Sampling and estimation
Urvalsmetoder och estimation

Master course

Daniel Thorburn

Winter semester 2011

Stockholm University

Sharon Lohr, *Sampling design and analysis*

- This course will cover chapters 2-8 and parts of 9,12 and 15, but also much material not in the book
- Chapter 1 is included but should be known from a course on survey methods
- The rest of chapters will be covered in the course Analysis of Survey Data
- Reference literature: *Urval - från teori till praktik*, Handbok 2008:1, SCB. (Can be down-loaded from the home page of Stat Sweden http://www.scb.se/Pages/PublishingCalendarViewInfo___259923.aspx?PublObjId=8482&lang=SV

Preliminary contents

- 17/1 Introduction, Paradigms, SRS Model-based, SRS Design-based, Auxiliary variables, Some simple sampling designs
- 24/1 Regression-type estimators, Difference, ratio, regression, GREG- and prediction estimators
- 31/1 Stratified sampling, Optimal allocation, number of strata, post stratification; Quota sampling
- 7/2 Sampling with varying inclusion probabilities, HT and SYG estimators, Combination of design and estimation techniques, Cluster sampling, Indirect sampling, sampling from an unknown frame
- 14/2 Estimation with non-response, Missing data, MAR, MCAR, post stratification, some reweighting and imputation techniques

Preliminary contents

- 21/2 Large sample properties, Central limit theorem for independent r.v. and sampling from finite populations, Berry Esseen type central limit theorems
- 28/2 Coordinated samples, Longitudinal studies, Negative-Positive coordination, Permanent random numbers, Rotating panels, Estimates of level and/or change.
- 7/3 Miscellaneous topics, Gauss-approximation, resampling, quantiles, other functions, Quality and Evaluation studies, embedded experiments
- 14/3 Bayesian probability and sampling, Paper presentations

- Some time: Guest lecturerer Frida Videll on the Swedish Labour Force Survey

Forms of examination and grading criteria

The course is examined through both individual assignments, and a written take home test.

The test will be handed out at the last lecture and it must be handed back within one week. No help from any person in any form is allowed (except from the examiner) but computers, computer programs, textbooks and literature are allowed. It can give 100 points

One of the individual assignments is reported in the form of an oral presentation where attendance is mandatory. All assignments are graded and can together provide a maximum of 100 credit points.. In addition to points they are also graded as unapproved or approved.

To pass the moment (grades A – E) it is required that all assignments are approved, the written test is given at least 40 points and the total score exceeds 100 points.

The final grade on the moment is based on a total count of the points which thus can be a maximum of 200 points

Written test is a home exam for one week. Handed out March 14. Next opportunity March 18.

- Grades are given on a seven-point rating scale
  A. Excellent - 180-200 points
  B. Very good - 160-179 points
  C. Good - 140-159 points
  D. Satisfactory - 120-139 points
  E. Enough - 100-119 points
  Fx. Insufficient - to 60-99 credits or more than 100 points but all assignments have not been approved or the exam gave less than 40 points
  F. Completely insufficient - equivalent to less than 59 points

Next semester the students who did not pass the course must make all assignments of that semester and rewrite the written test.

Urvalsmetoder och Estimation 1

Sampling and Estimation 1

2011-01-17

1.Introduction

1.1 Notations and background

What is sampling?

- Given a finite population
  - the ”frame” often contains auxiliary information
- How to draw a sample
  - where something will be observed
- Goal: This sample will be used to describe the full population (e.g to estimate the mean, the median or the variation).
  - E.g. a representative sample

Notations

- Population U of size N
  $U = \{e_1, \ldots, e_N\} = \{1, \ldots, N\}$
  (for simplicity often represented by an index number)
- Known auxiliary variables (may be a vectors)
  - $\{X_i ; i \in U\} = \{X_1, \ldots, X_N\}$
- Sample S of size n
  $S = \{e_{s1}, \ldots, e_{sn}\} = \{s_1, \ldots, s_n\}$
  (for simplicity sometimes numbered $(1, \ldots, n\}$)
- Unknown values
  (which we are interested in, may be vectors,)
  - $\{Y_i ; i \in U\} = \{Y_1, \ldots, Y_N\}$
- Sample values
  - $\{Y_i ; i \in S\} = \{Y_1, \ldots, Y_n\}$
    (to be measured in the sample)

What is estimation?

