Calibrated approximations for L-statistics

Andrius Čiginas

(joint work with Dalius Pumputis)

Vilnius University, Lithuania

The 5th Baltic-Nordic Conference on Survey Statistics
16-20 June, 2019 | Örebro, Sweden

Survey variable, sample and L-statistics

- A survey variable x with real values $\mathcal{X} = \{x_1, \ldots, x_N\}$ in the population $\mathcal{U} = \{1, \ldots, N\}$.
- Let $\mathbb{X} = \{X_1, \ldots, X_n\}$ be the measurements of the simple random sample units $\{1, \ldots, n\}$, $n < N$, drawn without replacement from $\mathcal{U}$.
- The L-statistic

$$L = L_n(\mathbb{X}) = \frac{1}{n} \sum_{j=1}^{n} c_{j,n} X_{j:n}$$

is a combination of the order statistics $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ of $\mathbb{X}$ with real coefficients

$$c_{j,n} = J\left(\frac{j}{n+1}\right), \quad J\colon (0,1) \to \mathbb{R},$$

called weights.

Examples of L-statistics

1. $J \equiv 1$ means the sample mean.
2. $J(s) = 4s - 2$ is used to define Gini's mean difference statistic.
3. $J(s) = 6s(1 - s)$.
4. The trimmed mean

$$M_{n;a;b}(\mathbb{X}) = \frac{1}{[bn] - [an]} \sum_{j=[an]+1}^{[bn]} X_{j:n}$$

is represented asymptotically by

$$J(s) = (b - a)^{-1} \mathbb{I}\{a < s < b\}$$

with the fixed trimming proportions $0 \leqslant a < b \leqslant 1$, where $[\cdot]$ and $\mathbb{I}\{\cdot\}$ are the greatest integer and the indicator functions.

Estimation of quality of L-statistic

- One can calculate the jackknife estimator

$$\hat{\sigma}_{\mathrm{J}}^2 = \hat{\sigma}_{\mathrm{J}}^2(\mathbb{X}) = (1-f)\frac{n-1}{n}\sum_{k=1}^{n}\bigl(L_{(k)} - \overline{L}\bigr)^2, \quad \overline{L} = \frac{1}{n}\sum_{k=1}^{n} L_{(k)},$$

  of the variance $\sigma^2 = \operatorname{Var} L$. Here $f = n/N$ is the sampling fraction, and $L_{(k)} = L_{n-1}(\mathbb{X}\backslash\{X_k\})$ are L-statistics with weights $c_{j,n-1} = J(j/n)$, $1 \leqslant j \leqslant n-1$.

- To make further inferences about the quality of the statistic, the distribution function

$$F_{\mathrm{S}}(y) = \mathrm{P}\{\hat{\sigma}_{\mathrm{J}}^{-1}(L - \mathrm{E}\, L) \leqslant y\}$$

  of the Studentized L-statistic is estimated.

Question: how to approximate $F_{\mathrm{S}}(y)$ if the sample size n is not large enough to apply the normal approximation $\Phi(y)$?

Empirical one-term Edgeworth expansions

The Edgeworth approximation to $F_{\mathrm{S}}(y)$ is (Bloznelis, 2003)

$$\begin{aligned}\widehat{G}_{\mathrm{S}}(y) &= G(y;\hat{\alpha},\hat{\kappa}) \\ &= \Phi(y) + \frac{(1-2f+(2-f)y^2)\hat{\alpha} + 3(y^2+1)\hat{\kappa}}{6\sqrt{(1-f)n}}\Phi'(y),\end{aligned}$$

where $\Phi'(y)$ is the derivative of $\Phi(y)$, and $\hat{\alpha}$ and $\hat{\kappa}$ are estimators of population parameters $\alpha = \alpha(J,\mathcal{X})$ and $\kappa = \kappa(J,\mathcal{X})$.

