Ekonomisk statistik
Economic statistics

Master course

Daniel Thorburn

Autumn semester 2010

Stockholm University

1.5 Sampling and estimation – coordinated samples, rotating panels, etc

Daniel Thorburn

Ekonomic statistics

Autumn semester 2010

Coordinated samples

- Types of coordination
    - Positive coordination with large overlap between the samples
    - Negative coordination with small or no overlap between surveys
- Why coordinate samples?
    - Respondent ”burden” or contributions
        - Destribute it more evenly/fairly
        - Decrease it
    - Get information on relations
        - Over time, longitudinal studies
        - Between surveys. If the same enterprise participates in two different surveys it is possible to study the relation between these variables, between profitability and working environment, say.

Distribute the burden more fairly

- Large enterprises are often included in surveys and they believe it is OK. They have routines to handle statististical surveys, even special employees
- Small enterprises are seldome selected and does not complain so often. Sometimes they are even glad that someone is interested in their business.
- But inbetween are firms that are so small that they do not have routines, but so large that they may be selected too often by chance. They will be the most irritated responders.
  - Important to design the surveys so that they are not included in several big surveys at the same time.
  - Often put them in quarantine after a survey.
  - But change the sampling weights of the remaining enterprises in the same group accordingly.

1.5.1 Subsampling

- Sometimes samples are coordinated by studying a subsample
    - E.g. Every fifth enterprise are asked to respond to some extra questions. This is good if the questions in the second step require a smaller sample size. In the estimation phase you must use that this is a subsample. You may for example calibrate after some important base questions to improve the precision

Screening

- Screening. You want in particular to study enterprises with a special property
    - e.g. newly made investments for environmental reasons, have a female managing director, have disabled employees or have used special EU-money.
- In the main survey you can ask about if the the firm belongs to this group. Later you return with a more detailed survey on that issue.
    - Sometimes the last step can be a planned subsample in the selected group. E.g. For the second step choose equally many with male and female managing director – or even matched pairs. This 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).

One example: Rotating samples

Every firm is included in the survey a predetermined number of times (waves) e.g. four years and a fourth of all firms are replaced every year.

- The respondent burden is decreased in this way
  - Since it is always most work the first time you are in the study
  - Basic questions are asked only once. Or you ask only if there is a change since last time
  - But no firm will always be in the study. This would be considered unfair and there may be an effect on the behaviour of the firm if you are in such a study for a long time. Many questions on e.g. education of the staff education
- Smaller tracking and contacting costs after the first wave. (The interviewers may know whom to phone)
- You can study the development over the years in another way,
- E.g short period salary statistics, LFS (the labour force survey).

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.

