Poststratification and complex survey analyses

Workshop: ‘Explaining and Promoting Health Equity’, Trondheim.

Martinez-Beneito, MA

This talk is part of the project MCIN/AEI/10.13039/501100011033/FEDER, UE, funded by Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

UNIÓN EUROPEA

FONDO EUROPEO DE DESARROLLO REGIONAL

"Una manera de hacer Europa"

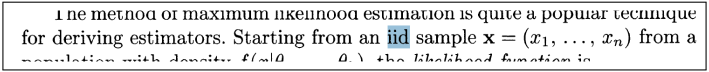
AGENCIA ESTATAL DE INVESTIGACIÓN

Statistics

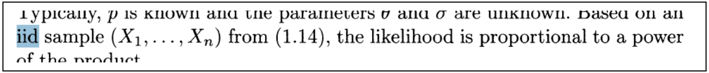
!! (the basic textbook version of) **Statistics is easy!!**

Statistics

!! (the basic textbook version of) **Statistics is easy!!**



The method of maximum likelihood estimation is quite a popular technique for deriving estimators. Starting from an iid sample $\mathbf{x} = (x_1, \ldots, x_n)$ from a population with density $f(x|\theta_1, \ldots, \theta_k)$, the *likelihood function* is



Typically, p is known and the parameters θ and σ are unknown. Based on an iid sample $(X_1, \ldots, X_n)$ from (1.14), the likelihood is proportional to a power of the product

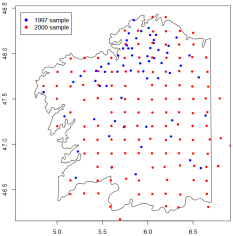
- Just draw an **iid sample and do** something.

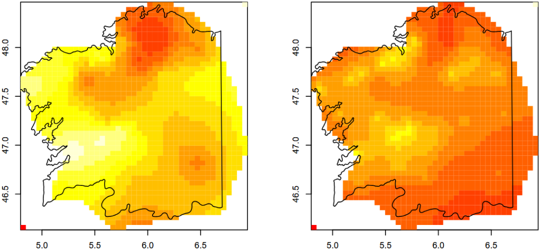
However . . .

- In practice, sampling is often **not iid**.

Preferential sampling:

- Study of **lead polution** in Galicia (Spain).
- Individuals are sampled with **different probabilities**: In 1997, sampling was focused to those regions with higher lead concentrations.





Analyses assuming **preferential and random** sampling.

- Taking the **sampling design** into account really **matters**.

Survey studies:

- **Survey** data **rarely** follow simple **random sampling** (SRS).
- Individuals usually have **different probabilities** of being selected, for either logistical, economic or efficiency reasons.
- In that case, we **must account** for the sampling **design** in the analysis, otherwise, estimates can be **biased**.
- Sampling designs consider **different tools** for data collection: stratification, clustering, PPS sampling, . . .
- From now on, we will **focus** on **stratification** and its treatment.

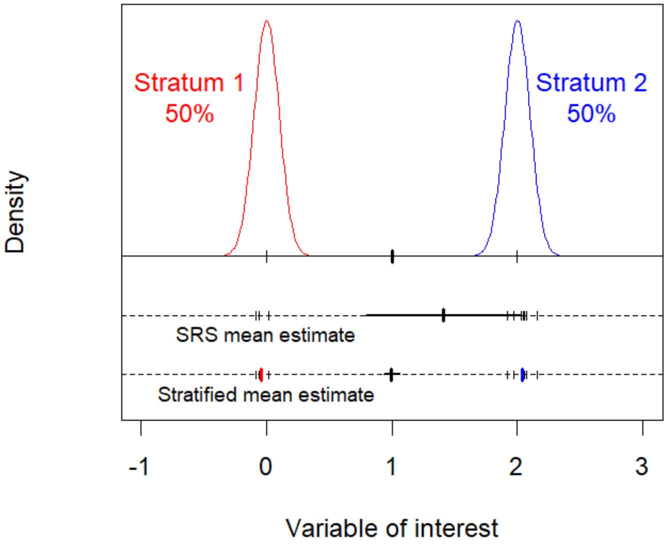
Stratified sampling

- Its goal is **increasing sampling efficiency** as compared to SRS.
- The **idea** is simple:
  - **Split the population** into strata and **estimate** each stratum **separately**.
  - **Combine** those estimates using some **known** population **proportions**.

$$\bar{Y}_{str} = \sum_{i=1}^{\#strata} P_i \cdot \bar{Y}_i$$

where P_i is the **(known) proportion** of people belonging to stratum i.