The examples of sample $\mathbb{X}$ based estimators of α and κ are:

- jackknife estimators $\hat{\alpha}_{\mathrm{J}} = \hat{\alpha}_{\mathrm{J}}(J,\mathbb{X})$ and $\hat{\kappa}_{\mathrm{J}} = \hat{\kappa}_{\mathrm{J}}(J,\mathbb{X})$ (Bloznelis, 2001);
- bootstrap estimators $\hat{\alpha}_{\mathrm{B}} = \hat{\alpha}_{\mathrm{B}}(J,\mathbb{X})$ and $\hat{\kappa}_{\mathrm{B}} = \hat{\kappa}_{\mathrm{B}}(J,\mathbb{X})$ (Čiginas, 2013).

Nonparametric bootstrap approximations

The Monte–Carlo approximation to one of bootstraps proposed by Booth et al. (1994) is

$$\widehat{F}_{\mathrm{SB}}(y) = \frac{1}{BR} \sum_{b=1}^{B} \sum_{r=1}^{R} \mathbb{I}\{\hat{\sigma}_{\mathrm{J}}^{-1}(\widetilde{\mathbb{X}}^{(b,r)})(L_n(\widetilde{\mathbb{X}}^{(b,r)}) - \mu(\widetilde{\mathcal{X}}^{(b)})) \leqslant y\},$$

where $\widetilde{\mathcal{X}}^{(b)}$, $1 \leqslant b \leqslant B$, are empirical (bootstrap) populations of size N reconstructed from $\mathbb{X}$, and $\widetilde{\mathbb{X}}^{(b,r)} = \{\widetilde{X}_1^{(b,r)}, \ldots, \widetilde{X}_n^{(b,r)}\}$, $1 \leqslant r \leqslant R$, are the samples without replacement from $\widetilde{\mathcal{X}}^{(b)}$. Here $\mu(\widetilde{\mathcal{X}}^{(b)})$ is the expectation of L under the fixed population $\widetilde{\mathcal{X}}^{(b)}$.

Saddlepoint approximations

The idea is to apply saddlepoint approximations to the distribution function of a suitably Studentized linear part of the L-statistic (Easton and Ronchetti, 1986).

We use Hoeffding's decomposition (Bloznelis and Götze, 2001)

$$L - \mathrm{E}\, L = H + R, \quad \text{where} \quad H = H_n(\mathbb{X}) = \frac{1}{n}\sum_{j=1}^{n} h(X_j)$$

is a linear statistic, and $R = R_n(\mathbb{X})$ is a remainder term.

The jackknife estimator of the variance $\sigma_H^2 = \mathrm{Var}\, H$ of H reduces from $\hat{\sigma}_{\mathrm{J}}^2$ to

$$\hat{\sigma}_{H\mathrm{J}}^2 = \hat{\sigma}_{H\mathrm{J}}^2(\mathbb{X}) = \frac{1-f}{n(n-1)}\sum_{j=1}^{n}\bigl(h(X_j) - H\bigr)^2.$$

First, we approximate the distribution function of interest,

$$F_{\mathrm{S}}(y) \approx \widetilde{F}_{\mathrm{S}}(y) = \mathrm{P}\{\hat{\sigma}_{H\mathrm{J}}^{-1} H \leqslant y\}.$$

Second, for $\widetilde{F}_{\mathrm{S}}(y)$, we apply the saddlepoint approximation results for the Studentized mean by Dai and Robinson (2001).

The formula of "true" saddlepoint approximation depends on unknown values $h(x_k)$, $1 \leqslant k \leqslant N$, of the function $h(\cdot)$. We obtain the sample $\mathbb{X}$ based empirical saddlepoint approximation by replacing these values by their bootstrap (Booth et al., 1994) estimators.

Auxiliary information and naive approximation

Denote by z the variable with known real values $\mathcal{Z} = \{z_1, \ldots, z_N\}$ in the population $\mathcal{U}$. Let $\mathbb{Z} = \{Z_1, \ldots, Z_n\}$ be the corresponding values of the sample units.

If the variables x and z are well-correlated, calibration techniques (Deville and Särndal, 1992) can be applied to derive efficient approximations to $F_{\mathrm{S}}(y)$.

