ECON3150/4150 Spring 2015

Lecture 1

Siv-Elisabeth Skjelbred

University of Oslo

January 19, 2015

This lecture

This lecture will cover:

- An introduction to econometrics.
- A repetition of the probability theory necessary for this course.

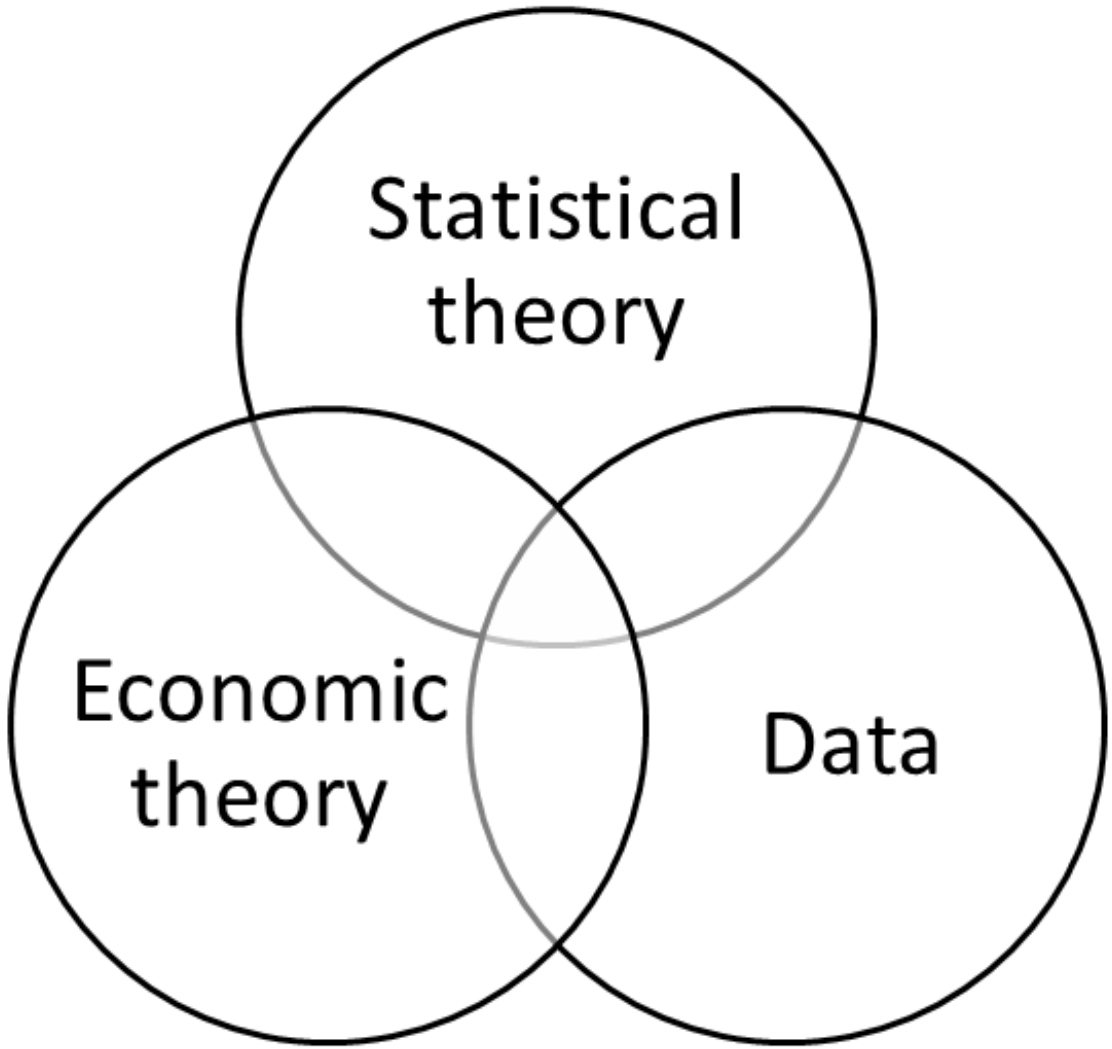
This course

After the end of this course you should be able to:

- Conduct empirical analysis.
  - Be able to forecast using time series data.
  - Be able to estimate causal effects using observational data.
  - Be able to explain the theoretical background of the standard methods used for conducting empirical analysis.
  - Perform statistical tests.
- Interpret and critically evaluate the outcomes of empirical analysis.
  - Read and understand the regression output from Stata.
  - Are the underlying assumptions of the regression satisfied?
  - Are the output externally and internally valid?
- Be able to read and understand (and potentially criticize) papers that make use of the concepts and methods introduced in this course.

