LECTURE 1

Introduction to Econometrics

Gega Todua

September 22, 2017

What is econometrics?

To beginning students, it may seem as if econometrics is an overly complex obstacle to an otherwise useful education. (...) To professionals in the field, econometric is a fascinating set of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends.

Studenmund (*Using Econometrics: A Practical Guide*)

What is econometrics?

- Econometrics is a set of statistical tools and techniques for quantitative measurement of actual economic and business phenomena
- It attempts to
  - quantify economic reality
  - bridge the gap between the abstract world of economic theory and the real world of human activity
- It has three major uses:
  1. describing economic reality
  2. testing hypotheses about economic theory
  3. forecasting future economic activity

"Are you just pissing and moaning, or can you verify what you're saying with data?"

Example

- Consumer demand for a particular commodity can be thought of as a relationship between
  - quantity demanded (Q)
  - commodity's price (P)
  - price of substitute good (P_s)
  - disposable income (Y)
- Theoretical functional relationship:

$$Q = f(P, P_s, Y)$$

- Econometrics allows us to specify:

$$Q = 31.50 - 0.73P + 0.11P_s + 0.23Y$$

Introductory econometrics course

- **Lecturer:** Gega Todua (CERGE-EI, Prague)
  gega.todua@cerge-ei.cz
- **Lectures:** Friday, 9,20-10,05, room VT 203
  Friday, 10,15-11,50, room VT 203
- **Office hours:** Friday, after Seminar by appointment
- **Web:** `https://is.muni.cz/auth/course/econ/podzim2017/BPE_AIEC?lang=en`

INTRODUCTORY ECONOMETRICS COURSE

- **Course requirements:**
  - NO EXAMS! :)
  - 3 home assignments (account for $3 \times 20 = 60$ points)
  - written Empirical Project (accounts for 40 points). Details will be announced during following weeks
  - to pass the course, student has to achieve at least 20 points in the project and 50 points in total

- **Recommended literature:**
  - Studenmund, A. H., *Using Econometrics: A Practical Guide*
  - Adkins, L., *Using gretl for Principles of Econometrics*
  - Wooldridge, J. M., *Introductory Econometrics: A Modern Approach*

Important Dates

- 24.11.2017: Last Lecture
- 15.12.2017 00:00 The deadline for the Empirical Project
- 17.11.2017: Public Holiday
- 29.09.2017: No Lectures (away for the conference)

Course content

- **Lectures:**
  - Lecture 1: Introduction, repetition of statistical background, non-technical introduction to regression
  - Lectures 2 - 4: Linear regression models
  - Lectures 5 - 12: Violations of standard assumptions
- **In-class exercises:**
  - Will serve to clarify and apply concepts presented on lectures
  - We will use statistical software (Gretl) to solve the exercises

LECTURE 1.

- **Introduction, repetition of statistical background**
  - probability theory
  - statistical inference
- Readings:
  - Studenmund, A. H., Using Econometrics: A Practical Guide, Chapter 17
  - Wooldridge, J. M., Introductory Econometrics: A Modern Approach, Appendix B and C

Random variables

- A **random variable** X is a variable whose numerical value is determined by chance. It is a quantification of the outcome of a random phenomenon.
- **Discrete random variable**: has a countable number of possible values
  - Example: the number of times that a coin will be flipped before a heads is obtained
- **Continuous random variable**: can take on any value in an interval
  - Example: time until the first goal is shot in a football match between FC Barcelona and Real Madrid

Discrete random variables

- Described by listing the possible values and the associated probability that it takes on each value
- **Probability distribution** of a variable X that can take values $x_1, x_2, x_3, \dots$:

$$\begin{aligned} P(X = x_1) &= p_1 \\ P(X = x_2) &= p_2 \\ P(X = x_3) &= p_3 \\ &\vdots \end{aligned}$$

- **Cumulative distribution function** (CDF) :

