Estimation of Complex Population Functionals & Randomized Response Theory

Perri Pier Francesco

University of Calabria - DESF

Kick-off Meeting, PRIN-2014, 2012F42NS8

Roma, February 12-13, 2014

Complex estimation

Sampling surveys frequently involve the estimation of target parameters, e.g.

- totals and mean - easy to obtain
- ratios, variances, regression coefficients, cdf, quantiles, inequality indicators - harder to find

Diana and Perri (2007, 2010, 2011, 2013) derived class of estimators for the population mean using auxiliary variables (also in the presence of nonrespondents).

Usually, the parameters of interest are nonlinear functions of population totals. Consequently, variance estimation is not a trivial matter and requires specific procedures. Methods for variance estimation can be classified in

1. resampling methods (jackknife, bootstrap, balanced repeated replication)
2. linearization methods

Current interest: *linearization of complex parameters and variance estimation in the design-based approach*

Linearization

Under the usual design-based approach

- $P = \{1, \ldots, N\}$ is a fixed population
- y_i is the value of the study variable Y on the i-th individual
- $\pi_i > 0$ and $\pi_{ij} > 0$ are the first and second-order inclusion probabilities under a sampling design $P(S = s) = p(s)$
- θ is a nonlinear function of population totals to be estimated

The rationale behind linearization is to obtain a linearized variable v_i for each observation y_i such that

$$\widehat{\theta} - \theta \approx \sum_{i \in S} \frac{v_i}{\pi_i} - \sum_{i \in P} v_i \Rightarrow \text{Var}(\widehat{\theta}) \approx \text{Var}(\sum_{i \in S} \frac{v_i}{\pi_i})$$

Variance estimation may be achieved by means of

$$\widehat{\text{Var}}(\widehat{\theta}) = \sum_{i,j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \widehat{v}_i \widehat{v}_j$$

Linearization via influence function

Several techniques have been proposed to provide the v_i's: Woodruff (J Am Stat Assoc, 1971); Deville (Surv Methodol, 1991); Kovačević and Binder (J Off Stat, 1997); Demnati and Rao (Surv Methodol, 2004): Goga et al (Biometrika, 2009).

Deville's approach

- The finite and discrete measure

$$M = \sum_{i \in P} \delta_{y_i}$$

  and its empirical counterpart

$$\widehat{M} = \sum_{i \in S} \frac{\delta_{y_i}}{\pi_i}$$

  are considered
- θ is rephrased as a functional F with respect to M, $\theta = F(M)$

Linearization via influence function

Under this framework:

- plug-in estimator of $\theta \Rightarrow \widehat{\theta} = F(\widehat{M})$
- $v_i = \mathsf{IF}_F(y_i; M)$ and $\widehat{v}_i = \mathsf{IF}_F(y_i; \widehat{M})$ where

$$\mathsf{IF}_F(u; M) = \lim_{t \to 0} \frac{1}{t} \left(F(M + t\delta_u) - F(M) \right)$$

represents the influence function in the design-based approach.

Some computational rules for IF are given in Deville (1991) and Langel and Tillé, (J R Stat Soc A Stat, 2013). Recently, Barabesi-Diana-Perri (2014) have worked out a simple rule for obtaining the influence function.

Barabesi-Diana-Perri differential rule

Let us consider a functional which may be expressed as

$$F(M) = \int \psi_y(U_y(M))\, \mathrm{d}M(y),$$

where $U_y(M) = (U_{1,y}(M), \ldots, U_{k,y}(M))^{\mathsf{T}}$ is a vector of further functionals (eventually) indexed by y and $\psi_y : \mathbb{R}^k \mapsto \mathbb{R}$ is a function family assumed to be differentiable and regularly indexed by y.
Let us suppose that θ is a member of the functional family F, or may be expressed as $\varphi(F(M)) = (\varphi \circ F)(M)$ where $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a differentiable function.