1.6 Permanent random numbers

Daniel Thorburn

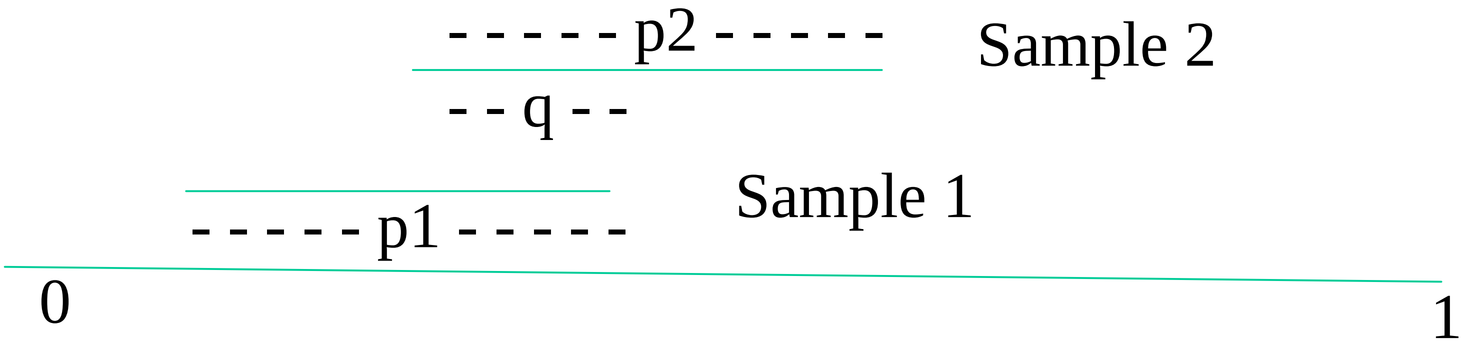
Economic statistics

Autumn semester 2010

Permanent random numbers
How to draw samples using them

All firms (units) in a frame register (e.g. the central enterprise register) is given a random number uniformly distributed in the interval (0,1) as soon as it enters the register, U_i. This number is kept as long as the firm remains in the register.

A simple way of drawing a sample is to take all firms with random numbers in an interval (a,b), say. This sample will then be a simple random sample with inclusion probability b-a and have (approximately) size N(b-a)



- - - - - p2 - - - - -

Sample 2

- - q - -

Sample 1

- - - - - p1 - - - - -

0

1

Coordinated samples using permanent random numbers

- Problem: Sometimes you want to draw two samples with inclusion probabilities p_1 and p_2. Sometimes you also want that a specified proportion should be in both samples (0 or a larger proportion q).
- Solution: Construct two intervals $(a, a+p_1)$ and $(a+p_1-q, a+p_1-q+p_2)$. Take all firms with numbers in the first interval for the first sample and all in the second interval for the second sample (If any boundary is larger than 1 do it modulo 1. E.g. The interval 0.9, 1.05 is replaced by the union of (0.9, 1) and (0, 0.05))
- This system is in Sweden called the SAMU system (Previously JALES after its inventors)

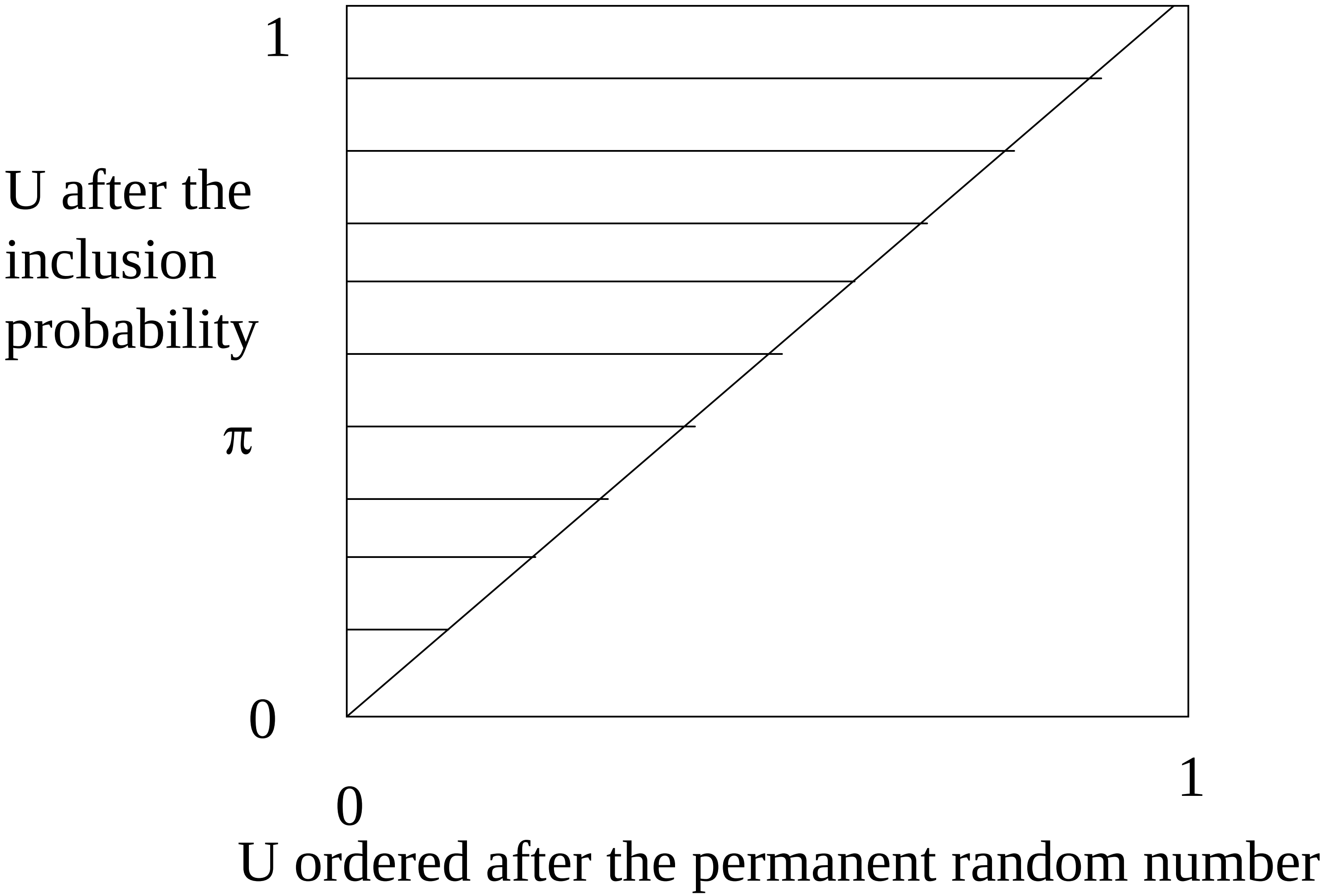
Permanent random numbers in rotating samples