- Let T_Y be a quantity defined by all Y_i ; i € U
  - usually the total of Y-values in U; $T_Y = \Sigma_U Y_i$.
- Find a value, (interval or density) based on Y_i; i € S and X_i; i € U, which describes U or estimates T_Y.
  - ususally N times the mean; $t_Y = N \Sigma_S Y_i/n$.
- This value should be good in some sense
  - eg small mean square error

- The goal is to describe the finite population. If the full finite population had been known, no uncertainty would remain.
- In most statistical problems outside sampling the goal is to estimate a parameter, which will be valid in the future outside the frame.

Sampling and estimation

- The sampling scheme and the estimation procedure should be determined simultaneously. It is the combination of them that decides the (design-based) precision/efficiency
- For some sampling schemes the choice of estimator is not so important.
  These schemes are usually better if the data will be used for secondary analyses, by people not familiar with statistical methods.
- (In practice the estimation formulas are often decided after the sample is taken. It is philosophically OK, if you follow the Bayesian paradigm, but not otherwise.)

Some large sample surveys

- LFS, Labour Force survey, ILO, AKU, Arbetskraftsundersökningen SCB
- EU-SILC Survey of Income and Living Conditions, Eurostat, SCB
- ULF Undersökningen av levnadsförhållanden, SCB
- LNU Levnadsnivåundersökningen, SOFI, Stockholms Universitet
- HBS Household Budget Survey EU HEK Hushållets ekonomi SCB
- Riksskogstaxeringen, Swedish forest survey, Lantbruksuniversitetet, Umeå
- PSU, Party preferences, Partisympatiundersökningen, SCB
- NCVS, National Crime Victimisation Survey, EU, BRÅ
- EHIS European health interview survey, EU socialstyrelsen
- TUS Time use survey, EU, tidsanvändningen, SIKA, SCB
- Företagens investeringsplaner, Investment plans, SCB
- World value survey, Statsvetenskap, Umeå
- Consumer Price Index, OECD, EU, SCB
- Purchasing managers´index, Inköpschefsindex
- Consumer satisfaction surveys, Nöjd Kund Index
- Traffic flow measurements, Trafikflödesmätningar, Vägverket
- Omnibus surveys, SIFO and some other opinion/market research institutes
- PISA, Programme for International Student Assessments, OECD
- …

1.2 Finite or infinite populations

- Compare!
  - Draw a sample of individuals. Ask about their income according to their self assessments.

    Goal: To estimate the total amount of tax income the municipality will get; i.e. $T_Y = \Sigma_U\, g(Y_i)$
  - Draw a sample of individuals. Give them a vaccine and look at the number getting sick.

    Goal: To predict the side effects of the vaccination, P(side effect | vaccinated).

    (For some reason usually not treated in most sampling literature)

Remember!

- A finite population does not necessarily imply that we are interested in the finite population
  - and thus not that ”a finite population” correction should be used.
- Example: The number of killed in traffic accidents was 480 in 2004 and 440 in 2005. Is the decrease statistically significant?
  (Here we have a 100% sample)

Answers

- Viewed as a finite population quantity.

  It is a clear decrease 440 < 480.

- Viewed as observations from a random variable $Po(\lambda_{2004})$ and $Po(\lambda_{2005})$, resp.

