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
  - 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 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

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. (eady 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 2011

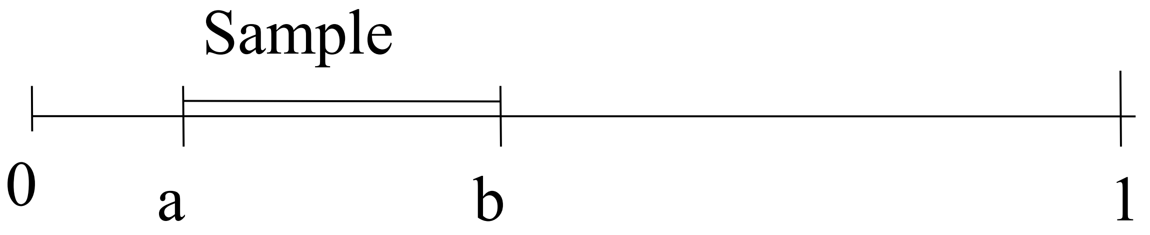
Permanent Random Numbers, PRN

- 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).

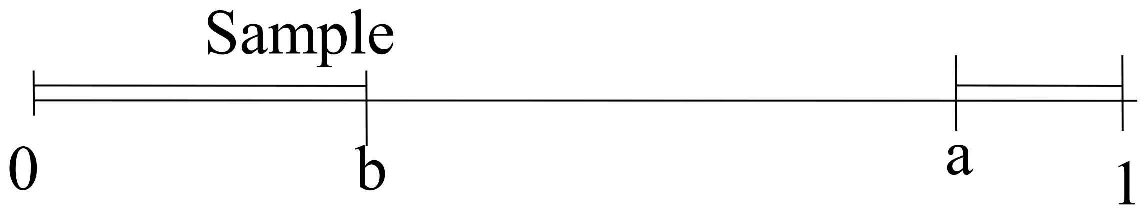
- (Formally the PRN is only given in SAMU not in the BR, i.e. the enterprises and the persons at Statistics, Sweden do not know them. They will only see the result of the sampling)

Sampling wirh permanent random numbers



The elements' permanent random numbers

Another sample – interval over the end

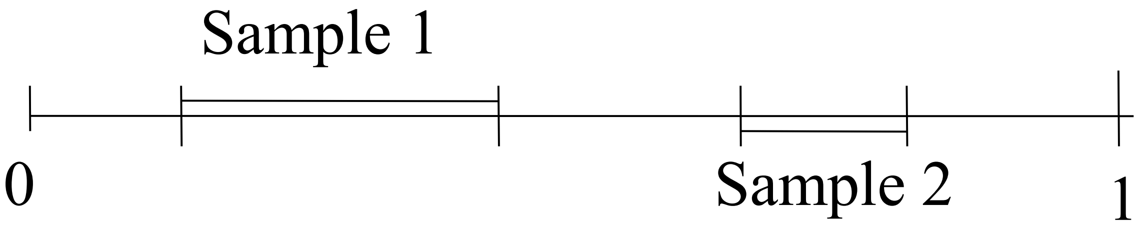


The elements' permanent random numbers

- 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.

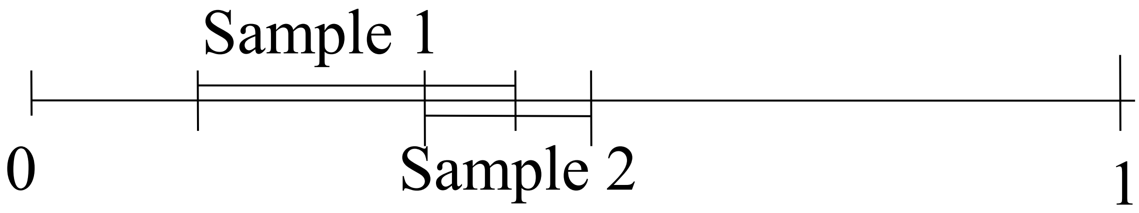
- Alternatively
- Take exactly n units starting from a. This is called a sequential SRSWOR
- If there are less than n units above a, restart at zero and take all units in the intervals (a,1) and (0,b).
- It is also possible to take the n units before a predefined point instead of after.

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..

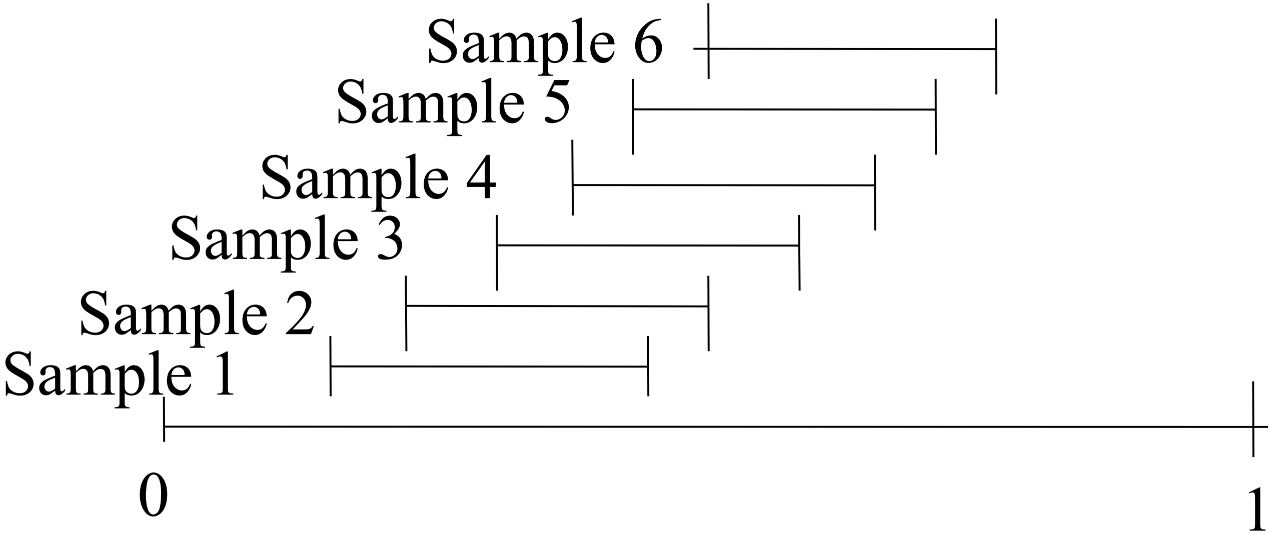
- If you choose two disjoint intervals every unit partakes in at most one survey. (possible to do as long as the sum of the inclusion probabilities does not exceed 1)
- By taking another interval for instance ((a+b)/2, c) on gets about half of the companies from the first study will be included in the 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.

