Ekonomisk statistik
Economic statistics

Master course

Daniel Thorburn

Autumn 2012

Stockholm University

Contents

1 Sampling

Daniel Thorburn

Ekonomic statistics

Autumn 2012

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 or 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 sampling
  - E. g. stratified sampling
- Balanced sampling
  - E.g. systematic samping
- Convenience sampling
  - Good or bad. e.g. Cut off limits or only those that are willing to participate
- Representative sampling
  - Inexact word for good samples that give good estimates with small errors (sometimes unweighted averages give good estimates)
- Purposive sampling
  - Select some special units e.g. typical farms with only the usual crops and animalss or enterprises that are expected to be cyclical react fast to changes in the economic cycle
- Scientific sampling
  - A collective name for e.g. probability samples but also for other designs like matched pairs

- 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 properties e.g. median or mode – at least consistent)
  - Sometimes also: every pair of elements in the sample must have a known and positive probability to be included both.

  Now also the variance may 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 use the companies listed on the Stockholm stock exchange. Is this really a good frame?

  - I recently saw a project in Lund (with Ph D theses) studying ”Corporate management in Sweden” and had used this frame (Large cap and Midcap).
  - Mention some large Swedish enterprises not included.

2. Business registers

Daniel Thorburn

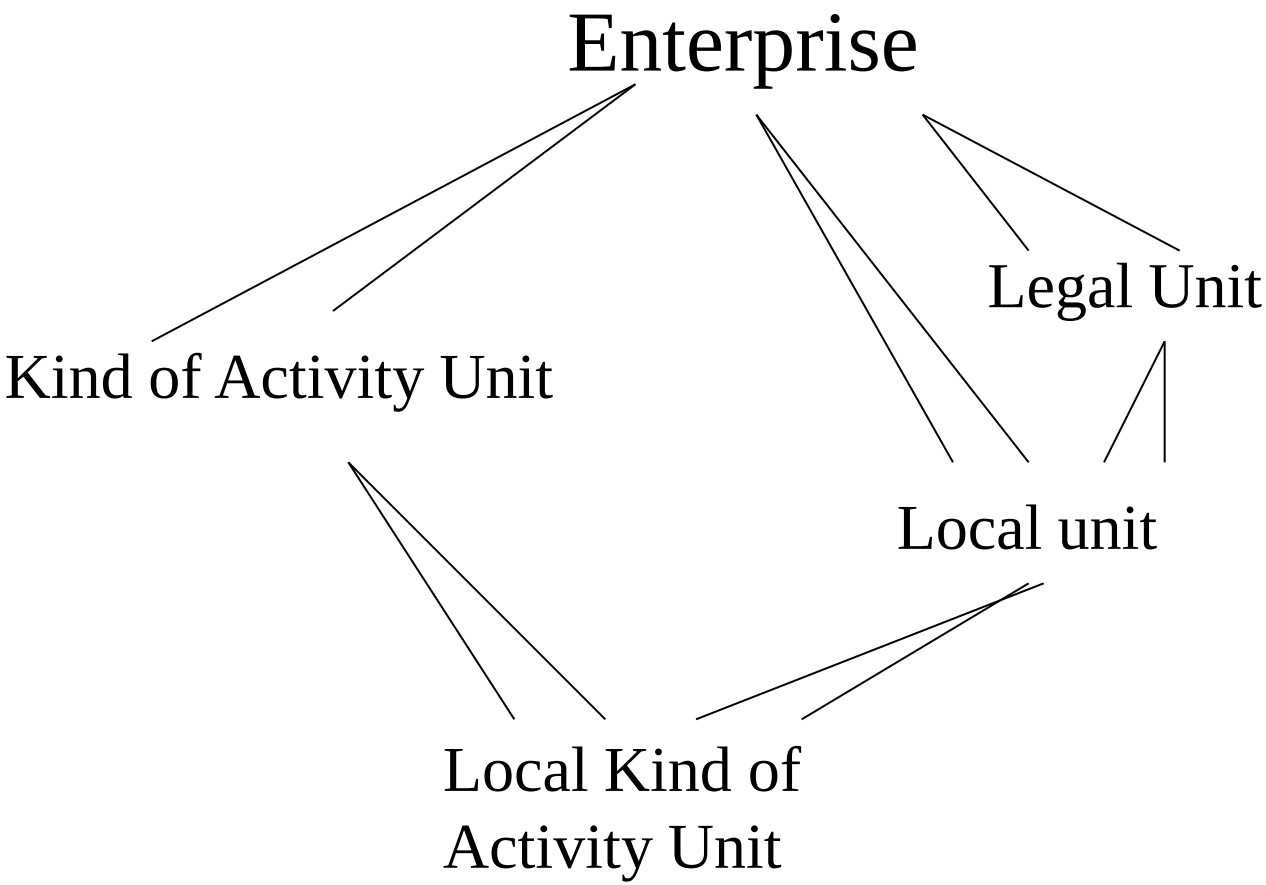
Ekonomic statistics

Autumn 2012

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 resten is a mixture of partnerships, voluntary associations, housing associations, municipalities, foundations, economic associations, foreign legal entities, etc. (handelsbolag, ideella föreningar, bostadsrättsföreningar, kommuner, stiftelser, ekonomiska föreningar,, utländska juridiska personer 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 some other types

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 (ISCO previously ISIC). It is recommended to use a Maximicriterion 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 air plane and maritime motors.
- Updating E.g. you want to investigate all businesses in the IT-industry and take a sample from those in the ITsector in the frame. But if you get business who no longer should belong to the IT-industry they are removed from the sample (overcoverage). But what to do with those who are not selected and have changed to that sector?

The Swedish BR is updated regularly. If, you in a study finds e.g. a change of industry (activity) , 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

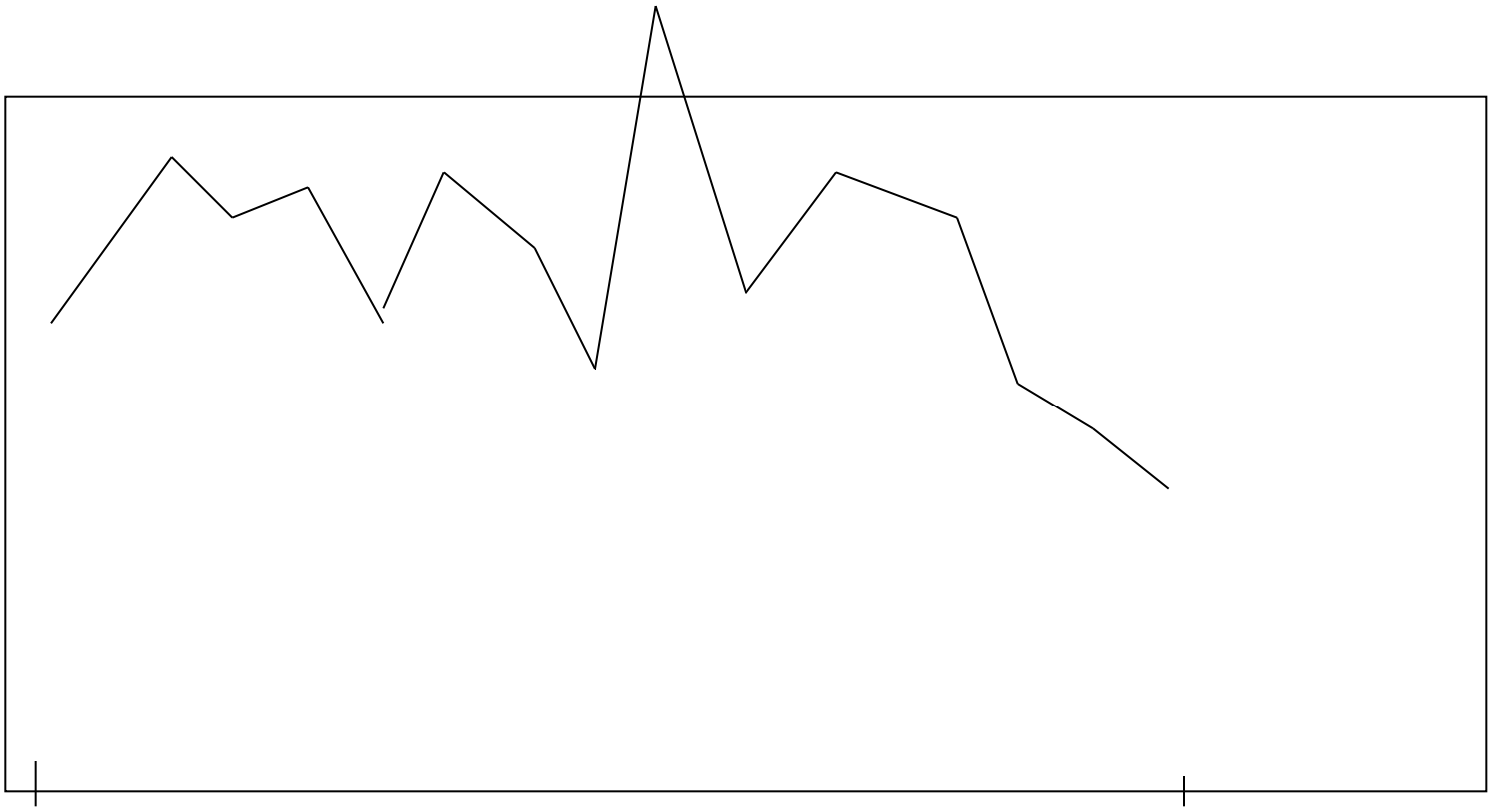
This means that the data becomes better for large companies.