$$F_X(x) = P(X \leq x) = \sum_{i=1, x_i \leq x} P(X = x_i)$$

Six-sided die: probability density function

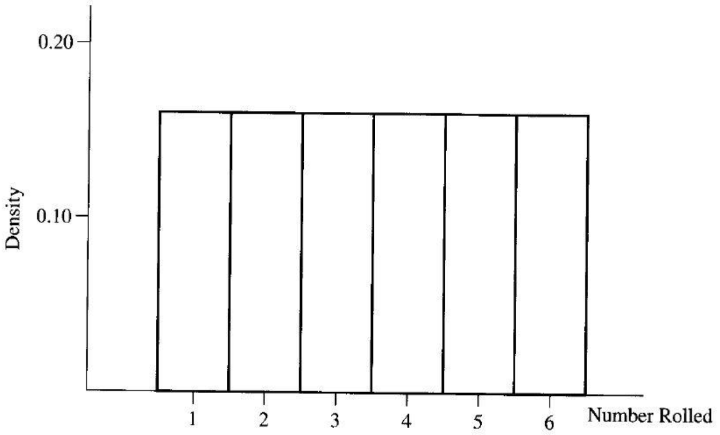
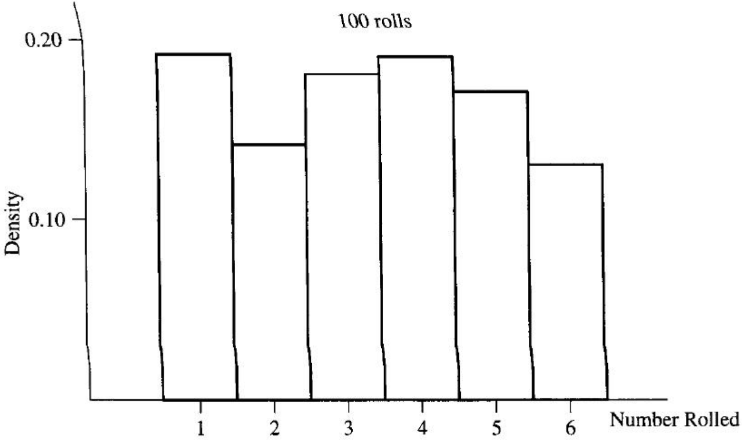
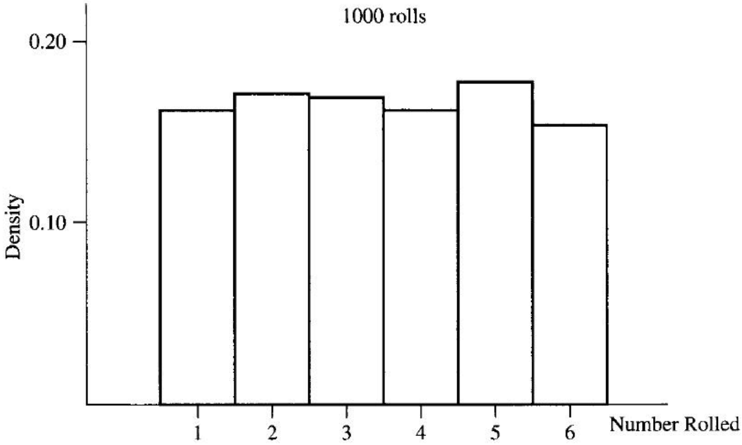


Figure 16.3 Probability Distribution for a Six-Sided Die

Six-sided die: histogram of data (100 rolls)



SIX-SIDED DIE: HISTOGRAM OF DATA (1000 ROLLS)



Continuous random variables

- **Probability density function** $f_X(x)$ (PDF) describes the relative likelihood for the random variable X to take on a particular value x

- **Cumulative distribution function** (CDF) :

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

- **Computational rule:**

$$P(X \geq x) = 1 - P(X \leq x)$$

Expected value and median

- **Expected value (mean)** :
  - Mean is the (long-run) average value of random variable

Discrete variable

$$E[X] = \sum_{i=1} x_i P(X = x_i)$$

Continuous variable

$$E[X] = \int_{-\infty}^{+\infty} x\, f_X(x) \mathrm{d}x$$

  - Example: calculating mean of six-sided die

- **Median** : "the value in the middle"

Exercise 1

- A researcher is analyzing data on financial wealth of 100 professors at a small liberal arts college. The values of their wealth range from \$400 to \$400,000, with a mean of \$40,000, and a median of \$25,000.
- However, when entering these data into a statistical software package, the researcher mistakenly enters \$4,000,000 for the person with \$400,000 wealth.
- How much does this error affect the mean and median?

Variance and standard deviation

- **Variance** :
  - Measures the extent to which the values of a random variable are dispersed from the mean.
  - If values (outcomes) are far away from the mean, variance is high. If they are close to the mean, variance is low.

$$Var[X] = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$$

- **Standard deviation** : $\sigma_X = \sqrt{Var[X]}$

Dancing statistics

Watch the video ”Dancing statistics: Explaining the statistical concept of variance through dance”:

- https://www.youtube.com/watch?v=pGfwj4GrUlA&list=PLEzw67WWDg82xKriFiOoixGpNLXK2GNs9&index=4

Use the ’dancing’ terminology to answer these questions:

1. How do we define variance?
2. How can we tell if variance is large or small?
3. What does it mean to evaluate variance within a set?
4. What does it mean to evaluate variance between sets?
5. What is the homogeneity of variance?
6. What is the heterogeneity of variance?

