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

Andrius Ėiggins¹ and Dalius Pumputis²

¹Vilnius University Institute of Mathematics and Informatics, LT-08663 Vilnius, Lithuania
²Lithuanian University of Educational Sciences, LT-08106 Vilnius, Lithuania

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 under possible presence of outliers in the data.

The paper presents also a wide comparative survey of properties of the Gini 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. An estimation of the variances and other parameters of the statistics is also in the study, where we exploit an auxiliary information on the population elements in the case of its availability. Theoretical results are illustrated with a simulation study.

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

MSC classes: 62E20

1 Introduction

Together with a location parameter, a spread (or scale) of a survey population are usually the parameters of interest. If a statistician assumes the classical model of independent and identically distributed (i.i.d.) observations, then, at least, he has at his disposal the number of parametric distributions families, e.g., Gaussian, Cauchy, etc. Assume that, he chooses the particular family for a further analysis of the data. This family comes with its own measures of location and scale, 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 has the answers how to get efficient estimates of locations and scales under commonly used populations models. However, parametric statistics models, being comparatively convenient, are known also as non-robust, i.e., deviations from their assumptions may lead to misleading conclusions. As it is often an instance, an

---

¹ The research of the first author is supported by European Union Structural Funds project "Postdoctoral Fellowship Implementation in Lithuania".
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, e.g., measurement errors, robust estimation methods can be a treatment of the problem. The pioneering (in formalization of the robust estimation) book of [Huber, 1981] starts with the normal distribution scale estimation example showing an inefficiency of the empirical variance compared to the mean absolute deviation under the presence of outliers in the sample data. In a sense similar to the latter statistic is Gini’s mean difference (GMD) statistic. This estimator, its properties, connections and comparisons with the sample variance is the aim of the present paper.

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

\[
G = \left(\frac{N}{2}\right)^{-1} \sum_{1 \leq i < j \leq N} |x_i - x_j|
\]

and

\[
V = \left(\frac{N}{2}\right)^{-1} \sum_{1 \leq i < j \leq N} (x_i - x_j)^2 / 2
\]

are two candidates to measure a scale of \(X\), only the latter seems more natural because of \(\text{Var} X_1 = (N - 1)V / N\). The corresponding unbiased estimators of these parameters are the GMD statistic

\[
U_G = \left(\frac{n}{2}\right)^{-1} \sum_{1 \leq i < j \leq n} |X_i - X_j|
\]

and the empirical variance

\[
U_V = \left(\frac{n}{2}\right)^{-1} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 / 2.
\]

As an alternative to (1.4), the GMD statistic, known better since Gini [1912], is widely used in economics. Now it is an ordinary measure of a dispersion of a distribution of income and also in cases of similar variables, see monograph of [Yitzhaki and Schechtman, 2013], where, by words of the authors, the commonly used variance-based analyses are ‘translated’ into Gini-based. A use of the GMD is not restricted with measurements of an economic inequality. As in problems of economists, where data deviate from the normality, the parameter \(G\) and its estimator \(U_G\) can be used as dispersion’s measures for many kinds of statistical data. Our choice of the finite populations setting has a motivation from the side of economics too, because, in economical surveys, the number \(N\) of surveyed objects or subjects is not necessarily so large (compared to the sample size) that to ignore a dependence between the observations in the set \(X\).

In Section 2 we consider three estimation of the finite population scale strategies related to the alternative \(U_G\). We exploit two assumptions, which are usually possible in the finite population context: the so-called superpopulation assumption, and the availability of an auxiliary information about the population elements. We perform also 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 as, e.g., the median absolute deviation, interquartile range, which are less sensitive to outliers than the sample variance. Looking from the side of the robust estimation theory, if we can link data to a parametric population model, then, in the particular situations, there are more effective robust estimators of scale than those common ones, see Huber (1981). But we focus here on an unified improvement of the empirical variance.

The next premium, which should be paid, is a relatively complex access to 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 it, 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 (1980). As the L-statistic, U_G is smooth in the sense that its weight function is smooth, see ibidem.

To be consistent with already known results, first, we mark that expressions of the variance of U_G and its approximations are known since Nair (1936) and Lomnicki (1952) in the case of i.i.d. observations, and since Glasser (1962) for the simple random samples without replacement. Second, a strong method to study the variances and the asymptotic normality of the statistics U_G and U_V is Hoeffding’s decomposition for U-statistics in Hoeffding (1948). Much latter, in Zhao and Chen (1990), the analogous decomposition was used in the case of finite population. Third, similarly, second-order approximations theory for samples without replacement has been realized after the case of i.i.d. observations: Kokic and Weber (1990), and Bloznelis and Götze (1999) follow Bickel et al. (1986) on one-term Edgeworth approximations to the distributions of standardized U-statistics; papers of Bloznelis (2003), and Bloznelis (2007), on an one-term Edgeworth expansion for Studentized U-statistics and bootstrap approximations, appeared after Helmers (1991).

Since the true values of variances of the statistics are almost always unknown, we prefer to consider 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 in Bloznelis and Götze (2001), Bloznelis (2003), and Bloznelis (2007) (with without-replacement bootstrap of Booth et al. (1994)), where, to ensure a validity of the approximations, quite general smoothness conditions are imposed on parts of the Hoeffding decomposition of U-statistics. Theorems of Section 3 let to compare distributional properties of U_G and U_V much easy.

A successful application of the one-term Edgeworth expansion requires to have good estimators (in the sense of an asymptotic consistency or a small mean square error) of the expansion’s 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 (1998), and Bloznelis (2001). In the separate cases of statistics, for example, for U_G and U_V, there are more ways to construct estimators of the Edgeworth expansions parameters, e.g., for L-statistics including U_G, the bootstrap was used in Čiginas (2013a) and, assuming that the auxiliary information is available, calibration methods were applied in Pumputis and Čiginas (2013). In Section 4 we propose simple and also efficient estimators of the parameters, without the auxiliary information and also using it. Similar estimators of the variances of U_G and U_V are also considered. In Section 5 we discuss empirical Edgeworth expansions, based on the estimators of the 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 interested also in a
role of outliers in populations. Conclusions of the paper are given in Section 7.

2 Estimation of scale

2.1 Outliers and estimation strategies

In the i.i.d. setup, for many common parametric models of populations, the sample variance is an efficient estimator under ideal or close to ideal conditions. But assume that 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 thus lowering their impact on the estimation.