What is econometrics?

Econometrics is a "combined discipline"



What is econometrics?

Definition: No clear agreement, S&W use: "Econometrics is the art of using economic theory and statistical techniques to analyze economic data." Which includes:

- Testing economic theories.
- Fitting mathematical economic models to real-world data.
- Using historical data to give policy recommendations.
- Using data to forecast future values of economic variables.
- Estimating causal effects.

Our goal with econometrics

- Economic theory suggests important relationships between factors, but tends to be satisfied with specifying the sign of the correlation.
- In econometrics we will use tools in order to try to estimate the quantitative magnitude of these relationships and establish causal effects.

Steps in an econometric analysis

1. Formulate en economic model - formulate a theoretical model, or use economic theory and economic reasoning to informally formulate a relationship between the variables of interest.
2. From the economic model to an econometric model - specify the functional form of the relationship (linear, log-linear...) General: $y = \beta_0 + \beta_1 x + u$. We call the left side variable the dependent variable and the right side independent variable or explanatory variable.
3. Collect data for the problem at hand
4. Estimate the econometric model
5. Use the estimates for statistical inference

In this course we will focus on step four and five.

Data types

Data types:

- Cross-sectional: data on different entities for a single time period.
- Time series: data for a single entity collected at multiple time periods.
- Panel data: data for multiple entities in which each entity is observed at two or more time periods.
- (Repeated cross section: A collection of cross-sectional data sets, where each cross-sectional data set corresponds to a different time period).

Data sources:

- Experiment
- Observational data, administrative records or surveys

Denotation

- In general we denote a variable with the subscript i (ex X_i) where i is either the time period or the entity number depending on the data type.
- When we need to be precise about using time series data we use the subscript t instead of i.
- When we use panel data we use both subscripts (Y_{it}, X_{it}) where the first subscript is the entity and the second the time period.

Step 4: Estimate the econometric model

Choose an estimator to produce estimates of the relationship we are interested in.

An **estimator** a mathematical procedure (rule) used on sample data. The estimate is the actual value taken by the estimator in a specific sample.

- Linear regression with single or multiple regressors (ch 4-6)
- Non-linear regression functions (ch 8)
- Regression with panel data (ch 10)
- Regressions with binary dependent variable (ch 11)
- Instrumental variable regression (ch 12)

Quality of the estimate

You will also learn to asses the quality of the estimate. Is it:

- Unbiased
- Consistent
- Efficient

Step 5: Statistical inference

Use the estimates to:

- Draw conclusions about the size of economic parameters, ex demand elasticities
- Predict economic outcomes, macroeconomic forecasting
- Test hypotheses, do class size matter for student learning
- Evaluate policy, will the new limit on toll free goods harm Norwegian firms?

But are the estimates reliable?