- It is easy to move the interval a suitable distance to the left each year. Starting with (a,b) you can e.g. change the interval the distance (b-a)/4) every year.
  - This will give a rotating sample with k=4
- A classical problems with rotating samples is that the oldest panel is drawn k years ago and is a sample from a k years old frame.
  - If the frame has changed new firms will be underrepresented in the old panels and thus in the whole sample..
  - Less than one year old firms will only be selected in the last panel
- But we want the sample to be representative at each occasion. This is solved with permanent random numbers
  - If a new firm gets a random number saying that it should be in an existing panel then it should. The new enterprises will be in the sample for a shorter period than four years. Since the frame is updated you will at each time have a sample drawn from the actual frame.

Permanent Random Numbers

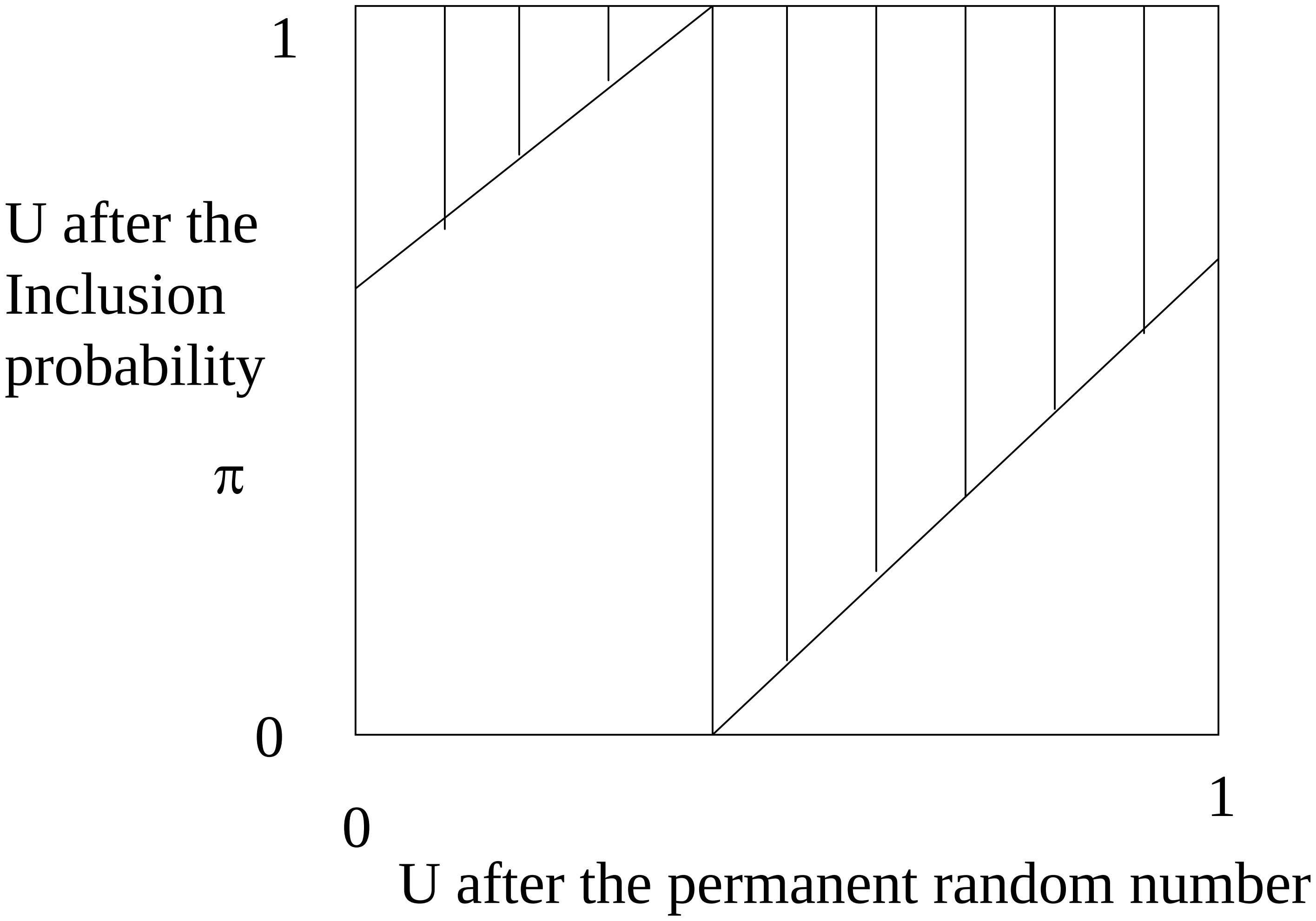
- This discussion concerned SRS. But what about πps-sampling?