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 leater.
- 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 business units or they may be old businesses with new owners.

- This means that 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 greowing firms).
- Their data base is larger than Statistics Sweden.
  - For sure they buy the database from Statistics Sweden
  - A division converting all annual financial reports to a computerised version (all annual reports to ”Patentverket”, including chairman of the board and managingdirector)) and
  - All credit fallacies reported to ”kronofogden”,
  - 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 within credit ratings
  - They produce statistics over number of bankruptcies that you may see in the Swedish media

More complicated frames

- Multiframe (you have several lists and use a combination of them. One list may contain legal units and another local units (arbetsställen)

- 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 without having to first list all shops in Sweden in the frame. …

3. Sampling and estimation – businesses, πps

Daniel Thorburn

Ekonomic statistics

Autumn 2012

Sampling from enterprise populations

- Enterprises and other economic units usually vary quite a lot in size. Three consequences:
  - 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
  - Larger problems to keep an up to date frame

πps-sampling notations

- Inclusion probabilities π are central in design-based inference. Inclusionsindikator 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:

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

  - Easily seen that the variance is zero if inclusion probability is proportional to y (last parenthesis = 0).
  - If ratios 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

Ekonomic statistics

Autumn 2012

- 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 partake 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 fair. They have routines and sometimes even special persons employed for participating in surveys.
- Small enterprises are seldom selected and does not complain much
- Inbetween companies, 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 not any of them 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 for instance calibrate after 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. In the main study have a question about this. Then return to a sample of those saying yes in the second phase. .
    - Sometimes planned selection like selecting equally many with male and female presidents. Will lead to more efficient comparisons

A simple example

- 500 enterprises are studied in a SRS-sample from 5000 firms in the first round
- In one important respect they can be classified into four classes with 250, 150, 75 and 25 firms after a variable observed in the first round
- 100 firms are selected for the second round, 25 in each group.
- Observed stratum means and variances in the second round are 5, 25, 30, 145 and 5, 4, 20, 200
- Now estimate the total
  - Mean (250*5+150*25+75*30+25*145)/500=21.75
  - Ita variance is more complicated (see next page).
- This approach can also be used for comparing the means in different groups in an efficient way

Variance estimation

- First consider the variance if the value of all units in the same group had been the same (i.e. 250 units with value 5, 150 with 25, …). Standard SRS-formulas give 1.65
- Next compute the variances in the second step wihin each group (drawing 25 from 250 and … with SRS). 0.18, 0.133, 0.533, 0. Weighting them together gives 0.069.
- The sum of the two components gives 1.72.
- The variance if only one SRS-sample with 100 units had been drawn would have been 8.39. The two stage sampling procedure has improved precision considerably.