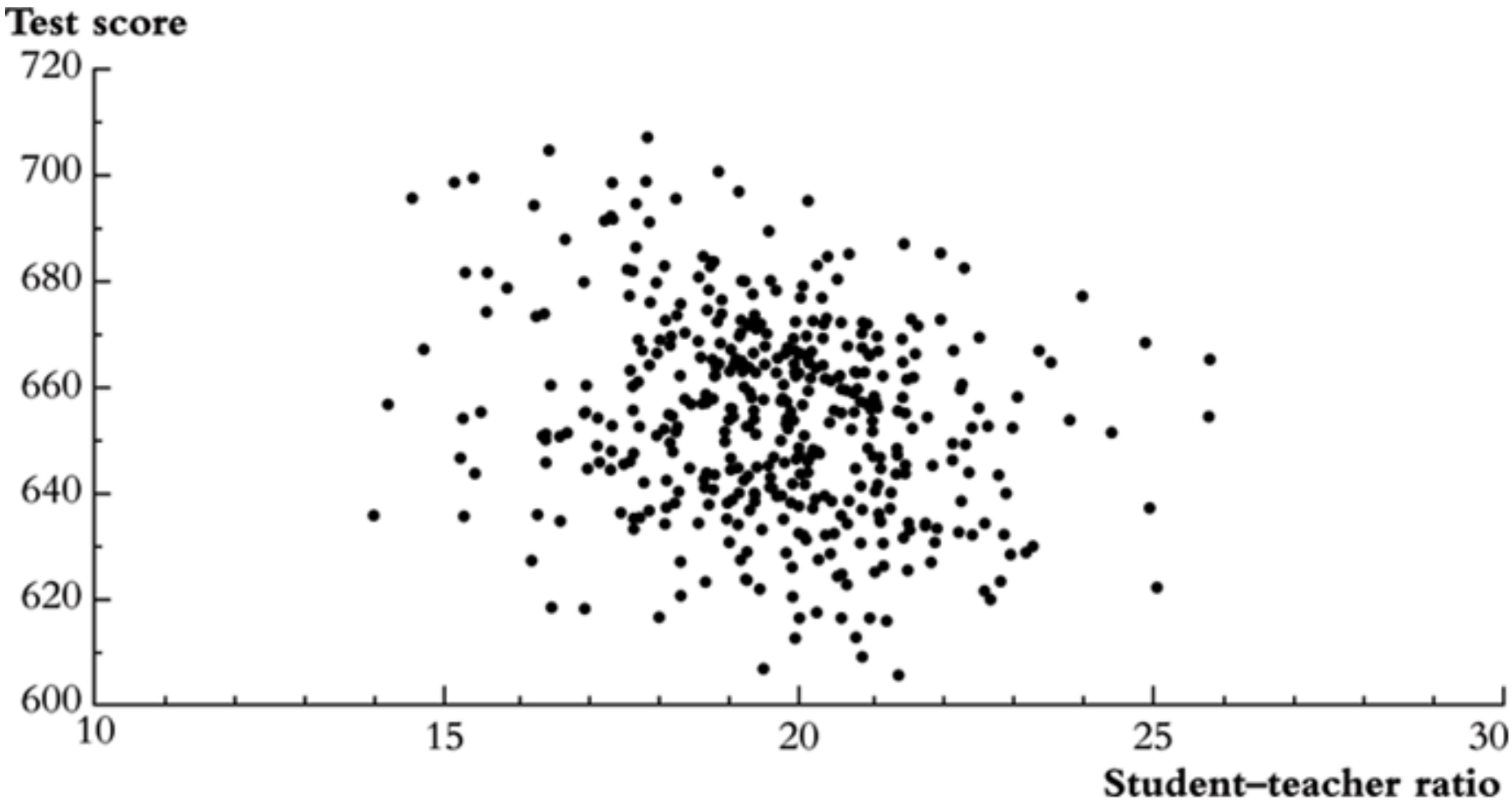
Review of probability

References to Lecture 1

- Stock and Watson (SW) Chapter 1 and 2

A pre-requisite for this course is an introductory statistics course. Thus this is considered repetition. Thus consult your statistics textbook if you need more information than provide in this lecture and the textbook.

Scatterplot



Random variables and probability distribution

- A random variable attaches a value to each possible outcome of a random process.
- **Outcomes** are the mutually exclusive results of the random process and the set of all potential outcomes is called the **sample space**.
- The **probability** of an outcome is the proportion of the time that the outcome occurs in the long run.
- The (marginal) **probability distribution** is the set of all possible outcomes and their associated probabilities.
- The cumulative probability distribution is the probability that the random variable is less than or equal to a particular value.