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

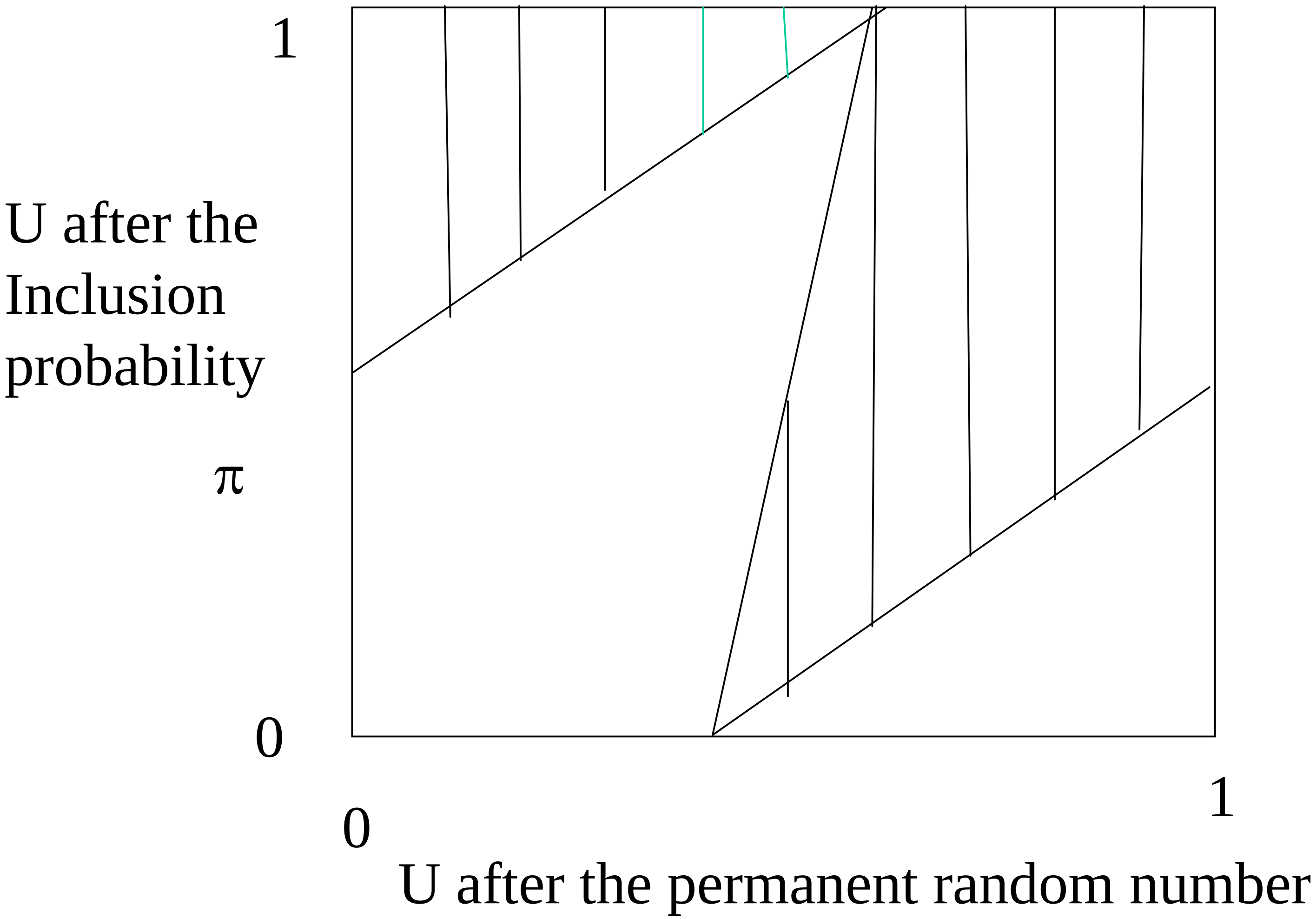
Permanent random number