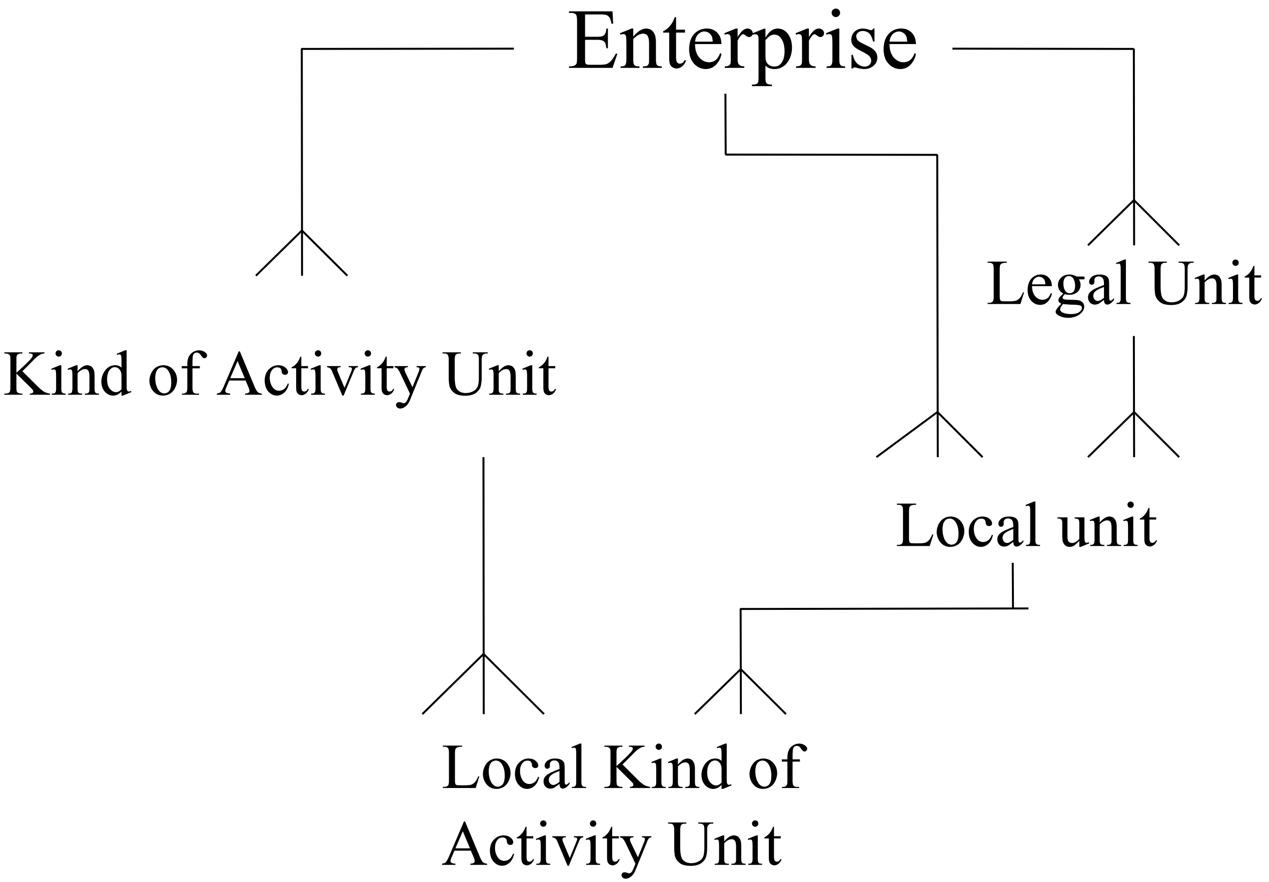
Longitudinal studies - rotating samples

- (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 (seven) types:
  - Enterprise,
  - Local units,
  - Legal units,
  - Kind of activity units,
  - Local kind of activity units.
  - (Firm)
  - (AST)

Types of units in Swedish BR



Enterprise

Legal Unit

Kind of Activity Unit

Local unit

Local Kind of Activity Unit

Other types
Firms,
AST.
Concerns

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) 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.