In the finite population case, outliers are less influential too, when $U_G$ is applied. To see it, let us write parameters (1.1) and (1.2) in the different form. Assume (here and further in the paper), without loss of generality, that $x_1 \leq \ldots \leq x_N$, and denote

$I_{1} = \sum_{i=1}^{N} \Delta_i = x_{i+1} - x_i, \quad i = 1, \ldots, N - 1.$

Then, taking $x_j - x_i = \sum_{k=i}^{j-1} \Delta_k$, one can obtain

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

and

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

These expressions are connected via the formal transformation

$$\Delta'_i \Delta'_j = \frac{4 j(N-i)}{N(N-1)} \Delta_i \Delta_j, \quad 1 \leq i \leq j \leq N - 1$$

of $X$, where $\Delta'_i = x'_{i+1} - x'_i, \quad i = 1, \ldots, N - 1$, which explains the assertion. We note that system of equations (2.1) has not a solution $X'$ except in cases of very simple $X$.

**Outliers model.** For the simple random samples without replacement, we assume an existence of so-called representative outliers. This notion was introduced in Chambers (1986). It means the assumptions that: outlying observations are not errors of a measurement; the unsampled population part should contain outliers too. If these assumptions do not hold, then, in sample surveys, the problem of outliers is treated usually as a different from the estimation.

More formally, denote by $0 \leq p \leq N$ 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 while the sample $X$ was not obtained. Then the corresponding values from $x_1, \ldots, x_N$ are treated as outliers. In the random sample $X$, the number of outliers is random and equals to 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 in Huber (1981), a part of gross errors (outliers) in samples usually is not larger than $10\%$. An interesting note on this issue is given in Chhikara and Feiveson (1980): “... 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
Estimation strategies. Specific for the finite population ways to apply the GMD statistic as the alternative to the sample variance are the following.

1. Assume that the fixed numbers $x_1, \ldots, x_N$ are the realizations of i.i.d. random variables $X_1^*, \ldots, X_N^*$ (superpopulation model) from a parametric family of distributions with the scale parameter which is an one-argument function of $\sqrt{\text{Var} X_1^*}$. Then the scale of $X$ 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.

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

3. The parameter $G$ is itself treated as the scale of $X$, and the GMD statistic $U_G$ is its estimator.

Case (S1) is close to the parametric statistics. In the i.i.d. settings, the multipliers $a$, which ensure that $aU_G$ is the unbiased estimator of $\sqrt{V}$, are known for commonly used parametric families: $a = \sqrt{\pi}/2$ for the normal distributions; $a = 1$ for the exponential distributions; etc. Therefore, assuming an existence of the superpopulation, we use the same constants for the estimation in the finite population. If a good auxiliary information $Z$ is available, then these theoretical $a$ should not so much differ from the corresponding values obtained by case (S2), where $a > 0$ is evaluated from $aG = \sqrt{V}$ using $Z$ instead of $X$. If the scatter of $X$ can not be linked to a distributions family, e.g., it is a mixture of two unknown distributions, and there is no other additional information, then we suggest strategy (S3).

2.2 Numerical analysis

We compare efficiencies of strategies (S1) and (S2) in respect of the common estimation by $\sqrt{U_V}$ under presence of outliers. We consider two populations which values of the variable $x$ are generated respectively from two different parametric families: the normal distributions $N(\mu, \sigma^2)$, and the gamma distributions $G(k, \theta)$ with the shape $k$ and scale $\theta$, where variance is equal to $k\theta^2$. In the case of gamma distribution, the correction $a = k^{-1/2}(2 - 4I_{0.5}(k + 1, k))^{-1}$ depends on $k$, where $I_{u}(u, v)$ is the regularized incomplete Beta function. For each of these populations, we consecutively increase the part of outliers in the population as follows. Firstly, we select some particular population elements randomly without replacement. Secondly, we replace their values by new 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 increased by selecting from those which still not belong to the outliers.

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

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

To understand better a role of the auxiliary information in strategy (S2), we simulate different cor-
relations $\rho_{zx}$ between $Z$ and $X$. The correlation is controlled with the variance $\theta^2$ in the linear model. Thus we choose the variance in order to have $\rho_{zx} = 0.9, 0.7, 0.5$ approximately. Tables 1, 2 present the comparison of the estimation methods by means of mean square errors and biases.

| $p/N$ | $\sqrt{U}$ | $N(0,1)$ | $\rho_{zx} = 0.9; (S_1)$ | $\rho_{zx} = 0.7; (S_2)$ | $\rho_{zx} = 0.5; (S_2)$ |
|-------|-------------|----------|------------------------|------------------------|------------------------|
| 0.00  | (-0.01, 0.48) | (-0.06, 0.47) | (-0.02, 0.47) | (0.00, 0.47) | (-0.05, 0.47) |
| 0.02  | (-0.03, 0.86) | (-0.41, 0.73) | (-0.13, 0.63) | (-0.32, 0.68) | (-0.40, 0.72) |
| 0.04  | (-0.03, 0.99) | (-0.67, 0.97) | (-0.16, 0.75) | (-0.44, 0.84) | (-0.63, 0.95) |
| 0.06  | (-0.03, 1.10) | (-0.92, 1.22) | (-0.24, 0.87) | (-0.71, 1.07) | (-0.86, 1.17) |
| 0.08  | (-0.03, 1.10) | (-0.95, 1.26) | (-0.28, 0.91) | (-0.58, 1.02) | (-0.92, 1.23) |
| 0.10  | (-0.03, 1.22) | (-1.18, 1.48) | (-0.21, 0.98) | (-0.91, 1.29) | (-1.06, 1.39) |

Table 2: $G(3, \sqrt{3})$ with outliers $G(3, \sqrt{3})$. Accuracy by $10 \times (BIAS(\cdot), \sqrt{MSE(\cdot)})$.