  A test of the null-hypothesis that $\lambda_{2004} = \lambda_{2005}$ on the 5% level gives the test quantity $z = (480 - 440)/(480 + 440)^{1/2} = 1.32$, which is not significant. You are not allowed to draw the conclusion that the traffic was less risky 2005 compared to 2004. The decrease may have occurred by chance

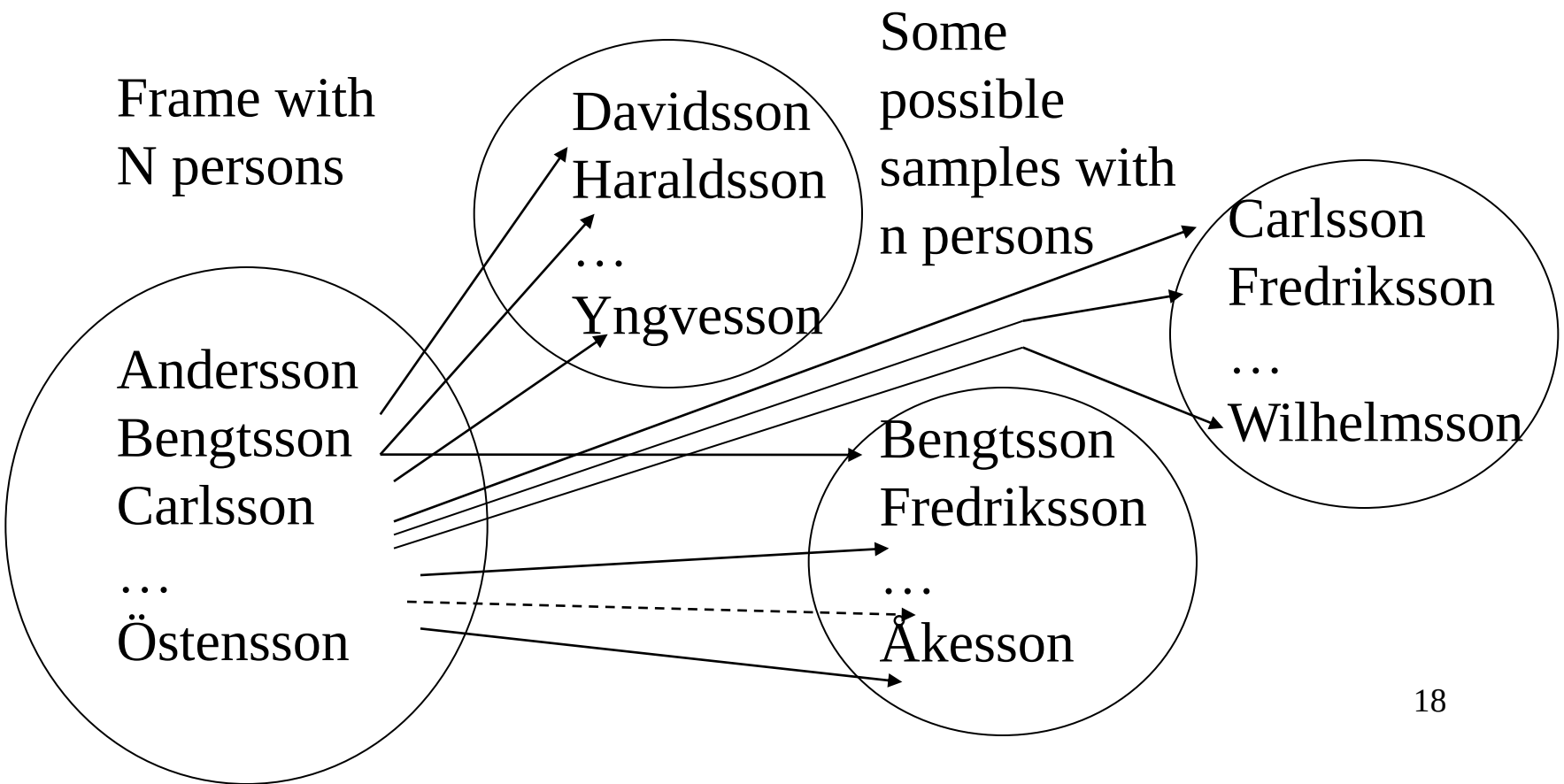
1.2 Model or Design-based

- Design-based
  The important randomness lies in the selection of the sample. Standard approach to sampling. (The main view in this course)
- Model-based
  The important randomness lies in a statistical model of nature. Standard approach to statistics
  - (We shall discuss them shortly in a very simple case i.e. under SRS and no auxiliary informnation)

- Bayesian approach
  The randomness describes the uncertainty (will not be discussed here). Correct approach.