EXERCISE 2

- Which has a higher expected value and which has a higher standard deviation:
  - a standard six-sided die or
  - a four-sided die with the numbers 1 through 4 printed on the sides?
- Explain your reasoning, without doing any calculations, then verify, doing the calculations.

Covariance, correlation, independence

- **Covariance** :
  - How, on average, two random variables vary with one another.
  - Do the two variables move in the same or opposite direction?
  - Measures the amount of linear dependence between two variables.

$$Cov(X,Y) = E\left[(X - E[X])\,(Y - E[Y])\right] = E\left[XY\right] - E[X]E[Y]$$

- **Correlation** :
  - Similar concept to covariance, but easier to interpret.
  - It has values between -1 and 1.

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Independence of variables

- **Independence** : X and Y are independent if the conditional probability distribution of X given the observed value of Y is the same as if the value of Y had not been observed.
- If X and Y are independent, then $Cov(X, Y) = 0$ (not the other way round in general)
- Dancing statistics: explaining the statistical concept of correlation through dance
  - https://www.youtube.com/watch?v=VFjaBh12C6s&index=3&list=PLEzw67WWDg82xKriFiOoixGpNLXK2GNs9

Computational rules

$$
\begin{aligned}
E\left(aX+b\right) &= aE(X)+b \\
Var(aX+b) &= a^2 Var(X) \\
Var(X+Y) &= Var(X)+Var(Y)+2Cov(X,Y) \\
Cov(aX,bY) &= Cov(bY,aX) = abCov(X,Y) \\
Cov(X+Z,Y) &= Cov(X,Y)+Cov(Z,Y) \\
Cov(X,X) &= Var[X]
\end{aligned}
$$

Random vectors

- Sometimes, we deal with vectors of random variables
- Example: $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$
- Expected value: $E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ E[X_3] \end{pmatrix}$
- Variance/covariance matrix:

$$Var[\mathbf{X}] = \begin{pmatrix} Var[X_1] & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var[X_2] & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var[X_3] \end{pmatrix}$$

STANDARDIZED RANDOM VARIABLES

- Standardization is used for better comparison of different variables
- Define Z to be the standardized variable of X:

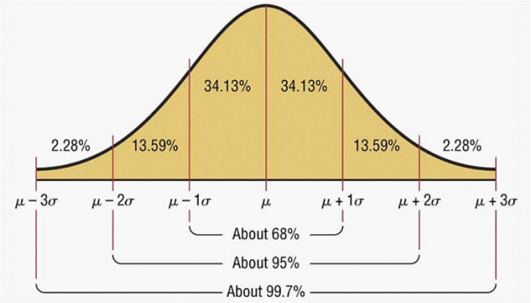
$$Z = \frac{X - \mu_X}{\sigma_X}$$

- The standardized variable Z measures how many standard deviations X is below or above its mean
- No matter what are the expected value and variance of X, it always holds that

$$E[Z] = 0 \quad \text{and} \quad Var[Z] = \sigma_Z = 1$$

NORMAL (GAUSSIAN) DISTRIBUTION

- Notation : $X \sim N(\mu, \sigma^2)$
- $E[X] = \mu$
- $Var[X] = \sigma^2$



- Dancing statistics
  - https://www.youtube.com/watch?v=dr1DynUzjq0&index=2&list=PLEzw67WWDg82xKriFiOoixGpNLXK2GNs9

Exercise 3

- The heights of U.S. females between age 25 and 34 are approximately normally distributed with a mean of 66 inches and a standard deviation of 2.5 inches.
- What fraction of U.S. female population in this age bracket is taller than 70 inches, the height of average adult U.S. male of this age?

Exercise 4

- A woman wrote to Dear Abby, saying that she had been pregnant for 310 days before giving birth.
- Completed pregnancies are normally distributed with a mean of 266 days and a standard deviation of 16 days.
- Use statistical tables to determine the probability that a completed pregnancy lasts
  - at least 270 days
  - at least 310 days

SUMMARY

- Today, we revised some concepts from statistics that we will use throughout our econometrics classes
- It was a very brief overview, serving only for information what students are expected to know already
- The focus was on properties of statistical distributions and on work with normal distribution tables

Next lecture

- We will go through terminology of sampling and estimation
- We will start with regression analysis and introduce the Ordinary Least Squares estimator