6. Permanent Random Numbers πps-sampling?

Daniel Thorburn

Ekonomic statistics

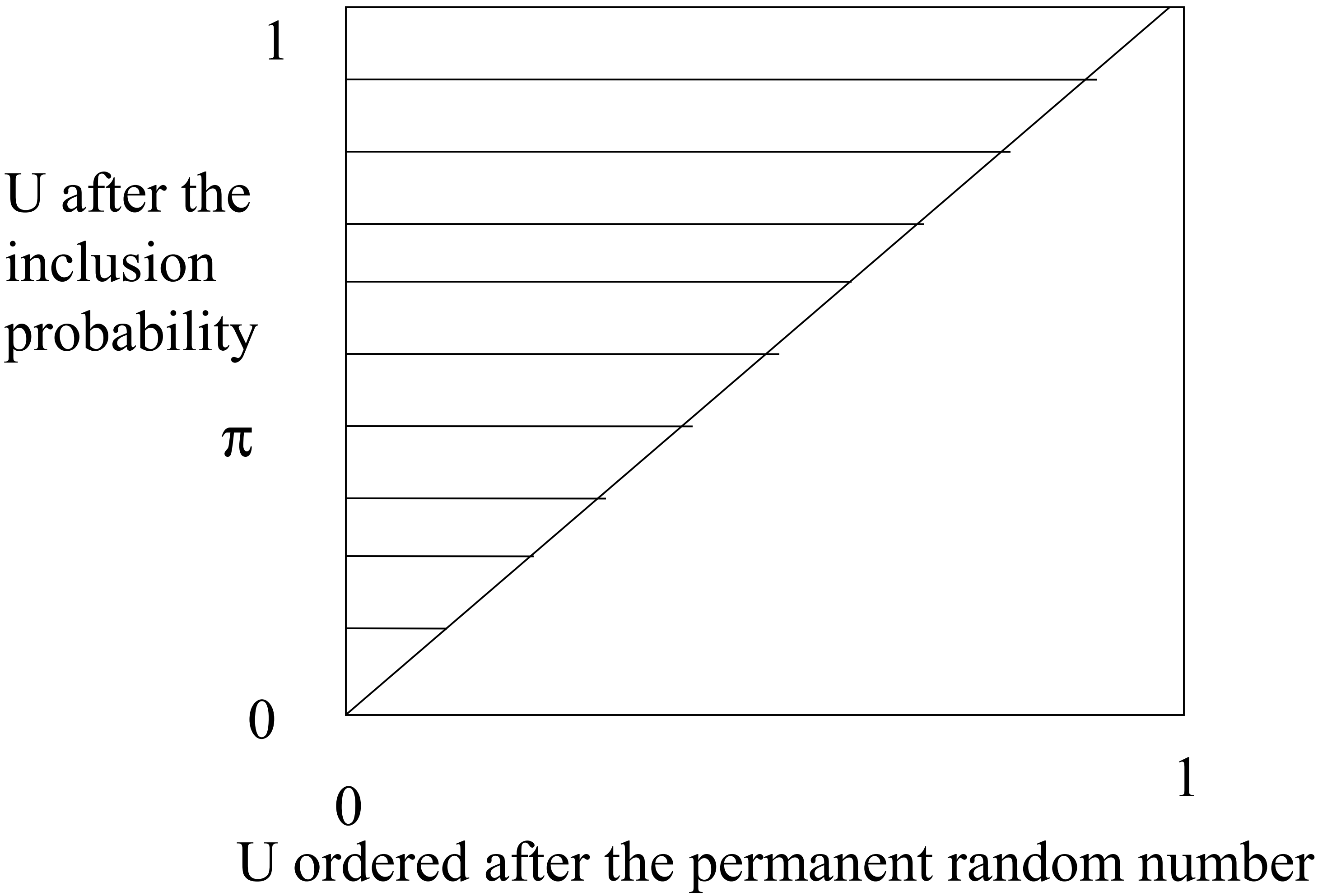
Autumn 2011

Permanent Random Numbers with πps-sampling?

- The previous discussion was about SRS. But what about πps-sampling?

- One way is to stratify after the inclusion probability and then handle the strata independently with different intervals.
- This is the most common approach but we shall discuss another way

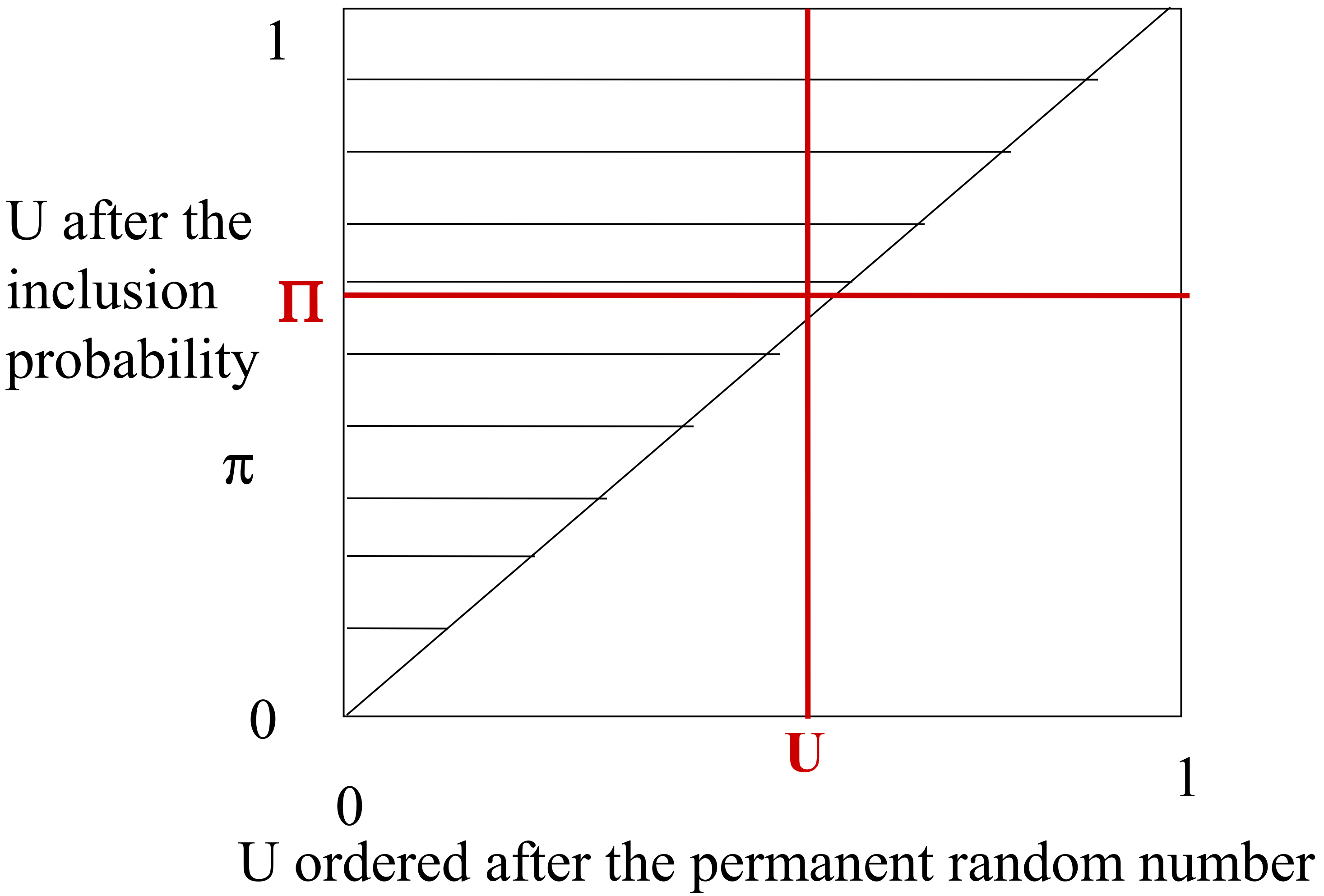
Permanent random number Selection with varying inclusion probabilities


1
U after the inclusion probability
π
0
0
1
U ordered after the permanent random number

Example: Should an enterprise be included?