Simple Random Sampling (SRSwor)
(Enkelt slumpmässigt urval OSUuå)

Sample S of size n, drawn so that all subsets with this size has the same chance to be drawn. Which?



SRSwor (SI), P(S = s) =

$$\binom{N}{n}^{-1}$$

First order inclusion probabilities: P(i € S) = π_i = n/N,

Second order inclusion probabilities:

P(i,j € S) = π_{ij} = n(n-1)/(N(N-1))

1.2.1 Model-based

- Goal: to estimate population total $T_y = \Sigma_U Y_i$
- Model: Let $\{Y_i; i = 1, 2, \ldots, n, \ldots, N\}$ be iid random variables with distribution F and mean θ.
- Estimate the parameter by the sample mean $\theta^* = \Sigma_S Y_i /n = \overline{Y_S}$
- Predict the remaining (N-n) units by the mean $Y_i^* = \theta^*; i \in U\text{-}S$
- Predict the total $T_y = \Sigma_U Y_i$ with the sum $T_y^* = t_y = \Sigma_S Y_i + \Sigma_{U-S} Y_i^* = (N/n) \Sigma_S Y_i$

Properties of this model-based estimator

- Unbiased $E(T_y - T_y^*) = 0$,
  - since
    $E(\Sigma_U Y_i - \Sigma_S Y_i - \Sigma_{U-S} Y_i^*) = \Sigma_{U-S} E(Y_i - Y_i^*) =$
    $\Sigma_{U-S} (E(Y_i) - E(Y_i^*)) = \Sigma_{U-S} (\theta - \theta) = 0$

- Variance: $Var(T_y - T_y^*) = N(N-n)/n\ \sigma^2$,
  - since
    $Var\ \Sigma_{U-S} (Y_i - Y_i^*) = Var\ (\Sigma_{U-S} Y_i - ((N-n)/n)\Sigma_S Y_i) =$
    $(N-n)\sigma^2 + ((N-n)/n)^2 n\ \sigma^2 = N(N-n)/n\ \sigma^2$

- Variance estimator = ?

$$\frac{N^2}{n}\,\frac{N-n}{N}\,\frac{1}{(n-1)}\sum\nolimits_S (Y_i-\bar{Y})^2$$

The second factor is called correction for finite population

Why? What are the others factors?

- Note that the model-based estimator does not depend on the sampling mechanism (the inclusion probabilities).
- Which individuals that are included in the sample does not contain information in itself, and is an ”ancillary statistic”.
- The model we used here assumed that all variables were iid.

A general way to obtain the variance of an estimator:

$$"Reverse Martingale Property"$$

$$E(\bar{Y}_S \mid \bar{Y}_U, \theta) = E(\bar{Y}_S \mid \bar{Y}_U) = \bar{Y}_U$$

$$E(\bar{Y}_S \mid \theta) = \theta,$$

$$where \overline{Y_U} = T_y / N \ and \ \overline{Y_S} = t_y / n$$

”Independent increments”

θ and Y_S are independent given Y_U .

$$E(\bar{Y}_n - \theta)^2 =$$
$$E(\bar{Y}_n - \bar{Y}_N + \bar{Y}_N - \theta)^2 =$$
$$E(\bar{Y}_n - \bar{Y}_N)^2 + E(\bar{Y}_N - \theta)^2$$

