Ekonomisk statistik
Economic statistics

Master course

Daniel Thorburn

Autumn 2011

Stockholm University

(Preliminary) Contents

1 Sampling
Repetition

Daniel Thorburn

Ekonomic statistics

Autumn 2011

Sampling

- You have a register ”a frame” containing all interesting units (e.g. enterprises) in a population.
- You may have background variables (X) (e. g. last year’s turnover according to accounting, number of employed according to preliminary tax payments)
- You are interested in other figures (Y). (e.g. order volume or investments in environmental protection). The goal is to estimate the total of Y for all units in the frame.
- Select a sample from the frame (using X) and find their values on Y.
- How should the sample be drawn to get the best estimate? How should the data be used (Y and X)?

- Probability samples
  - E. g. stratified samples
- Balanced samples
  - E.g. systematic samples
- Convenience samples
  - E.g. Cut off limits or only those that are willing to participate
- Representative samples
  - Inexact word for ”good samples” samples that give good estimates with small errors (sometimes unweighted averages give good estimates)
- Scientific samples
  - E.g. probability samples but also other designs like matched pairs etc

- Probability samples
  - Every unit in the frame must have a positive probability to be included in the sample
  - Every element in the sample has a known inclusion probability.

Totals and means can now be estimated with an unbiased estimate (and often other parameters e.g. median or mode)

  - Sometimes: every pair of elements in the sample must have known and positive probability to be included.

Also the variance can be estimated

- Think about the frame! Estimates are good only if the frame is good, since the estimates relate to the parameters of the population.

- Economists at the university often uses the companies listed at the Stockholm stock exchange. Is this really a good frame?

  I recently saw a project in Lund (Ph D theses) studying ”Corporate management in Sweden” and they used as frame Large cap and Midcap.

  – Mention some large Swedish enterprises not included.