Two-stratum population



When is stratification a good option?

- It works well when the population can be divided into groups with **clear differences for the outcome**, as for example, age groups in health surveys.
- If groups are **not** very **different**, stratification does **not** return hardly any **improvement**.

- Stratified sampling takes advantage of **additional information** for inference: the **proportion of people** belonging to each stratum.

- It also allows us to **oversample** the most convenient groups.

However,

- It **requires** an **appropriate** (stratified) data analysis, **otherwise** estimates may be **biased**.

Heart problems in The Netherlands (ESS11).

- The 11th edition of the European Social Survey (**ESS11**) of 2023 collected data from **24 countries**.
- We focus on the study of "**Heart or circulation problems**" in The **Netherlands**.
- **Sampling design** in The Netherlands: **48 strata**, the combination of 4 regions, 2 genders and 6 age groups.
- **SRS** proportion estimate: 10.27% [8.82, 11.71] (sd=0.74)
- **Stratified** proportion estimate: 9.30% [8.04, 10.55] (sd=0.64)
- In addition to the **increased efficiency**, the stratified estimate is known to be **unbiased**.

Poststratification

- Although **stratification** may be used for sampling design, it may have **more uses**.
- Poststratification uses the **same logic**, but **after** the data have been **collected** (just for data analysis).

- Sometimes the **sample does not** perfectly **match** the population.
- Poststratification **adjusts the estimates** so that they reflect the true **population composition**.
    - **Draw** your sample.
    - **Split** it according some groups (that were not consider in the design).
    - **Calculate** the mean for each group.
    - **Combine** them according to known population proportions.

Poststratifying heart problems in The Netherlands.

- The **response rate** for The Netherlands in ESS11 is **34%**, which could make the **sample to deviate** from the population.
- **ESS11 poststratifies** estimates, for every country, according to **gender**, **age** (3 groups), **education** (3 groups) and **region**, so their `estimates` are `balanced` to the population according to **these 4 factors**.
- For The Netherlands, the effect of **gender, region and age** is **already stratified** by taking the sample design into account; however, **education should be poststratified** for guaranteeing that the estimates meet the population also in that sense.
- **Poststratified** proportion estimate: 9.40% [8.13, 10.67] (sd=0.65)
- This estimate **does not differ** from the stratified one, at the end heart problems **do not depend** so much on **education**.

Drawbacks of (pre and post)stratification

- It **requires population proportions** for each stratum, which are particularly hard to know as the number of **stratification variables grows**.
- They are **direct estimates**, based on just the **information of each** particular unit **(domain)**.

- When domains are **small, information becomes scarce**, so direct estimates are problematic (empty strata), and therefore stratification.
- For small areas, **sharing information** between domains is a **good idea** for increasing the amount of information per unit.

- Multilevel regression with Poststratification **(MRP) does** exactly **that**.

How does MRP work?

- MRP works in **two steps**.
- **Regression:**
  - Fit a **regression model** with the **variables** used in the **sampling design**.
  - Controlling the **stratification variables** makes the **sampling design ignorable** for inference.
- **Poststratification:**
  - **Poststratify** the fitted values using the known **population proportions**.
  - Poststratification **restores representativeness** of the sample.
- **MRP**, under Normality and with as many covariates as strata (saturated model), is **equivalent** to direct **poststratified estimation**.

And what is the real advantage of MRP?

- When working with **small areas**, the **number of strata** is usually quite **large** → some **strata combinations** will have **no data** → direct estimates **cannot** be even **calculated**.
- Due to the equivalence of **poststratification** and **MRP** with a saturated model, **MRP** poses **similar problems** for small areas.
- However, **in MRP !!we do have a model!!**, not just simple estimates, and its assumptions could be **modified/relaxed**.
  - If some strata combinations have **no data**, some **interactions** can be **excluded** from the model.
  - If some stratification covariate has **lots of levels** use **random effects**.
  - If those levels correspond to **temporal or geographical** data, include **that dependence** in the model.
  - . . .

