MEAN ESTIMATION WITH ROBUST CALIBRATED ESTIMATORS

Taras Shevchenko National University of Kyiv

I. Rozora and O. Lukovych

BaNoCoss, Helsinki,

24-28 August, 2015

Outline

- Introduction
- Asymptotic normality
  - Sample mean
  - Sample median
  - Trimmed mean
  - Median of Walsh Averages
- Calibration approach to estimation
- Calibrated Trimmed Mean
- Calibrated Median of Walsh Averages
- Example
- References

Asymptotic normality

Let $(X_1,\ldots,X_n)$ be a sample. (Xi are i.i.d.r.v. with cdf F)

Definition. T is called a *statistics (estimator)* if it is an arbitrary borelean function of sample $(X_1,\ldots,X_n)$.

Definition. The statistic $T=T_n$ is called *asymptotically normal* if there exists such numeric sequences a_n and b_n that

$$\frac{T_n - a_n}{b_n} \xrightarrow{d} Z \sim N(0,1)$$

Asymptotic normality

In general case $b_n = \frac{1}{\sqrt{n}}$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim N(0, \sigma^2(\theta)),\ n \to \infty$$

Definition. The value $\sigma^2(\theta)$ is called an ***asymptotic variance*** of asymptotically normal estimator $\hat{\theta}_n$.

Asymptotic normality

- **SAMPLE MEAN**

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

From CLT follows that (in this case $EX_1=\theta$, $\mathrm{Var}\, X_1<\infty$)

$$\sqrt{n}(\hat{\mu}-\theta) \xrightarrow{d} Z \sim N(0, \mathrm{Var}\, X_1), \quad n \to \infty.$$

Asymptotic normality

- **Definition.** CDF F belongs to the *class of symmetric continuously differentiable distributions* (Ω_s) if there exists such constant c: 0<c≤ ∞ that

  F(-c)=0, F(c)=1 and

  on (-c;c) has even continuous and positive density function p(x).

- **Definition.** A relative asymptotic efficiency of as. normal estimator $\hat{\theta}_1$ to as. normal estimator $\hat{\theta}_2$ is called the value

$$e_{\hat{\theta}_1,\hat{\theta}_2} = \frac{\sigma_2^2}{\sigma_1^2}$$

Asymptotic normality

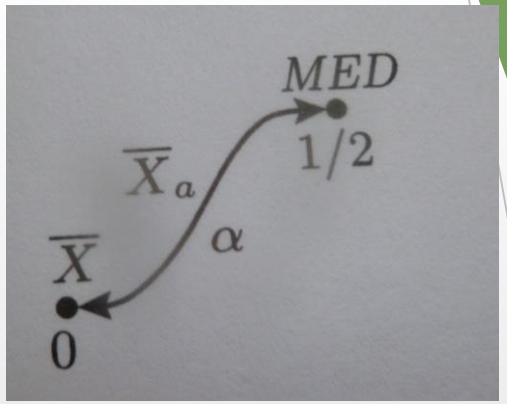
- **SAMPLE MEDIAN**

$$MED = \begin{cases} X_{(k+1)}, & n = 2k+1, \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k. \end{cases}$$

- **Theorem.** Let the elements of sample X_i have cdf $F(x-\theta)$ with density function $p(x)$, where $F \in \Omega_s$. Then

$$\sqrt{n}(\mathrm{MED} - \theta) \to \xi \sim \mathrm{N}\left(0, \frac{1}{4p^2(\theta)}\right)$$

Asymptotic normality



- **TRIMMED MEAN**

$$\alpha \in (0,1/2), \quad k = [\alpha N] \qquad \bar{X}_\alpha = \frac{1}{N-2k}\left(X_{(k+1)} + \ldots + X_{(N-k)}\right)$$

- **Theorem.** Let the elements of sample X_i have cdf F(x-θ) with density function p(x), where F ϵ Ω_s. Then

$$\sqrt{n}(\bar{X}_\alpha - \theta) \to \xi \sim N\left(0, \sigma_\alpha^2\right), \quad n \to \infty,$$

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{x_{1-\alpha}} t^2 p(t)\,dt + \alpha x_{1-\alpha}^2\right],$$

$x_{1-\alpha}$ is a (1-α)-quantile of F.