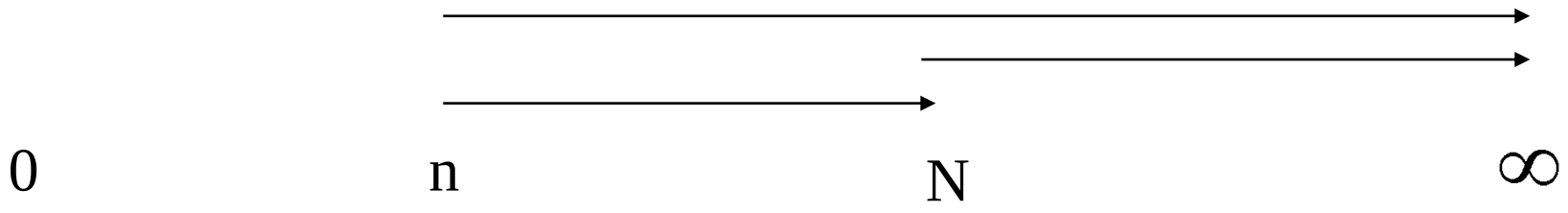
From this follows

$$E(\bar{Y}_n - \bar{Y}_N)^2 = Var(\bar{Y}_n) - Var(\bar{Y}_N)$$

This is a fairly general formula.
It holds for most sampling schemes,
not only nodelbased and SRS

Interpretation

- Suppose that you know how to estimate the parameter infinite population best (UMVU) from the total and similarly from the sample and what the variance will be then
  - (Most standard statistical textbooks are full of this).
- Then the prediction variance of the finite population parameter is just the difference between these variances.



The distance (variance) from n to N is the difference between the distances (variances) from n to infinity and from N to infinity

1.2.2 Design-based

- No model – no θ – the only probability is home-made when selecting the sample

- Goal: to estimate population total $T_y = \Sigma_U Y_i$
- Estimate the population mean by the sample mean $\theta^* = \Sigma_S Y_i /n = \overline{Y_S}$
- The estimator of the population total will thus be $T_y^* = (N/n) \Sigma_S Y_i$

(The same expression as for model-based)

Properties

- Unbiased $E(T_y^*) = \Sigma_U Y_i$
- The distribution here is over the sampling:
- Let I_i be an indicator function = 1 if i € S and = 0 otherwise, then
  $T_y^* = (t_y) = (N/n) \Sigma_S Y_i = (N/n) \Sigma_U I_i Y_i$
- Proof of unbiasedness:
  $E(T_y^*) = E(\Sigma_U (N/n) I_i Y_i) = \Sigma_U (N/n) E(I_i) Y_i =$
  $\Sigma_U (N/n)(n/N) Y_i = \Sigma_U Y_i$
- Note that Y_i are considered fixed but that all randomness lies in the ”home-made” sampling I_i.

Properties

Variance. Assume for simplicity mean 0.

$$E(T_y^{*2}) =$$
$$E((\Sigma_U (N/n)\, I_i Y_i)^2) =$$
$$E(\Sigma\Sigma_{UxU;i\#j} (N/n)^2\, I_i I_j Y_i Y_j) + E(\Sigma_U (N/n)^2\, I_i Y^2_i) =$$
$$\Sigma_U \Sigma_{U;i\#j} (N/n)(n-1)/(N-1) Y_i Y_j + \Sigma_U (N/n)\, Y^2_i =$$
$$\Sigma_U \Sigma_U (N/n)(n-1)/(N-1) Y_i Y_j + N(N-n)/(n(N-1))\, \Sigma_U\, Y^2_i$$
$$= N(N-n)/(n(N-1))\, \Sigma_U\, Y^2_i$$

$= N(N-n)/n\, \sigma^2$, where $\sigma^2 = \Sigma_U Y^2_i - (\Sigma_U Y_i)^2/(N-1)$

i.e. the same expression as above for model-based (but another interpretation of σ^2).

Estimation of population variance

- $E(s^2) = E((\Sigma_S Y^2_i - (\Sigma_S Y_i)^2/n)/(n-1))$
  $= E((\Sigma_U I_i Y^2_i - nT_y{}^{*2}/N^2)/(n-1))$
  $=((n/N)\Sigma_U Y^2_i - (N-n)/(N(N-1))\Sigma_U Y^2_i /(n-1))$
  $= \Sigma_U Y^2_i /(N-1)$

Properties

The variance estimator is the same as for model-based:

$$\frac{N^2}{n}\frac{N-n}{N}\frac{1}{(n-1)}\sum_S (Y_i-\bar{Y})^2$$

1.2.3 Differences model-design

- Not always the same result
- Usually one has $\text{MSE} = \text{Bias}^2 + \text{Variance}$
- With a good model one may decrease the variance but to the price of a bias if the model does not hold.
- Design-based can be compared to nonparametric methods. Works for any distribution but may give higher variances
- In many sampling situations the samples are so large that the variance is small compared to the possible bias.
- Thus the most common practice is to use a design-based methods.
- We will mostly deal with that in this course.