Geographical distribution of self-rated health in the ESS11

- We are interested in the distribution of **self-rated health** ($1 = $ Poor; $0 =$ Good) across **subnational regions** (mostly NUTS2) of ESS11, a total of **193 regions**.
- 4 **poststratification factors**: gender, age (3 groups), education (3 groups) and region $\approx$ **3000 strata** → **11.5 individuals** per stratum.
- The **regression model**:

  $$\texttt{Self.rated} \sim \texttt{gender * age * education * region}$$

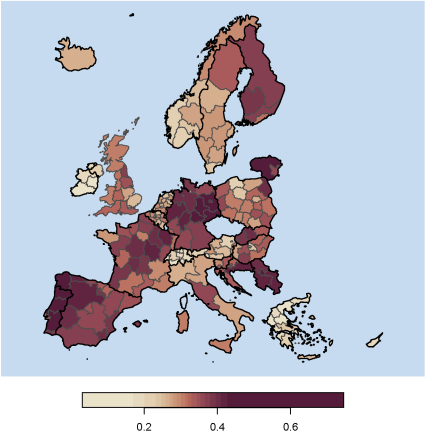
  with a subsequent **poststratification** of gender, age and education produces (direct) poststratified **region estimates**.
- However, these **estimates** are **limited/unfeasible** by:
  - the small **sample size** of some strata.
  - using **direct** estimates.

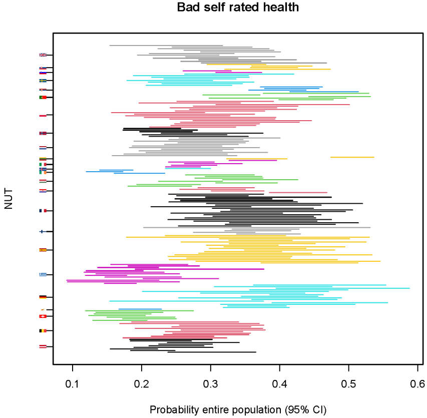
- By using MRP, **we could modify** the previous regression model as follows:

  ```
  Self.rated ~ gender * age * education + region + region:gender
                   + region:age + region:education
  ```

- Now, **information is shared across** regions and strata (fewer parameters than strata).
- If we modelled `region` as **spatial random effects, information** would be **shared** across **neighbouring** units.
- This reduces the number of **stratum** estimates **from 3474 to 1176** (spatially dependent) estimates.

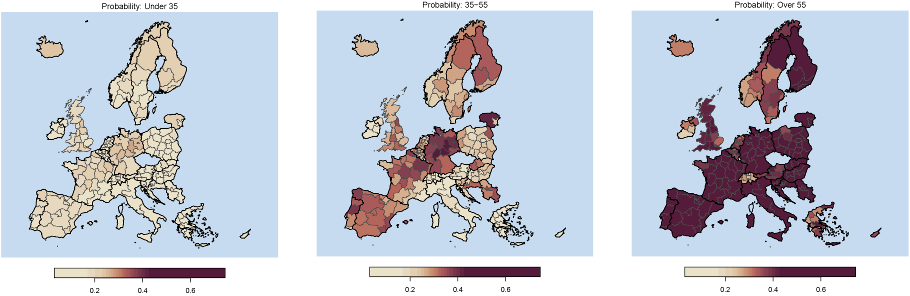
- **Poor health** proportion estimates:


Probability: Entire population


Bad self rated health

- Spatial random effects account for **spatial correlation**.
- Important **variability both between and within** countries.

- If some covariate is **not poststratified** we obtain **level-specific patterns**, such as for age:



- **Italy** goes from **very low** probabilities in the two youngest groups **to very high** probabilities **in the oldest** group.
- **Belgium and the Netherlands** show **much milder changes** as a function of age.

In summary

- **Stratification** is a valuable sampling tool, leading to more **efficient designs**.
- **Poststratification** can adjust for **factors** that were **ignored**, or arised, during the sampling design.
- **Model-based estimates**, particularly MRP, make it possible to **borrow strength** across strata and the incorporation of **spatial and temporal dependence** between units.

- Using **MRP** with ESS data enables analyses at a **higher** level of **detail than current official** estimates.

In summary

- **Stratification** is a valuable sampling tool, leading to more **efficient designs**.
- **Poststratification** can adjust for **factors** that were **ignored**, or arised, during the sampling design.
- **Model-based estimates**, particularly MRP, make it possible to **borrow strength** across strata and the incorporation of **spatial and temporal dependence** between units.

- Using **MRP** with ESS data enables analyses at a **higher** level of **detail than current official** estimates.

!!Thanks for your attention!!