| $p/N$ | $\sqrt{U}$ | $N(0,1)$ | $\rho_{zx} = 0.9; (S_1)$ | $\rho_{zx} = 0.7; (S_2)$ | $\rho_{zx} = 0.5; (S_2)$ |
|-------|-------------|----------|------------------------|------------------------|------------------------|
| 0.00  | (-0.03, 0.71) | (-0.18, 0.64) | (-0.21, 0.65) | (-0.42, 0.73) | (-0.60, 0.84) |
| 0.02  | (-0.09, 1.40) | (-1.09, 1.40) | (-0.48, 1.04) | (-1.05, 1.36) | (-1.48, 1.70) |
| 0.04  | (-0.08, 1.42) | (-1.24, 1.55) | (-0.66, 1.18) | (-1.21, 1.53) | (-1.69, 1.91) |
| 0.06  | (-0.07, 1.47) | (-1.56, 1.87) | (-0.76, 1.35) | (-1.71, 1.99) | (-1.98, 2.23) |
| 0.08  | (-0.12, 1.93) | (-2.24, 2.56) | (-1.08, 1.71) | (-2.06, 2.41) | (-2.70, 2.94) |
| 0.10  | (-0.13, 1.99) | (-2.56, 2.89) | (-1.31, 1.97) | (-2.54, 2.88) | (-2.85, 3.14) |

It is seen from Table 1 that strategy $(S_1)$ improves the estimator $\sqrt{U}$ where the proportion $p/N$ is smaller. For $p/N$ larger than 0.04, $(S_1)$ becomes inefficient (by $MSE(\cdot)$) because its bias is large, since the fixed correction $a$ is too approximate for the mix of the normal distributions. Strategy $(S_2)$ is the best under strong correlation between $x$ and $z$, because the estimation bias is well-corrected. The efficiency of $(S_2)$ decreases with the decrease of the correlation $\rho_{zx}$.

Table 2 shows similar results for the asymmetric gamma distributions. Here outliers affect the estimators stronger because the distribution of outliers has 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 thus the GMD statistic, are efficient, in respect of $\sqrt{U}$, 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.

### 3 Theoretical properties of the statistics

#### 3.1 Hoeffding’s decompositions and variances

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

$$h(X_1, X_2) = \binom{n}{2}^{-1} |X_1 - X_2|$$
structures of the variances of the statistics linear part in (3.1) dominates against the remainder in the sense of variance size. In particular, we consider

The so-called first- and second-order influence functions to the variance of

where

and, for 1 ≤ k ≤ N,

and, for 1 ≤ k ≠ l ≤ N,

The so-called first- and second-order influence functions g₁(·) and g₂(·, ·) have usually a different impact to the variance of U-statistic. As in cases of any other linearization techniques, it is expected that the linear part in (3.1) dominates against the remainder in the sense of variance size. In particular, we consider structures of the variances of the statistics U_G and U_Ω by formula (2.6) in Bloznelis and Götze (2001):

\[
\text{Var} U = \frac{n(n−n)}{N−1} \sigma_1^2 + \left(\frac{n}{2}\right) \left(\frac{N−n}{2}\right) \left(\frac{N−2}{2}\right) \sigma_2^2,
\]

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

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

\[
U_G = \left(\frac{n}{2}\right)^{-1} \sum_{j=1}^{n} (2j−n−1)X_{j:n},
\]

where \(X_{1:n} ≤ \cdots ≤ X_{n:n}\) are the order statistics of the observations \(X\), and apply the Hoeffding decomposition results for L-statistics from Ciginas (2012). Denote \(a_i = (2i−N)/N\), 1 ≤ i ≤ N − 1. Then, for 1 ≤ k ≤ N,

\[
g_1(x_k) = -\frac{2}{n N−2} \sum_{i=1}^{N−1} \left(\mathbb{I}\{i ≥ k\} − \frac{i}{N}\right) a_i \Delta_i,
\]

where \(\mathbb{I}\{·\}\) is the indicator function, and, for 1 ≤ k < l ≤ N,

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

where

\[
\phi_{k,l}(i) = \begin{cases} 
   i(i−1)/A, & \text{if } 1 ≤ i < k, \\
   -(i−1)(N−i−1)/A, & \text{if } k ≤ i < l, \\
   (N−i−1)(N−i)/A, & \text{if } l ≤ i < N,
\end{cases}
\]
with $A = (N - 1)(N - 2)$. Next, direct calculations give the expressions of variance decomposition (3.2) components:

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

(3.3)

and

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

(3.4)

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

$$g_1(x_k) = \frac{1}{nN-2} \left[ (x_k - b_1)^2 - \mu_2 \right],$$

(3.5)

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

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

(3.6)

After straightforward calculations, we obtain the following formulas:

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

(3.7)

and

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

(3.8)

In fact, various expressions of $\text{Var} UV$ are known in the literature. For a comparison, we mention just that appeared in [Irwin and Kendall (1944)].

### 3.2 Asymptotic normality

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

$$F_{nS}(y) = \mathbb{P} \{ U - \mathbb{E}U \leq yS \}$$

(3.9)

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

$$S^2 = S^2(\mathbb{X}) = \left( 1 - \frac{n}{N} \right) \frac{n-1}{n} \sum_{i=1}^{n} (U_{n-1}(\mathbb{X} \setminus X_i) - \bar{U})^2,$$

where $\bar{U} = \frac{1}{n} \sum_{i=1}^{n} U_{n-1}(\mathbb{X} \setminus X_i),$

(3.10)
is the jackknife estimator of the variance for any $U$-statistic.

In the finite populations 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 $X_r = \{x_{r,1}, \ldots, x_{r,N_r}\}$ in the populations, with $N_r \to \infty$ as $r \to \infty$, and a sequence of statistics $U_{n_r}(X_r)$, where $X_r = \{x_{r,1}, \ldots, x_{r,n_r}\}$ is a sample drawn without replacement from $X_r$, should be considered. Further, we omit the subscript $r$ for these and other quantities for notational simplicity.

Denote $\tau^2 = n(1 - n/N)$ for short. Erdős and Rényi (1959), and Hájek (1960) Lindeberg-type condition: for every $\varepsilon > 0$,

$$\sigma_1^{-2} \mathbb{E} g_1^2(X_1) \mathbb{I}\{|g_1(X_1)| > \varepsilon \tau \sigma_1\} = o(1) \quad \text{as} \quad n_* \to \infty,$$

imposed on the linear part of $U$-statistic, is necessary for the normality of asymptotically linear statistics as the size $n_*$ grows. This condition, together with moments 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) for $U_G$, $\mathbb{E} X_1^2 \leq C_1 < \infty$ holds; (ii) for $U_V$, $\mathbb{E} X_1^4 \leq C_2 < \infty$ holds. Then, for $U_G$ and $U_V$, (3.9) tends to the standard normal distribution function $\Phi(y)$ for every $y \in \mathbb{R}$, respectively.