One can apply the approximation

$$F_{\mathrm{S}z}(y) = \mathrm{P}\{\hat{\sigma}_{\mathrm{J}}^{-1}(\mathbb{Z})(L_n(\mathbb{Z}) - \mathrm{E}\, L_n(\mathbb{Z})) \leqslant y\},$$

which is very efficient if the shapes of distributions of the variables x and z are similar. However, it yields misleading results in practical situations (Čiginas and Pumputis, 2019a).

Calibrating Edgeworth and saddlepoint

Using ideas of Deville and Särndal (1992):

- Pumputis and Čiginas (2013) calibrated the bootstrap estimators $\hat{\alpha}_{\mathrm{B}}$ and $\hat{\kappa}_{\mathrm{B}}$ of α and κ. The constructed estimators

$$\hat{\alpha}_{\mathrm{B}w} = \hat{\alpha}_{\mathrm{B}w}(J, \mathbb{X}, \mathcal{Z}) \quad \text{and} \quad \hat{\kappa}_{\mathrm{B}w} = \hat{\kappa}_{\mathrm{B}w}(J, \mathbb{X}, \mathcal{Z})$$

  are plugged into the one-term Edgeworth expansion;
- Čiginas and Pumputis (2019b) calibrated the bootstrap estimators of the values (population parameters) $h(x_k)$, $1 \leqslant k \leqslant N$, to obtain the sample $\mathbb{X}$ and the auxiliary data $\mathcal{Z}$ based empirical saddlepoint approximation.

Calibrating nonparametric bootstrap

Using the Monte–Carlo representation $\widehat{F}_{\mathrm{SB}}(y)$, we define the calibrated nonparametric bootstrap approximation

$$\widehat{F}_{\mathrm{SB}w}(y) = \frac{1}{BR}\sum_{b=1}^{B}\sum_{r=1}^{R} w_{br}\mathbb{I}\{\hat{\sigma}_{\mathrm{J}}^{-1}(\widetilde{\mathbb{X}}^{(b,r)})(L_n(\widetilde{\mathbb{X}}^{(b,r)})-\mu(\widetilde{\mathcal{X}}^{(b)})) \leqslant y\},$$

where the weights $\mathbf{W} = (w_{br}) \in \mathbb{R}^{B\times R}$ minimize the function

$$d(\mathbf{W}) = \frac{1}{BR}\sum_{b=1}^{B}\sum_{r=1}^{R}(w_{br}-1)^2$$

and, for chosen points $y_1, \ldots, y_T$, satisfy the calibration equations

$$\frac{1}{BR}\sum_{b=1}^{B}\sum_{r=1}^{R} w_{br}\mathbb{I}\{Z_{b,r} \leqslant y_i\} = F_{\mathrm{S}z}(y_i), \qquad 1 \leqslant i \leqslant T,$$

where $Z_{b,r} = \hat{\sigma}_{\mathrm{J}}^{-1}(\widetilde{\mathbb{Z}}^{(b,r)})(L_n(\widetilde{\mathbb{Z}}^{(b,r)}) - \mu(\widetilde{\mathcal{Z}}^{(b)}))$. Here the sets $\widetilde{\mathcal{Z}}^{(b)}$ and $\widetilde{\mathbb{Z}}^{(b,r)}$, constructed from $\mathbb{Z}$, represent exactly the same sample units as the sets $\widetilde{\mathcal{X}}^{(b)}$ and $\widetilde{\mathbb{X}}^{(b,r)}$ selected from the given $\mathbb{X}$.

