UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007 - 2013

Counterfactual Impact Assessment

Day 1

Klara Major

PROJECT COFINANCED BY EFRD THROUGH TAOP 2007-2013

Introduction

- Counterfactual impact analysis vs. theory-based impact analysis
- Econometrics
- The programme of the training

.

Impact evaluation

- Quantifying and explaining the effects of interventions
  - For policy makers to make informed decision
- Questions
  - does it make a difference? ⟶ quantifying the impact
  - why it works? ⟶ explaining the impact
- Counterfactual impact analysis (CIA)
  - Is the change due to the policy or would it have occurred anyway?
  - need to find a credible approximation to what would have occurred in *the absence* of the intervention
  - compare it with what actually happened.
  - The difference is the estimated effect, or impact, of the intervention, on the particular outcome of interest

Impact evaluation

- Theory-based impact evaluation (TBIE)
  - CIA produces answers that are typically *numbers*
  - Is the difference *observed* in the outcome after the implementation of the intervention*caused* by the intervention itself, or by something else?
  - great deal of other information
  - why a set of interventions produces effects, intended as well as unintended, for whom and in which context
  - this approach does not produce a number, it produces a narrative

- This training gives introduction to Counterfactual impact analysis
- However CIA uses econometric techniques to quantify the impacts…

- Application of statistical methods in the analysis of economic data
- Four steps of econometric analysis
  - Economic model → empirically testable model
  - Data collection (cleaning and transformation of data)
  - Estimation of the model and its verification on observed data
  - Forecasting, decision support

Research questions in econometrics

- Searching for causal relationships
  - More than the statistics discipline
    - correlation ≠ causal relationship
    - There may be a common unobservable cause in the background
  - Thought experiment
  - Natural experiment vs. Non-experimental situation
- Forecasting.
  - A causal relationship is not necessarily searched
- Impact analysis, decision support.

- Cross-sectional data
  - Aggregate data, e.g. data for different countries in a given year
  - Individual level data (microdata), e.g.
    - Labour force survey of Hungarian Statistical Office (economic activity, employment)
    - Wage survey of National Employment Office (wages)
    - Household budget survey of the Hungarian Statistical Office (consumption, income)
    - Administrative databases (NAV, ONYF, OEP)
- Time series (e.g. evolution of macrodata in time)
- Panel: cross section and time series as well
  - (country panel, panel from microdata)

Example 1

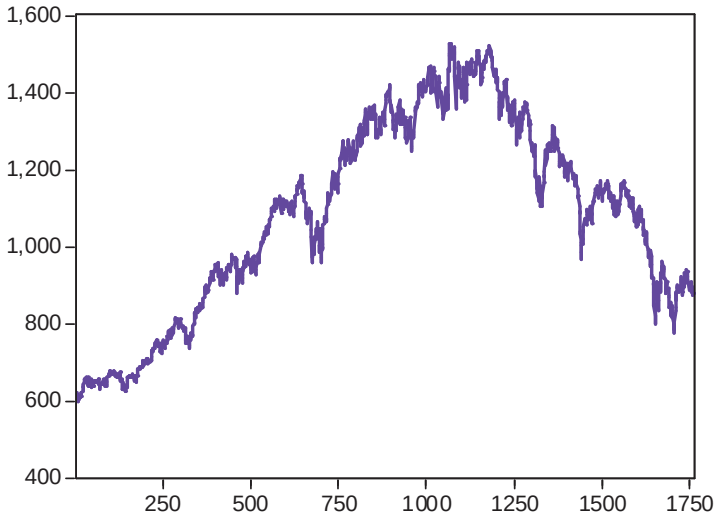
- minimum wage and unemployment
- Standard model: employment is reduced by increasing the minimum wage
- but: Card-Krueger (1994)
- Time series: the trend of employment growth was reduced after the minwage increase
- Causal relationship?
- Meanwhile:
  - Decreasing external demand,
  - Stronger real exchange rate
- individual / company level wage / employment data are needed
  - More data, personal characteristics, easier identification