**Proof.** To be consistent with conditions imposed on symmetric (and thus $U$-) statistics in Bloznelis and Götze (2001), consider normalized versions of the statistics of interest: $\sqrt{n}U_G$ and $\sqrt{n}U_V$. Then the variances of linear parts from the 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 Theorem 1 in Čiginas (2013b) through Proposition 3 in Bloznelis and Götze (2001). In the case of $U_V$, by Theorem 1 and Proposition 3 in Bloznelis and Götze (2001), 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, it follows easily from the explicit formulas above.

### 3.3 True one-term Edgeworth expansions

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

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

for Studentized $U$-statistics, constructed in Bloznelis (2003), can be an improvement. Here $\varphi(y)$ is the standard normal density function, and

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

are the population characteristics. Next, we give detailed expressions of these parameters for both statistics of interest.
GMD statistic. Routine but tedious combinatorial calculations give
\[
\alpha = -\sigma_1^{-3} \frac{8}{n^3 (N-2)^3} \left[ \sum_{i=1}^{N-1} i(N-2i)(N-i)a_i^3 \Delta_i^3 + 3 \sum_{1 \leq i < j \leq N-1} i(N-2i)(N-j)a_i^2 a_j^2 \Delta_i \Delta_j + 3 \sum_{1 \leq i < j \leq N-1} i(N-2i)(N-j)a_i^2 a_j^2 \Delta_i \Delta_j \right] \\
+ 6 \sum_{1 \leq i < j < m \leq N-1} i(N-2j)(N-m)a_i a_j a_m \Delta_i \Delta_j \Delta_m, \tag{3.13}
\]
and
\[
\kappa = -\sigma_1^{-3} \tau^2 \frac{16}{n^3 (n-1) (N-1)^2 (N-2)^3} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \sum_{m=1}^{N-1} c_{i jm} a_i a_j a_m \Delta_i \Delta_j \Delta_m, \tag{3.14}
\]
where
\[
c_{ijm} = \begin{cases} 
    i(i-1)(N-m)[N-j-1+N^{-1} j(m-j)], & \text{if } i \leq j \leq m, \\
    i(i-1)(N-j)[N-m-1+N^{-1} m(m-j)], & \text{if } i \leq m < j, \\
    j(N-m)[(i-1)(N-i-1)+N^{-1}(N-i)(i-j)+i(i-1)(m-i)], & \text{if } j < i < m, \\
    m(N-j)[(i-1)(N-i-1)+N^{-1} i(i-1)(i-j)+(N-i-1)(N-i)(m-i)], & \text{if } m < i < j, \\
    j(N-i-1)(N-i)[m-1+N^{-1}(N-m)(m-j)], & \text{if } j < m \leq i, \\
    m(N-i-1)(N-i)[j-1+N^{-1}(N-j)(m-j)], & \text{if } m \leq j \leq i.
\end{cases}
\]

These formulas are new in the literature.

Sample variance. With straightforward calculations one can arrive to the following results:
\[
\alpha = \sigma_1^{-3} \frac{1}{n^3} \left( \frac{N}{N-2} \right)^3 \left( 2\mu_2^3 - 3\mu_2 \mu_4 + \mu_6 \right) \tag{3.15}
\]
and
\[
\kappa = \sigma_1^{-3} \tau^2 \frac{2}{n^3 (n-1)} \left( \frac{N}{N-2} \right)^3 \frac{1}{N-1} \left( -(N-2)\mu_2^3 - \frac{2N-1}{N-1} \mu_4 \mu_2 + \frac{N}{N-1} \mu_2^3 + \mu_6 \right) \tag{3.16}
\]

Note that (3.16) can be simplified (approximated) by leaving the term with \( \mu_2^3 \) in the brackets only. For comparison, expressions similar to these can be identified in the Edgeworth approximation given by Kokic and Weber (1990) for standardized sample variance.

While an error of the normal approximation is typically of the order \( O(n_{-1/2}^-) \), see, e.g., Zhao and Chen (1990) for the case of standardized \( U \)-statistics, the error of the true (with known parameters \( \alpha \) and \( \kappa \)) one-term Edgeworth approximation (3.12) is of the order \( o(n_{-1/2}^-) \) under certain conditions. The first condition, from those, is the asymptotical nonlatticeness of the linear part of \( U \)-statistic: for every \( \varepsilon > 0 \) and every \( B > 0 \),
\[
\liminf_{n_{-1/2} \to \infty} \sup_{\varepsilon < |t| < B} \left| \mathbb{E} \exp \left\{ \frac{it}{\sigma_1^{-1}} g_1(X_1) \right\} \right| < 1, \tag{3.17}
\]
see Bloznelis and Götze (2001). This and other specific sufficient conditions 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) for \( U_G \), \( E|X_1|^{6+\delta} \leq C_1 < \infty \) holds; (ii) for \( U_V \), \( E|X_1|^{12+\delta} \leq C_2 < \infty \) holds. Then, we have
\[
\sup_{y \in \mathbb{R}} |F_{NH}(y) - H_{NH}(y)| = o(n_*^{-1/2}) \quad \text{as} \quad n_* \to \infty,
\]
for \( U_G \) and \( U_V \), respectively.