Another sample



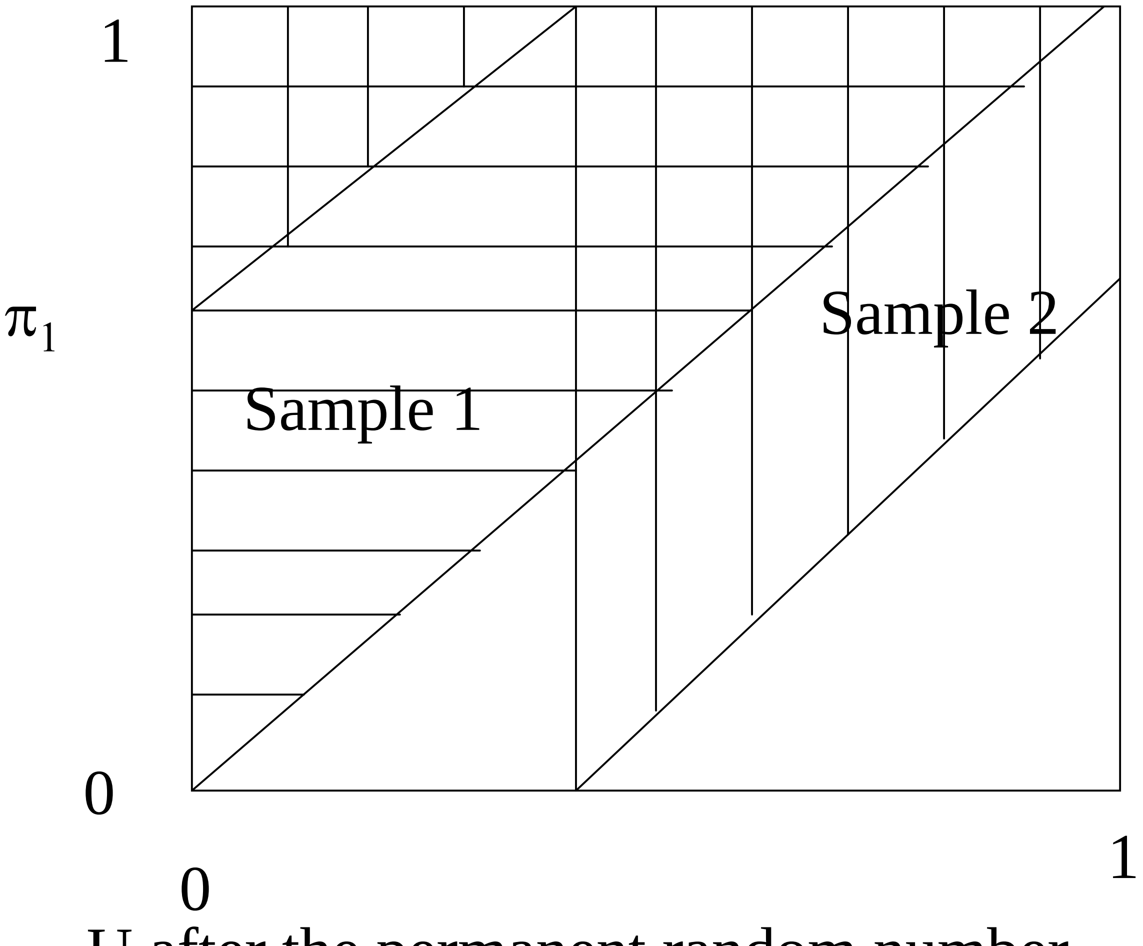
Permanent random number

And yet another



Permanent random numbers