Example 2

- do more policemen reduce crime?
- Thought experiment: two identical towns, more policemen in one of them. Is the crime rate smaller there?
- Problem: endogeneity (or simultaneity)
  - A simple regression is not enough, there may be a reverse relationship: more policemen if the crime rate is higher
- Possible exogenous shock (natural experiment)
  - More policemen in election years → is the crime rate lower?
- but: common sense and a knowledge of institutional details is always important!
- Other example for endogeneity: effect of education on wages

Example 3

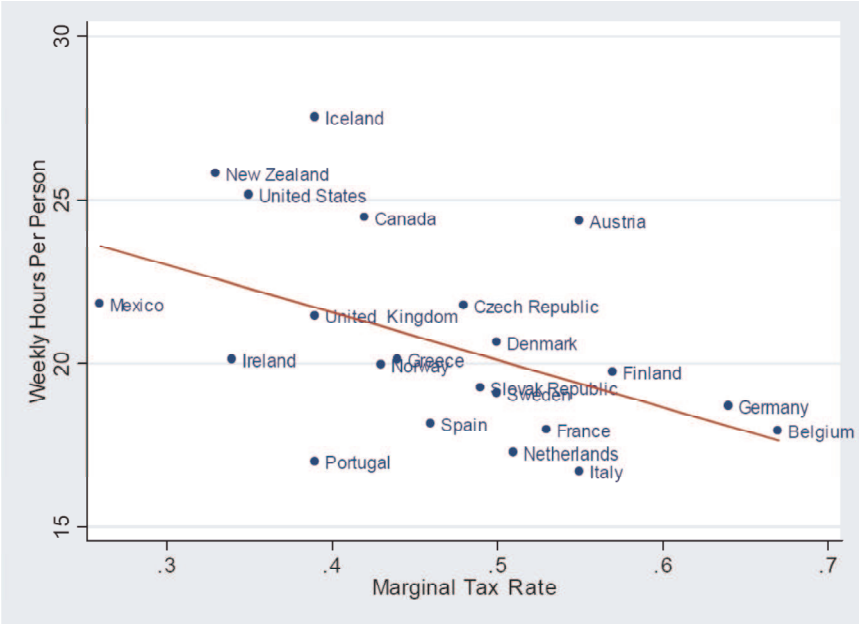
- forecasting stock prices
- Can tomorrow's stock price be predicted by today's one?
  - Approximately a random walk
- Can the forecast be improved by using other information?
- Is the volatility predictable?
- Not necessarily a structural model, a purely statistical description is useful
  - But be careful with them during a crisis!



Example 4

- working hours and marginal tax rates
- Marginal tax rate
  - By how much is the tax increasing when the gross income is increasing by 1 unit?
  - Quite high in Hungary
- Questions
  - Does it reduce labour supply?
  - To what extent does the marginal tax gap between the EU and USA explain the working hours gap between the two regions?

- Cross sectional sample of countries
  - but: causal relationship?
- Relevance for Hungary
  - Elasticity of taxable income
  - Incentive effects of sick-pay



Relationship with other disciplines

- Probability and statistics
  - Multidimensional probability theory.
  - Theory of estimation, hypothesis testing
  - Gaussian, t- and F-distribution
- Economics
  - interpretations of the results
- Revision of the necessary statistical results
  - throught examples
  - Focus is put on the intuitive statements

The aim of the training

- Counterfactual impact analysis is an econometrics technique
- Econometrics is used in these analysis
- It requires
  - data collection,
  - data handling,
  - econometrics software
  - the choice of the model/methods applied in the analysis
- In order to have useful results of the analysis we need
  - correctly chose data, method
  - understand the limitations and boundaries these methods have
  - put on clear questions *before* the analysis
- Conclusion: preparation (ie. TOR) matters!