Asymptotic normality

$$\bar{X}_{\alpha} = \frac{1}{N-2k}\left(X_{(k+1)} + \ldots + X_{(N-k)}\right)$$

- **Example.** Consider N(0,1). Then

12,5% data protection

α	0	1/20	1/8	1/4	3/8	1/2
$e_{\bar{X}_{\alpha},\bar{X}}$	1,00	0,99	0,94	0,84	0,74	0,64

Loss of efficiency 6%

- **Theorem.** For any F ϵ Ω_s

$$(1-2\alpha)^2 \le e_{\bar{X}_{\alpha},\bar{X}}(\mathrm{F}) \le \infty.$$

α	0	1/20	1/8	1/4	3/8	1/2
$(1-2\alpha)^2$	1,00	0,81	0,56	0,25	0,06	0,00

Loss of efficiency 44%

Asymptotic normality

- **MEDIAN OF WALSH AVERAGES**

Consider M=n(n-1)/2 new r.v. $Z_k = \frac{1}{2}\left(X_i + X_j\right), \quad i \le j.$

$$W = MED\left\{Z_1, ..., Z_M\right\}.$$

- **Theorem.** Let the elements of sample X_i have cdf F(x-θ) with density function p(x), where F ϵ Ω_s. Then

$$\sqrt{n}(\mathrm{W} - \theta) \to \xi \square \mathrm{N}\left(0, \sigma_F^2\right)$$

$$\sigma_F^2 = \frac{1}{E(\mathrm{F})}, \quad E(\mathrm{F}) = 12\left(\int_R p^2(\mathrm{t})\,\mathrm{d}t\right)^2$$

Asymptotic normality

$$W = MED\{Z_1, ..., Z_M\}.$$

- **Example.** Consider N(0,1). Then $e_{W,\bar{X}} \approx 0,955$

Loss of efficiency 4,5%

- **Theorem.** For any F ϵ Ω_s

$$e_{W,\bar{X}}(F) \geq 108/125 \approx 0,864.$$

Loss of efficiency only 14%

y

Calibration approach to estimation

- Finite population: $U = \{1, 2, ..., k, ..., N\}$
- Variable of interest: y
- Parameter of interest: mean $\mu_y = \frac{\sum_U y_k}{N}$

Calibration approach to estimation

Definition. (Deville and Särndal (1992))

$$\hat{\mu}_{CAL,y} = \frac{\sum_{k \in s} w_k y_k}{N}$$

is called calibrated estimator of μ_y if

- it estimates the known mean μ_x without error: $E\hat{\mu}_{CAL,x} = \mu_x$
- the distance between the weights d_k and weights w_k is minimal according to the loss function

$$L(w,d) = L(w_k, d_k, k \in s)$$

The minimum distance method to find calibrated mean

Considering the weights in such form

$$w_k = d_k(1 + \mathrm{q}_k \mathrm{x}_k^{'} \lambda)$$

and solving calibration equation with respect to λ

obtain

$$\hat{\mu}_{CAL,y} = \frac{\sum_{k \in s} w_k y_k}{N} = \hat{\mu}_y^{HT} + \hat{\beta}(\mu_x - \hat{\mu}_x^{HT}),$$

where

$\hat{\beta} = (\tilde{x}\tilde{x}^{'})^{-1}\tilde{x}^{'}\tilde{y}, \quad \tilde{x} = (\tilde{x}_k)_{k \in s}$ is a matrix with corrected vectors $\tilde{x}_k = \sqrt{q_k d_k}\, x_k$

$\tilde{y}_k = \sqrt{q_k d_k}\, y_k$

Calibrated cdf and median

Consider the Heaviside function $$I(t)=\begin{cases}1, & t\geq 0,\\ 0, & t<0.\end{cases}$$

The cumulative distribution function of y is $$\mathrm{F}_y(t)=\sum_{k\in U}\frac{I(\mathrm{t}-\mathrm{y}_k)}{N}$$

The α-quantile of the finite population $$\mathrm{Q}_{y\alpha}=\inf\{\mathrm{t}\in\mathrm{R}\,|\mathrm{F}_y(\mathrm{t})\geq\alpha\}$$