Model-based methods usually give the same or smaller variances –
if the model is true

Design-based methods are more robust – works even if when the model does not hold

Sometimes model-based methods give larger variances. This should be a warning sign!

Example

- Suppose that the population was created by drawing random number from the standard Laplace distribution ($f(y) = \frac{1}{2} \exp(-|y-m|)$) (with variance 2)
- The design-based estimator of the population mean is the sample mean and the estimation variance is the population variance which is around $2(N-n)/(Nn)$
- The MVUE estimator of the parameter is the median. The model-based estimator is a weighted sum of the median and the mean $(n\overline{Y_n} + (N-n)\tilde{Y}_n)/N$
- Its prediction variance is $(1/N^2)(Var(\sum_{U-S} Y_i) + Var((N-n)\tilde{Y}_n)) \approx (1/N^2)((2(N-n))) + (N-n)^2/N) = (1+n/N)(N-n)/(nN)$, which is smaller than the design-based variance

  (The variance of the p:th quantile, $F_n^{-1}(p)$, is, for large n, approximately $(p(1-p))/(nf^2(F^{-1}(p)))$)
- But if the model does not hold there is also a model bias

1.3 Auxiliary information

- If there is no information at all about the units in the population except their identity (which is supposed to carry no information) simple random sampling is the only possibility
- But almost always there is some information.
- And one should use it.
- Take into account what you know. Randomise to safeguard against what you do not know.

  (Old Jungle Proverb)

Auxiliary information

- Data known in advance, which are not of primary interest for the study, but can be used to facilitate the survey or to improve the precision
- Sometimes used for presentation purposes too
- Sometimes also data that are not of primary interest but which is collected in order to improve the precision

- Auxiliary variables are often known for the whole population, but sometimes
  - known only for a part of the population (can be used in the part where it is known) or
  - only the population total is known (can be used only at estimation phase)

Typical auxiliary information

- Number of males and females in the population
- Sex, age, marital status, address, … for all units in population
- Order in register (sometimes informative; may be chronological, geographical or …)
- Values from previous years or register (e.g. turnover and number of employees, self assessed income)
- Value on the invoice audited
- More examples ? ...

How to use auxiliary data?

- In the sampling phase!
- In the estimation phase!
- Often both in combination. Usually best!

How to use auxiliary data in the sampling phase?

- Let the probability of sampling units (full sample) depend on the auxiliary information
- Define inclusion probabilities
  $\pi_i = P(i \in S)$; for all $i \in U$ [$\pi_i(X_i, X_{j \in U})$] using the auxiliary information
- Second order inclusion probabilities π_{ij}, third order π_{ijk}, ...
- Will be discussed later: πps-sampling

2. Some sampling methods

2.1 Definition of a probability sample :

- It should be possible to compute the probability that a unit is included in the sample
  $\pi_i = P(i \in S)$ for all $i \in S$

  (some people say for all $i \in U$, but that rules out many cluster sampling schemes. Some people require that all higher order inclusion probability can be calculated)
- The probability $\pi_i = P(i \in S) > 0$ for all $i \in U$
- Sometimes also $\pi_{ij} = P(i \in S, j \in S) > 0$ for all pairs $i,j \in U$

Why probability samples

?

(There are alternative answers)

Take into account what you know. Randomise to safeguard against what you do not know. (Old Jungle Proverb)

- Avoids convenience sampling. (Taking those that are most easy to get (e.g. persons at home, ”stugsittarurval”, Quota sample).
- Avoids the influence of periods or other systematic sortings in the frame.

