Estimation of Quantiles based on Fay-Herriot Models

Ann-Kristin Kreutzmann\textsuperscript{1}, Nicola Salvati\textsuperscript{2}
Timo Schmid\textsuperscript{1}

\textsuperscript{1} Freie Universität Berlin
\textsuperscript{2} University of Pisa

ITAlian Conference on Survey Methodology (ITACOSM)
June, 2017
Overview

Motivation

Direct Quantile Estimation

Simulation Results

Conclusion
Why Estimating Wealth

- In many countries wealth inequality is higher than labour income inequality
- For instance for the year 2014 in Germany, the Gini coefficient of net income is 30.7 whereas the Gini coefficient of net wealth is 76
- Furthermore, it is assumed that wealth inequality will increase critically in the following decades in many developed countries
Estimation of Wealth in Germany

Figure: Density of net wealth in Germany with the median equal to €59500 (red), the 75% quantile equal to €221700 (blue) and the 90% quantile equal to €470320 (black) from the data source Panel on Household Finances
Synthetic Wealth Data

In order to evaluate the quantile estimators, we use a synthetic wealth distribution.

Figure: Density of the pareto distribution with shape 8 and scale 400000 (up to the 99% quantile)
Simulation Results with Large Sample Size

Figure: Relative bias (left) and bias (right) of the median, the 75% quantile and the 90% quantile when sample sizes are between 250 and 7500 observations.
Estimation of Wealth in Planning Regions in Germany

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>22</td>
<td>37.5</td>
<td>59</td>
<td>340</td>
</tr>
</tbody>
</table>

Table: Planning Regions in Germany (left) and the sample sizes of the Panel on Household Finances in these regions (above)

Research question: Is it still possible to get unbiased estimates for the quantiles when sample sizes get smaller?
Overview

Motivation

Direct Quantile Estimation

Simulation Results

Conclusion
Direct Estimators for Quantiles

- A wide range of definitions for sample quantiles exists for the estimation of the theoretical quantile:

$$Q(p) = F^{-1}(p) = \inf \{ x \in \mathbb{R} : F(x) \geq p \} \quad \text{for} \quad p \in (0, 1)$$

where $F(x)$ is the distribution function
- For instance, the statistical software SAS provides at least 5 definitions and R even at least 12
- Generally, most definitions can be classified in two different categories
Direct Estimator for Quantiles

1. Order statistics or weighted averages of one or two order statistics:

\[ Q_p = (1 - \gamma)X_{[i]} + \gamma X_{[i+1]}, \quad (1) \]

where \( X_{[i]} \) is the ith order statistic, \( n \) is the sample size, the value of \( \gamma \) is a weighting factor (often a function of \( i \))

2. Weighted average of all order statistics with different weights:

\[ Q_p = \sum_{i=1}^{n} W_{n,i}X_{[i]} \quad (2) \]

where \( X_{[i]} \) is the ith order statistic and \( W_{n,i} \) a weighting factor depending on the sample size \( n \) and \( i \)
Weighted Quantiles in R

- Standard definitions do not include sampling weights but there are implementations that enable the incorporation of sampling weights
- Two popular functions in the statistical software R are function `weightedQuantile` from the package `laeken` and function `wtd.quantile` from package `Hmisc` (both category 1)
- Package `Hmisc` also provides the Harrel-Davis quantile estimator but without sampling weights (category 2)
- In order to include a quantile estimator that weights every order statistic we incorporated sampling weights to the Harrel-Davis estimator
According to Alfons et al. (2013) sampling weights can be incorporated to the inverse of the empirical distribution function as follows,

\[
Q_p = \begin{cases} \frac{1}{2}(X_{[j]} + X_{[j+1]}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n} w_i \\ X_{[j+1]} & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n} w_i < \sum_{i=1}^{j+1} w_i \end{cases}
\]

where \(X_{[j]}\) is the jth order statistic of sample observations \(X = (X_1, \ldots, X_n)\), \(w = (w_1, \ldots, w_n)\) are the corresponding sampling weights and \(p \in (0, 1)\)
The standard quantile estimator in R is a linear interpolation of the modes for the order statistics for the uniform distribution on [0,1] (Gumbel 1939).

\[ Q_p = (1 - \gamma)X[i] + \gamma X[i+1] \]  

(3)

where \( i = [1 + (n - 1)p] \), \( \gamma = 1 + (n - 1)p - i \) and \( p \in (0, 1) \). When \( p = 1 \), use the highest value \( X[n] \).

