Ekonomisk statistik
Economic statistics

Master course

Daniel Thorburn

Autumn semester 2012

Stockholm University

Exam

- October 31st, 2012, 9.00-14.00. Värtasalen. Examination time is 5 hours. Registration is compulsary.
-
- Approved aids: Pen, pencil, calculator, dictionary. Cell phones must be shut off.
-
- Each question gives at 20 credits if correctly answered. Hence a maximum of 100 credits is possible. Answers may be given in Swedish or English. Answers to the five problems must be given on separate sheets.
-
- All notations must be carefully defined or explained (even those notations that are introduced during the course). Proofs, explanations and arguments must be so clear and detailed that they are easy to understand. Incomprehensible and badly motivated solutions will give zero points or a deduction. Write clean and properly. No text in concept.

- Date of handing back – not known yet – you will be notified by e-mail

Sampling part 3

Daniel Thorburn

Economic statistics

Autumn 2012

Contents

1. (Sampling – repetition,
2. Frames - Business registers
3. Sampling and estimation of businesses, πps
4. Coordinated samples
5. SAMU – Permanent random numbers
6. Permanent random numbers, πps-sampling?
7. Data collection of enterprises, Good advise)

8. Data collection, A cognitive model,

9. Sampling and estimation with outliers,

9. A model for the response process

Economic statistics

Autumn semester 2012

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 always 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 often 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.

Time series part 2

Daniel Thorburn

Economic statistics

Autumn 2012

Linkage

Economic statistics

Autumn 2012

Stockholm University

When changes are made in the statistics collection there will often be jumps. For example, when a classification like NACE or ISCO is modernised or a questionnaire is changed or when a cut-off lumit is changed. If the change is noticable you have to recalculate the series. Un linkage you try to reconstruct what the old statistics would have looked like if the new system had been used also before the change. People who wark with the statistics often wants to be able to campare to earlier time points (In particular econometicians often want long time series for their analyses.

Same problem when their are changes in the rules in the society. Changes in the rules of sick leaves e.g. the number of qualifying days means that the number of days of paid sickleaves change or the rules of inventory evaluation pr depreciation will change the profits and thus the National accounts. These types of changes are usually handled by econometricians whaly linkage is handled by the producer of statistics.

At changes eg in the classification system, statisticians should plan for linkage. At reforms in society politicians should plan so that reform can be evaluated.

Linkage ?

Imagine that you have a series that looks like this (Percentage of positive government investment guarantees%.)

Date	1	2	3	4	5	6	7	8	9	10	11	12	13
	23	24	27	27	26	28	30	42	43	45	44	43	47

In the year 8 the question was reworded. If you want to analyze the performance, you want a series, without a jump.

What should you do?

- Answer 1 o 2: Easiest The jump between 7 and 8 is 12 (or 40 %). Correct all values before

Year	1	2	3	4	5	6	7	8	9	10	11	12	13
Original	23	24	27	27	26	28	30	42	43	45	44	43	47
add	35	36	39	39	38	40	42	42	43	45	44	43	47
Mult	32.2	33,6	37.8	37.8	36.4	39.3	42	42	43	45	44	43	47

But it does not say anything about the change between the years 7 and 8

- Answer 3: Estimate the jump by a linear regression $y_t = a + b*I(t<8)+c*t$ and adjust with b^ (here additive and rounded)

32 33 36 36 35 37 39 42 43 45 44 43 47

(other models might have been better. e.g. multiplicative)

Problem: You want to be able to adjust already year 8 when the new statistics is published for the first time

- Answer 4: Use this principle for a shorter period, say the years 3-8 to estimate the jump

33 34 37 37 36 38 40 42 43 45 44 43 47

Gives higher uncertainty (variance) but less bias than answer 3

- Answer 5: Even more complex models can be used, for example, an ARIMA process or the Kalman filter. eg
  - $X_t = a*X_{t-1} + \varepsilon_t$ X is a latent class
  - $Y_t = X_t + \eta_t$ before break;
  - $Y_t = X_t + b + \eta_t$ after break;
  - Estimate the parameters and predicted the value of Y+b before the break.

  But in general try to choose as simple models as possible

- This was mechanical linkage methods for a single series in which it was assumed that it had no more information.

- A common recommendation is that if it is possible to do the survey in both ways in parallel at one time or make any other methodological studies. For example, when SILC (Survey on Living Conditions) changed from PAPI to CATI. Half the sample was inreviesed by visiting interviewers and the other half by telephone during the transition years. Random halves.

- When major changes are made, for example, a new NACE code in the national accounts one tries to encode the data in both ways during a period or to look at old statistical data (forms) and recode companies according to the new template.

- These methods provide a good estimates of the jump - but not how to change further back.