The aim of the training

- GOAL: be able to prepare a TOR such that it maximizes the chance of getting useful results from the analysis
- To reach the goal, TOR needs to
  - specifie exactly the questions
  - define clearly the database of the analysis
  - define the appropriate methods that are relevant in the problem
- Evaluation commisioners needs to know
  - what are the methods that can be applied
  - what are the data requirements and limitations of these methods
  - what are the possible outcomes of that method
  - how to interpret the results of such a method
  - how to put on questions to be able to address by these methods
- This training aims to get these skills and the necessary knowledge

The programme of the training

- How to handle econometric softwares?
- How to handle data?
- What are the basic concepts in econometrics (simple regression analysis)?
  - sample, how to choose a sample
  - „the results”: regression coefficients
  - „the goodness of these results”: the test-statistics
- What are those features that make counterfactual setting so special?
- What are those special methods that are relevant for the counterfactual settings?
- How to prepare a TOR for counterfactual analysis?

Exercise

1. Form groups of 2-3 persons. Set up a list of 3 government development project in which you had experience. Collect information on each of the programme from the following aspects:
   - aim of the policy action,
   - target group of the policy action,
   - tools of the policy action (subsidy, guarantee, interest subsidy, etc.)
2. Prepare a TOR (or an outline of a TOR) to tender the evaluation one of the previously listed policy action. Try to be as specific, as you can.

Techniques of data handling

- Software needs
- EXCEL
- Gretl

.

Data processing

- Quantitave analysis requires data
- Size of the data set determines its software needs
- Small size
  - cc. hunderds of observations, a few variables
  - EXCEL for illustrating and estimating purposes (simple methods)
- Medium size
  - cc. thousands of observations, cc. max 100 variables
  - EXCEL for illustrating, data collecting, cleaning purposes, probably pilot estimation
  - Gretl (Eviews, Stata) for estimation
- Large sample size
  - millions of data: like microcensus, firm level dataset
  - Database management systems (like Oracle) for collecting and cleaning, ordering the data

- caluclations with data: built-in functions
  - mathematical functions
  - statistical functions
  - Analysis Toolpak
- illustrating data
  - XYplot (scatter diagram)
- exporting/importing data
  - supported data types

- Import data from excel:
  - File/Open data/Import/Excel…
- Graphs:
  - View/Graphs specified vars…
- Descriptive statistics for each variable:
  - View/Summary Statistics
- Estimation:
  - Model (later)
- Export data
  - File/Export data…

Working with data

- Data types
- Graphical methods
- Descriptive statistics

.

- Time series:
  - Variables ordered in time
  - Freqency of observations (e.g. monthly, yearly)
  - Notation: Y_t
  - Examples (macroeconomic, financial – individual?)
- Cross sectional:
  - Sample of economic agents at a given time point
  - Examples (individuals, enterprises, countries)
  - Notation: Y_i
  - Random sample

- Panel:
  - Time series + cross sectional jointly
  - Observation of the cross sectional sample throughout more time periods
  - Notation: Y_{it}
  - Examples (GDP of European countries, panel of individual households)

Data types, cont.

- Quantitative and qualitative
  - Quantitative: e.g. inflation, income
  - Qualitative: e.g. male/female, education level – code as numbers
- Level of measurement
  - nominal
  - ordinal
  - interval
  - ratio
- Level and dynamics