Wanted inclusion probability **Π**

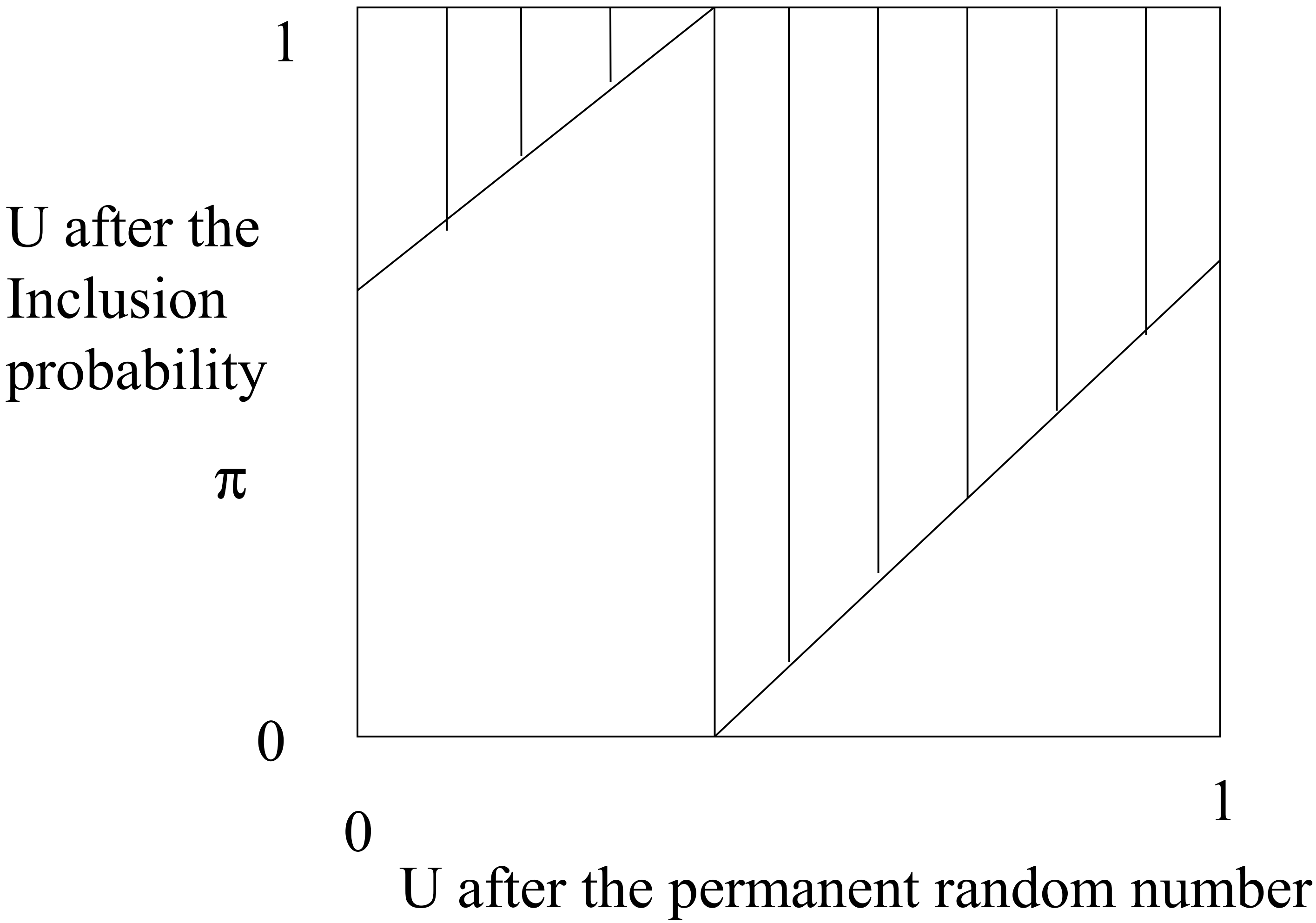
Permanent random number **U**



1
U after the inclusion probability π
Π
0
U
0
1
U ordered after the permanent random number

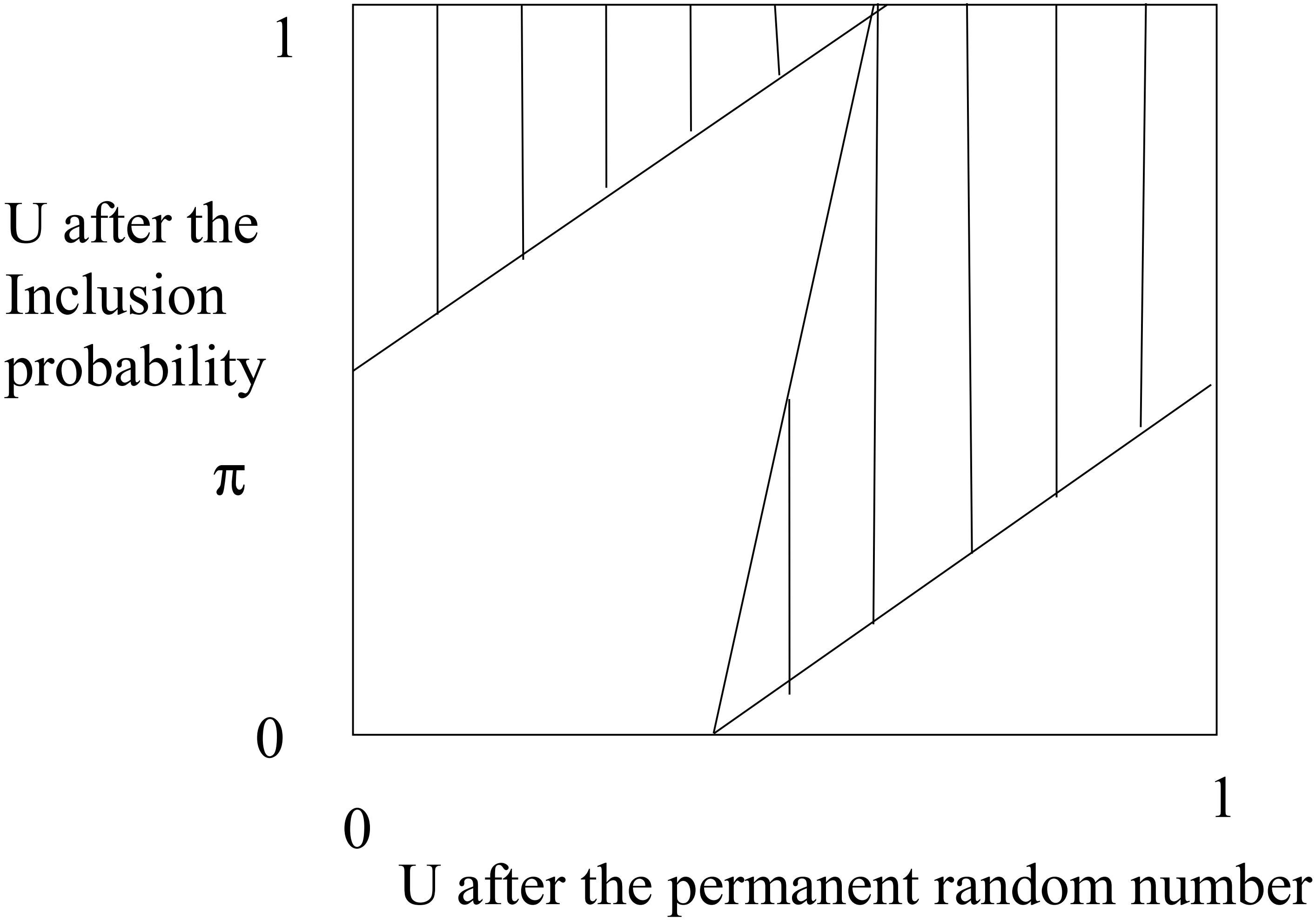
Permanent random number

Another sample


1
U after the Inclusion probability
π
0
0
1
U after the permanent random number

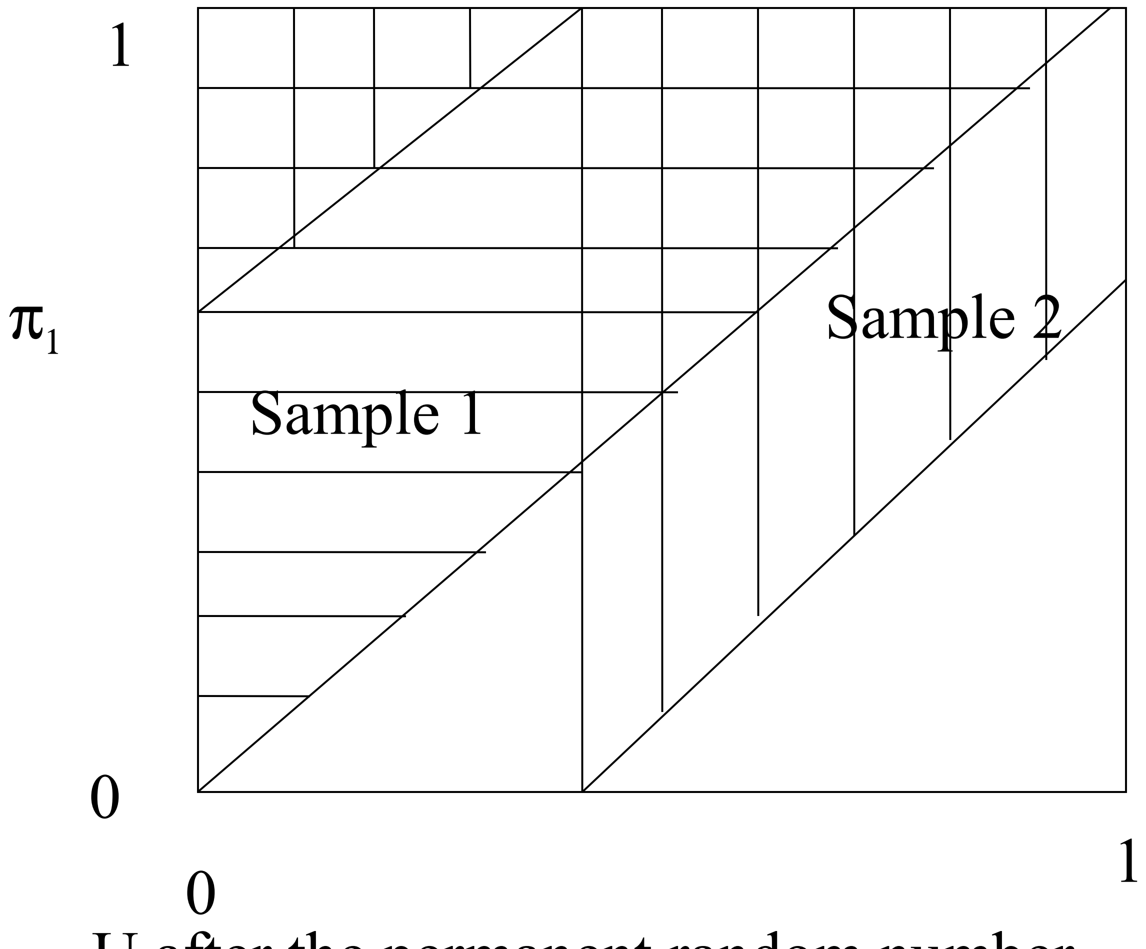
Permanent random number

And yet another



Permanent random numbers

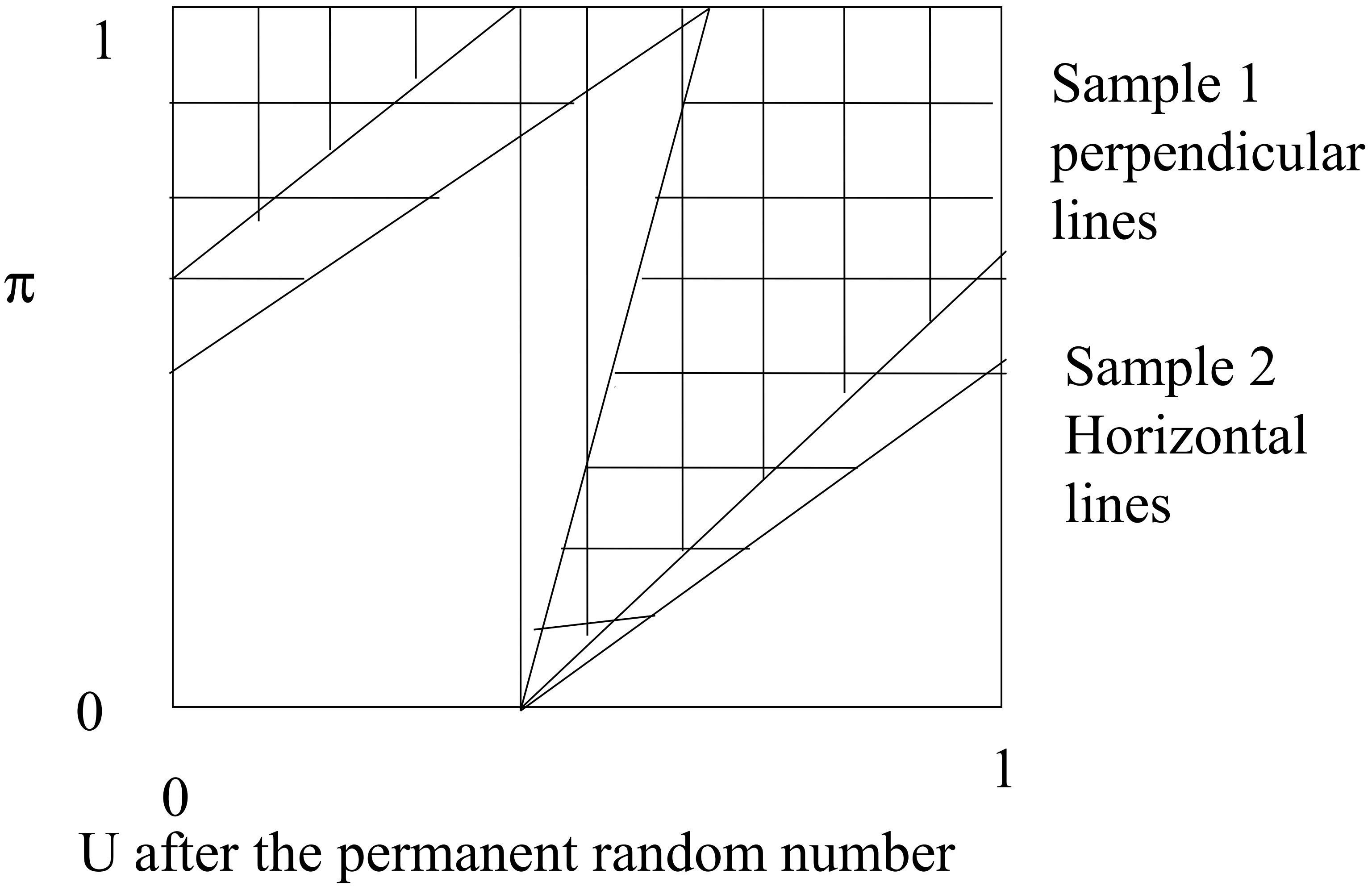
Two coordinated samples /with the same inclusion probabilities



U after the permanent random number

Permanent random numbers

Two coordinated samples /with the same inclusion probabilities. Rotating samples with ¾ overlap if possible



Sample 1
perpendicular
lines

Sample 2
Horizontal
lines

Permanent Random Numbers

- When the inclusion probabilities are different in the two samples, you cannot illustrate this in a two-dimensional figure (with π at the y-axis)
- It is simple to use Poisson sampling or Pareto πps
  - Poisson: Compute $Z_i = \pi_i/U_i$ and take the n largest (Roughly, those larger than one)
  - Pareto: Compute $Z_i = \pi_i(1-U_i)/ (U_i(1-\pi_i))$ and take the n largest (Roughly, those larger than one)
  - Check that $P(Z_i > 1) = \pi_i$ (in both cases)

Coordinated samples