Proposition

Let $y_1 < \cdots < y_T$, and there is at least one value from the set $\{Z_{b,r}, 1 \leqslant b \leqslant B, 1 \leqslant r \leqslant R\}$ between each pair of these points. Then the weights $\mathbf{W}$ minimizing the distance function $d(\mathbf{W})$ and satisfying the calibration equations are unique and expressed by

$$w_{br} = 1 + \frac{1}{2}\sum_{j=1}^{T} \lambda_j \mathbb{I}\{Z_{b,r} \leqslant y_j\}, \qquad 1 \leqslant b \leqslant B, \quad 1 \leqslant r \leqslant R,$$

where the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)^\intercal = \mathbf{A}^{-1}\mathbf{b}$ is defined by $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{T \times T}$ and $\mathbf{b} = (b_1, \ldots, b_T)^\intercal$ that have the values

$$a_{ij} = \frac{1}{2BR}\sum_{b=1}^{B}\sum_{r=1}^{R} \mathbb{I}\{Z_{b,r} \leqslant y_i\}\mathbb{I}\{Z_{b,r} \leqslant y_j\}$$

and

$$b_i = F_{\mathrm{S}z}(y_i) - \frac{1}{BR}\sum_{b=1}^{B}\sum_{r=1}^{R} \mathbb{I}\{Z_{b,r} \leqslant y_i\}.$$

Remark

The arbitrarily chosen points $y_1 < \cdots < y_{T-1}$ are, for example, uniformly spaced quantiles of the distribution function of values $\{Z_{b,r}, 1 \leqslant b \leqslant B, 1 \leqslant r \leqslant R\}$, and the last point $y_T = 10^3$.
In our practice, the choice $T = 10^2$ is better than $T = 10$, but $T = 10^3$ gives no significant further improvement.

Remark

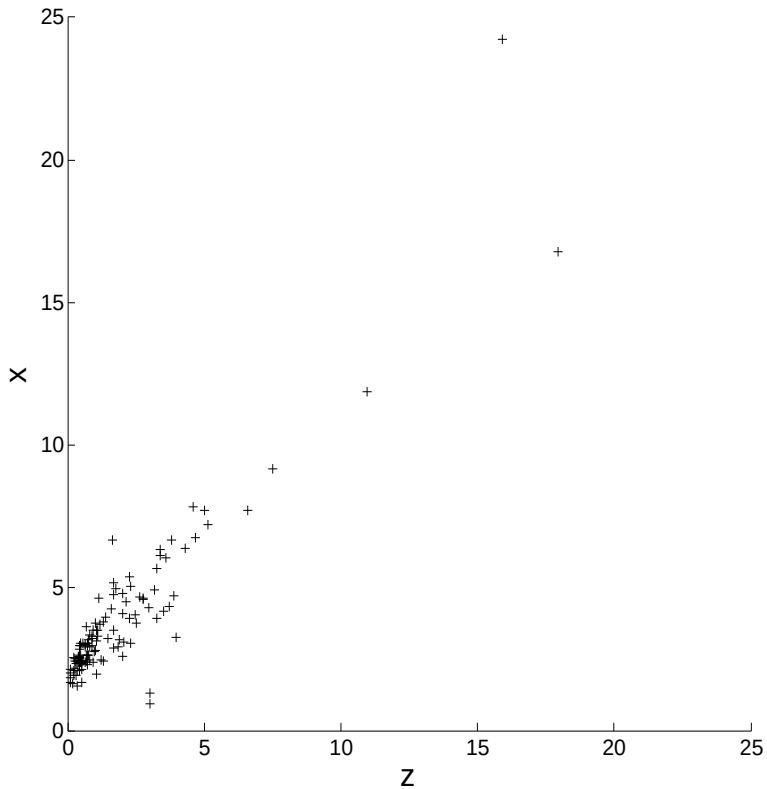
Replacing the minimization of distance $d(\mathbf{W})$ by the maximization of function