  E.g. number of employed

  vs. change in employment

$$\%\text{change} = \frac{(Y_{t+1} - Y_t)}{Y_t} \cdot 100$$

Exercise

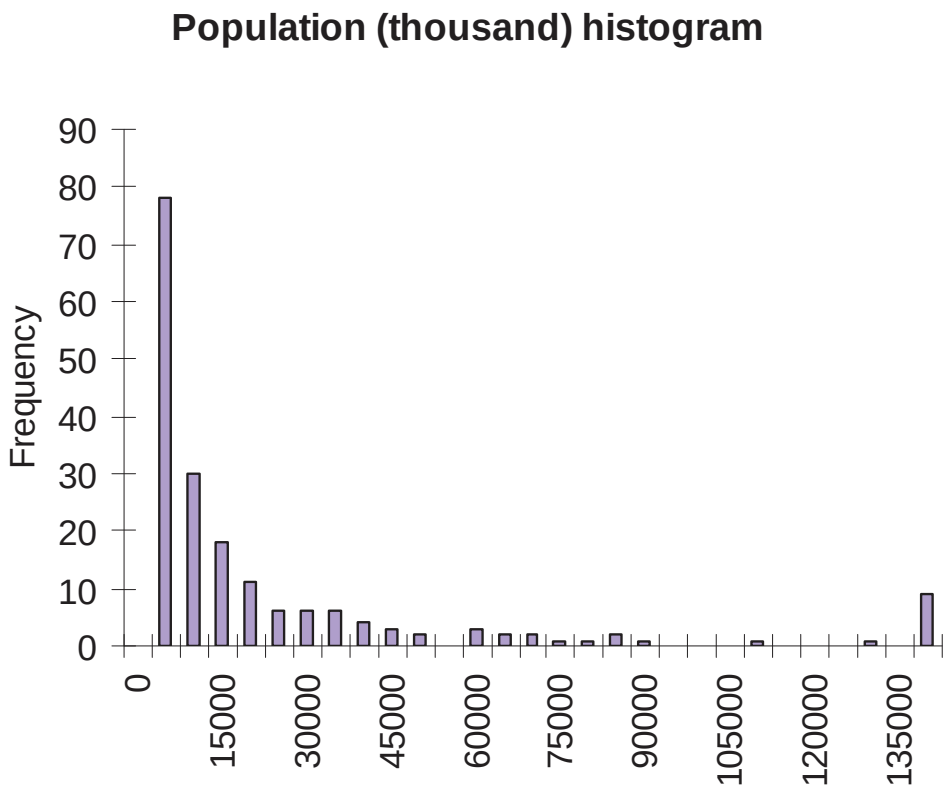
- Create a time series graph.
- File INCOME.XLS contains data on the natural logarithm of personal income and consumption in the US from 1954Q1 to 1994Q2. Make one time series graph that contains both of these variables. (Note that 1954Q1 means the first quarter (i.e. January, February and March) of 1954.)
- Transform the logged personal income data to growth rates. Note that the percentage change in personal income between period t - 1 and t is approximately $100 \times [\ln(Y_t) - \ln(Y_{t-1})]$ and the data provided in INCOME.XLS is already logged. Make a time series graph of the series you have created.

Histograms

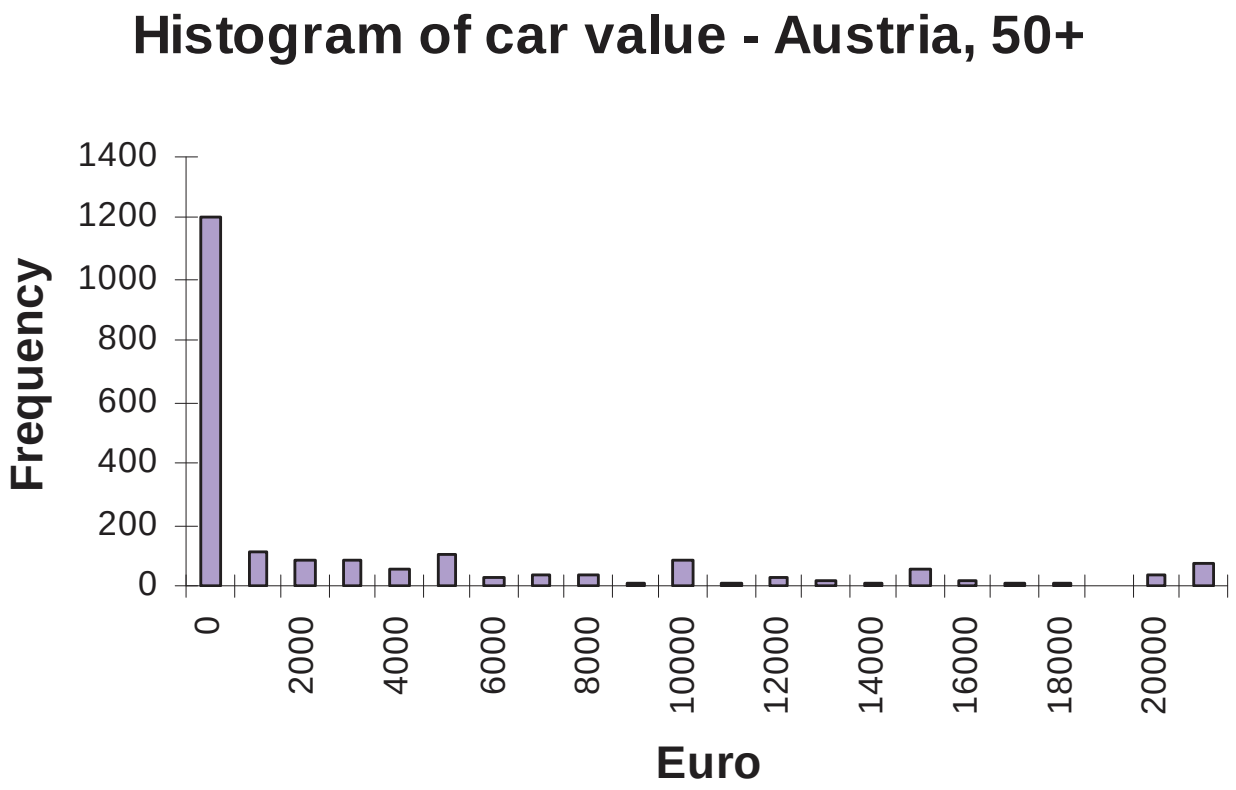
- Plotting cross sectional data
- Example: distribution of income per capita
- Equal intervals (brackets) – determine it in Excel according to the data
- Frequency within the intervals

