in Section 2.2. Thus, we also consider how outliers affect these approximations.

For the statistics U_G and U_V, we denote their "exact" distribution functions by $F_{nSG}(y)$ and $F_{nSV}(y)$, respectively. In the simulation experiments, these functions were evaluated by the Monte Carlo method, drawing independently 10^6 samples without replacement from the population and using all the values $\mathcal{X}$, as well as their bootstrap approximations, based on the one (since $N = kn$) empirical population $\widetilde{\mathcal{X}}$ constructed from a particular sample $\mathbb{X}$. We denote true Edgeworth approximations (3.12) of the statistics by $H_{nSG}(y)$ and $H_{nSV}(y)$, respectively. To measure the efficiency of empirical Edgeworth approximations $\widehat{H}_{nSG}(y)$ and $\widehat{H}_{nSV}(y)$ and of the bootstrap approximations $\widetilde{F}_{nSG}(y)$ and $\widetilde{F}_{nSV}(y)$, 10^3 samples without replacement were drawn from the population independently.

More specifically, in Tables 3–10, the "exact" distribution functions of the statistics, their normal approximations, the true one-term Edgeworth expansions, the corresponding estimated Edgeworth approximations of two types, and the bootstrap approximations are represented by several commonly used q-quantiles, $q = 0.01, 0.05, 0.10, 0.90, 0.95, 0.99$. For the approximations with the quantiles depending on the sample, we present two characteristics of the efficiency, the empirical expectations $\hat{\mathbf{E}}(\cdot)$ and standard errors $\hat{\mathbf{S}}(\cdot)$ from the realizations of these quantiles.

Tables 3–6 present the results of approximations where there are no outliers (the case of $p/N = 0$) in the same underlying populations generated from the normal and gamma distributions in Section 2.2. The correlation is $\rho_{zx} = 0.7$.

In Table 3, the true Edgeworth approximation $H_{nSG}(y)$ improves substantially the normal approximation to $F_{nSG}(y)$. With the help of the auxiliary information, $H_{nSG}(y)$ is estimated well by ${}_z\widehat{H}_{nSG}(y)$. The bias of this estimate is small in comparison with a possible error of the estimator $\widehat{H}_{nSG}(y)$. But the latter improves the normal approximation to the distribution of U_G too. Unlike all other approximations, the bootstrap

Table 3. Approximations to $F_{nSG}(y)$ under $\mathcal{N}(0,1)$ with 0% outliers from $\mathcal{N}(0,9)$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F_{nSG}^{-1}(q)\approx$	−2.592	−1.779	−1.363	1.223	1.546	2.157
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H_{nSG}^{-1}(q)\approx$	−2.528	−1.762	−1.357	1.214	1.536	2.096
${}_z\widehat{H}_{nSG}^{-1}(q)\approx$	−2.513	−1.752	−1.350	1.220	1.545	2.116
$\hat{\mathbf{E}}\,\widehat{H}_{nSG}^{-1}(q)\approx$	−2.519	−1.756	−1.353	1.217	1.541	2.107
$\hat{\mathbf{S}}\,\widehat{H}_{nSG}^{-1}(q)\approx$	0.034	0.023	0.016	0.013	0.021	0.044
$\hat{\mathbf{E}}\,\widetilde{F}_{nSG}^{-1}(q)\approx$	−2.600	−1.776	−1.360	1.222	1.555	2.167
$\hat{\mathbf{S}}\,\widetilde{F}_{nSG}^{-1}(q)\approx$	0.075	0.038	0.027	0.020	0.026	0.044

Table 4. Approximations to $F_{nSV}(y)$ under $\mathcal{N}(0,1)$ with 0% outliers from $\mathcal{N}(0,9)$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F_{nSV}^{-1}(q)\approx$	−2.918	−1.962	−1.477	1.160	1.461	2.008
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H_{nSV}^{-1}(q)\approx$	−2.680	−1.882	−1.447	1.145	1.432	1.878
${}_z\widehat{H}_{nSV}^{-1}(q)\approx$	−2.658	−1.864	−1.433	1.155	1.447	1.910
$\hat{\mathbf{E}}\,\widehat{H}_{nSV}^{-1}(q)\approx$	−2.653	−1.861	−1.431	1.157	1.450	1.917
$\hat{\mathbf{S}}\,\widehat{H}_{nSV}^{-1}(q)\approx$	0.046	0.038	0.029	0.020	0.032	0.068
$\hat{\mathbf{E}}\,\widetilde{F}_{nSV}^{-1}(q)\approx$	−2.914	−1.932	−1.460	1.166	1.473	2.027
$\hat{\mathbf{S}}\,\widetilde{F}_{nSV}^{-1}(q)\approx$	0.147	0.074	0.046	0.023	0.031	0.050