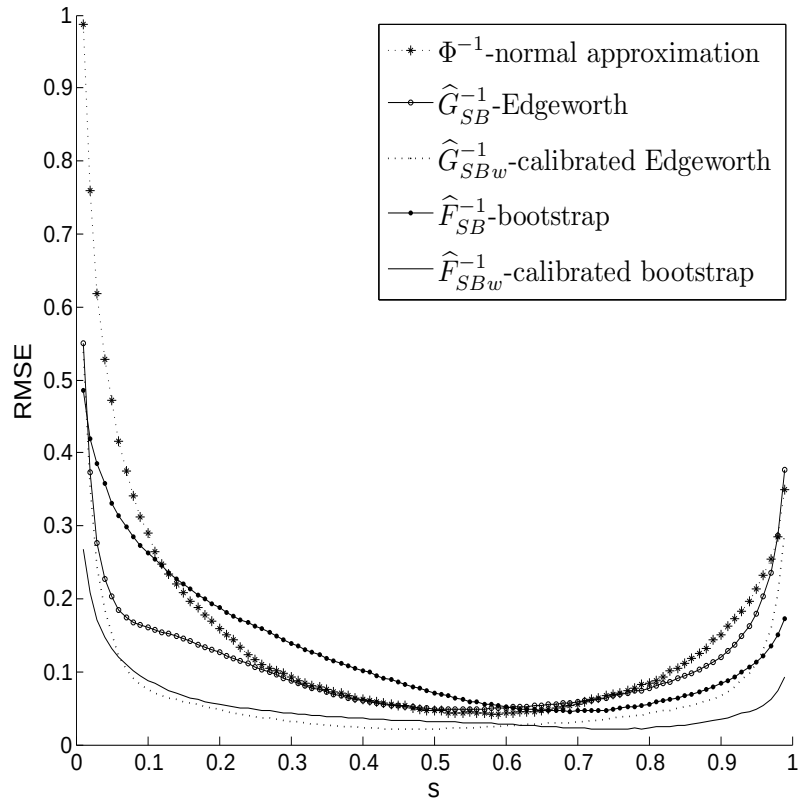
$$g(\mathbf{W}) = \sum_{b=1}^{B} \sum_{r=1}^{R} \log(w_{br}),$$

the calibrated estimation becomes a finite population version of the empirical likelihood (EL) method from Chen and Qin (1993). The calibration and EL yield very similar results, but the weights of EL method cannot be written explicitly.

Simulation (A)

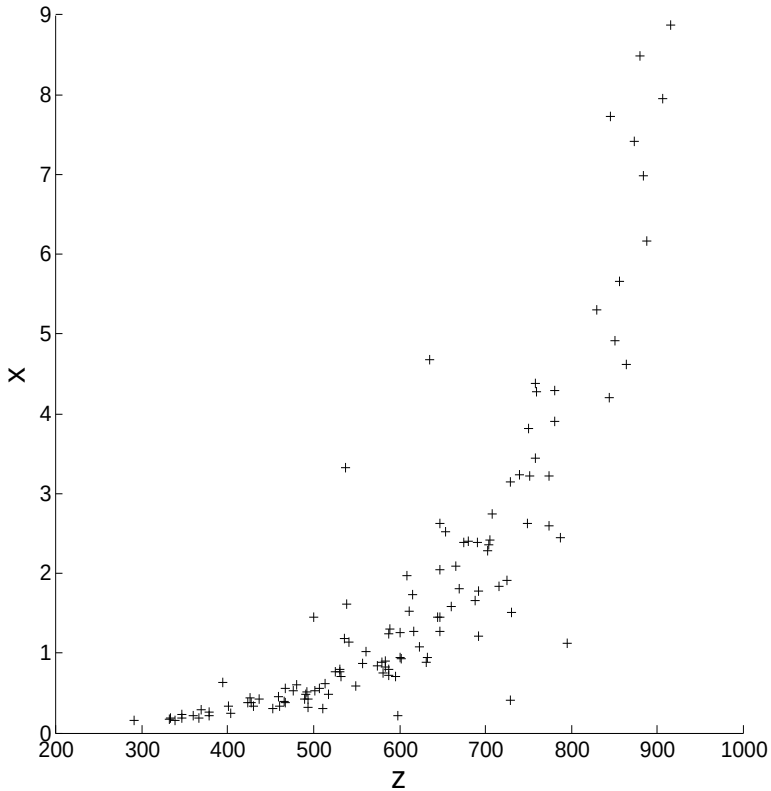
The variables x and z in the population $\mathcal{U}_1$ ($N = 120$ and $n = 40$), and the approximations for the trimmed mean (two largest observations are trimmed).