1.5.2 Longitudinal studies

Longitudinal studies is a term for studies where you follow the same units over time (The opposite is cross-sectional studies)

A typical example if if you want to follow up what happens to those firms that were reconstructed during the financial crises or were fined from environmental reasons or have female members of the board

To follow units over time means that you must be able to know what is meant by the same unit. What to do at takeovers, fusions, bankruptcies with a following reconstruction and spin offs. (easy for persons but not enterprises or households).

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 noone 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

– E.g. The short-periodic wage-study (kortperiodisk lönestatistik)

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

Estimation with rotating studies with k active and equally large panels

Composite estimators

- Let X_{ti} be the estimate of the mean from the i:the panel
- Suppose that all these estimates have the same variance, σ^2, and that the correlation decreases exponentially between times within panels $\rho^{|t1-t2|}$ (Large firms are usually large also next year)
- A simple estimate of the meam at time t is then the mean of all panels $\Sigma_i X_{ti}/k$ with the variance σ^2/k (no correlation between panels)
- The variance for the difference between two time points, t och t+1, will then be $2(1 - ((k-1)/k)\,\rho)\,\sigma^2/k$ (Prove it!)
- The random error decreases with the number of panels i.e. the period of rotation, k. The variance without any overlap (two independent samples) would have been $2\,\sigma^2/k$
- E.g. with k = 4 and $\rho = 0.9$ the gain is a factor 0.325

- But it is possible to do something even better (but it is seldom done)
- The difference between the first and second time point can be estimated in two ways:
  - The difference betweethe common panels
    $D_1 = \Sigma_{i=2}^{k} (X_{2i} - X_{1i})/(k-1)$ with variance $2(1-\rho)\ \sigma^2/(k-1)$
  - The difference between the new and old panel
    $D_2 = (X_{2k+1} - X_{11})$ with variances $2\sigma^2$
  - If these are weighted together with optimal weights (inversely proportional to their variance) one gets
    $(D_1 + (1-\rho)/(k-1)\ D_2)/(1 + (1-\rho)/(k-1))$
    with the variance $2\sigma^2/(1 + (k-1)/(1-\rho))$ (Prove it!)
- With k = 4 och = 0.9 the gain will be a factor 0.129

- Can you explain why this is seldom used?

- One does not want to change already published estimates.
- And it is natural (but not optimal) to estimate the level one year with the average of all the values observed that year
- One wants to have consistency, the estimate of the change should be the difference between the two level estimates. But as we saw one looses precision by requiring this.

5. SAMU – Permanent random numbers

Daniel Thorburn

Ekonomic statistics

Autumn 2012

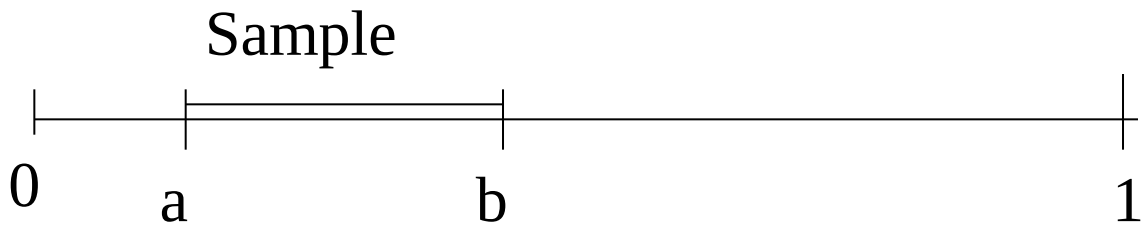
Permanent Random Numbers, PRN

- Every unit in a frame (eg the Business Register) is given a uniformly distributed random number, as soon as it comes into the register, U_i, (in the interval (0,1)). It is solely used for sampling purposes. This random number is (in principle) retained as long as the unit remains in the frame.