Table 5. Approximations to $F_{nSG}(y)$ under $\mathcal{G}(3,1/\sqrt{3})$ with 0% outliers from $\mathcal{G}(3,\sqrt{3})$, and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F_{nSG}^{-1}(q)\approx$	−2.888	−1.903	−1.443	1.188	1.503	2.062
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H_{nSG}^{-1}(q)\approx$	−2.638	−1.843	−1.413	1.172	1.468	1.946
${}_z\widehat{H}_{nSG}^{-1}(q)\approx$	−2.572	−1.793	−1.378	1.198	1.510	2.038
$\hat{\mathbf{E}}\,\widehat{H}_{nSG}^{-1}(q)\approx$	−2.624	−1.833	−1.407	1.177	1.476	1.965
$\hat{\mathbf{S}}\,\widehat{H}_{nSG}^{-1}(q)\approx$	0.055	0.045	0.033	0.024	0.037	0.079
$\hat{\mathbf{E}}\,\widetilde{F}_{nSG}^{-1}(q)\approx$	−2.864	−1.899	−1.436	1.190	1.506	2.079
$\hat{\mathbf{S}}\,\widetilde{F}_{nSG}^{-1}(q)\approx$	0.156	0.077	0.048	0.025	0.035	0.060

Table 6. Approximations to $F_{nSV}(y)$ under $\mathcal{G}(3,1/\sqrt{3})$ with 0% outliers from $\mathcal{G}(3,\sqrt{3})$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F_{nSV}^{-1}(q)\approx$	−3.744	−2.310	−1.699	1.109	1.391	1.876
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H_{nSV}^{-1}(q)\approx$	−2.829	−2.025	−1.561	1.077	1.324	1.656
${}_z\widehat{H}_{nSV}^{-1}(q)\approx$	−2.796	−1.991	−1.533	1.091	1.347	1.704
$\hat{\mathbf{E}}\,\widehat{H}_{nSV}^{-1}(q)\approx$	−2.788	−1.985	−1.529	1.097	1.355	1.719
$\hat{\mathbf{S}}\,\widehat{H}_{nSV}^{-1}(q)\approx$	0.075	0.077	0.067	0.040	0.060	0.114
$\hat{\mathbf{E}}\,\widetilde{F}_{nSV}^{-1}(q)\approx$	−3.642	−2.267	−1.655	1.120	1.406	1.909
$\hat{\mathbf{S}}\,\widetilde{F}_{nSV}^{-1}(q)\approx$	0.487	0.277	0.166	0.033	0.047	0.082

Table 7. Approximations to $F_{nSG}(y)$ under $\mathcal{N}(0,1)$ with 6% outliers from $\mathcal{N}(0,9)$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F^{-1}_{nSG}(q)\approx$	−3.133	−2.061	−1.553	1.143	1.434	1.953
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H^{-1}_{nSG}(q)\approx$	−2.745	−1.940	−1.490	1.118	1.388	1.783
${}_z\widehat{H}^{-1}_{nSG}(q)\approx$	−2.602	−1.816	−1.395	1.184	1.489	1.996
$\hat{\mathbf{E}}\,\widehat{H}^{-1}_{nSG}(q)\approx$	−2.713	−1.913	−1.470	1.132	1.410	1.831
$\hat{\mathbf{S}}\,\widehat{H}^{-1}_{nSG}(q)\approx$	0.069	0.063	0.050	0.033	0.051	0.103
$\hat{\mathbf{E}}\,\widetilde{F}^{-1}_{nSG}(q)\approx$	−3.124	−2.039	−1.524	1.152	1.449	1.982
$\hat{\mathbf{S}}\,\widetilde{F}^{-1}_{nSG}(q)\approx$	0.267	0.139	0.087	0.030	0.043	0.076