**Proof.** In the case of \( U_G \), the proof is the corollary of Theorem 1 in Čiginas (2012) following technique in the proof of Theorem 1 in Bloznelis (2003). In particular, by these theorems, the boundedness of the characteristics \( \beta_s = \sigma_1^{-s} E|g_1(X_1)|^s \) and \( \gamma_s = \sigma_1^{-s} \tau^{2s} E|g_2(X_1, X_2)|^s \), 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 \( s \geq 1 \), applying inequalities \(|a - b|^s \leq 2^{s-1}(a^s + b^s)\) where \( a, b \geq 0 \), and \( \mu_2^s \leq \mu_{2s} \), we get
\[
E|g_1(X_1)|^s = \frac{1}{N} \sum_{k=1}^{N} |g_1(x_k)|^s \leq \frac{2^{s-1}}{n^s} \left( \frac{N}{N-2} \right)^s \frac{1}{N} \sum_{k=1}^{N} ((x_k - b_1)^{2s} + \mu_2^s) \leq \left( \frac{2N}{n(N-2)} \right)^s \mu_{2s}. \tag{3.18}
\]
By (3.6), for \( 1 \leq k < l \leq N \), applying \((x_k - x_l)^2 \leq 2 ((x_k - b_1)^2 + (x_l - b_1)^2)\), we have
\[
|g_2(x_k, x_l)| \leq \frac{1}{n(n-1)} \left( \frac{3N-4}{N-2} \right) ((x_k - b_1)^2 + (x_l - b_1)^2) + \frac{2N}{(N-1)(N-2)} \mu_2^s \leq \frac{3}{n(n-1)N-2} \left( (x_k - b_1)^2 + (x_l - b_1)^2 + \mu_2^s \right).
\]
Then, for \( s \geq 1 \), similarly as in (3.18), applying \((a + b)^s \leq 2^{s-1}(a^s + b^s)\) twice, where \( a, b \geq 0 \), and noting that \( \sum_{1 \leq k < l \leq N} ((x_k - b_1)^{2s} + (x_l - b_1)^{2s}) = N(N-1)\mu_{2s} \), we obtain
\[
E|g_2(X_1, X_2)|^s = \frac{N}{2} \left( \frac{N}{2} \right)^{-1} \sum_{1 \leq k < l \leq N} |g_2(x_k, x_l)|^s \leq \frac{3^s}{n^s(n-1)^s} \left( \frac{N}{N-2} \right)^s \left( \frac{N}{2} \right)^{-1} \sum_{1 \leq k < l \leq N} ((x_k - b_1)^2 + (x_l - b_1)^2 + \mu_2^s)^s \tag{3.19}
\]
\[
\leq \frac{3^s 2^{s-1}}{n^s(n-1)^s} \left( \frac{N}{N-2} \right)^s \left( \frac{N}{2} \right)^{-1} \sum_{1 \leq k < l \leq N} (2^{s-1}((x_k - b_1)^{2s} + (x_l - b_1)^{2s}) + \mu_2^s) \leq \frac{3^s 2^{s-1}(2^s + 1)}{n^s(n-1)^s} \left( \frac{N}{N-2} \right)^s \mu_{2s}.
\]
Then we get from (3.18), (3.19) and (3.7) that
\[
\beta_s \leq \frac{2^s \mu_{2s}}{(\mu_4 - \mu_2^2)^{s/2}} \quad \text{and} \quad \gamma_s \leq 3^s 2^{2s-1}(2^s + 1) \left( 1 - \frac{n}{N} \right)^s \left( \mu_{2s} / (\mu_4 - \mu_2^2)^{s/2} \right).
\]
The proof is completed. \( \square \)
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 for the particular ones. In Pumputis and Čiginas (2013), 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 as well as simple estimators of the variances are possible. We give here, in fact, plug-in estimators of the variances, replacing population moments by their empirical counterparts in the parameters \(\sigma_1^2\) and \(\sigma_2^2\) defining variance (3.2).

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

\[
\hat{\sigma}_{1G}^2 = \frac{4}{n^2} \left( \frac{N}{N - 2} \right)^2 \left[ \frac{1}{n-1} \sum_{i=1}^{n-1} i(n-i)A_i^2 \Delta_{i:n}^2 + 2 \sum_{1 \leq i < j \leq n-1} i(n-j)A_iA_j \Delta_{i:n} \Delta_{j:n} \right]
\]

and

\[
\hat{\sigma}_{2G}^2 = \frac{16}{n^4(n-1)^4} \frac{N}{N - 2} \left[ \frac{1}{n-1} \sum_{i=1}^{n-1} i(i-1)(n-i-1)(n-i)\Delta_{i:n}^2 \right. \\
\left. + 2 \sum_{1 \leq i < j \leq n-1} i(i-1)(n-j-1)(n-j)\Delta_{i:n} \Delta_{j:n} \right].
\]

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 \(m_k = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^k\), for \(k = 2, \ldots, 6\). Replacing the central moments in (3.7) and (3.8) by the corresponding empirical moments, we get

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

and

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

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 the statistics, the parameters \(\alpha\) and \(\kappa\) must be evaluated. Firstly, case (A), analogously to the variance estimation case, we construct estimators of the parameters directly from the explicit expressions available. Secondly, case (B), we assume that the auxiliary variable \(z\) is at our disposal with the known values \(\{z_1, \ldots, z_N\}\) 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 the parameters.
**GMD statistic.** Case (A). With the notations used for the variance estimator, by formulas (3.13) and (3.14), the estimators are

\[
\hat{\sigma}_G = -\hat{\sigma}_{1G}^{-3} \frac{8}{n^6} \left( \frac{N}{N-2} \right)^3 \sum_{i=1}^{n-1} i(n-2i)(n-i)A_i^3 \Delta_{i,n}^3 + 3 \sum_{1 \leq i < j \leq n-1} i(n-2j)(n-j)A_i^2 A_j^2 \Delta_{i,n}^2 \Delta_{j,n}^2 + 3 \sum_{1 \leq i < j \leq n-1} i(n-2j)(n-j)A_i^2 A_j^2 \Delta_{i,n}^2 \Delta_{j,n}^2 \\
+ 6 \sum_{1 \leq i < j < m \leq n-1} i(n-2j)(n-m)A_i A_j A_m \Delta_{i,n}^2 \Delta_{j,n}^2 \Delta_{m,n},
\]

(4.5)

and

\[
\hat{\kappa}_G = -\hat{\sigma}_{1G}^{-3} \frac{16}{n^3(n-1)^3} \frac{N}{N-2} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \sum_{m=1}^{n-1} C_{ijm} A_i A_j A_m \Delta_{i,n}^2 \Delta_{j,n}^2 \Delta_{m,n},
\]

(4.6)

with the case function

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

Case (B). Having the additional information, the ordered sequence of the values \(z_1, \ldots, z_N\) is used instead of \(x_1 \leq \cdots \leq x_N\) in the expressions (3.13) and (3.14) of the true parameters \(\alpha\) and \(\kappa\). Denote the resulting estimates by \(\hat{\alpha}_G\) and \(\hat{\kappa}_G\).