- Excel: Analysis ToolPak extension
- Equal intervals (brackets) – determine it in Excel according to the
- Frequency within the brackets

- Penn World: distribution of countries according to population (bracket size: 5000)



- SHARE: cross sectional sample of people aged 50+
- Value of cars, Austrian subsample (bracket size: 1000)

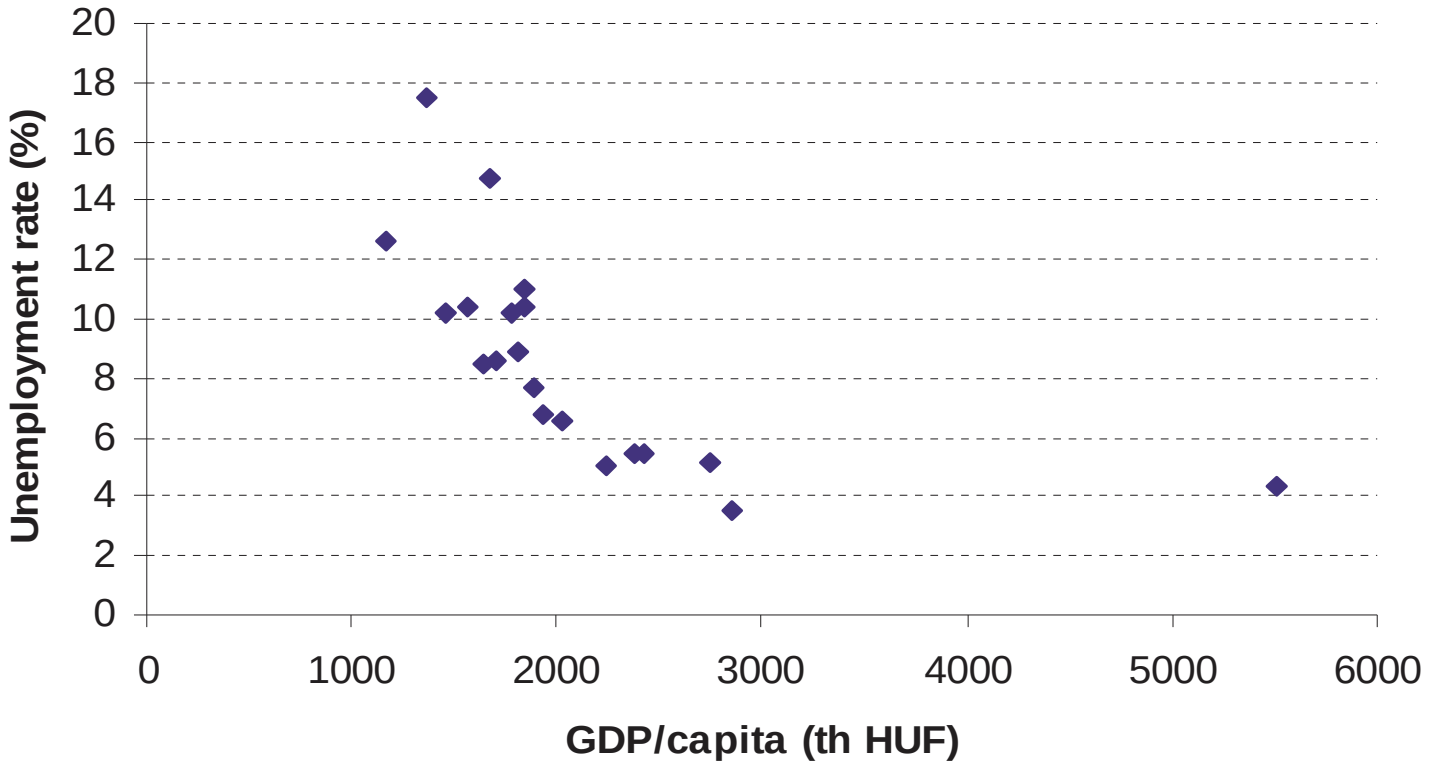


Exercise

- Create a histogram.
- Excel file GDPPC.XLS contains cross-sectional data on real GDP per capita in 1992 for 90 countries from the PWT. Create histograms using different class intervals. For instance, begin by letting your software package choose default values and see what you get, then try values of your own.

XY-plot (scatter diagram)

- Relationship between two variables
- KSH (Hungarian Central Statistical Office): data on counties



- Create graph on deforestation (i.e. the average annual forest loss over the period 1981–90 expressed as a percentage of total forested area) for 70 tropical countries, along with data on population density (i.e. number of people per thousand hectares). (This data is available in Excel file FOREST.XLS.)