Example - Coin toss

Consider the random process of flipping two coins:

- Four combinations: two heads, first is head and second is tail, first is tail and second is heads, two tails.
- If variable of interest is number of heads the potential outcomes are [0,1,2]

Number of heads	0	1	2
Probability	0.25	0.5	0.25
Cumulative probability	0.25	0.75	1

Joint and conditional distribution

- The joint distribution is the probability that two (or more) random variables take on certain values simultaneously.
- Conditional distribution is the distribution of a random variable Y conditional on another random variable X taking on a specific value.

$$Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)}$$

Maybe more commonly known as: $Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$

Distributions

The most common probability distributions in econometrics is:

- Normal distribution. Properties:
  - Bell-shaped with (μ) and variance σ^2, written as: N(μ, σ^2)
  - Symmetric around the mean
  - 95% of its probability between $\mu + / - 1.96\sigma$.
  - Note: A sum of n normally distributed random variables is itself normally distributed.
- Standard normal distribution. Properties:
  - N(0,1)
  - Typically the variable is denoted Z and the standard normal cumulative distribution function is denoted with ϕ and $PR(Z \leq c) = \phi(c)$
  - A normal distributed variable can be standardized using: $Z = \frac{X - \mu_x}{\sigma_x}$
- Chi-square distribution is used for comparing estimated variance values to the values based on theoretical assumptions.
- Student t distribution - used to calculate confidence intervals (using the critical t-value)
- The F distribution we will use to compute F-tests

Bernoulli distribution

- A Bernoulli random variable is a binary random variable, which means that the outcome is either zero or one
- The Bernoulli distribution of variable G is then:

$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}$$

- The simplicity of the Bernoulli distribution makes the variance and mean simple to calculate

Measures of a distribution

- Mean = expected value = $E(Y) = \mu_Y$
- Variance is the measure of the square spread of the distribution
- Standard deviation is the square root of the variance
- Skewness - measures the asymmetry of a distribution
- Kurtosis - measures the mass in tails, i.e. probability of large values
- Covariance is the measure of the linear association between two random variables

$$corr(X, Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Z)}}$$

Expectations

If X is a discrete random variable, the expectation is:

$$\mu_x = E(X) = \sum_{i=1}^{k} x_i f_X(x_i) \text{ and } Var(X) = E(X^2) - \mu_x^2$$

Rules for the expectation:

1. $E(a) = a$, for a constant a
2. $E(aX) = aE(X)$ for a constant a
3. $E(X + Y) = E(X) + E(Y)$

Rules for variance:

1. $Var(a) = 0$, for a constant a
2. $Var(aX) = a^2 Var(X) = a^2\sigma_X^2$ for a constant a
3. $Var(aX + bY) = a^2\sigma_x^2 + 2ab\sigma_{xy} + b^2\sigma_y^2$

See key concept 2.3 for more details.

Random sampling

- Simple random sampling: n objects $(Y_1, Y_2, ...Y_n)$ are selected at random from a population. At random means that each member of the population is equally likely to be included in the sample.
- The observations are independently and identically distributed (i.i.d)
  - Same marginal distribution
  - The value of Y_1 provides no information about the value of Y_2

Sample average

Suppose that the observations $Y_1, Y_2, ... Y_n$ are i.i.d. with mean μ_y and variance σ_y^2 then the sample average is:

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + ... Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

which is itself a random variable with a probability distribution called the sampling distribution. Furthermore:

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_y \text{ and } Var(\bar{Y}) = var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{\sigma_y^2}{n}$$

Large sample approximations

- Unless the distribution of Y is normal the exact sampling distribution of the sample average ($\bar{Y}$) is complicated
- However, when the sample size is large we can impose the central limit theorem and the law of large numbers

The law of large numbers

Law of large numbers

Under general conditions, the sample average will be close to the population mean with very high probability when the sample is large.

I.e. when n is large $\bar{Y}$ is close to μ_y with high probability.
When a large number of random variables with the same mean are averaged together, the large values balance the small values and the sample averages is close to the common mean.