Table 8. Approximations to $F_{nSV}(y)$ under $\mathcal{N}(0,1)$ with 6% outliers from $\mathcal{N}(0,9)$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F^{-1}_{nSV}(q)\approx$	−4.724	−2.924	−2.100	1.037	1.280	1.699
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H^{-1}_{nSV}(q)\approx$	−2.985	−2.207	−1.740	0.979	1.181	1.415
${}_z\widehat{H}^{-1}_{nSV}(q)\approx$	−2.861	−2.062	−1.596	1.054	1.291	1.602
$\hat{\mathbf{E}}\,\widehat{H}^{-1}_{nSV}(q)\approx$	−2.889	−2.097	−1.634	1.036	1.266	1.560
$\hat{\mathbf{S}}\,\widehat{H}^{-1}_{nSV}(q)\approx$	0.086	0.098	0.095	0.050	0.075	0.130
$\hat{\mathbf{E}}\,\widetilde{F}^{-1}_{nSV}(q)\approx$	−4.600	−2.792	−1.974	1.069	1.333	1.787
$\hat{\mathbf{S}}\,\widetilde{F}^{-1}_{nSV}(q)\approx$	1.064	0.663	0.430	0.042	0.057	0.095

Table 9. Approximations to $F_{nSG}(y)$ under $\mathcal{G}(3,1/\sqrt{3})$ with 6% outliers from $\mathcal{G}(3,\sqrt{3})$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F^{-1}_{nSG}(q)\approx$	−3.224	−2.068	−1.546	1.148	1.445	1.966
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H^{-1}_{nSG}(q)\approx$	−2.740	−1.933	−1.483	1.124	1.395	1.795
${}_z\widehat{H}^{-1}_{nSG}(q)\approx$	−2.601	−1.815	−1.394	1.186	1.491	1.997
$\hat{\mathbf{E}}\,\widehat{H}^{-1}_{nSG}(q)\approx$	−2.717	−1.915	−1.470	1.134	1.410	1.827
$\hat{\mathbf{S}}\,\widehat{H}^{-1}_{nSG}(q)\approx$	0.064	0.060	0.049	0.032	0.048	0.097
$\hat{\mathbf{E}}\,\widetilde{F}^{-1}_{nSG}(q)\approx$	−3.213	−2.057	−1.531	1.154	1.453	1.990
$\hat{\mathbf{S}}\,\widetilde{F}^{-1}_{nSG}(q)\approx$	0.268	0.133	0.083	0.029	0.043	0.075

Table 10. Approximations to $F_{nSV}(y)$ under $\mathcal{G}(3,1/\sqrt{3})$ with 6% outliers from $\mathcal{G}(3,\sqrt{3})$ and $\rho_{zx}=0.7$

$q=$	0.01	0.05	0.10	0.90	0.95	0.99
$F^{-1}_{nSV}(q)\approx$	−4.895	−2.890	−2.042	1.045	1.296	1.725
$\Phi^{-1}(q)\approx$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$H^{-1}_{nSV}(q)\approx$	−2.978	−2.198	−1.728	0.988	1.193	1.430
${}_z\widehat{H}^{-1}_{nSV}(q)\approx$	−2.864	−2.064	−1.597	1.055	1.292	1.600
$\hat{\mathbf{E}}\,\widehat{H}^{-1}_{nSV}(q)\approx$	−2.882	−2.087	−1.622	1.045	1.277	1.577
$\hat{\mathbf{S}}\,\widehat{H}^{-1}_{nSV}(q)\approx$	0.083	0.095	0.092	0.050	0.073	0.127
$\hat{\mathbf{E}}\,\widetilde{F}^{-1}_{nSV}(q)\approx$	−4.782	−2.769	−1.944	1.079	1.347	1.809
$\hat{\mathbf{S}}\,\widetilde{F}^{-1}_{nSV}(q)\approx$	1.184	0.651	0.416	0.041	0.058	0.097

approximation $\widetilde{F}_{nSG}(y)$ is almost unbiased, but its empirical quantiles have larger standard errors as compared to the empirical Edgeworth approximation. In Table 4, tendencies of the approximations to the distribution function of U_V are the same. In Tables 5–6, for the population from the gamma distribution, the results are analogous to those in Tables 3–4, but all the corresponding approximations are less accurate. This is because of the asymmetry of the gamma distribution.