- This file contains data on both the percentage increase in cropland (the column labeled “Crop ch”) from 1980 to 1990 and on the percentage increase in permanent pasture (the column labeled “Pasture ch”) over the same period. Construct and interpret XY-plots of these two variables (one at a time) against deforestation.
- Does there seem to be a positive relationship between deforestation and expansion of pasture land? How about between deforestation and the expansion of cropland?

Descriptive statistics

- Up to now: graphical methods
- Descriptive statistics: numerical summary of some characteristics of the variables
  - Level? – mean, median, mode
  - Variability? – standard deviation, range

- *N*: number of observations
- Example: mean of country population (Penn World Table) – ca. 34 million

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

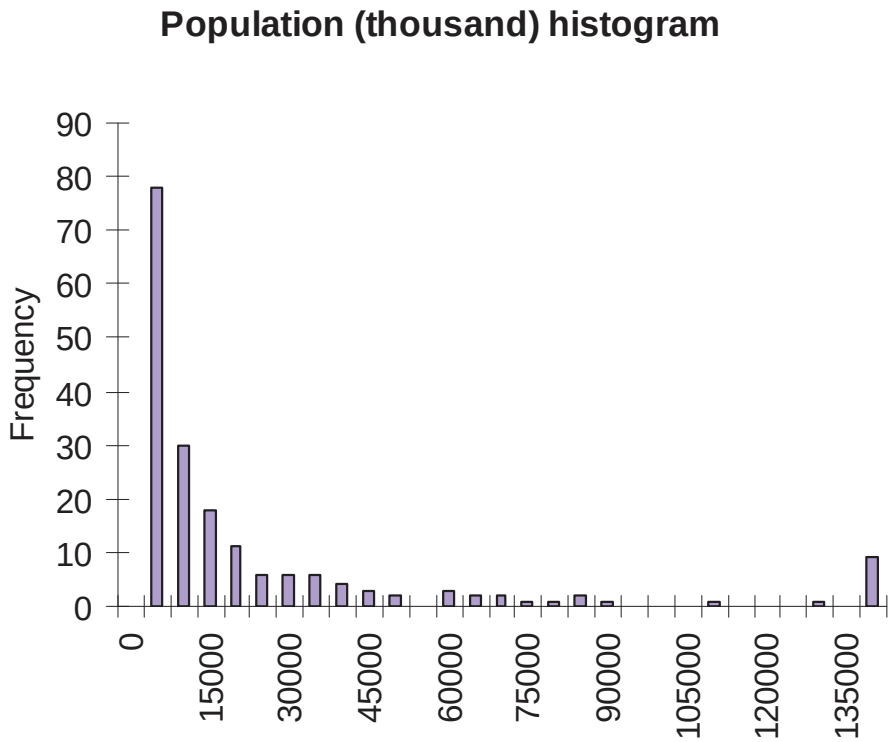
- Mode: most frequent observation
  - Problem: does not always exist (e.g. one from each value), there can be more modes
  - Possible solution: highest point of the histogram (depends on brackets) – center of the interval