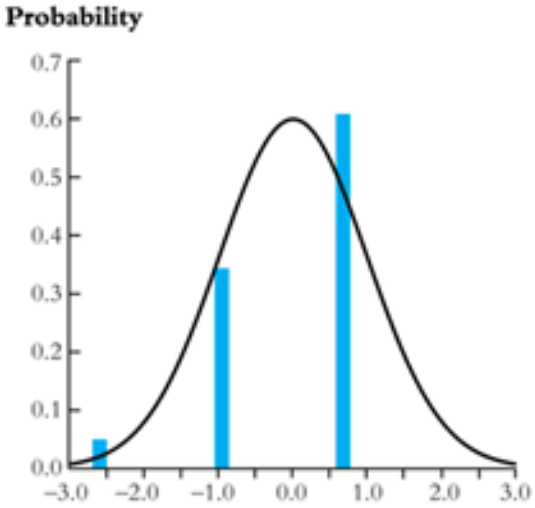
Central limit theorem

Central limit theorem

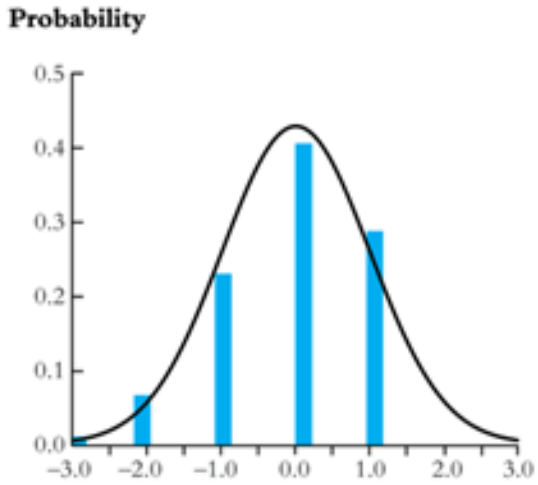
Under general conditions, the sampling distribution of the standardized sample average is well approximated by a standard normal distribution when the sample size is large.

When n is large the distribution $\bar{Y}$ of converges to the normal distribution. That is: $\bar{Y}$ is approximately $\mathrm{N}(\mu_y, \sigma^2_{\bar{Y}})$

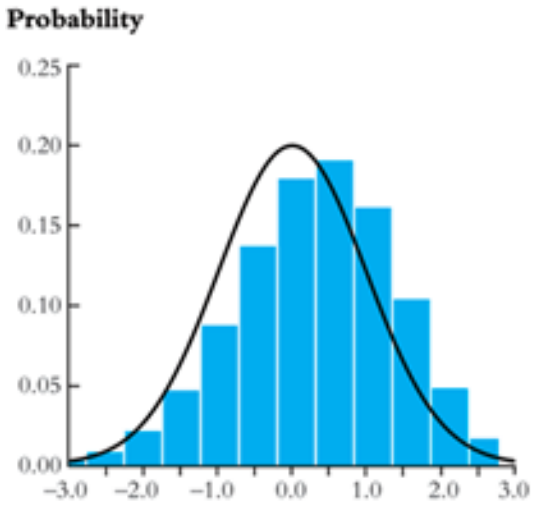
Example: CLT



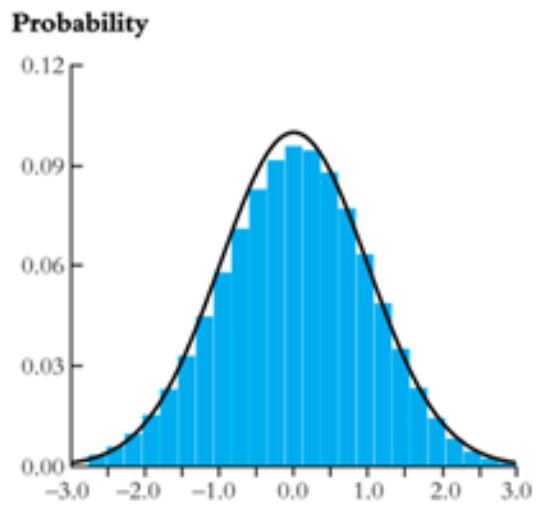
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

Properties of the sample average

- An estimator is consistent if the probability that it falls within an interval of the true population value tends to one as the sample size increases.
- The law of large numbers specifies the conditions under which the sample average is a consistent estimate of the population mean.
- We say that $\bar{Y}$ converges in probability to μ_y or that $\bar{Y}$ is consistent for μ_y.

Sampling distribution of $\bar{Y}$

For small sample sizes the distribution of $\bar{Y}$ is complicated but if n is large then:

- As n increases the distribution of $\bar{Y}$ becomes more tightly centered around μ_y (The law of large numbers)
- The distribution of $\bar{Y} - \mu_y$ becomes normal (The central limit theorem)