- A simple way to draw a sample is then to take all businesses with PRN:s in the interval (a, b). The selection will account for (roughly) the proportion b-a of the population, and is an SRS regardless of when we draw the sample. (Check that the inclusion probabilities is correct).

- (Formally the PRN is only given in SAMU not in the BR, i.e. the enterprises and the persons at Statistics, Sweden do not know them. They will only see the result of the sampling)

Sampling wirh permanent random numbers



The elements' permanent random numbers

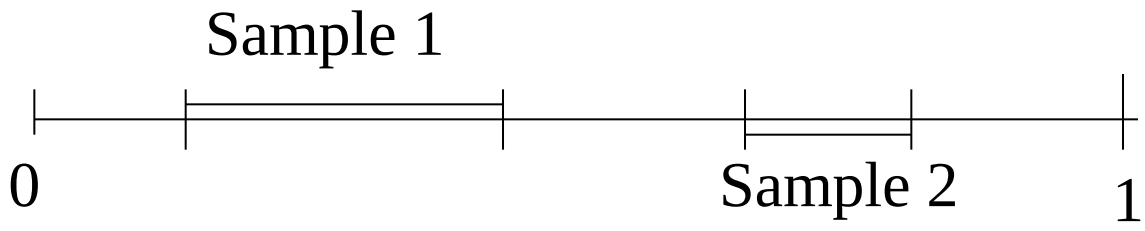
Another sample – interval over the end

The elements' permanent random numbers

- This idea was first suggested by Johan Atmer and Lars-Eric Strandberg at Statistics, Sweden and was called JALES
- It is an integrated part of the selection system that is now used at Statistics Sweden called SAMU (for ”SAMordnade Urval” = Coordinated samples)
- The idea is exported and is used at many statistical bureaus (NSIs) around the world.

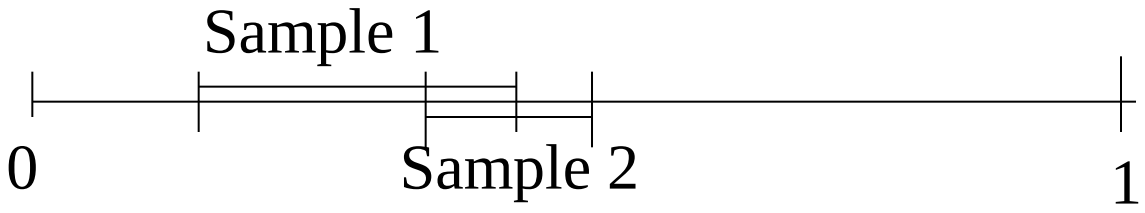
- Alternatively
- Take exactly n units starting from a. This is called a sequential SRSWOR
- If there are less than n units above a, restart at zero and take all units in the intervals (a,1) and (0,b).
- It is also possible to take the n units before a predefined point instead of after.

Negative coordination – No elements in common



The elements' permanent random numbers

Positive coordination – 50 of % of sample 2 in common



The elements' permanent random numbers

It is simple to take another sample with a chosen degree of overlap. Just take the intervals overlapping to a certain degree..

