Topics in Social Statistics
University of Helsinki
Part I C

Seppo Laaksonen

Sampling Principles
And
Statistical Editing
2013

Content

What is survey?

Key concepts in surveys

From Survey Data Collection to Cleaned Survey Data

Through

- Designing the survey
- Designing the questionnaire
- Designing the sample(s)
- Data collection with alternative single and mixed modes
- Data entry
- Editing the raw data
- Imputing the data
- Weighting the data
- Adding other features into the data file

Part 3 extends some to Part 1, especially sampling techniques and analysing cleaned data. Part 2 is focused on measurement questions in data collection and analysis.

Sampling design

You saw in the scheme of Part 1 A that a sample survey data have some missingness, just due to sampling. If the population is big, sampling is a natural tool to work on. I present here a compact framework for sampling. This is called sampling design. Often a narrower framework has been given.

My framework is for probability sampling, not for quota or other non-probability sampling. Voluntary sampling is nowadays becoming more (too) common especially using web arsenals. These are also non-probability methods for sampling.

Moreover, so-called access panels are created. These are attempted to make as probability based as possible but voluntariness means that they are completely such ones. Their idea is to construct a ‘sample’ of target population people who are voluntary to continue toward other surveys. E.g. market research institutes are using this approach much.

Sampling design 2

First some basic concepts:

Cluster = e.g.

- small area where residents, birds, students (enumeration areas)
- school where students
- household where its members
- address where residents or employees
- enterprise where employees

Inclusion probability: probability that a frame (target population) unit will be included in the (gross) sample. In probability sampling this probability must be >0 (maximum=1 is accepted naturally). Otherwise some units cannot be drawn in the sample. This is the product of the selection probability and the desired sample size.

Primary sampling unit = *psu*: the unit that has been included in the sample in the first step (stage) in sampling. This is a cluster in two stage sampling.

Stratum: group or sub-population that will be included definitely in the sample, thus its inclusion probability = 100%. The strata are independent of each other.

Sampling design 3

A. Frame	Study units are explicitly in the frame or they are not there.
B. Sampling unit	The sampling unit is the study unit as well, or not.
C. Stage	Hierarchy to approach to the study units by using probability sampling. First going to the first-stage units (=*psu*’s), and then to the second stage units, …Terms: one-stage sampling, two-stage sampling, three-stage sampling
D. Phase	First a probability sampling applied for drawing a first-phase sample, and afterwards a new sample has been drawn at the second phase from the first sample.
E. Stratification	The population divided into several independent sub-populations.
F. Allocation of the sample	How a desired gross sample has been shared into each stratum.

Sampling design 4

G. Panel vs. cross-sectional study	If a panel is desired, it is needed to design also how to follow up the first sample units, and how to maintain the sample. Whereas a cross-sectional study is desired, it is good to design it so that a possible repeated survey can be conducted (thus getting a correct time series).
H. Selection method	How to select the study units - probability equal to all (srs, equi-distance, Bernoulli) or - probability varies unequally typically by size (pps =probability proportional to size)
I. Missingness anticipation	Trying to anticipate response rates and allocate a gross sample so that the net sample is as optimal as possible in order to get as accurate results as possible.

Thus: choose an optimal alternative from each A to I alternatives, and you will have a gross sample.