Let us take the populations of Section 2.2 with $p/N = 0.06$. In this case of outliers, the results corresponding to Tables 3–6 are given in Tables 7–10. The behavior of approximations to the distributions is very similar to that in the case of no outliers, but the errors of the approximations are now greater. We can also notice that the estimates of the true Edgeworth expansions that use the auxiliary information are much more biased. This holds for the alternative empirical Edgeworth approximations too, but only in the case of the statistic U_V (Tables 8 and 10). The sensitivity to outliers is the smallest one, as is seen by comparing Table 9 with Table 5.

7 Summary

The estimation strategies for scales are presented for simple random samples without replacement. Strategies (S_1) and (S_2) are consistent with the scale measurement by the variance of the population because they combine the use of the GMD statistic and its bias correction, evaluated applying specific assumptions used for finite populations. It is a known fact, confirmed by the new connection (2.1), that the GMD statistic is less sensitive to extreme sample observations than the empirical variance. However, the GMD statistic is not the best robust choice, in general, but the proposed strategies improve upon the sample variance, where the part of representative outliers is not large in the population. As indicated by the numerical modeling, under the normal population with no outliers, where the empirical variance is known as the most efficient estimator, the efficiency of the strategies is not worse. It is an important property of robustness.

Asymptotic properties of the statistics are collected applying recent general results on the U- and L-statistics. In addition, if to compare conditions sufficient for the asymptotic normality of the statistics U_G and $\sqrt{U_V}$, then, by Theorem 1 and analogously to the i.i.d. case, finite $\mathbf{E}\, X_1^4$ should be sufficient for $\sqrt{U_V}$, but it suffices to have bounded $\mathbf{E}\, X_1^2$ in the case of U_G. This is one more advantage of the GMD statistic.

Using the detailed decompositions of the statistics, explicit formulas of variances and Edgeworth expansion parameters of the statistics are derived. These formulas allow us to obtain new estimators of the plug-in type. The resulting empirical Edgeworth expansions and bootstrap approximations improve the normal approximation as stated in Theorems 2 and 3 and shown in the simulations.

The paper spotlights the use of the GMD statistic instead of the common empirical variance. In surveys of finite populations, the estimation of scales is also topical under more complex sampling schemes than the simple random sampling. However, then analysis of distributional properties of statistics is much more complicated, even for stratified simple random samples.

Appendix A: Proof of Theorem 1

To be consistent with conditions imposed on symmetric (and thereby U-) statistics by Bloznelis and Götze [8], consider the normalized versions of the statistics of interest, $\sqrt{n}U_G$ and $\sqrt{n}U_V$. Then the variances of linear parts of decompositions of these statistics are bounded away from zero and are finite if the corresponding conditions (i) and (ii) are satisfied. Therefore, in the case of U_G, the normality proof follows immediately from [15, Thm. 1] through [8, Prop. 3]. In the case of U_V, by [8, Thm. 1 and Prop. 3], it suffices to verify that the variance of quadratic part of $\sqrt{n}U_V$ tends to zero as $n_* \to \infty$. If (ii) is satisfied, then it easily follows from the explicit formulas of Section 3.1.

Appendix B: Proof of Theorem 2

In the case of U_G, the proof is a corollary of Theorem 1 in [5], following the technique in the proof of Theorem 1 in [13]. In particular, by these theorems, the boundedness of the characteristics $\beta_s = \sigma_1^{-s}\, \mathbf{E}\, |g_1(X_1)|^s$ and $\gamma_s = \sigma_1^{-s}\tau^{2s}\, \mathbf{E}\, |g_2(X_1, X_2)|^s$ as $n_* \to \infty$ must be verified for $s > 6$ only.