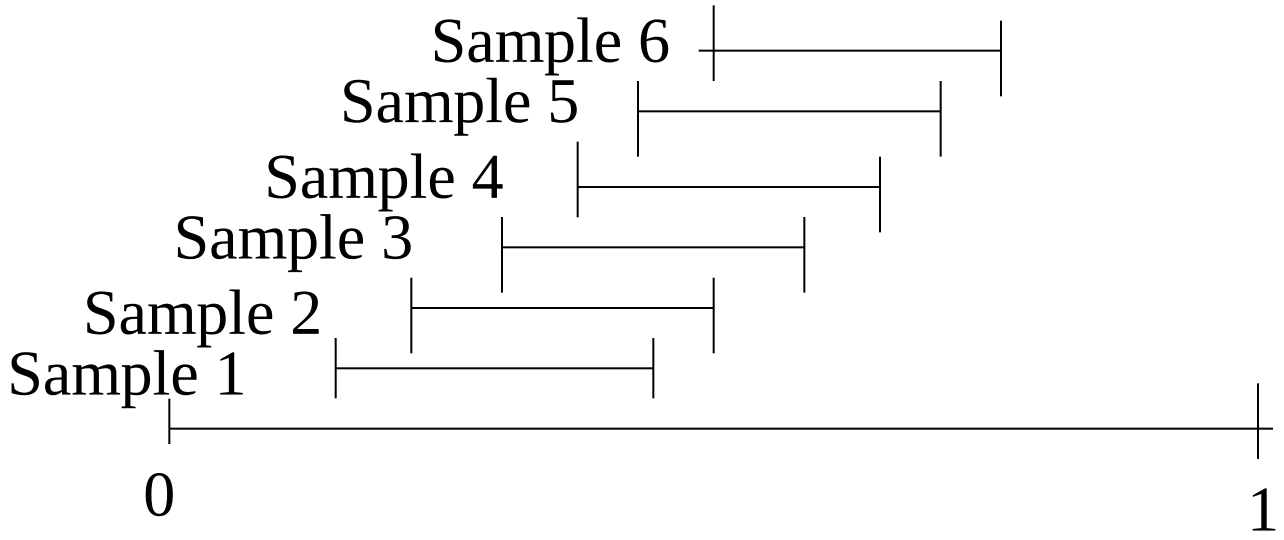
- If you choose two disjoint intervals every unit partakes in at most one survey. (possible to do as long as the sum of the inclusion probabilities does not exceed 1)
- By taking another interval for instance ((a+b)/2, c) on gets about half of the companies from the first study will be included in the second. Or take (2b-c, c) half of those in the second study can also be found in the first study (if 2b-c>a, otherwise impossible)
- Note that any interval above 1 or below 0 should be taken at the other end. E.g. c>1 => (0,c-1) is added.

Permanent Random Numbers

- Since the numbers are permanent it is easy to use them in longitudinal studies.

Rotating studies, Panel studies,

4 years rotation period



The elements' permanent random numbers

Longitudinal studies - rotating samples

- One problem with longitudinal samples is usually that the remaining sample is drawn from an old frame.
  - Less than one year old companies can only be found in the last panel not in the old part.
  - But in order to get unbiased estimates the sample must be drawn from an up-to-date frame.
  - This is solved by keeping the same interval but applying it to the latest frame. All new companies getting a PRN within that interval will be selected for the study.