- You have two intended inclusion probabilities $\pi1_i$ and $\pi2_i$.
- Take the first sample with Poisson or Pareto. Compute $\pi1_i / U_i$ resp. $\pi1_i(1-U_i)/(U_i(1-\pi1_i))$ and take the n largest
  - Or use another starting point (i.e. use U_i–a modulo 1)
- Next sample. Replace a with another number b (where b-a is chosen larger than most of the inclusion probablilities if you want samples with no overlap or b=a if you want maximal overlap) $\pi2_i/(U_i-b)$ resp. $\pi2_i(1-U_i+b)/((U_i-b)(1-\pi2_i))$ and take the n largest

Changing populations

- Are always difficult.
- The register is usually updated continuously. This means that if you make two different surveys at two different times you get two different populations too.
- If you use auxiliary variable in both surveys (say the turnover last year). Since the population has changed the value on last years total will thus be different in the two surveys. The readers will be confused, when the same quantity is presented with two different values. (You may sometimes see this in official statistics)
- To solve this different solutions are suggested
  - draw the samples at the same time from the same frame for different surveys at different time points
  - Adjust the weights (by calibration) so that the estimatess always will be the same figure
  - But these methods usually give slightly larger random errors

7. Sampling – outliers

Economic statistics

Autumn semester 2011

Stockholm University

Outliers

- Outliers are values which deviate much from what is expected.
  - They may be wrong (good ”editing” procedures are needed)
  - But they are often correct – and in that case they affect the estimate unreasonable
- In ordinary statistics you are advised to use robust estimators, when there may be outliers (like the median or the trimmed mean)
- In a sample survey this is usually not possible. You are interested in the total turnover of an industry not the median, the Hodges Lehmann estimate or any other robust location estimator. You cannot just trim away Volvo or Microsoft just because they are extreme outliers

Outliers

- Sometimes odd things happen. A dormant desktop corporation with   5 000 € in consolidated capital and no employees makes a new emission of shares, builds a new paper mill for 200 miljons Euros and employs 1500 persons.
- If it had the inclusion probability 1 in 10 000 it represents other enterprises and the standard estimate would say that 15 miljon Swedes are employed in the paper industry which has invested 200 billion euros.
- Thus one often has to construct rules for how to handle these problems.