The median of y is $$med_y=Q_{y/2}$$

Calibrated cdf and median

An estimator of cdf of y is $$\hat{\mathrm{F}}_y(t) = \frac{\sum_{k \in s} d_k I(\mathrm{t} - \mathrm{y}_k)}{\sum_{k \in s} d_k}$$

An estimator of median of y is $$\hat{med}_y = \inf\{\mathrm{t} \in \mathrm{R} \,|\, \hat{\mathrm{F}}_y(\mathrm{t}) \geq 1/2\}$$

An estimator of α-quantile $$\hat{Q}_{\alpha y} = \inf\{\mathrm{t} \in \mathrm{R} \,|\, \hat{\mathrm{F}}_y(\mathrm{t}) \geq \alpha\}$$

Calibrated cdf and median

Calibrated estimators

$$\hat{\mathrm{F}}_{yCAL}(t) = \frac{\sum_{k \in s} w_k I(\mathrm{t} - \mathrm{y}_k)}{\sum_{k \in s} w_k},$$

$$\hat{med}_{CAL,y} = \inf\{\mathrm{t} \in \mathrm{R} \,|\, \hat{\mathrm{F}}_{yCAL}(\mathrm{t}) \geq 1/2\}$$

To find weights w_k consider 2 different approaches:

- ✓ (Harms and Duchesne (2006)) Complete auxiliary information is not required;
- ✓ (Rueda et al. (2007)) Complete auxiliary information is required.

Calibrated median
Harms and Duchesne (2006)

Known N, and known medians med_{x_j} for j = 1, 2, ..., J .

The calibration equations: $$\sum_{k \in s} w_k = N, \quad \hat{med}_{x_j} = med_{x_j}, \quad j = 1, ..., J$$

minimization the chi-square distance $$L(w,d) = \sum_s (w_k - d_k)^2 / (q_k d_k) \to \min$$

Calibrated median

Rueda et al. (2007)

Model calibration

Using the known x_k, compute first predictions $\hat{y}_k = \hat{\beta}' x_k$ for $k \in U$,

$$\hat{\beta} = (\tilde{x}\tilde{x}')^{-1}\tilde{x}'\tilde{y}, \qquad \tilde{x} = (\tilde{x}_k)_{k\in s}, \qquad \tilde{x}_k = \sqrt{q_k d_k}\, x_k, \qquad \tilde{y}_k = \sqrt{q_k d_k}\, y_k.$$

The calibration equations:

$$\frac{\sum_{k\in s} w_k I(\mathrm{t}_j - \hat{\mathrm{y}}_k)}{N} = F_{\hat{y}}(\mathrm{t}_j), \quad \mathrm{j} = 1,...,\mathrm{J}$$

:

minimization the chi-square distance

$$\mathrm{L}(w,d) = \sum_s (w_k - d_k)^2 / (q_k d_k) \to \min$$

Calibrated trimmed mean

Use calibrated cdf

$$\hat{\mathrm{F}}_{yCAL}(t) = \frac{\sum_{k \in s} w_k I(\mathrm{t} - \mathrm{y}_k)}{\sum_{k \in s} w_k}.$$

Construct a sequence

$$\hat{y}_{(\mathrm{k})}^{Cal} = \inf\{\mathrm{t} \in \mathrm{R} \,|\, \hat{\mathrm{F}}_{yCAL}(\mathrm{t}) \geq 1/\mathrm{k}\}, \quad \mathrm{k} = \overline{1,\mathrm{n}}.$$

$$\alpha \in (0,1/2), \quad k = [\alpha n]$$

$$\hat{Y}_{\alpha}^{Cal} = \frac{1}{n-2k}\left(\hat{y}_{(\mathrm{k}+1)}^{Cal} + \ldots + \hat{y}_{(\mathrm{n}-\mathrm{k})}^{Cal}\right)$$

Calibrated Median of Walsh Averages

Consider M=n(n-1)/2 new r.v.