- When reclassification, such as rescheduling of NACE, you may know the sum of the two groups, but there will be a new breakdown.
  - For example, when IT consultants were broken out from the category other consultancy.
  - You may do a good estimate at the time of the change by parallell coding, say that you get 30% IT.
  - But it is not reasonable to assume that this percentage held for every year earkier.
  - Then you add a subjective guess. For example, that the proportion was 0 % ten years earlier, and then a linear increase in proportion to the first observed level. (i.e. 0, 3, 6, 8, …, 27, 30 %)

- If you have linked a data series, the user should always be able to read about what happened. The reason for the leap, and how it is eliminated.
- Akways document what you do

Final remarks

Daniel Thorburn

Economic statistcs

Autumn 2012

Quality is important

- SCB is working on becoming ISO certificied

- Much of what you have learned earlier, concerns the quality of statistical surveys (sampling and total).
-
- Quality of registers is something else. E.g. How should one measure and report the uncertainty in the number of restaurants in a register?
- Dark numbers (mörkertal). Incidences or units who ought to be in the register but are not.

Relevance, validity, reliability

- Notions that you should know the definitions of and the difference between

- Many aspects of quality - some of which are more important in the economic statistics
  - Variable definition +,
  - Questionnaire design
  - Editing +,
  - Imputation +, The main way of handling non respoinse
  - Data collection
  - The result - the statistics should be understandable and easy to access (if it is produced by public organisations).

Relevance, validity, reliability

- Relevance - The information must be relevant for the intended use

- Validity - Small Errors – the statistics should measure what we think it measures. Roughly a small bias.

- Reliability - If the measurement is repeated one should get the same results. Eg it should not depend on the interviewed persons managers mood or the weather, or the interviewers' gender. Small random errors.

Quality in registers

- Quality of registers has a significant impact. In principle, it has been a census, but how good is the information in the registry. In many cases, control studies are justified. How they should be implemented is a problem in itself.

- The requirements for quality of administrative records are other than for statistical production. In an administrative records, as many as possible must be accurate. In statistical register, the proportion of a particular property should be correct.

- Previously the statistical registers were used once. Nowadays, most of SCB studies and registers are used many times. It should e.g. be easy to merge two registers. There should be good identity keys.

- Previously students were taught descriptive statistics for the convenience of readers to interpret the results. Today it is easy to use registers and students should learn how to construct and manage statistical registers like the register of enterprises

Editing

- 40 % of all costs for data collection in the business statistics is for editing
- Micro editing – The computer check individual items. Every unexpected value is checked. E.g. the company employees but no cost for wages or an unexpectedly high change since last year.
- To decreas costs only a statistical sample may be contacted for smaller deviationsa. There will be more incorrect figures in the register but the statistics may be not be influenced very much.
- Macro editing …

Macro editing

- Edit the final statistics. (You may find computational errors, unreasonable effects of outliers or systematic errors in that way.
- One problem is what to do if you find something weird- You know that there is something wrong but not what.
  - Felet i KPI upptäckte man efter en månad (Dock ej i formell granskning. Makrogranskning fanns inte på delindex), men när man gick igenom data fann man inga fel. Det tog tre månader innan felet hittades
  - För några år sedan inträffade något liknande. Volvo hade lagt om datorsystemet så att fabriken i Gents siffror inkluderades i Volvo Sveriges produktion. Vid kontrollringning verifierade Volvo att den redovisade siffran var korrekt. Och Sveriges bilproduktion redovisades med kanske 30% för höga värden.

Documentation

- Always dokumentera what you have done. Eg all changes
- Five years ago Statistics Sweden mad a large error in the estimation of CPI. But when they realised that the figures were suspicious thay could not find the reason due to missing documentation of changes in the computer programs. So they decided to publish the incorrect figures. (Which resulted in higher pensions and that the head of Stat Sweden had to report their quality work every month to the ministry)
- Definitions, questionnaires aso ought always to be available

- Vid reformer i samhället bör politikerna planera så att reformen kan utvärderas. Att ge underlag för det är statistikerns uppgift.

- Vid omläggningar bör statistikern själv planera för länkning.

- Samanvändbarhet är viktigt. Samma definitioner i olika undersökningar. Undersökningar görs på samma sätt i olika länder. Jämförbarhet mellan år, grupper, regioner etc.

Timeliness

- Aktuality The data should be recent when the statistics are published. It is important to work fast.
- Punctuality is är extremeluy important. Promised publications dates should be keoteringstider skall hållas. In that case the users e.g. the ministries can schedule their meetings to the day after to be able to react fast. If the date is not known it may be difficult to find a date when all the involved persons are free. The decisions may become postponed for a month due to a one day delay in publication-

7. Sampling – outliers

Economic statistics

Autumn semester 2012

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/p) 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.

- $(Y_i|Y_i > c) \in$ 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 $\Sigma\ (1/\pi_t - 1)$. where the sum is over all observations larger than c.
    - With an approximate model variance around this value

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

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

Thank you for your attention