What to do about (true) outliers?

- The most common approach is trimming i.e. change the design weight to 1. (i.e. it only represents itself. In order to estimate the total number of firms correctly, you should also increase the weights of the other in the sample. E.g. if the weights (1/π) 30, 20, 25 50, 5, 8, 2, 10 -> 1, 24,9, 31.1, 62.3, 6.2, 9.9, 2.5, 12.2)
  - One problem is to decide when this rule should apply.
  - Often this is decided by how much of the stratum total the enterprise represents. (Such a rule may say that if it represents more than 5% of the total or more than half the stratum total the company should only represent itself)
  - The most common way is to make the decision ad hoc.

- Another way is to use Winsorizing i.e. divide the company into two parts
  - One with the observed values and the weight 1
  - One with the remaining weight (w-1) with the study variable equal to a certain percentile (90%, say) in the stratum ((1, 29), 20, 25 50, 5, 8, 2, 10.

Other suggestions

- A third way is to take the weight the company had received if it had looked like this in the frame (and adjust the other weights)

- All these ways gives at the average too small estimates (bias) since all adjustments are made downwards. (Economic outliers are almost always too large). In simulations the mean square error seems to be smallest with the first version. Note, however, that a correction may be motivated at the regional level but not on the country level

- There are also suggestions to smooth over the years or regions. Do this correction each year for a ten year period and sum and compare with the unadjusted 10 year average. Increase all year estimates with this amount. (The level will then be corrected upwards if there are no outliers in the sample (there might be in the population) but downwards when there is not). In this way the estimate will be unbiased in the long run but not every year.

- A fifth way is to use model-based estimation in the tail and design-based in the rest (upper 5 or 10 percent or about 10-20 observations). The expected value of a unit above the 5% limit, c*, is estimated using a model.

- When all parameters are estimated the total is estimated by giving all observations
  - Less than c* their usual weight
  - Larger than c* the weight 1
  - The remaining weights is the sum of all weights for observations larger than c* minus the number of observations larger than c*. These will be given the expected value under the model.

- (The description here is to handle a skew distribution and fairly constant weights). Assume that the observations in the tail are from a skew distribution e.g. Pareto, Weibull or lognormal, which all can be used to describe skew distributions. For economic variables lognormal is often used). This procedure should be followed if there is an outlier or not.

- $(Y_i|Y_i > c)$ € Pareto(a,b,c). Density

$$f(x) = b\frac{(c-a)^b}{(x-a)^{b+1}} \quad x > c > a$$

- c is the cutoff limit (or can be estimated by min(x_i)). If a is known the ML-estimate of b is

$$\frac{n}{\Sigma\,(\ln(x_i - a) - \ln(c-a))}$$

(equal inclusion probabilities, for simplicity).

- Also a can easily be estimated numerically with the ML-method (or even be put to 0)

-

- The remaining outliers se will be described as independent drawings from Pareto(a*,b*,c*),.

  - In the example with pareto distribution they will be replaced by their expected value
    - E(Y|a*, b*, c*) = c* + (c*- a*)/(b*-1)

  - And the weights Σ (1/π_ι – 1). where the sum is over all observations larger than c.
  - With an approximate model variance around this value

    Var(Y|a*,b*,c*) =

$$(\frac{1}{n'} - \frac{1}{\pi n'})\frac{b*(c*-a*)^2}{(b*-2)(b*-1)^2}$$

8. Good advice on data collection of enterprises

Economic statistics

Autumn semester 2011

Stockholm University

The response process

The theory on how enterprises behave when they respond is not so well developed as for individuals

It is mostly a set of incoherent rules learned by experience

But for small businesses (farms, artisans, shopkeepers …) the same rules as for persons apply to a large extent

We will thus mostly discuss larger enterprises

- Enterprises are in some senses special.
- For many of them, time is money. They are used to get paid for their time
  - Avoid offering ridiculously low compensation (a lottery ticket for instance). It is better with no compensation at all. Personal compensation directly to employees who respond during their paid working hours is unethical/unsuitable.
- Data collection directly by phone is usually difficult.
  - The responder has seldom all data immediately at hand or is often busy with something else. (But efter editing, for additional information or clarification. Then the responding person is often named on the questionnaire with a telephone number)

Data collection

- Enterprises are more used to dull forms than persons. The questionnaires to companies should not be made too glossy. They must look serious. Funny creatures who point on the place where a figure should be entered are not considered to be serious and will usually lead to less response. (But may work for sole proprietors. And also lead to less editing afterwards)
- Statistics Sweden has a certain person for each of the 50 biggest enterprises Fixed contact persons for all studies for large companies.

Data collection

- Almost all data collection from larger companies are nowadays made electronically in some way or another.
- Most accounting software is designed so that it should be easy to give all the information Statistics Sweden and other government agencies ask for
- In many studies the questions are formulated so that the answer easily can be found in the ordinary accounting system. (But special new questions are not supported (like research issues) and the the responders often complain for that type of questions).
- Those who complain most are the medium large companies. (The big firms are used to official statistics and they have often one person employed whose task is to respond. The small companies are seldom selected and the amount of information asked for is limited.)

Data collection

- It is quite common with preprinted figures also on electronic questionnaires. E.g. Last year you told us that the sales area was 350 square meters. If the area is changed please enter the new figure here”. This makes the responding much simpler but has a conserving effect.
- It is difficult to motivate the firms. Many of them does not understand what the study is good for. (It is often compulsary to participate in official statistical studies. If you do not you can get fined).
- Feedback is a way to increase the interest. ”In this study last year we found that the average turnover per employee in your industry was 132 000 SEK/month”. If the study is longitudinal one may also say things like ”87 % of all firms sell les per employee than you”. This is most popular among small and medium sized firms. The large ones have there own department of analysis and business intelligence. They know it anyhow.