$$Z_k^{Cal} = \frac{1}{2}\left(\hat{y}_{(i)}^{Cal} + \hat{y}_{(j)}^{Cal} \right), \quad i \le j.$$

$$W^{Cal} = MED\left\{ Z_1^{Cal}, \ldots, Z_M^{Cal} \right\}.$$

Example

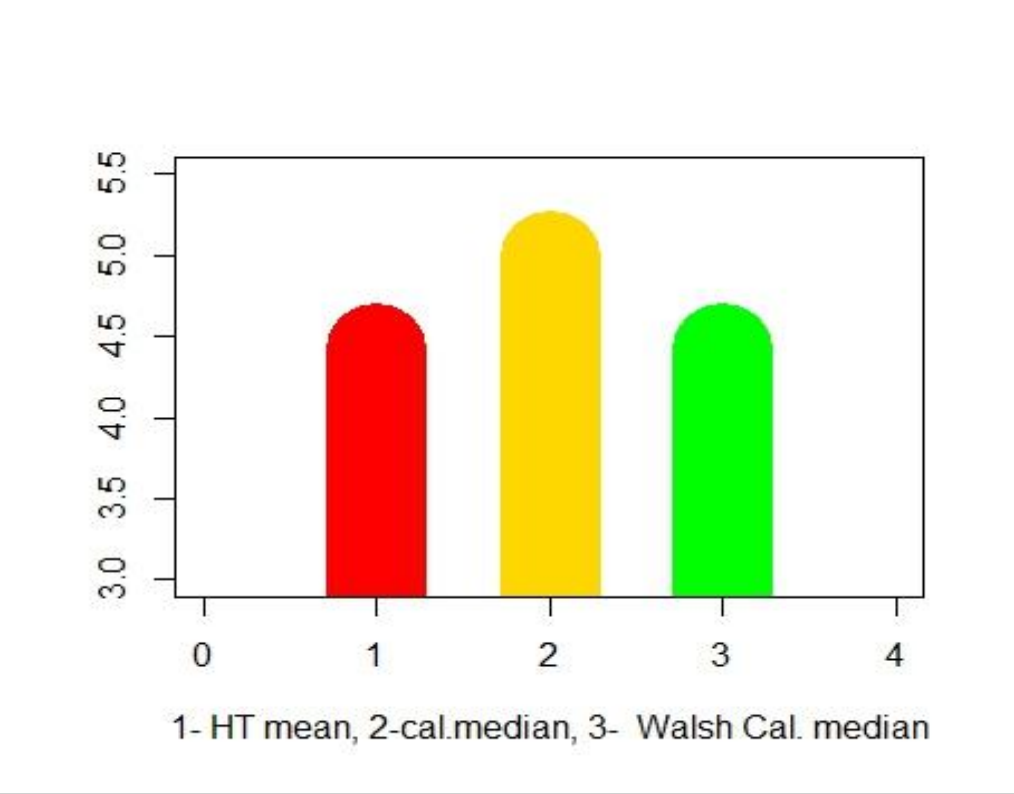
Interested variable

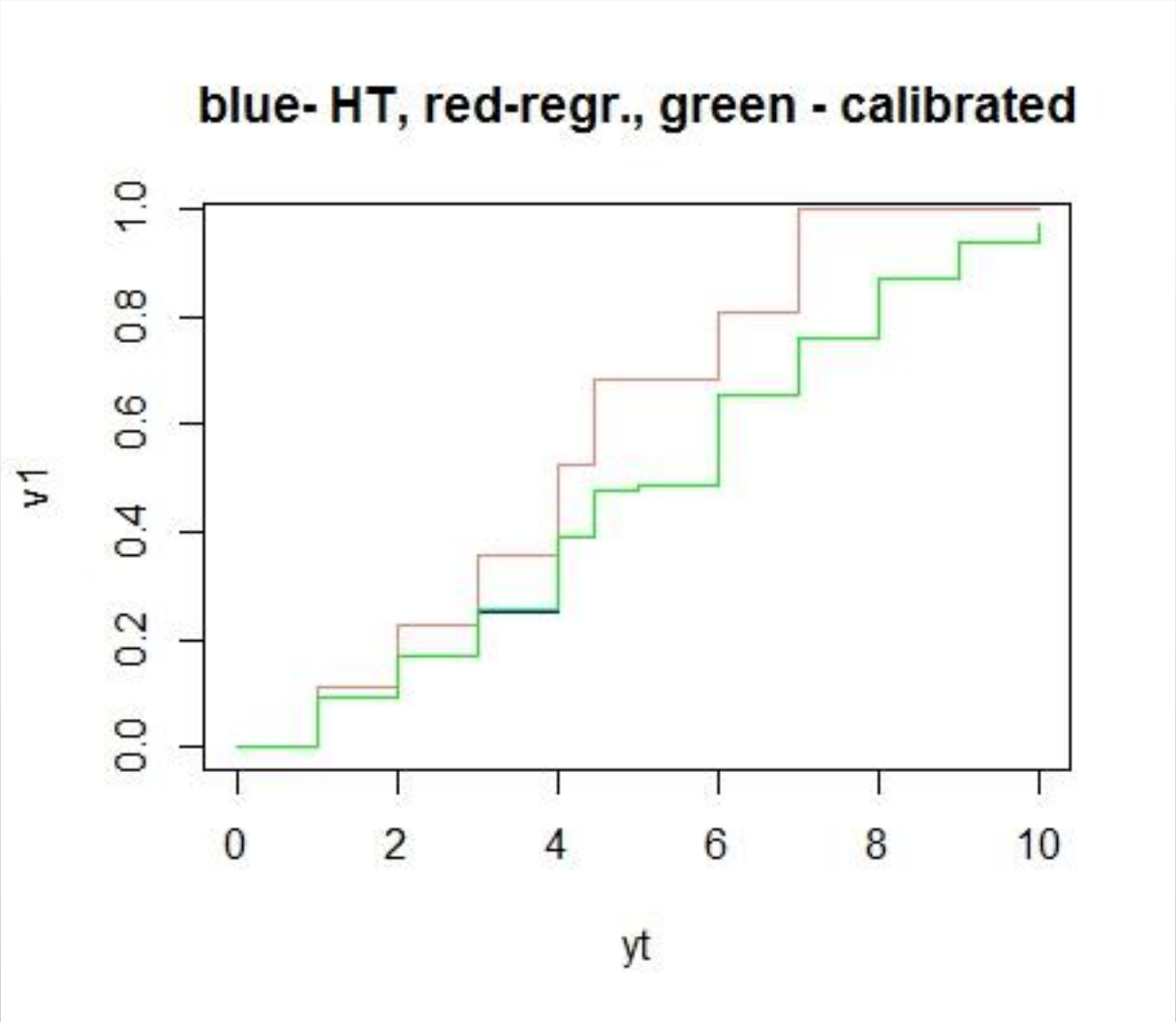
Most people can be trusted

or you can't be too careful!

Example

Auxiliary information variable – “Tv watching, total time on average weekday”




blue- HT, red-regr., green - calibrated
v1
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
yt

References

- Alexander, C.H. (1987). A class of methods for using person controls in household weighting. Survey Methodology, 13, 183-198.
- Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working paper, Census Operations Section, Social Surveys Methods Division, Statistics Canada.
- Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. Journal of Official Statistics, 3, 141-153.
- Chambers, R.L. (1996). Robust caseweighting for multipurpose establishment surveys. Journal of Official Statistics, 12, 332.
- Deville, J. C., Särndal, C. E. (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide Food Consumption Survey. Survey Methodology, 20, 75-85.
- Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14, 333-362.
- Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.
- Kalton, G., and Flores-Cervantes, I. (1998). Weighting methods. In New Methods for Survey Research (Eds. A.Westlake, J. Martin, M. Rigg and C. Skinner),. Berkeley, U.K.: Association for Survey Computing.
- Kovaĉević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. Proceedings of the Survey Methods Section, Statistical Society of Canada, 139-144.

References

- Lehtonen, R., Särndal, C.E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. Survey Methodology, 29, 3344.
- Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.
- Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. Journal of Statistical Planning and Inference, 137, 435-448.
- C.-E. Särndal (2007). The calibration approach in survey theory and practice. Survey methodology, 33(2), 99-119.
- Théberge, A. (2000). Calibration and restricted weights. Survey Methodology, 26, 99-107.
- Särndal, C., Swensson, B. & Wretman, J. (2003). Model assisted survey sampling. Springer Verlag.
- Wu, C., and Sitter, R.R. (2001). A modelcalibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Thank you!!!