2. Business registers

Daniel Thorburn

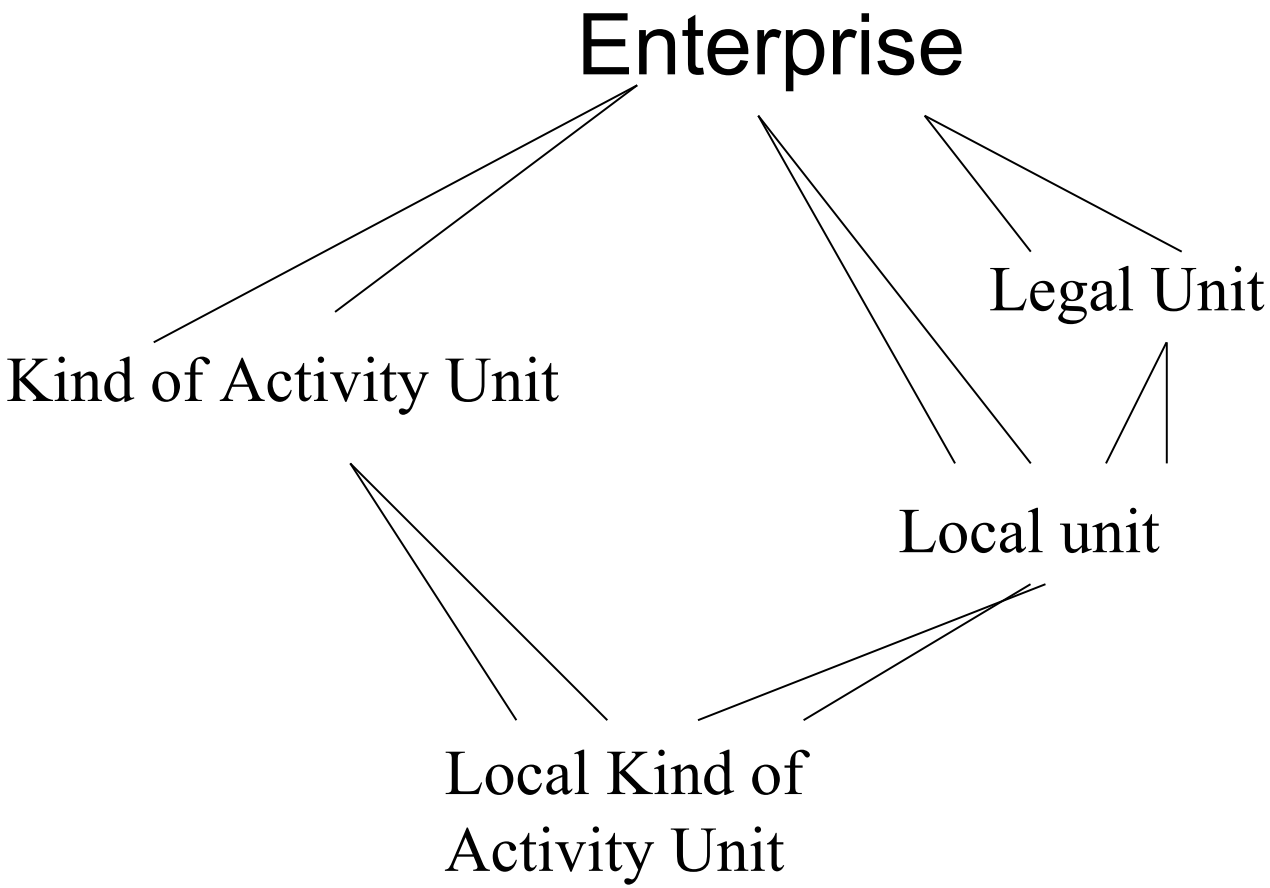
Economic statistics

Autumn 2011

Business registers, BR

- 2007 there were 945801 enterprises in Sweden with 1021083 local units (in Sweden).
- 538 101 were physical persons and 270 084 Swedish companies owned by share holders
- The rest is a mixture of partnerships, voluntary associations, housing associations, municipalities, foundations, economic associations, foreign legal entities, etc.
- The main source to the business register of Statistics Sweden is the tax legislation
- When the legislation changes, the number of businesses is changed.
  - In 1996 the limit to be registered in the VAT-register was lowered to at least one SEK of VAT.
  - The number of businesses then increased by 200 000.

Types of units in Swedish BR



Also
Firms,
AST

Data sources

- Register
- VAT registrations
- Salary Data (payroll taxes, withholding tax)
- Tax assessments, financial statements, annual reports
- Intrastat (foreign trade)
- Sampling and other special studies

Problems with business registers

- What is a business? Ekonomic unit, Legal unit, Local unit or …
- How to detect new businesses? To know when a company is discontinued. When a company is bought by another, is it discontinued or are there still two companies? When two companies merge are both discontinued or only one of them?
- Businesses are classified after activity (ISIC). It is recommended to use a maximum criteria after turnover. If the main part of the activity is trucks, the company is classified as belonging to the car industry even if it has a large IT-department and it is the biggest company in Sweden for both aircrafts and maritime engines.
- Updating. E.g. You want to investigate all businesses in the IT-sector and take a sample from the IT-frame. If you get a company who no longer belong to the IT-sector then it is removed from the sample (overcoverage). But what to do with new companies in the IT-frame?

The Swedish BR is updated regularly. If in a study a change of industry (activity) is found then it will immediately be included in the register and used in the next version.

But for comparison reasons only four versions of the register are used as a frames every year. But those doing the interviews must have access to the latest data