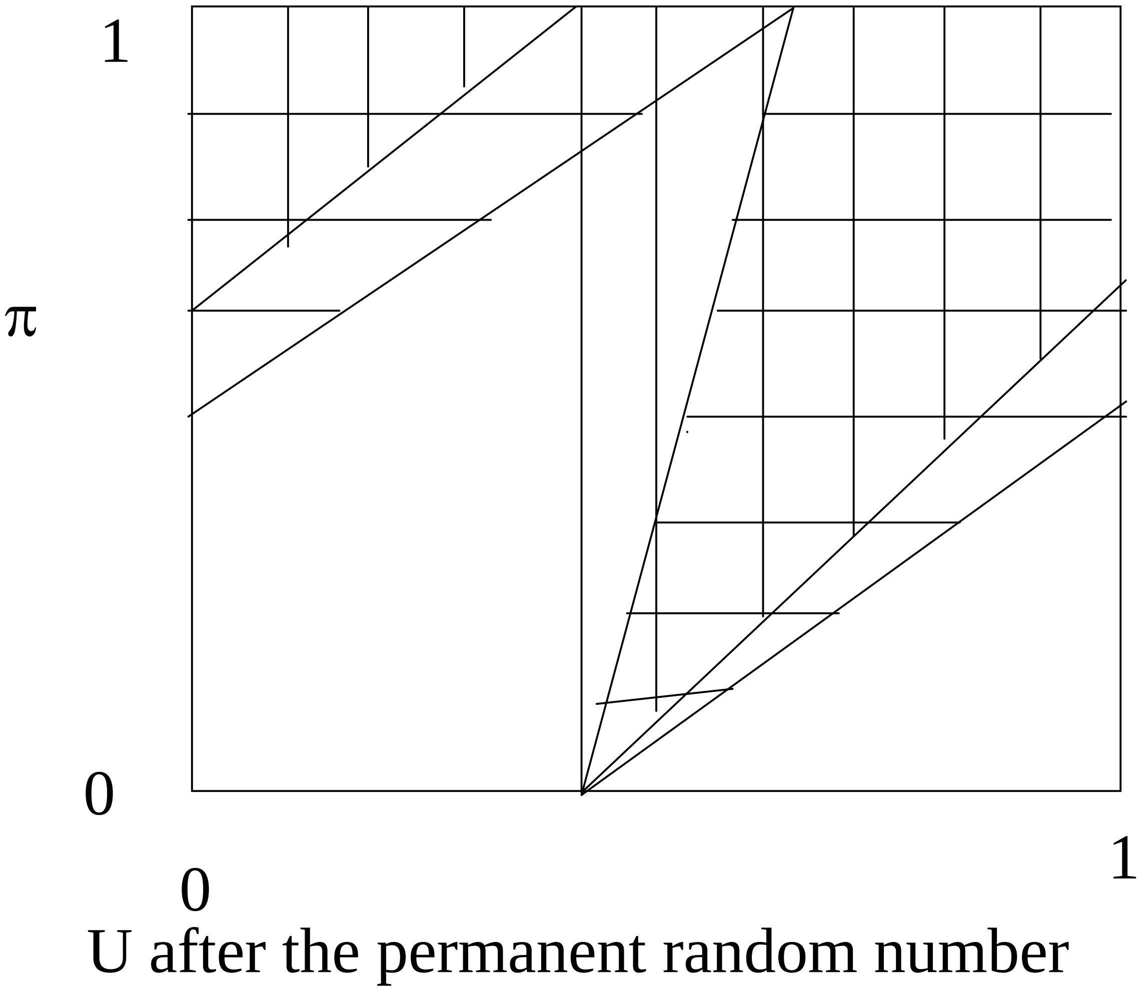
Two coordinated samples /with the same inclusion probabilities