Proposition. If $U_{j,y}$ is Fréchet differentiable for each j, the IF of F is given by

$$\mathrm{IF}_F(u; M) = \psi_u(U_u(M)) + \int \nabla \psi_y(U_y(M))^{\mathsf{T}} \mathrm{IF}_{U_y}(u; M)\, \mathrm{d}M(y),$$

where $\mathrm{IF}_{U_y}(u; M) = (\mathrm{IF}_{U_{1,y}}(u; M), \ldots, \mathrm{IF}_{U_{k,y}}(u; M))^{\mathsf{T}}$. The IF of $(\varphi \circ F)$ follows as

$$\mathrm{IF}_{\varphi \circ F}(u; M) = \varphi'(F) \mathrm{IF}_F(u; M).$$

Application to inequality indexes

- The result stated in the previous Proposition has been employed to obtain the IF of some of the most popular inequality indexes (Cowell, *Measuring Inequality*, 2011)
  - Gini concentration index
  - Generalized entropy family
  - Atkinson family
  - Amato index (Arnold, Stat Probabil Lett, 2012)
  - Zenga (Statistica & Applicazioni, 2007; Lagel and Tillé, Metrika, 2012)
- The study of indicators based on quantiles is currently under investigation

The Gini index

Let us focus on the Gini index (Berger, J Off Stat, 2008):

$$G(M) = \int \frac{2yH_y(M)}{N(M)T(M)}\, \mathrm{d}M(y) - 1\ ,$$

where

$$N(M) = \int \mathrm{d}M(x) \quad \text{and} \quad T(M) = \int x\, \mathrm{d}M(x).$$

Let

$$H_y(M) = \int I_{[x,\infty[}(y)\, \mathrm{d}M(x) \quad \text{and} \quad K_y(M) = \int xI_{[y,\infty[}(x)\, \mathrm{d}M(x).$$

We have $U_y(M) = (H_y(M), N(M), T(M))^\mathsf{T}$, while $\psi_y(U_y(M)) = \frac{2yH_y(M)}{N(M)T(M)}$ and $\varphi(F) = F - 1$. Moreover

$$\nabla\psi_y(U_y(M)) = \frac{2y}{NT}\left(1, -\frac{H_y}{N}, -\frac{H_y}{T}\right)^\mathsf{T} \quad \text{and} \quad \mathsf{IF}_{U_y}(u;M) = (I_{[u,\infty[}(y), 1, u)^\mathsf{T}$$

After some algebra it follows that

$$\mathsf{IF}_G(u;M) = \frac{2}{NT}\left(uH_u + K_u\right) - (G+1)\left(\frac{1}{N} + \frac{u}{T}\right)\ .$$

Asking sensitive questions

Surveys on sensitive or highly personal issues (e.g., income, wealth, living conditions) are likely to meet with:

1. refusal to cooperate (*unit-non-response*)
2. refusal to answer specific questions (*item-non-response*)
3. untruthful or misleading responses (*measurement error*)

... this is particularly true when sensitive data are collected by means of direct questioning methods (e.g., f2f)

Asking sensitive questions

Income is notoriously considered a sensitive character in the sense that people are reluctant to disclose it, mostly in the case of income from self-employment, property and financial assets (Neri and Zizza, 2010 Economic WP, 777, Bank of Italy, 2010)

"There appears to be considerable evidence in the literature on economic surveys that respondents to direct questions tend to understate income." (Greeberg et al., J Am Stat Assoc, 1971)

- Consequently, this issue may result in seriously-biased estimates of inequality indicators
- To alleviate this problem, the respondent cooperation has to be increased
- Survey modes which ensure anonymity may improve confidentiality and, consequently, ensure more reliable information

Increasing respondent cooperation

The Randomized Response Theory (Warner, J Am Stat Assoc, 1965) is a useful tool for tackling sensitive questions

- *A randomizing device (deck of cards, die, coin,...) is used to hide the answer in the sense that respondents reply to one of two or more questions depending on the result of the device*
- *Privacy is protected since respondents do not reveal to anyone the question that has been selected, and both the interviewer and the researcher are unaware of the randomizer outcome*