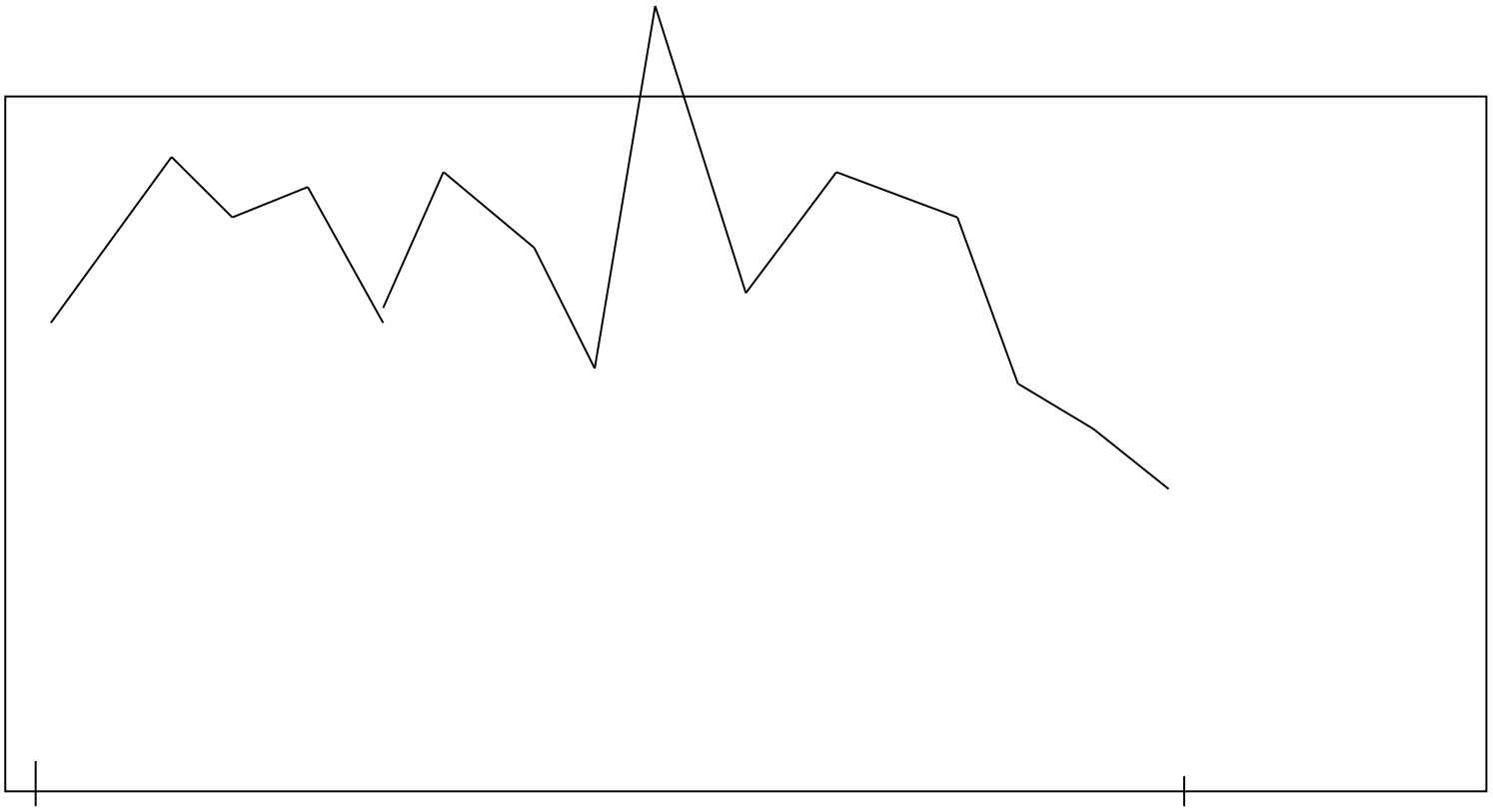
This means that the data becomes better for large companies.

It also means that the order of the studies may affect the results. A study of the chemical technology industry first and then of the engineering industry. The second study will include those that changed from chemical engineering. While those who switched to chemical engineering will not appear in any.

Note that the register at Sept 1 2011 may have at least two meanings

- Statistics based on all units in the register at that date
- But the register is alive and updated. Businesses started in August will not be entered until later. Change of industries, discontinuations, fusions a.s.o will also be entered later.
- Statistics made in December for all units in the register at that time active in August will give another figure since many changes have been registered at that date.

A typical diagram over the number of new started businesses by month Why this decline at the end?



- ”Large companies are seldom new”. They are usually spin off from business units or they may be old businesses with new owners.

- Hence Statistics Sweden investigates all large units declaring themselves being new.

- Statistics for new large businesses can thus have even larger delays (up to three years). (The company is included in the register but with unknown starting year)

Other Business Registers in Sweden

- Other actors, among the largest in Sweden are,
- RATOS an investment company that owns many companies in the information sector
  - Dunn och Bradstreet (credit ratings), MM-analys (e.g identify potential customers and growing firms).
- Their data base is larger than Statistics Sweden.
  - they buy Statistics Swedens business register
  - a special division inputs
    - all annual financial reports to the database (including chairman of the board and managing director)
    - all credit fallacies reported to ”kronofogden” (bailiff),
  - Historically not so good. Has to follow the Swedish laws KUL / PUL, (credit ratings and data integrity laws). SCB has larger possibilities to keep data over time
- UC (Upplysningscentralen) the most well known credit ratings
  - They produce statistics over all bankruptcies in Sweden

More complicated frames

- Multiframes (you have several lists and use a combination of them. One list may contain legal units and other establishment (local units))

- Hierarchical sampling (Cluster sampling). The object is to select local units. This is done by first selecting legal units and then constructing a subframe of all local units within that unit. A way of finding shops by first selecting chains (H&M, ICA, Hemtex, Kappahl, Jysk and Plantagen …) and then shops from them.
  - You need not a list of all shops in Sweden.