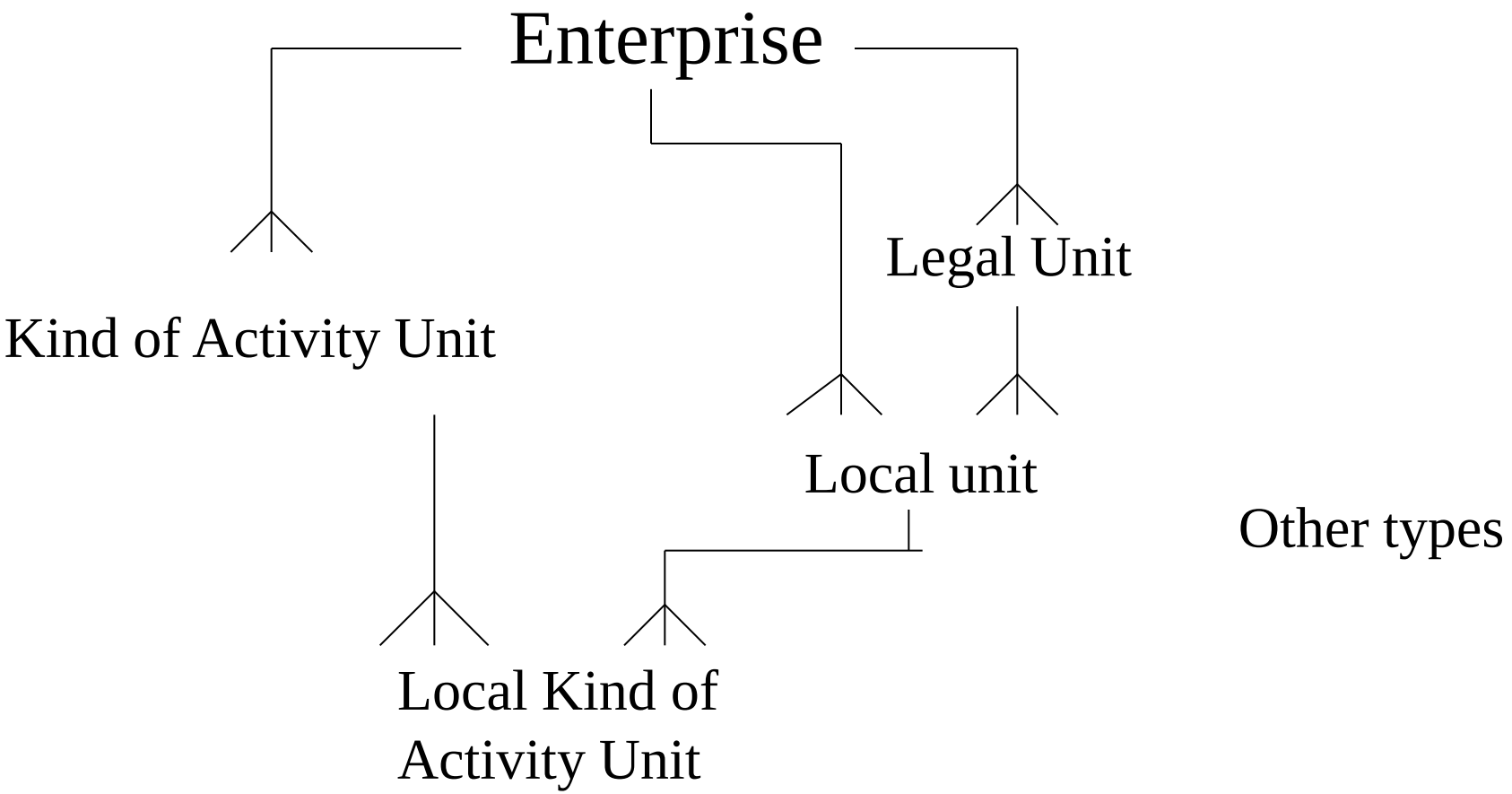
Longitudinal studies - rotating samples

- (Technical points:)
- For sequential Poisson sampling one moves the left point in the same way and counts n firms at each time from there.
  - (Since the right end of the interval may depend on the addition and removal of enterprises this may mean that a few companies may be included three ore five times)
- For practical reasons Statistics Sweden changes all PRN:s with the same amount to the left instead of moving the interval to the right. Gives the same effect but the same computer programmes can be used.

Different types of enterprises – how to handle PRN?

- The previous discussion handled objects of one type. But in the BR there are five (and more) types:
  - Enterprise,
  - Local units,
  - Legal units,
  - Kind of activity units,
  - Local kind of activity units.

Types of units in Swedish BR



Assessing PRN:s to all types

- Start by giving all local kinds of activity units PRNs. Let the higher levels get one these according to som fixed rules. (Usually the largest one when the number is given) aso
- In this way one may coordinate studies based on the enterprise level with those at the local unit level aso.
- The coordination may become inefficient with many local units.
- Probems with changing activities or starting new local units keeping a small unit at the previous chief unit.

Thank you for your attention