Median, percentile

- Median: value in the middle – half of the observations below the median, other half is above
- Xth percentile: X% of the observations below X
- Quartile: cuts the observations into four
  - 1st quartile: 25% below, 2nd quartile = median

Skewness

- Example: mean > median
- Some large values – mean is large
- Skewed to the left
- Long right tail



- Range: difference between maximum and minimum
  - Not reliable (outlier values)
- Variance: mean of squared differences from the mean
- Standard deviation:

$$s = \sqrt{Var} = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N-1}}$$

- Construct and interpret descriptive statistics for the pasture change and cropland change variables in FOREST.XLS.

- caluclations with data: built-in functions
  - mathematical functions
  - statistical functions
  - Analysis Toolpak
- illustrating data
  - XYplot (scatter diagram)
- exporting/importing data
  - supported data types

Data Sampling

- Introduction
- Simple Random Sampling
- Stratified Sampling

.

- goal: estimate of the characteristics of the whole population
- two approaches
  - census of the whole population
  - sampling
- Statistical tools help us to draw conclusions on the characteristics of the whole population *based on our observations of the sample*
- area of application
  - statistical survey
  - quality control
- sampling is important, see next example

Number of passengers at Ferihegy Airport in 1994 by airlines		
No	Airline	No of passengers (1000s)
1	AEROFLOT	42,5
2	Air France	31,6
3	ALITALIA	34,9
4	AUA	32,0
5	British Airways	72,6
6	Delta Airlines	48,8
7	ELAL	21,3
8	KLM	57,4
9	Lufthansa	110,5
10	SABENA	17,4
	Mean	46,9
	Standard deviation	26,4

Example 1

- Population: foreign airlines fly to Budapest
- Observed variable: number of passengers in a year
- How can we make inference on the average number of passengers if we only have a chance to observe a sample from the above population?
- Select samples of 2,3 or 5 elements from the above population. For simplicity assume that the subsets of population contain consecutive airlines.
- The result are summarized in the next table.

Sample means and standard deviation

Mean and standard deviation of sample means							
Sample	Mean		Sample	Mean		Sample	Mean
1,2	37,05		1,2,3	36,33		1,2,3,4,5	42,72
2,3	33,25		2,3,4	32,83		2,3,4,5,6	43,98
3,4	33,45		3,4,5	46,50		3,4,5,6,7	41,92
4,5	52,30		4,5,6	51,13		4,5,6,7,8	46,42
5,6	60,70		5,6,7	47,57		5,6,7,8,9	62,12
6,7	35,05		6,7,8	42,50		6,7,8,9,10	51,08
7,8	39,35		7,8,9	63,07			
8,9	83,95		8,9,10	61,77			
9,10	63,95						
Mean	48,78		Mean	47,47		Mean	48,04
Std.dev.	16,74		Std.dev.	10,20		Std.dev.	6,98

Conclusions

1. Sample means change sample to sample
2. Sample means varies around the population's mean
3. Standard deviation of sample means is a function of the sample size
   - the higher the sample size the smaller the standard deviaton of the sample mean
   - the smaller the sample size the higher the standard deviation of the sample mean

