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. (\(X_i\) 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

\[
\frac{T_n - a_n}{b_n} \xrightarrow{d} Z \equiv N(0,1)
\]
Asymptotic normality

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

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

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 \( E X_1 = \theta, \ Var X_1 < \infty \))

\[ \sqrt{n} (\hat{\mu} - \theta) \xrightarrow{d} N(0, \ Var X_1), \ n \to \infty. \]
Asymptotic normality

**Definition.** CDF $F$ belongs to the *class of symmetric continuously differentiable distributions* $(\Omega_s)$ if there exists such constant $c$: $0 < c \leq \infty$ that

$$F(-c) = 0, \quad F(c) = 1$$

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

$$\varepsilon_{\hat{\theta}_1, \hat{\theta}_2} = \frac{\sigma_2^2}{\sigma_1^2}$$
Asymptotic normality

**SAMPLE MEDIAN**

\[
MED = \begin{cases} 
X_{(k+1)}, & n = 2k+1, \\
\frac{X_k + X_{(k+1)}}{2}, & n = 2k.
\end{cases}
\]

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

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

**TRIMMED MEAN**

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

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

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

\[ \sqrt{n}(\overline{X}_\alpha - \theta) \rightarrow \xi \quad N(0, \sigma^2_\alpha), \quad n \rightarrow \infty, \]

\[ \sigma^2_\alpha = \frac{2}{(1-2\alpha)^2} \left[ \int_{x_{1-\alpha}}^{x_{1-\alpha}} t^2 p(t) \, dt + \alpha x_{1-\alpha}^2 \right], \]

\( x_{1-\alpha} \) is a \((1-\alpha)\)-quantile of \( F \).
Asymptotic normality

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

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

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0</th>
<th>1/20</th>
<th>1/8</th>
<th>1/4</th>
<th>3/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{X_{k}}$</td>
<td>1,00</td>
<td>0,99</td>
<td>0,94</td>
<td>0,84</td>
<td>0,74</td>
<td>0,64</td>
</tr>
</tbody>
</table>

- **Theorem.** For any $F \in \Omega$, $e_{X_{k}} \leq \frac{1}{1-2\alpha} \leq \infty$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0</th>
<th>1/20</th>
<th>1/8</th>
<th>1/4</th>
<th>3/8</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1-2\alpha)^2$</td>
<td>1,00</td>
<td>0,81</td>
<td>0,56</td>
<td>0,25</td>
<td>0,06</td>
<td>0,00</td>
</tr>
</tbody>
</table>

- **12.5% data protection**
- **Loss of efficiency 6%**

- **Loss of efficiency 44%**
Asymptotic normality

**MEDIAN OF WALSH AVERAGES**

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

\[
Z_k = \frac{1}{2} (X_i + X_j), \quad i \leq j.
\]

\[
W = MED \{ Z_1, \ldots, Z_M \}.
\]

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

\[
\sqrt{n}(W - \theta) \rightarrow \xi \square N \left( 0, \sigma_F^2 \right)
\]

\[
\sigma_F^2 = \frac{1}{E(F)}, \quad E(F) = 12 \left( \int_R p^2(t) \, dt \right)^2
\]
Asymptotic normality

\[ W = MED\{Z_1, \ldots, Z_M\}. \]

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

\[ e_{W,\bar{X}} \approx 0.955 \]

Loss of efficiency 4.5%

- **Theorem.** For any \( F \in \Omega_s \)

\[ e_{W,\bar{X}}(F) \geq 108/125 \approx 0.864. \]

Loss of efficiency only 14%
Calibration approach to estimation

- Finite population: \( U = \{1, 2, \ldots, k, \ldots, N\} \)

- 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))

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

is called calibrated estimator of \( \mu_y \) if

- it estimates the known mean \( \mu_x \) without error: \( E\hat{\mu}_{\text{CAL},x} = \mu_x \)
- 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

\[ w_k = d_k (1 + q_k x_k' \lambda) \]

and solving calibration equation with respect to \( \lambda \)

give

\[ \hat{\lambda} = \frac{\sum_{k \in S} w_k y_k}{N} = \hat{\mu}_y^{HT} + \hat{\beta}(\mu_x - \hat{\mu}_x^{HT}), \]

where

\[ \hat{\beta} = (\tilde{x}\tilde{x}'^{-1}\tilde{x}' \tilde{y}, \; \tilde{x} = (\tilde{x}_k)_{k \in S} \] 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

\[ F_y(t) = \sum_{k \in U} \frac{I(t - y_k)}{N} \]

The \( \alpha \)-quantile of the finite population

\[ Q_{y\alpha} = \inf\{ t \in \mathbb{R} \mid F_y(t) \geq \alpha \} \]

The median of \( y \) is

\[ med_y = Q_{y/2} \]
Calibrated cdf and median

An estimator of cdf of $y$ is

$$\hat{F}_y(t) = \frac{\sum_{k \in s} d_k I(t - y_k)}{\sum_{k \in s} d_k}$$

An estimator of median of $y$ is

$$\text{med}_y = \inf \{ t \in \mathbb{R} | \hat{F}_y(t) \geq 1/2 \}$$

An estimator of $\alpha$-quantile

$$\hat{Q}_{\alpha \cdot y} = \inf \{ t \in \mathbb{R} | \hat{F}_y(t) \geq \alpha \}$$
Calibrated cdf and median

Calibrated estimators

\[ \hat{F}_{y_{\text{CAL}}} (t) = \frac{\sum_{k \in s} w_k I(t - y_k)}{\sum_{k \in s} w_k}, \]

\[ \text{med}_{\text{CAL}, y} = \inf \{ t \in \mathbb{R} \mid \hat{F}_{y_{\text{CAL}}} (t) \geq 1/2 \} \]

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 $\text{med}_{x_j}$ for $j = 1, 2, ..., J$.

The calibration equations:

$$\sum_{k \in s} w_k = N, \quad \hat{\text{med}}_{x_j} = \text{med}_{x_j}, \quad j = 1, ..., J$$

minimization the chi-square distance

$$L(w, d) = \sum_{s} (w_k - d_k)^2 / (q_k d_k) \rightarrow \min$$
Calibrated median
Rueda et al. (2007)

Model calibration

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

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

The calibration equations:

\[
\sum_{k \in s} w_k I(t_j - \hat{y}_k) \quad \sum_{k \in s} w_k \quad = F_{\hat{y}}(t_j), \quad j = 1, ..., J
\]

minimization the chi-square distance

\[
L(w, d) = \sum_s (w_k - d_k)^2 / (q_k d_k) \to \min
\]
Calibrated trimmed mean

Use calibrated cdf

$$\hat{F}_{y_{\text{CAL}}} (t) = \frac{\sum_{k \in s} w_k I(t - y_k)}{\sum_{k \in s} w_k}.$$

Construct a sequence

$$\hat{y}_{(k)}^{\text{Cal}} = \inf \{ t \in \mathbb{R} | \hat{F}_{y_{\text{CAL}}} (t) \geq 1/k \}, \ k = 1, n.$$

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

Given $$\alpha \in (0, 1/2), \ k = [\alpha n]$$
Calibrated Median of Walsh Averages

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

$$Z^\text{Cal}_k = \frac{1}{2} (\hat{y}^\text{Cal}_{(i)} + \hat{y}^\text{Cal}_{(j)}) , \ i \leq j.$$

$$W^\text{Cal} = MED\{Z^\text{Cal}_1, \ldots, Z^\text{Cal}_M\}.$$
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”
References


References

Thank you!!!