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

\[
\hat{\alpha}_V = \hat{\sigma}_{1V}^{-3} \frac{1}{n^3} \left( \frac{N}{N-2} \right)^3 \left( 2m_2^3 - 3m_2m_4 + m_6 \right)
\]

(4.7)

and

\[
\hat{\kappa}_V = \hat{\sigma}_{1V}^{-3} \frac{2}{n^3(n-1)} \left( \frac{N}{N-2} \right)^3 \frac{1}{N-1} \left( -(N-2)m_3^2 - \frac{2N-1}{N-1}m_4m_2 + \frac{N}{N-1}m_2^3 + m_6 \right).
\]

(4.8)

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

### 5 Empirical Edgeworth and bootstrap approximations

Replacing the population parameters \(\alpha\) and \(\kappa\) in Edgeworth expansion (3.12) by their estimators, we obtain the so-called empirical Edgeworth expansion. If the 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. In [Bloznelis 2001], consistent jackknife estimators of the parameters were constructed. Bootstrap and calibrated estimators of the parameters were considered in Ciginas (2013a), and Pumputis and Ciginas (2013), 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

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

(5.1)

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

$$
\hat{z}_{nSG}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)}{6\tau} \hat{G}_G + 3 (y^2 + 1) \hat{\kappa}_G \Phi(y),
$$

(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

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

(5.3)

and

$$
\hat{z}_{nSV}(y) = \Phi(y) + \frac{(1 - 2n/N + (2 - n/N)y^2)}{6\tau} \hat{G}_V + 3 (y^2 + 1) \hat{\kappa}_V \Phi(y),
$$

(5.4)

where the later does not depend on the sample.

Estimators of the parameters $\alpha$ and $\kappa$ in expansion (5.3) are asymptotically consistent under conditions of Theorem 2. Efficiency of the other empirical Edgeworth expansions is examined in the simulation study in Section 6.

It is known that, in general, non-parametric 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 in [Booth et al. 1994]. We apply the results of Bloznelis (2007) 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 $0 \leq l < n$. Then, given the sample $\mathbb{X}$, the empirical population $\tilde{X}$ 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 $\mathbb{Y} = \{Y_1, \ldots, Y_l\}$ drawn without replacement from the set $\mathbb{X}$. With this particular bootstrap population $\tilde{X}$, one can turn already to an estimator of (3.9), despite that it is only the one of $\binom{n}{l}$ empirical populations. Next, we draw the simple random sample $\tilde{\mathbb{X}} = \{\tilde{X}_1, \ldots, \tilde{X}_n\}$ without replacement from $\tilde{X}$. Denote by $\tilde{U} = U_n(\tilde{\mathbb{X}})$ the bootstrap estimator for the statistic of interest, and introduce the corresponding jackknife estimator $\tilde{S}^2 = S^2(\tilde{\mathbb{X}})$ of the variance of $\tilde{U}$ under given population $\tilde{X}$. Then the bootstrap approximation to (3.9) is

$$
\tilde{F}_{nS}(y) = P\{\tilde{U} - E(\tilde{U} \mid \mathbb{X}, \mathbb{Y}) \leq y \tilde{S} \mid \mathbb{X}\},
$$

(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 $U_V$. 

14
Theorem 3. Assume that the conditions of Theorem 2 are satisfied. Then, we have
\[ \sup_{y \in \mathbb{R}} |F_{nS}(y) - \tilde{F}_{nS}(y)| = o_p(n_*^{-1/2}) \quad \text{as} \quad n_* \to \infty, \]
for \( U_G \) and \( U_V \), respectively.

Proof. It follows from condition (8) in Bloznelis (2007), that it suffices to verify that, for the statistics \( U_G \) and \( U_V \), the moments \( \mathbb{E}(X_1 - X_2)^6 \) and \( \mathbb{E}(X_1 - X_2)^{12} \) are bounded for all \( n_* \), respectively. By the conditions of theorem, this requirement holds.

Denote by \( \tilde{F}_{nSG}(y) \) and \( \tilde{F}_{nSV}(y) \) 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 consider also how outliers affect these approximations.

For the statistics \( U_G \) and \( U_V \), 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 values \( X \), as well as their bootstrap approximations based on the one (because of \( N = kn \)) empirical population \( \tilde{X} \) constructed from the particular sample \( X \). Denote true Edgeworth approximations (3.12) of the statistics by \( H_{nSG}(y) \) and \( H_{nSV}(y) \), respectively. To measure an efficiency of the empirical Edgeworth approximations \( H_{nSG}(y) \) and \( H_{nSV}(y) \), and the bootstrap approximations \( \tilde{F}_{nSG}(y) \) and \( \tilde{F}_{nSV}(y) \), \( 10^3 \) samples without replacement were drawn independently from the population.

More specifically, in the tables below, the ‘exact’ distribution functions of the statistics, their normal approximation, the true one-term Edgeworth expansions, the corresponding estimated Edgeworth approximations of two types, and the bootstrap approximations are represented by the several commonly used \( q \)-quantiles, \( q = 0.01, 0.05, 0.10, 0.90, 0.95, 0.99 \). For the approximations, with the quantiles dependent on the sample, we give two characteristics of the efficiency: the empirical expectations \( \hat{\mathbb{E}}(\cdot) \) and standard errors \( \hat{\mathbb{V}}(\cdot) \) from the realizations of these quantiles.

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

By 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 \( \hat{H}_{nSG}(y) \). The bias of this estimate is small in comparison to a possible error of the estimator \( \hat{H}_{nSG}(y) \). But the later improves the normal approximation to the distribution of \( U_G \) too. Differently from all other, the bootstrap approximation \( \tilde{F}_{nSG}(y) \) is almost unbiased, but its empirical quantiles have larger standard errors 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 an asymmetry of the gamma distribution.
Table 3: Approximations to $F_{nSG}(y)$ under $\mathcal{N}(0, 1)$ with 0% outliers from $\mathcal{N}(0, 9)$, and $\rho_{cx} = 0.7$.

| $q$   | 0.01   | 0.05   | 0.10   | 0.90   | 0.95   | 0.99   |
|-------|--------|--------|--------|--------|--------|--------|
| $F_{nSG}^{-1}(q)$ | -2.592 | -1.779 | -1.363 | 1.223  | 1.546  | 2.157  |
| $\Phi^{-1}(q)$    | -2.326 | -1.645 | -1.282 | 1.282  | 1.645  | 2.326  |
| $H_{nSG}^{-1}(q)$ | -2.528 | -1.762 | -1.357 | 1.214  | 1.536  | 2.096  |
| $\tilde{H}_{nSG}^{-1}(q)$ | -2.513 | -1.755 | -1.350 | 1.220  | 1.545  | 2.116  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | -2.519 | -1.756 | -1.353 | 1.217  | 1.541  | 2.107  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | 0.034  | 0.023  | 0.016  | 0.013  | 0.021  | 0.044  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | -2.600 | -1.776 | -1.360 | 1.222  | 1.555  | 2.167  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | 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 $\rho_{cx} = 0.7$.