The sampling procedure

- Defining the population of concern
- Specifying a sampling list, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting

Population definiton

- defining the population from which our sample is drawn
- sometimes it is obvious
- other cases
  - sometimes we need to sample over time, space, or some combination of these dimensions
    - periods or different occasion
  - population' may be even less tangible
    - taking repeated measurements of some physical characteristic such as the electrical conductivity of copper
  - the population from which the sample is drawn may not be the same as the population about which we actually want information

Sampling approaches

- probability sampling scheme
  - every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.
  - Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling.
  - common characteristics of these techniques:
    1. Every element has a known nonzero probability of being sampled and
    2. involves random selection at some point.
- nonprobability sampling
  - some elements of the population have *no* chance of selection
  - or the probability of selection can't be accurately determined.

Simple random sampling

- all items are chosen to the sample with equal probability
- moreover, any given subset of the population (list) has equal probabilty to chose than any other – equally sized – subset
- variance between individualy within the sample is a good indicator of the variance in the overall population
- therefore it is relatively easy to estimate the accuracy of the estimated population mean
- sampling error might occur if the composition of the sample differs from that of the population

Systematic sampling

- arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list
  - random start, then choose every k-th element
- drawbacks of systematic samplings:
  - vulnerable to periodicities in the list
  - its theoretical properties are very cumbersome, hard to quantify the accuracy

Stratified sampling

- elements in the list can be organized by categories, into „strata”
- each stratum is considered as indipendent subpopulation, out of the subsample has randomly chosen
- Advantages:

1. possible to draw inferences about specific subgroups
2. more efficient statistical estimates
   - so long as the selection criteria is relevant in the given question
3. possibility of „post-stratifying”
4. each stratum is treated as independent population, therefore different sampling approaches can be applied to different strata
   - most cost-effective approach might be applied

- Suppose you wish to analyze the average salary in a firm where blue-collar workers and white-collar workers are also hired.
- We have an information that 82,5% of workers are blue-collars, whereas 17,5% of workers are white collars.
- You observe 8 salaries (4-4 in each group), namely:
  - blue-collar: 45, 41, 32, 38.
  - white-collar: 59, 67, 53, 58.
- What will be the best estimate for the average salary in this firm?
- See Excel MINE.XLS

Advantages of stratified sampling

- A stratified sampling approach is most effective when
  - Variability within strata are minimized
  - Variability between strata are maximized
  - The variables upon which the population is stratified are strongly correlated with the desired dependent variable.
- Advantages over other sampling methods
  - Focuses on important subpopulations and ignores irrelevant ones.
  - Allows use of different sampling techniques for different subpopulations.
  - Improves the accuracy/efficiency of estimation.
  - Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages of stratified sampling

- Disadvantages
  - Requires selection of relevant stratification variables which can be difficult.
  - Is not useful when there are no homogeneous subgroups.
  - Can be expensive to implement.

Sample distribution

- How to determine the size of the subsample for each stratum?
- uniform distribution
  - the same size for each stratum: n/M, where n is the sample size, M is the number of strata
- proportional to population share
  - the size of each stratum in the sample is proportional to its share in the whole population: nN_j/N, where
    - n is the sample size
    - N_j is the size of the j-th stratum in the population
    - N is the size of the population
  - simple
  - leads to minimal sampling error if standard deviation of all strata is the same.

Errors in sampling surveys

- Sampling errors and biases are induced by the sample design.
  - Selection bias: When the true selection probabilities differ from those assumed in calculating the results.
  - Random sampling error: Random variation in the results due to the elements in the sample being selected at random.
- Non-sampling errors are caused by other problems in data collection and processing.
  - Overcoverage: Inclusion of data from outside of the population.
  - Undercoverage: Sampling frame does not include elements in the population.
  - Measurement error: E.g. when respondents misunderstand a question, or find it difficult to answer.
  - Processing error: Mistakes in data coding.
  - Non-response: Failure to obtain complete data from all selected individuals.

AAMCONSULTING

Thank you for your attention!

Bucharest 30.10.2011