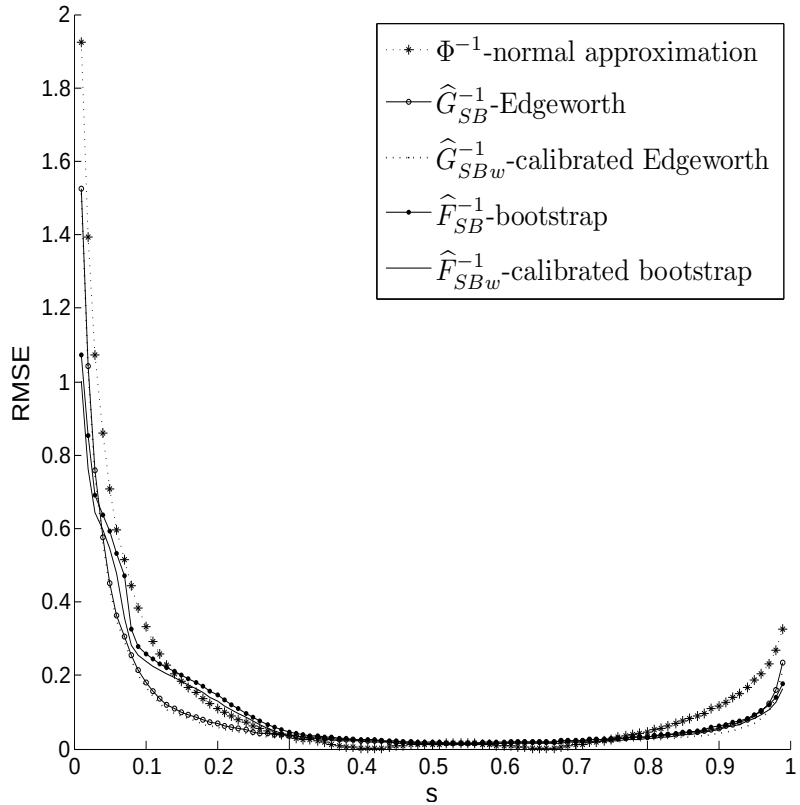




Simulation (B)

The variables x and z in the population $\mathcal{U}_2$ ($N = 120$ and $n = 40$), and the approximations for the Gini mean difference statistic.





Several conclusions

- If good auxiliary information is available, then the calibrated approximations improve the respective approximations based only on the sample data.
- The calibrated bootstrap and saddlepoint approximations adapt better to estimate extreme quantiles and are less biased than the calibrated Edgeworth.
- The calibrated saddlepoint approximation is slightly worse than the calibrated bootstrap. To get better saddlepoint approximations, use higher-order terms of the Hoeffding decomposition!

References

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

Bloznelis, M. (2003). Edgeworth expansions for Studentized versions of symmetric finite population statistics. *Lithuanian Mathematical Journal*, **43**, 221–240

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

Chen, J., Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**, 107–116

Čiginas, A. (2013). Second-order approximations of finite population L-statistics. *Statistics*, **47**, 954–965

Čiginas, A., Pumputis, D. (2019a). Calibrated Edgeworth expansions of finite population L-statistics. *Mathematical Population Studies*, pp. 1–22, http://dx.doi.org/10.1080/08898480.2018.1553408

Čiginas, A., Pumputis, D. (2019b). Calibrated bootstrap and saddlepoint approximations of finite population L-statistics. *Lithuanian Mathematical Journal* (to appear)

Dai, W., Robinson, J. (2001). Empirical saddlepoint approximations of the Studentized mean under simple random sampling. *Statistics and Probability Letters*, **53**, 331–337

Deville, J.C., Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382

Easton, G.S., Ronchetti, E. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association*, **81**, 420–430

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