| $q$   | 0.01   | 0.05   | 0.10   | 0.90   | 0.95   | 0.99   |
|-------|--------|--------|--------|--------|--------|--------|
| $F_{nSV}^{-1}(q)$ | -2.918 | -1.962 | -1.477 | 1.160  | 1.461  | 2.008  |
| $\Phi^{-1}(q)$    | -2.326 | -1.645 | -1.282 | 1.282  | 1.645  | 2.326  |
| $H_{nSV}^{-1}(q)$ | -2.680 | -1.882 | -1.447 | 1.145  | 1.432  | 1.878  |
| $\tilde{H}_{nSV}^{-1}(q)$ | -2.658 | -1.864 | -1.433 | 1.155  | 1.447  | 1.910  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | -2.653 | -1.861 | -1.431 | 1.157  | 1.450  | 1.917  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | 0.046  | 0.038  | 0.029  | 0.020  | 0.032  | 0.068  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | -2.914 | -1.932 | -1.460 | 1.166  | 1.473  | 2.027  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | 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 $\rho_{cx} = 0.7$.

| $q$   | 0.01   | 0.05   | 0.10   | 0.90   | 0.95   | 0.99   |
|-------|--------|--------|--------|--------|--------|--------|
| $F_{nSG}^{-1}(q)$ | -2.888 | -1.903 | -1.443 | 1.188  | 1.503  | 2.062  |
| $\Phi^{-1}(q)$    | -2.326 | -1.645 | -1.282 | 1.282  | 1.645  | 2.326  |
| $H_{nSG}^{-1}(q)$ | -2.638 | -1.843 | -1.413 | 1.172  | 1.468  | 1.946  |
| $\tilde{H}_{nSG}^{-1}(q)$ | -2.572 | -1.793 | -1.378 | 1.198  | 1.510  | 2.038  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | -2.624 | -1.833 | -1.407 | 1.177  | 1.476  | 1.965  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | 0.055  | 0.045  | 0.033  | 0.024  | 0.037  | 0.079  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | -2.864 | -1.899 | -1.436 | 1.190  | 1.506  | 2.079  |
| $\tilde{\Phi}_{nSG}^{-1}(q)$ | 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 $\rho_{cx} = 0.7$.

| $q$   | 0.01   | 0.05   | 0.10   | 0.90   | 0.95   | 0.99   |
|-------|--------|--------|--------|--------|--------|--------|
| $F_{nSV}^{-1}(q)$ | -3.744 | -2.310 | -1.699 | 1.109  | 1.391  | 1.876  |
| $\Phi^{-1}(q)$    | -2.326 | -1.645 | -1.282 | 1.282  | 1.645  | 2.326  |
| $H_{nSV}^{-1}(q)$ | -2.829 | -2.025 | -1.561 | 1.077  | 1.324  | 1.656  |
| $\tilde{H}_{nSV}^{-1}(q)$ | -2.796 | -1.991 | -1.533 | 1.091  | 1.347  | 1.704  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | -2.788 | -1.985 | -1.529 | 1.097  | 1.355  | 1.719  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | 0.075  | 0.077  | 0.067  | 0.040  | 0.060  | 0.114  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | -3.642 | -2.267 | -1.655 | 1.120  | 1.406  | 1.909  |
| $\tilde{\Phi}_{nSV}^{-1}(q)$ | 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 $\rho_{2x} = 0.7$.

| $q$   | 0.01  | 0.05  | 0.10  | 0.90  | 0.95  | 0.99  |
|-------|-------|-------|-------|-------|-------|-------|
| $F_{nSG}^{-1}(q) \approx$ | -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 |
| $H_{nSG}^{-1}(q) \approx$ | -2.745 | -1.940 | -1.490 | 1.118 | 1.388 | 1.783 |
| $\hat{H}_{nSG}^{-1}(q) \approx$ | -2.602 | -1.816 | -1.395 | 1.184 | 1.489 | 1.996 |
| $\hat{\Phi}_{nSG}(q) \approx$ | -2.713 | -1.913 | -1.470 | 1.132 | 1.410 | 1.831 |
| $\hat{S}_{nSG}(q) \approx$ | 0.069 | 0.063 | 0.050 | 0.033 | 0.051 | 0.103 |
| $\hat{F}_{nSG}^{-1}(q) \approx$ | -3.124 | -2.039 | -1.524 | 1.152 | 1.449 | 1.982 |
| $\hat{S}_{nSG}(q) \approx$ | 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 $\rho_{2x} = 0.7$.

| $q$   | 0.01  | 0.05  | 0.10  | 0.90  | 0.95  | 0.99  |
|-------|-------|-------|-------|-------|-------|-------|
| $F_{nSV}^{-1}(q) \approx$ | -4.724 | -2.924 | -2.190 | 1.037 | 1.280 | 1.699 |
| $\Phi^{-1}(q) \approx$ | -2.326 | -1.645 | -1.282 | 1.282 | 1.645 | 2.326 |
| $H_{nSV}^{-1}(q) \approx$ | -2.985 | -2.207 | -1.740 | 0.979 | 1.181 | 1.415 |
| $\hat{H}_{nSV}^{-1}(q) \approx$ | -2.861 | -2.062 | -1.596 | 1.054 | 1.291 | 1.602 |
| $\hat{\Phi}_{nSV}(q) \approx$ | -2.889 | -2.097 | -1.634 | 1.036 | 1.266 | 1.560 |
| $\hat{S}_{nSV}(q) \approx$ | 0.086 | 0.098 | 0.095 | 0.050 | 0.075 | 0.130 |
| $\hat{F}_{nSV}^{-1}(q) \approx$ | -4.600 | -2.792 | -1.974 | 1.069 | 1.333 | 1.787 |
| $\hat{S}_{nSV}(q) \approx$ | 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 $\rho_{2x} = 0.7$.

| $q$   | 0.01  | 0.05  | 0.10  | 0.90  | 0.95  | 0.99  |
|-------|-------|-------|-------|-------|-------|-------|
| $F_{nSG}^{-1}(q) \approx$ | -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 |
| $H_{nSG}^{-1}(q) \approx$ | -2.740 | -1.933 | -1.483 | 1.124 | 1.395 | 1.795 |
| $\hat{H}_{nSG}^{-1}(q) \approx$ | -2.601 | -1.815 | -1.394 | 1.186 | 1.491 | 1.997 |
| $\hat{\Phi}_{nSG}(q) \approx$ | -2.717 | -1.915 | -1.470 | 1.134 | 1.410 | 1.827 |
| $\hat{S}_{nSG}(q) \approx$ | 0.064 | 0.060 | 0.049 | 0.032 | 0.048 | 0.097 |
| $\hat{F}_{nSG}^{-1}(q) \approx$ | -3.213 | -2.057 | -1.531 | 1.154 | 1.453 | 1.990 |
| $\hat{S}_{nSG}(q) \approx$ | 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 $\rho_{2x} = 0.7$.