In the case of U_V, the task is the same. By (3.5), for $s \geqslant 1$, applying the inequalities $|a-b|^s \leqslant 2^{s-1}(a^s+b^s)$ for $a, b \geqslant 0$ and $\mu_2^s \leqslant \mu_{2s}$, we get

$$\begin{aligned}\mathbf{E}\,|g_1(X_1)|^s &= \frac{1}{N}\sum_{k=1}^{N}\big|g_1(x_k)\big|^s \leqslant \frac{2^{s-1}}{n^s}\bigg(\frac{N}{N-2}\bigg)^s\frac{1}{N}\sum_{k=1}^{N}\big((x_k-b_1)^{2s}+\mu_2^s\big)\\ &\leqslant \bigg(\frac{2N}{n(N-2)}\bigg)^s\mu_{2s}. \end{aligned} \tag{B.1}$$

By (3.6), for $1 \leqslant k < l \leqslant N$, applying $(x_k - x_l)^2 \leqslant 2((x_k-b_1)^2+(x_l-b_1)^2)$, we obtain

$$\begin{aligned}\big|g_2(x_k,x_l)\big| &\leqslant \frac{1}{n(n-1)}\bigg(\frac{3N-4}{N-2}\big((x_k-b_1)^2+(x_l-b_1)^2\big)+\frac{2N}{(N-1)(N-2)}\mu_2\bigg)\\ &\leqslant \frac{3}{n(n-1)}\frac{N}{N-2}\big((x_k-b_1)^2+(x_l-b_1)^2+\mu_2\big).\end{aligned}$$

Then, for $s \geqslant 1$, similarly as in (B.1), applying $(a+b)^s \leqslant 2^{s-1}(a^s+b^s)$ for $a, b \geqslant 0$ twice and noting that $\sum_{1\leqslant k<l\leqslant N}((x_k-b_1)^{2s}+(x_l-b_1)^{2s}) = N(N-1)\mu_{2s}$, we obtain

$$\begin{aligned}\mathbf{E}\,\big|g_2(X_1,X_2)\big|^s &= \binom{N}{2}^{-1}\sum_{1\leqslant k<l\leqslant N}|g_2(x_k,x_l)|^s\\ &\leqslant \frac{3^s}{n^s(n-1)^s}\bigg(\frac{N}{N-2}\bigg)^s\binom{N}{2}^{-1}\sum_{1\leqslant k<l\leqslant N}\big((x_k-b_1)^2+(x_l-b_1)^2+\mu_2\big)^s\\ &\leqslant \frac{3^s2^{s-1}}{n^s(n-1)^s}\bigg(\frac{N}{N-2}\bigg)^s\binom{N}{2}^{-1}\sum_{1\leqslant k<l\leqslant N}\big(2^{s-1}\big((x_k-b_1)^{2s}+(x_l-b_1)^{2s}\big)+\mu_2^s\big)\\ &\leqslant \frac{3^s2^{s-1}(2^s+1)}{n^s(n-1)^s}\bigg(\frac{N}{N-2}\bigg)^s\mu_{2s}.\end{aligned} \tag{B.2}$$

Then we derive from (B.1), (B.2), and (3.7) that

$$\beta_s \leqslant \frac{2^s\mu_{2s}}{(\mu_4-\mu_2^2)^{s/2}} \quad\text{and}\quad \gamma_s \leqslant 3^s2^{2s-1}\big(2^s+1\big)\bigg(1-\frac{n}{N}\bigg)^s\frac{\mu_{2s}}{(\mu_4-\mu_2^2)^{s/2}}.$$

The proof is completed.

Appendix C: Proof of Theorem 3

It follows from condition (8) in [6] that it suffices to verify that, for the statistics U_G and U_V, the moments $\mathbf{E}(X_1-X_2)^6$ and $\mathbf{E}(X_1-X_2)^{12}$ are bounded for all n_*, respectively. By the conditions of the theorem, this requirement holds.

References

1. F. Alqallaf, S. van Aelst, V.J. Yohai, and R.H. Zamar, Propagation of outliers in multivariate data, *Ann. Stat.*, **37**:311–331, 2009.

2. C. Béguin and B. Hulliger, The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data, *Survey Methodology*, **34**:91–103, 2008.