3. Sampling and estimation – businesses, πps

Daniel Thorburn

Economic statistics

Autumn 2011

Sampling from enterprise populations

- Enterprises and other economic units usually vary quite a lot in size.
    - You want to select some important (large) enterprises with very high probabilities (even one. Often called a total sampling stratum) and some with very small probabilities (even zero, cut off).
    - The populations are very skew with large true outliers. (It is important to have a good system for handling outliers)
- Enterprise populations change fast over time
    - large problem with keeping an up to date frame

πps-sampling notations

- Inclusion probabilities π are central in design-based inference. Inclusionindicator I=1 if the unit is in the sample (I = 0 otherwise)
- First order inclusion probabilities:

$$\pi_i = P(i \in S) = P(I_i = 1)$$

- Second order inclusion probabilities:

$$\pi_{ij} = P(i,j \in S) = P(I_{i,j} = 1)$$

specielly: $\pi_{ii} = \pi_i$

πps-sampling - formulas

- Horvitz-Thompson (HT) estimator:

$$t_{y,HT} = \Sigma_S\, y_i/\pi_i = \Sigma_U\, I_i y_i/\pi_i$$

Often: $t_{y,HT} = \Sigma_S\, \omega_i y_i$, (where $\omega_i = 1/\pi_i$ is called design weights)

Unbiased: $E(\Sigma_U\, I_i y_i/\pi_i) = \Sigma_U\, E(I_i) y_i/\pi_i =$

$\Sigma_U\, \pi_i\, y_i/\pi_i = \Sigma_U\, y_i$

- Variance: $\Sigma\Sigma_{UU} ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j))\, y_i\, y_j$
- Alternative form:

$$\tfrac{1}{2}\, \Sigma\Sigma_{UU} (\pi_i\pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2$$

  - Easy to see that the variance is zero if the inclusion probability is proportional to y (last parenthesis = 0).
  - If ratios are very different they should be independent (first parenthesis = 0). (Cf stratified sampling)
- Variance estimation: Sum over SS instead of UU and divide the terms by π_{ij}

Some methods to do πps

- Sampford
- Stratification (after inclusion probability, most common)
- Systematic
- Pareto
- Poisson sampling (only approximate)
- The SAMU-system (will be discussed below) allows several of these methods to be used (not Sampford nor conditional Poisson)

4. Coordinated samples

Daniel Thorburn

Economic statistics

Autumn 2011

- Type of coordination
  - Positive coordination with large overlap between samples
  - Negative coordination with small or no overlap
- Why coordinate samples?
  - Response burden
    - Can be distributed more equally
    - Can be decreased
  - Get information about relations
    - Over time, longitudinal studies
    - Between questions in different surveys. If the same companies participates in two different surveys on profits and occupational environment, say, you can see the relation between them

Distribute burden fairly

- Large enterprises are often selected and think that it is okey. They have routines and sometimes even special persons employed for participating in surveys.
- Small enterprises are seldom selected and does not complain much
- Companies in between, which have no routines but are nevertheless questioned often (and where it is not always so easy to reply) are most often irritated. It is important for the NSI that they do not any have to participate in too many surveys in a short time. Put them in quarantine after the first survey..

Table 4. Enterprises included in sample surveys 2001, Excerpt

	Number of employed					
No Surv	0	…	10-19	…	200 -	Total
1	1809		5384		15	28468
…						
5	0		80		135	1158
…						
13	0		0		344	354
Total in S	2409		9507		1366	45427
Total in U	613144		17763		1716	826787

Subsamples - Screening

- Sometimes surveys are coordinated by just studying a subsample of those from the first study in the second
  - For example, every fifth company may also answer some additional questions. The sample size in the second stage need not always be large. In the estimation phase you can do a calibration depending on some important basic questions in the first study.
  - Screening. We want to specially investigate companies with certain characteristics, such as those that have made work environment improvements or have female presidents. Have a question about that in the main study. Then in the second phase return to a sample of those saying yes.
    - Sometimes planned selection like selecting equally many with male and female presidents will lead to more efficient comparisons

Longitudinal studies

Longitudinal studies is a collection term for studies where the same units are followed over time.

Thus you must be able to follow the same units over time. Some work is needed on the definition of same unit. What to do with takeovers, mergers, reconstructions a.s.o.

Exemple: Rotating samples

Every selected enterprise is included a predetermined number of times, say four consecutive years.

The reponse burden decreases since

- The first time you are in a study is always the hardest
- Basic questions may be asked only once
- But rotation means that none is included forever which would be considered unfair

– Easier contacting costs. You know which employee who answered last time

– Possibility to follow the development over years

An example of such a study is the short-periodic study of salaries

Four active rotating panels

A simple example

Time Panel	2000	2001	2002	2003	2004	2004	2005	2006	2007
A	X								
B	X	X							
C	X	X	X						
D	X	X	X	X					
E		X	X	X	X				
F			X	X	X	X			
G				X	X	X	X		
H					X	X	X	X	
I						X	X	X	X
J							X	X	X
K								X	X
L									X