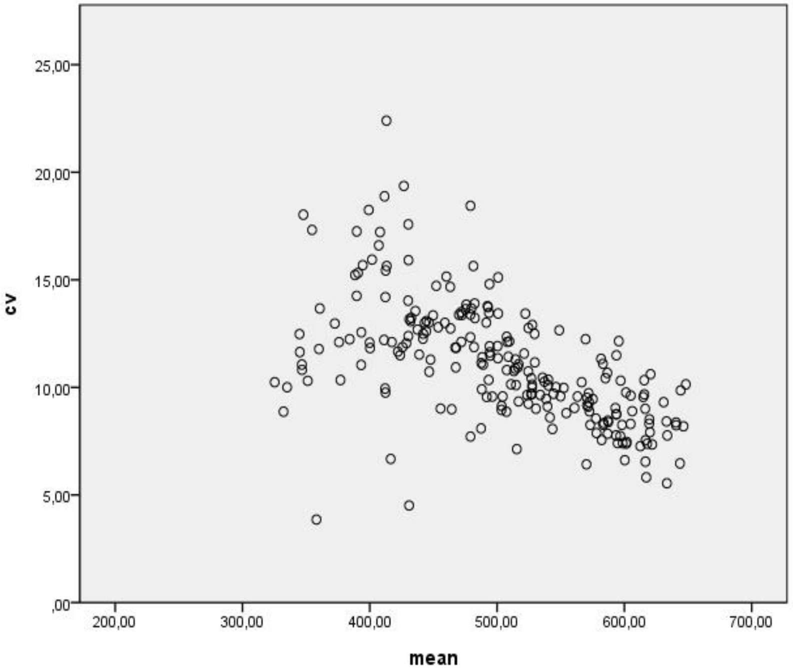
Of course this task is not easy, since you have to anticipate many things. One such thing is intra-class correlation *rho* given that your *psu* is a cluster.

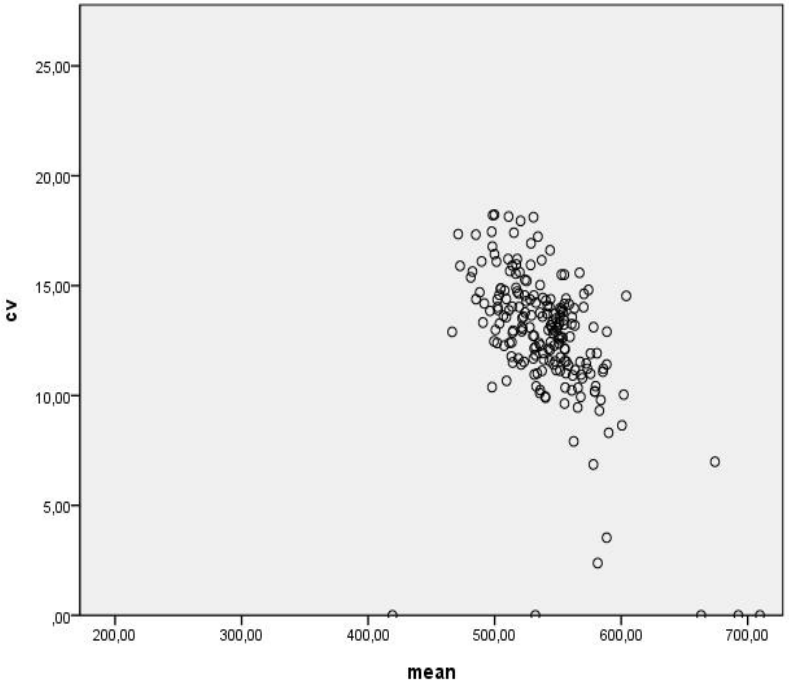
Sampling design 5

The intra-class correlation $rho = \frac{Between\_variation}{Total\_variation}$ is an indicator for (in)homogeneity of clusters. It varies a lot from a survey to the next. For example in PISA 2009, where the clusters are school classes and the variables are scores of literacy, it is for Finland around 0.1 but for Germany around 0.6. What this means?
The ESS *rho*'s are much smaller, since *psu*'s are small areas that are not as homogenous. Typically *rho* is around 0.02-0.04, in some countries higher (depending on the estimate). In Finland it is = 0 since clusters are not used.

So, when designing ESS samples we have to anticipate many things, also response rates. Unequal probabilities mean that the accuracy will worsen. Hence we also increase the gross sample size (analogously to cluster effect in which case a higher *rho* requires a larger gross sample). This is due to our target that all participating countries achieve an about same accuracy level. This has been measured at sampling design with *effective sample size* that should be 1500 at minimum.

Here are two scatter plots from the 2009 PISA so that the x-axis is the mean of success rate for mathematical-statistical literacy, and the y-axis the coefficient of the variation, respectively. The plots are the PISA sample schools. May you guess, which graph is for Germany, and which for Finland? Other observations from the graphs?