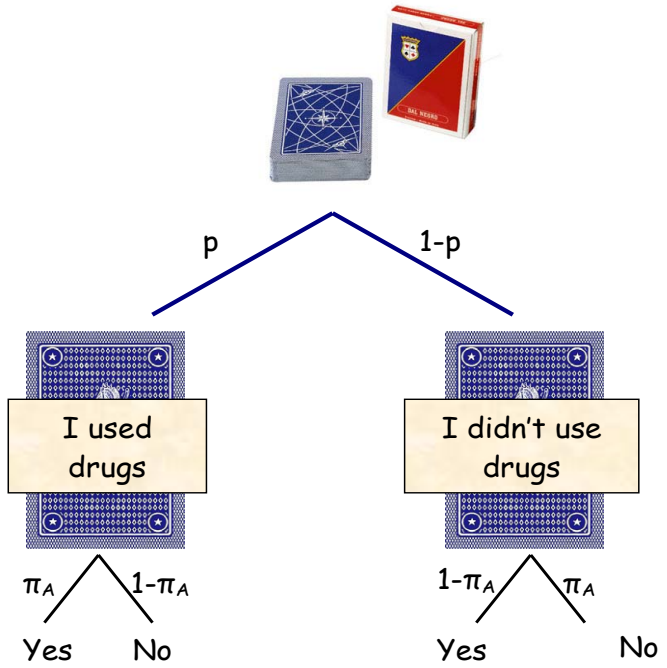
- Scrambling response methods for quantitative variables ...

Example: the Warner model

Have you ever used drugs?

Such direct questioning can be met with refusal to cooperate or untruthful answers. To induce greater cooperation while ensuring anonymity to respondents, the Warner strategy can be adopted

1. a deck of cards is considered as randomizer
2. on each card is written one of two statements "*I used drugs*" or "*I didn't use drugs*" in the proportions of p and $1-p$
3. respondents are asked to select at random a card and report a *yes* or *no* response depending on whether their true status matches the statement on the selected card or not



Respondents are assumed to answer truthfully and, since they are instructed to not reveal to anyone the question/card selected, their true status remains unknown and thus privacy is protected

RRT applications

- The validity of the RRT upon conventional data collecting methods has been assessed by many studies: van der Heijden et al. (Soc Meth R, 2000), Lara et al. (Soc Meth R, 2004), Lensvelt-Mulders et al. (Soc Meth R, 2005)

- **Some recent applications**

  - ✓ Lensvelt-Mulders et al. (J R Stat Soc A Stat, 2006): estimation of the prevalence of fraud in the area of disability benefits
  - ✓ Lara et al. (Soc Meth R, 2004, 2006): estimation of the prevalence of abortion in Mexico
  - ✓ Ostapczuk et al. (Eur J Soc Psychol, 2009), Krumpal (Soc Sci Res, 2012): the issue of xenophobia and anti-semitism in Germany
  - ✓ Arnab and Singh (J Stat Plan Infer, 2010): the impact of HIV/AIDS infection in Botswana

Personal contribution

- Much of the literature on the RRT focuses on the estimation of population proportions, totals and means (Giordano and Perri, 2012; Diana and Perri 2009-2013; Perri and van der Heijden, 2012)
- Recently, Barabesi-Diana-Perri
  - modified the HT-estimator in the presence of true response, non-response and randomized response (MASA 2014)
  - used the RRT to simultaneously estimate the prevalence of an hidden group (e.g. *tax evaders*) and the distribution function of a sensitive character (e.g. *income*) within the group (Metrika, 2013)
  - estimated the Gini index in a RRT framework and assessed the accuracy of the estimates through a simulation study based on income data from SHIW 2010 (under review)
- Barabesi-Diana-Perri are currently using the *differential rule* to derive the IF of indicators of poverty and inequality under a RRT setup

Conclusions

- Estimation of inequality indicators
- Sensitive items
- Optimal use of auxiliary information at the estimation stage
- Nonresponse