In real life one seldom has a true probability sample even if you have tried to get one.

There are almost always factors outside your control

- Formally no sample obtained after non-response is a probability sample,
  - since the probability of non response is not known
  - if the probability is modelled as 0 or 1 the inclusion probabilities are not positive

- Coverage errors, frame inconsistencies
  - The frame does not cover the population e.g. Random Digit Dialling. Some people have no phone or two phones. Or the frame is old and contains non-eligible units)
  - Doubles or multiple entries
  - With illegible elements the population size is not known

- Unknown selection probabilities.
  - Select customers by taking a random sample of invoices or visitors to a park. Some customers have more than one invoice or visits to the park.
- Real time sampling, like interviewing persons passing a counter or shop exit.
  - You are unable to interview more than one at a time so some persons will not be approached even if they should have been selected.
- Snowball sampling
- ...

2.2 Horvitz-Thomson estimator

Πps-sampling

- Sampling with varying probabilities is often called πps-sampling
- Originally: (Inclusion) Probability Proportional to Size
  - now to something instead of to size
  - Pps-sampling (selection) probability proportional to size. E.g. draw n units with probability $p_i = x_i / \Sigma x_j$. The inclusion probability will be $1 - (1 - p_i)^n$ and the sample size will usually be smaller than n.

HT-estimator

- If a unit is taken with probability 1/k it can be said to represent k units in the population
- If a unit is taken with probability π_i it can be said to represent $1/\pi_i$ in the population
- A good guess for an estimate of the full population is thus $T^*_{yHT} = \Sigma_S Y_i/\pi_i$
- Design-unbiased $E(T^*_{yHT}|U) = E(\Sigma_U I_i Y_i/\pi_i) = \Sigma_U \pi_i Y_i/\pi_i = \Sigma_U Y_i = T_y$

Probability sampling rev.

- This estimation approach can be used for most design-based methods
- ”Sometimes also π_{ij} = P(i € S, j € S) > 0 for all pairs i,j € U”
- This condition is needed to be able to estimate variance with an design-unbiased estimator. Will be discussed later

2.3 Some simple sampling schemes using auxiliary information

- Stratified sampling (Very common)
- Cluster sampling (Multi stage sampling)
- Systematic sampling
- And many others: like Area sampling, Quota sampling (Be careful! Many examples of biased results), RDD (ask about number in household, of telephones, time at home), Real time sampling…

2.3.1 Simple Systematic

- Order the units in some sensible way.
- Select every (N/n):th unit
- To make it a probability sample. Select the starting point randomly.
- Is almost always better than SRS (except when there is a period). Why?
- The variance can not be estimated (design-based) – but …

Simple systematic sampling

- The variance under SRS (or other schemes) can easily be estimated (even better than from a SRS-sample)
- And this is known to be an overestimate so every interval will be conservative
- Can be remedied – e.g. by taking two starting-points and then every (2N/n)th
- An alternative is stratified sampling with n/2 strata (in the same order) and 2 units per stratum. Why good?

2.3.2 Πps Systematic Sample

- Order the units in some sensible way
- Calculate the cumulative inclusion probabilities $\Pi_k = \Sigma_{i<k+1}\ \pi_i$
- Chose a random starting point in $(0, \Pi_N/n)$
- Chose all units at distance Π_N/n
- Same result as above – better than ordinary πps-sampling but for periodicity and that variance cannot be estimated unbiasedly (but a useful upper bound can be found)

2.3.3 Self weighted samples (Självvägda urval)

- Probability sampling, where all first-order inclusion probabilities are equal.
- For example select municipalities with probability proportional to size – select k individuals in each selected municipality
- All ordinary estimates (means) are unbiased. (Thus you can give the data to other persons who do not know as much statistics as you).
- The ordinary variances are not correct, though (usually substantial underestimates)