Sampling design 6

We use the concept *DEFF (design effect)* when planning the sample of the ESS and its gross sample size, in particular. This indicator is the ratio between the anticipated accuracy of this particular design and the corresponding design based on *srs* (although not used in most cases).

We have two *DEFF*'s:

- Due to unequal inclusion probabilities *DEFFp*
- Due to clustering *DEFFc = 1 + (b-1)rho (b*=average *net* cluster size*)*

The whole *DEFF* is the product of both *DEFF*'s

I give a theoretical example based on this whole strategy.

Sampling design 7

Operation	Example calculation (average-based, the figures may vary by stratum, cluster and another domain)
1. Target for the effective sample size (*neff*)	2000
2. Anticipated missingness due to unit nonresponse	30% i.e. 2000/.7 = 2857
3. Anticipated missingness due to in-eligibility	5% eli 2857/.95 = 3008
4. Anticipated Design Effect (DEFF) due to clustering including anticipated intra-class correlation (=0.025), average net cluster size (=5.3) and missingness (average gross cluster size = 8)	$DEFF_c$ = 1+(5.3-1)*.025 = 1.11 3008*1.11= 3338
5. Anticipated DEFF due to varying inclusion probabilities (calculated for anticipated respondents if possible)	$DEFF_p$ = 1.25 3338*1.25 = 4173
6. Risk factor, leading to increase the above Gross Sample Size Anticipated Net Sample Size	 4250 2826

Look at the sampling guidelines and other methodological documents of the European Social Survey (ESS)

http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=80&Itemid=365

Data editing or statistical editing

This is often made together with imputation since editing can lead to imputation e.g.

to replace missing or strange values with imputed ones.

Editing is basically a phase that can require a lot of time if entry data are dirty. This is the case more likely if

- Postal survey has been used i.e. a respondent has filled in his/her questionnaire.
- Pre-editing has not been tried in data entry (this can be the case in all modes but usually it is easy to give the limits for acceptable values in web, face-to-face and telephone surveys). Textual answers still need further editing usually.

However, a good survey provider always include a simple editing at minimum in data entry. Hence the initial 'raw' data file is slightly edited and it will easier to continue toward next editing steps.

Data editing or statistical editing 2

Nevertheless, editing is not only for correcting typing or other errors. It is also a development process:

- To learn about the whole survey process and

in particular

- Errors and not best practices of this current survey.

These learnings have been used in survey documentation since users/clients desire

to know the survey quality from its all aspects. Secondly, the experience from each

survey is beneficial to use in future surveys even though you have not in mind to

conduct new surveys. It is important to save all editing (as software program).

Data editing or statistical editing 3

Some concepts from the UNECE data editing glossary (search more as yourself):

CHECKING RULE

A logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct. In various connections other terms are used, e.g. edit rule.

CONSISTENCY CHECK

Detecting whether the value of two or more data items are not in contradiction.

ERROR LOCALIZATION

The (automatic) identification of the fields to impute in an edit-failing record. In most cases, an optimization algorithm is used to determine the minimal set of fields to impute so that the final (corrected) record will not fail edits.

GRAPHICAL EDITING

Using graphs to identify anomalies in data. While such graphical methods can employ paper, the more sophisticated use powerful interactive methods that interconnect groups of graphs automatically and retrieve detailed records for manual review and editing.

Data editing or statistical editing 4

I do not go to details in editing since it is specific for each survey, that is, it is required

to check all values both

- One-dimensionally (like checking whether all values are acceptable)
- Two-dimensionally as well as possible (such as whether the values is acceptable also by background variables such as gender, age, region, education. E.g. A male cannot be a mother, a very young human cannot have a child or cannot be married).
- Distributionally
- Multi-dimensionally that can be made also using a multivariate model.
- In graphs that often facilitates to see the previous options more concretely.

All errors or inconsistencies cannot be correctly definitely but it is not always needed, fortunately, unless their effect is fatal in estimates.

Just to do your best, and improve the data even during data analysis.