| $q$   | 0.01  | 0.05  | 0.10  | 0.90  | 0.95  | 0.99  |
|-------|-------|-------|-------|-------|-------|-------|
| $F_{nSV}^{-1}(q) \approx$ | -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 |
| $H_{nSV}^{-1}(q) \approx$ | -2.978 | -2.198 | -1.728 | 0.988 | 1.193 | 1.430 |
| $\hat{H}_{nSV}^{-1}(q) \approx$ | -2.864 | -2.064 | -1.597 | 1.055 | 1.292 | 1.600 |
| $\hat{\Phi}_{nSV}(q) \approx$ | -2.882 | -2.087 | -1.622 | 1.045 | 1.277 | 1.577 |
| $\hat{S}_{nSV}(q) \approx$ | 0.083 | 0.095 | 0.092 | 0.050 | 0.073 | 0.127 |
| $\hat{F}_{nSV}^{-1}(q) \approx$ | -4.782 | -2.769 | -1.944 | 1.079 | 1.347 | 1.809 |
| $\hat{S}_{nSV}(q) \approx$ | 1.184 | 0.651 | 0.416 | 0.041 | 0.058 | 0.097 |
Let us take the populations of Section 2.2 with \( p/N = 0.06 \). In this case of outliers, the corresponding to Tables 3–6 results are given in Tables 7–10. A behaviour of the approximations to the distributions is very similar to that in the case of no outliers, but errors of the approximations are larger now. One can observe also that the estimates of the true Edgeworth expansions, which use the auxiliary information, are much more biased. It holds for the alternative empirical Edgeworth approximations too but in the case of the statistic \( U_V \) only (Tables 8 and 10). A sensitivity to the outliers is the smallest comparing Table 9 with Table 5.

7 Summary

The specific estimation strategies for scales are considered under simple random samples without replacement. In a sense, they are consistent with the scale estimation by the sample variance. In particular, the proposed strategies (\( S_1 \)) and (\( S_2 \)) combine the use of the GMD statistic and its bias correction. This combination allows an improvement of the scale estimation in populations where the part of outliers is not large. As the numerical modeling indicates too, under ideal for the sample variance conditions (when there are no outliers), the efficiency of the strategies is not worse. It is important robustness property.

The new estimators of the parameters and also empirical Edgeworth expansions for the GMD statistic and the sample variance are proposed using the detailed decompositions of the statistics. In general, well-correlated auxiliary information leads to effective inferences about the statistics of interest.

References

Bickel, P.J., Götze, F., van Zwet, W.R., 1986. The Edgeworth expansion for \( U \)-statistics of degree two. The Annals of Statistics 14, 1463–1484.

Bloznelis, M., 2001. Empirical Edgeworth expansion for finite population statistics I. Lithuanian Mathematical Journal 41, 120–134.

Bloznelis, M., 2003. An Edgeworth expansion for Studentized finite population statistics. Acta Applicandae Mathematicae 78, 51–60.

Bloznelis, M., 2007. Bootstrap approximation to distributions of finite population \( U \)-statistics. Acta Applicandae Mathematicae 96, 71–86.

Bloznelis, M., Götze, F., 1999. One-term Edgeworth expansion for finite population \( U \)-statistics of degree two. Acta Applicandae Mathematicae 58, 75–90.

Bloznelis, M., Götze, F., 2001. Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. The Annals of Statistics 29, 899–917.

Booth, J.G., Butler, R.W., Hall, P., 1994. Bootstrap methods for finite populations. Journal of the American Statistical Association 89, 1282–1289.

Chambers, R.L., 1986. Outlier robust finite population estimation. Journal of the American Statistical Association 81, 1063–1069.
Chhikara, R.S., Feiveson, A.L., 1980. Extended critical values of extreme studentized deviate test statistics for detecting multiple outliers. Communications in Statistics - Simulation and Computation 9, 155–166.

Čiginas, A., 2012. An Edgeworth expansion for finite-population $L$-statistics. Lithuanian Mathematical Journal 52, 40–52.

Čiginas, A., 2013a. Second-order approximations of finite population $L$-statistics. Statistics 47, 954–965.

Čiginas, A., 2013b. On the asymptotic normality of finite population $L$-statistics. Statistical Papers, pp. 1–12, doi:10.1007/s00362-013-0553-7.

Erdős, P., Rényi, A., 1959. On the central limit theorem for samples from a finite population. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 4, 49–61.

Gini, C., 1912. Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. Cuppini, Bologna.

Glasser, G.J., 1962. Variance formulas for the mean difference and coefficient of concentration. Journal of the American Statistical Association 57, 648–654.

Hájek, J., 1960. Limiting distributions in simple random sampling from a finite population. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 361–374.

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

Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics 19, 293–325.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Irwin, J.O., Kendall, M.G., 1944. Sampling moments of moments for a finite population. Annals of Eugenics 12, 138–142.

Kokic, P.N., Weber, N.C., 1990. An Edgeworth expansion for $U$-statistics based on samples from finite populations. Annals of Probability 18, 390–404.

Lomnicki, Z.A., 1952. The standard error of Gini’s mean difference. The Annals of Mathematical Statistics 23, 635–637.

Nair, U.S., 1936. The standard error of Gini’s mean difference. Biometrika 28, 428–436.

Pumputis, D., Čiginas, A., 2013. Estimation of parameters of finite population $L$-statistics. Nonlinear Analysis: Modelling and Control 18, 327–343.

Putter, H., van Zwet, W.R., 1998. Empirical Edgeworth expansions for symmetric statistics. The Annals of Statistics 26, 1540–1569.

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

Yitzhaki, S., Schechtman, E., 2013. The Gini Methodology: A Primer on a Statistical Methodology. Springer, New York.
Zhao, L.C., Chen, X.R., 1990. Normal approximation for finite-population $U$-statistics. Acta Mathematicae Applicatae Sinica 6, 263–272.