2.4 Decision of the Sample size Standard recommended approach:

- Decide needed precision (denoted e) (eg level .95)
- Decide sampling scheme and estimation formula (eg SRS and mean)
- Compute the theoretic formula for the variance (eg $N(N-n)/n\ \sigma^2$)
- Solve for sample size (e.g. $n = 4N^2\sigma^2/(e^2 + 4N\sigma^2)$)
- Insert probable values of unknown quantities (here σ^2) and compute the needed size

Problems with the ordinary recommended approach:

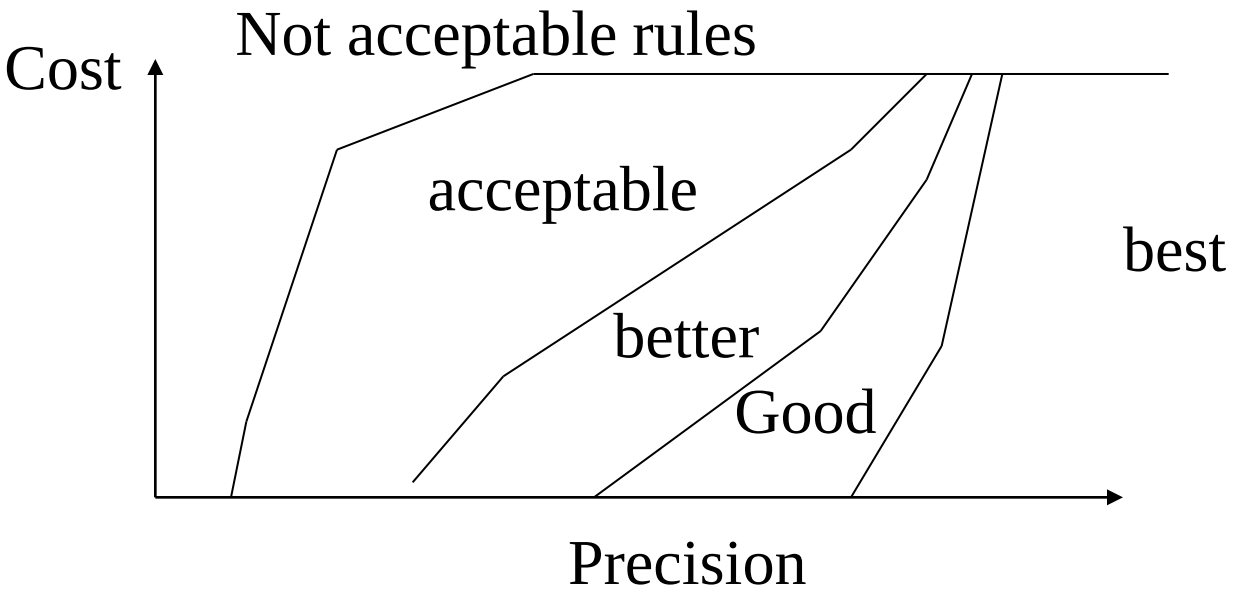
- Practitioners can seldom translate a practical problem into the needed precision
- A consulted statistician usually will and cannot help. He will say that it is up to the user.
- Random error is only one type of error. One must also weigh against other aspects
- If you cannot afford the study, you do not know what to do. Decrease the needed precision?
  - But in that case you could obviously not decide the precision correctly in the first place.
  - Or if you could the results of the study no use

In practice!

- Define a few resonable designs or total cost levels for the survey (below what you can afford)
- Compute the full design and corresponding sample sizes (take all costs into account. Consider also other design and quality variables probable non-response, questionnaire design a.s.o)
- Compute the possible precisions with these cost levels for some reasonable parameter values. (Decide on the full design, number of of reminders, interviewer education, amount of editing …)
- Consider also other data sources and whether to do the survey at all, which may help
- Decide if the precision is worth the cost for one of the designs. If not don't do any study. If the precision for all cost levels is to low ask yourself/financier if you/he can afford more (or decide not to do a study)

In practice!

- Sometimes you start with a sampling design and calculate both the expected cost and the expected precision
- The optimal choice is always a weighting between them.



In practice!

- Sometimes you start with a sampling design and calculate both the expected cost and the expected precision
- The optimal choice is always a weighting between them.