”The respondent burden”

- NUTEK (A former Swedish government agency) has estimated the total cost for all enterprises to fill and send froms and questionnaires to the government is about 100 billions SEK per year. (Mainly: self assessments, preliminary tax information, different permits like for building and environmental purposes)
- Statistics accounted for about 0,3 % of this. But statistics has come to be the symbol of unnecessary questionnaires.
- Out of these 0,3 % foreign trade accounts for (Intrastat) about three quarters. It is regulated by the EU like many other parts of the statistics.

9. A model for the response process

Economic statistics

Autumn semester 2011

Stockholm University

- There are cognitive models for the response process of individuals. These theories are used when constructing questionnaires and designing surveys (cf the course by Lars Lyberg).
- There is much less knowledge on enterprises
  - One reason is that the information are distributed over the firm and one person cannot answer all questions in one survey. You have to involve many people (Distributive cognition – deals with the way oganisations reason and work in particular with data handling).
  - Another reason is that the questions aften deal with a specific figure, which no person knows. You have to look in the computers or the books or to compute it using other items from the computers. Cf individuals who look in their memory.

A Cognitive Model

- Torangeau’s (1984) model for individuals’ response process:
  - Comprehension
    - Try ro interpret the text and to understand the question
  - Retrieval
    - Look in your memory and try to find an answer
  - Judgement
    - Judge if this is what is asked for and if you want to disclose it
  - Formatting
    - Formulate the response (e.g. give a sentence or choose the alternative that suits best)
- The process can go wrong in any of these stages. Design the questionnaire so that you have safeguarded against all four types af error. Go through them one at a time

A corresponding scheme for enterprises

- Encoding the information in memory or company records
- Selection of the respondent
- Assessment of priorities
- Comprehension of data request
- Retrieval of relevant information from memory or records
- Judgement
- Formatting, communication of the response
- Release of the data

Sudman, Willimack, Nicholls, Meusenbourg (2000)

Socially distributed cognition

- The knowledge is not in the brain of one person or in one place in the organisation
- Much of what was going on in the head in the model of Torangeaus occurs openly and can be observed
- Who asks whom, what files are opened, who approves the disclosure a.s.o
- Propagation of errors. Like the whispering game. What errors occur when the information is transferred from one medium or person to another
- Hutchins (1995) not with data collection but more general information handling within organisations. He was a psychologist but nowadays it is more popular within information technology.

Encoding the information in memory or company records

In what way can the statistical bureau influence how the data is saved and can be accessed in a firm?

Encoding the information in memory or company records

In what way can the statistical bureau influence how the data is saved and can be accessed in a firm?

- Ask about things that are saved. Chose definitions which are supported by acounting programs. Always ask yourself if the answer is possible to get
- Support the companies which develop accounting systems
- Ask about things that will happen in the future.

Selection of the respondent

How can you ensure that the questionnaire reaches those that can answer the questions? (in particular avoid supply personnel or trainees)

Selection of the respondent

How can you ensure that the questionnaire reaches those that can answer the questions? (in particular avoid supply personnel or trainees)

- Fixed contact persons, who knows the company and has responsability. They can forward the request to the correct person
- A post opening office is often a problem.
- Ask a person in charge high in the hierarchy. He will often delegate it to someone who knows or at least who will feel responsible since asked by the boss
- Adress the letter to a special department (e.g. the personell responsible, the purchasing manager, aso)
- Always ask who is responsible for these things and/or who filled in the questionnaire. (Good to know next time and also for editing)

Assessment of priorities

How do you make the organisations answer and answer with some care and within a reasonable time? How do you get the enterprise to prioritise the task

Assessment of priorities

Answers

- Compulsary, threat to fine them
- Motivate the respondent e.g.
  - In the cover letter
  - Get support from industry organisations and others who are trusted by the firms
  - Write/initiate positive but factual articles in the industry journals
  - Show that you yourself take the survey seriously. Sloppy or funny surveys will not instil confidence. React fast with editing and call-backs.
  - Today only a web survey does not instil confidence. Always give an alternative even if many eventually will choose the web as the best mode.
- Respondent/panel care
  - Feedback, Thank you card, Christmas card
  - In longitudinal surveys put much work to make people/firms cooperate from the beginning. If they understand the importance from the beginning the following waves often easy

Torangeau

- Comprehension of data request
- Retrieval of relevant information from memory or records
- Judgment
- Communication of the response

These points do not differ much from collection from individuals (e.g. test the questionnaires, use a question laboratory, pilot studies in the field, interviewer education aso) .

But sometimes:

Comprehension, Retrieval, Communication

How do you make enterprises pick out the correct information from the registers. And to communicate it to the statistical agency?

Comprehension, Retrieval, Communication

How do you make enterprises pick out the correct information from the registers. And to communicate it to the statistical agency?

- Help those that develop accounting programs, so that the correct information is easily accessible and even is simply found (by the respondent firm or by the statistical agency itself, ”Electronic Data Interchange”)
- Accept the information in different formats (e.g. most file types/computer languages, ASCII, Excel sheet, via Internet formulaires (HTML-coded), e-mail-attachments, paper, CDs, etc. Put the formatting trouble with the agency)

Release of the data

How do you get the enterprise to disclose the information; information that may be potential business secrets or could be harmful to the company in other ways if they were disclosed?

Release of the data

How do you get the enterprise to disclose ur the information; information that may be potential business secrets or could be harmful to the company in other ways if they were disclosed?

- Strict privacy trials for all statistics. It may sometimes be good to refuse to disclose (other) data. If this becomes known to other potential respondents they will be more positive to disclose their data
- Build a confidence in the statistical agency. Never make doubtful surveys or political opportunistic or market surveys. Avoid making errors. An article criticiseing the way a party preference study is done will affect respondents in all other studies.
- Create a sense of responsability with the responding firm. ”We will together work for a better Sweden”. The state needs money (taxes) and information (statistics) in that process. Everyone must participate with his share.

Not much of a model

- Much less theoretical underpinning than advice for data collection from individuals and households.
- Mostly a collection of good advice and recommendations
- Business/organisations differ much more than individuals. Questionnaire construction and contact strategies must be more varying.
- The best technique depends a lot on the subject of the survey.

Thank you for your attention