3. P.J. Bickel, F. Götze, and W.R. van Zwet, The Edgeworth expansion for U-statistics of degree two, *Ann. Stat.*, **14**:1463–1484, 1986.
4. M. Bloznelis, Empirical Edgeworth expansion for finite population statistics. I, *Lith. Math. J.*, **41**:120–134, 2001.
5. M. Bloznelis, An Edgeworth expansion for Studentized finite population statistics, *Acta Appl. Math.*, **78**:51–60, 2003.
6. M. Bloznelis, Bootstrap approximation to distributions of finite population U-statistics, *Acta Appl. Math.*, **96**:71–86, 2007.
7. M. Bloznelis and F. Götze, One-term Edgeworth expansion for finite population U-statistics of degree two, *Acta Appl. Math.*, **58**:75–90, 1999.
8. M. Bloznelis and F. Götze, Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics, *Ann. Stat.*, **29**:899–917, 2001.
9. J.G. Booth, R.W. Butler, and P. Hall, Bootstrap methods for finite populations, *J. Am. Stat. Assoc.*, **89**:1282–1289, 1994.
10. R.L. Chambers, Outlier robust finite population estimation, *J. Am. Stat. Assoc.*, **81**:1063–1069, 1986.
11. R.L. Chambers, A. Hentges, and X. Zhao, Robust automatic methods for outlier and error detection, *J. R. Stat. Soc., Ser. A, Stat. Soc*, **167**:323–339, 2004.
12. R.S. Chhikara and A.L. Feiveson, Extended critical values of extreme studentized deviate test statistics for detecting multiple outliers, *Commun. Stat., Simulation Comput.*, **9**:155–166, 1980.
13. A. Čiginas, An Edgeworth expansion for finite-population L-statistics, *Lith. Math. J.*, **52**:40–52, 2012.
14. A. Čiginas, Second-order approximations of finite population L-statistics, *Statistics*, **47**:954–965, 2013.
15. A. Čiginas, On the asymptotic normality of finite population L-statistics, *Stat. Pap.*, **55**:1047–1058, 2014.
16. P. Erdős and A. Rényi, On the central limit theorem for samples from a finite population, *Publ. Math. Inst. Hung. Acad. Sci.*, **4**:49–61, 1959.
17. C. Gerstenberger and D. Vogel, On the efficiency of Gini's mean difference, pp. 1–18, 2014, arXiv:1405.5027.
18. C. Gini, *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche*, Cuppini, Bologna, 1912.
19. G.J. Glasser, Variance formulas for the mean difference and coefficient of concentration, *J. Am. Stat. Assoc.*, **57**:648–654, 1962.
20. J. Hájek, Limiting distributions in simple random sampling from a finite population, *Publ. Math. Inst. Hung. Acad. Sci.*, **5**:361–374, 1960.
21. R. Helmers, On the Edgeworth expansion and the bootstrap approximation for a Studentized U-statistic, *Ann. Stat.*, **19**:470–484, 1991.
22. W. Hoeffding, A class of statistics with asymptotically normal distribution, *Ann. Math. Stat.*, **19**:293–325, 1948.
23. P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.
24. J.O. Irwin and M.G. Kendall, Sampling moments of moments for a finite population, *Annals of Eugenics*, **12**:138–142, 1944.
25. P.N. Kokic and N.C. Weber, An Edgeworth expansion for U-statistics based on samples from finite populations, *Ann. Probab.*, **18**:390–404, 1990.
26. Z.A. Lomnicki, The standard error of Gini's mean difference, *Ann. Math. Stat.*, **23**:635–637, 1952.

27. U.S. Nair, The standard error of Gini's mean difference, *Biometrika*, **28**:428–436, 1936.

28. D. Pumputis and A. Čiginas, Estimation of parameters of finite population L-statistics, *Nonlinear Anal. Model. Control*, **18**:327–343, 2013.

29. H. Putter and W.R. Zwet, Empirical Edgeworth expansions for symmetric statistics, *Ann. Stat.*, **26**:1540–1569, 1998.

30. R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.

31. S. Yitzhaki and E. Schechtman, *The Gini Methodology: A Primer on a Statistical Methodology*, Springer, New York, 2013.

32. L.C. Zhao and X.R. Chen, Normal approximation for finite-population U-statistics, *Acta Math. Appl. Sin.*, **6**:263–272, 1990.