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 it in a two-dimensional figure.
- It is simple to use Pareto πps when you want to draw with varying inclusion probabilities
  - Compute $\pi_i(1\text{-}U_i)/\ (U_i(1–\pi_i))$ and take the n largest (Roughly, those larger than one)
  - Next sample. Replace U_i with another number U_{2i} .Usually it is done by only changing the starting point with a fixed quantity $U_{2i}= U_i - b$, where b is chosen larger than most of the inclusion probablilities to get samples with no overlap
  - Sometimes you change signs instead $U_{2i}= 1\text{-}\ U_i$
  - You get rotating samples with more complicated expressions like $U_{2i}= U_i - \pi_{i1}$. ($U_i - \pi_{i1}$ +1 if $U_i - \pi_{i1}$ is negative) and use the formula $\pi_{i2}(1\text{-}U_{2i}\ )/\ (U_{2i}\ (1–\pi_{i2}))$. (Also $U_{2i}= U_i - \pi_{i1}$ is a uniformly distributed random number). This gives no overlap (if $\pi_{i1}+\pi_{i2} < 1$). With $U_{ki}= U_i – \pi_{i1}/4$, $U_i – 2\pi_{i1}/4$, $U_i – 3\pi_{i1}/4$, … you get a four year rotating panels.

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 estimate always will be the same figure
  - But these methods usually give slightly larger random errors

1.7 Sampling – outliers

Economic statistics

Autumn semester 2010

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 so simple. You are interested in the total turnover of an industry not the median times the number of firms. You cannot just trim Volvo or Microsoft away 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 miljons Swedes are employed in the paper industry which has invested 200 billions 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. 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 (95%, 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. 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 it 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 (ther might be in the population) but downwards when ther 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 kan easily be estimated numerically with the ML-metoden (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*) =

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

2. Data collection
2.1 Good advice on data collection

Economic statistics

Autumn semester 2010

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 business (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 meanings special.
- To 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. Compensation directly to employees who respond for the firm during their paid working hours is unethical/ unsuitable. If you give something give it for free.
- 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 given 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) better response.
- Statistics Sweden has a certain person for each of the 50 biggestfenterprises 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 new environmental 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 squaare meter. 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 it 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 foound that the average turnover per emplyee in your industry was 132 000 SEK/month. If the study is longitudinal one may also say things 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 learn 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 for environmental purposes)
- Statistics accoundts 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.

2.2 A model for the response process in enterprises and organisations

Economic statistics

Autumn semester 2010

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. Individuals look in their memory.

A Cognitive Model

- Torangeau’s (1984) model for individuals’ response process:
  - Comprehension
    - Try ro interpret the text and understand the question
  - Retrieval
    - Look in your memory and try to find an answer
  - Judgment
    - 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
- Judgment
- 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 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

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

- 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.
  - In longitudinal surveys put much work to make people/firms cooperate from the beginning. If they understand the importance from the beginning the following times are 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 do doubtful surveys or political opportunistic or market surveys. Avoid making errors.
- 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