→ Sampling weights are included using \( \tilde{n} = \sum_{j=1}^{n} w_j \) as n and the cumulative sum of the sampling weights to choose the order statistics.
Harrel-Davis Estimator with Sampling Weights

The Harrel-Davis (HD) (Harrel and Davis 1982) estimator with sampling weights can be expressed in terms of weights $W_{w_i,i}$

$$Q_p = \sum_{i=1}^{n} W_{w_i,i} X_{(i)}$$

(4)

with weight $W_{w_i,i}$ that depends on $i$ and the sampling weights $w_i$. 

---

Ann-Kristin Kreutzmann 14 (27)  
Small Area Estimation in R - NTTS 2017
Overview

Motivation

Direct Quantile Estimation

Simulation Results

Conclusion
Generation of Close-to-Reality Sample Sizes

1. Sample sizes of the planning regions in Germany in the Panel on Household Finances in 2014

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>22</td>
<td>37.5</td>
<td>59</td>
<td>340</td>
</tr>
</tbody>
</table>

2. Sample sizes generated for the model-based simulation study

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>25</td>
<td>40</td>
<td>60</td>
<td>65</td>
</tr>
</tbody>
</table>
Relative Bias and Bias Across Domains

In order to evaluate the quantile estimators two measures are considered for three different quantiles: 50%, 75% and 90%

1. Mean relative bias (RB)

\[ RB_i = \frac{1}{R} \sum_{r=1}^{R} \left( \frac{\hat{\theta}_{Smp}^{r, i} - \hat{\theta}_{Pop}^{r, i}}{\hat{\theta}_{Pop}^{r, i}} \right) \]  

(5)

2. Mean bias (Bias)

\[ \text{Bias}_i = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}_{Smp}^{r, i} - \hat{\theta}_{Pop}^{r, i} \right) \]  

(6)

with Monte Carlo iterations \( r = 1, \ldots, 5000 \) and domains \( i = 1, \ldots, 130 \)
RB for Close-to-Reality Sample Sizes

**Figure:** Relative biases in % of the median, the 75% quantile and the 90% quantile for the wtd, laeken and HD definition when sample sizes are between 5 and 65 observations
**RB for Moderate Sample Sizes**

**Figure:** Relative biases in % of the median, the 75% quantile and the 90% quantile for the wtd, laeken and HD definition when sample sizes are between 25 and 65 observations.
Bias for Moderate Sample Sizes

Figure: Biases of the median, the 75% quantile and the 90% quantile for the wtd, laeken and HD definition when sample sizes are between 25 and 65 observations.
Differences across Quantile Definitions

**Figure:** Bias using laeken and bias using HD of median, 75% and 90% quantiles when sample sizes are between 25 and 65
Overview

Motivation

Direct Quantile Estimation

Simulation Results

Conclusion
Conclusion

- The considered quantile definitions do not differ a lot under a certain pareto setting.
- When sample sizes are moderately large (between 25 and 65) the relative bias of the median, the 75% and the 90% quantile across domains is around 1% to 2%.
- When sample sizes are small (less than 25 observations) the relative bias is larger.
Further Research

- The estimation of quantiles when sample sizes are small (5 to 25 observations): is an unbiased estimation possible?  
  → Parametric approaches or Bayes estimation
- Even when sample sizes are moderate (25 to 65 observations) variances are assumed to be large
- In order to increase the efficiency of estimates we use a small area estimation model, particularly the Fay-Herriot model
The research is funded by the German Research Foundation within the project QUESSAMI.

Thank you very much for your attention.

Ann-Kristin Kreutzmann (Ann-Kristin.Kreutzmann@fu-berlin.de)
References

*Journal of Statistical Software* 54(15), 1-25.

*Monatsbericht 61*

*URL: https://www.destatis.de*

Eurosystem Household Finance and Consumption Network (2013), The Eurosystem Household Finance and Consumption Survey - Results from the first wave.
*Statistics Paper Series 2, European Central Bank*
References

Fay, R.E. & Herriot, R.A. (1979), 'Estimates of income for small places: An application of James-Stein procedures to census data'.

Gumbel, E.J. (1939), 'La Probabilité des Hypothèses'.

Biometrika 69(3), 635-640.

John Wiley & Sons