Fisher's noncentral hypergeometric distribution for the size estimation of graduated unemployed in Italy

Veronica Ballerini[1]
veronica.ballerini@unifi.it

a joint work with
Brunero Liseo[2]

[1]Now at DiSIA, University of Florence
[2]MEMOTEF, Sapienza University of Rome

XII Giornata della Ricerca MEMOTEF
June 1, 2022

During my PhD we have addressed the following questions:

1. How to estimate the population size when a single list of records is available?
2. How to estimate/take into account the bias given by a MNAR mechanisms when estimating a population size?

The motivating study

How many people are still unemployed one year after they graduated?

Dataset: Graduates' Employment Status Survey (GESS) by **AlmaLaurea**

- response rates $< 100\%$
  - we expect the propensity to participate in the survey for purely statistical purposes to be different between those employed and those who have not found a job yet

$$\Longrightarrow \textit{MNAR mechanism}!$$

Formally speaking...

In other words: we conduct a survey and observe x_1 still unemployed and x_2 already employed individuals. We assume

$$X_1 \sim \text{Binom}(M_1, p_1) \quad X_2 \sim \text{Binom}(M_2, p_2)$$

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = w \neq 1$$

How to estimate M_1 and M_2

- We need to estimate the odds ratio w, i.e., the unemployed's exposure in the AlmaLaurea survey
- To this aim we
  - **leverage the Fisher's noncentral hypergeometric model**, and
  - **exploit the information from another data source**: ISTAT (random) sample.

Fisher's NCH for population size estimation

NCH describe *biased urn* experiments

⇒ each color has its own weight

Fisher's NCH for population size estimation

NCH describe *biased urn* experiments

⇒ each color has its own weight

NCH are underused in the literature

- confusion around the existence of two different NCH distributions Fog (2008)
  - Wallenius' (see Wallenius (1963))
  - Fisher's (see Fisher (1935))
- computational complexity given by their probability mass functions

Fisher's NCH for population size estimation

NCH describe *biased urn* experiments

$\Rightarrow$ each color has its own weight

NCH are underused in the literature

- confusion around the existence of two different NCH distributions Fog (2008)
  - Wallenius' (see Wallenius (1963))
  - Fisher's (see Fisher (1935))
- computational complexity given by their probability mass functions

BUT

- Fisher's NCH has great potential in the official statistics field
- Wallenius NCH is potentially a flexible model when dealing with preferences

Fisher's NCH for population size estimation

Assume

$$X_1 \sim \text{Binom}(M_1, p_1) \quad X_2 \sim \text{Binom}(M_2, p_2)$$

Then,

$$X_1 | X_1 + X_2 = n \sim \text{FNCH}(M_1, M_2, n, w)$$

where w is the odds ratio

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Estimating the number of unemployed graduates in Italy

Three steps:

1. estimation of the number of unemployed graduates (M_1) among the 2011 cohort exploiting
   - the ISTAT sample (assume $w = 1$) and
   - the *Anagrafe Nazionale degli Studenti* registered values ($N = M_1 + M_2$ known);
2. estimation of the response bias in the Graduates' Employment Status Survey 2012 exploiting the results at step 1;
3. prediction of the size of unemployed graduates from the 2012 to the 2020 cohort, assuming the response bias to remain constant over years.

Results

Promptly employed people are generally more inclined to answer the questionnaire

- differences among disciplines: e.g., employed economists are about 10 times more exposed than unemployed ones, while *unemployed* jurists tend to respond more (about 2.5 times): practicum?
- differences between men and women

Results

Promptly employed people are generally more inclined to answer the questionnaire

- differences among disciplines: e.g., employed economists are about 10 times more exposed than unemployed ones, while *unemployed* jurists tend to respond more (about 2.5 times): practicum?
- differences between men and women

…what if we have more categories?

The multivariate case

In the multivariate case, things become more challenging: it is possible to evaluate the likelihood in principle, but as N and C increase, it becomes computationally costly.

$$\text{FNCH}(\boldsymbol{x}|\boldsymbol{M}, n, \boldsymbol{w}) = \frac{\prod_{c=1}^{C} \binom{M_c}{x_c} w_c^{x_c}}{\sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}} \prod_{c=1}^{C} \binom{M_c}{z_c} w_c^{z_c}}$$

where $\mathcal{Z} = \{\boldsymbol{x} \in \mathbb{N}_0^C : \sum_{c=1}^{C} x_c = n, 0 \leq x_c \leq M_c, c = 1, \ldots, C\}$

Two proposals:

- exploiting the conditional structure of the FNCH to implement still a MCMC
- Approximate Bayesian Computation

References

R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935.

A. Fog. Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Communications in Statistics—Simulation and Computation*, 37(2):241–257, 2008.

K. T. Wallenius. Biased sampling; the noncentral hypergeometric probability distribution. Technical report, Stanford University CA Applied Mathematics and Statistics Labs, 1963.