LECTURE #1

Introductory Econometrics

INTRODUCTION TO THE COURSE & RECAP OF STATISTICAL BACKGROUND

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, September 29

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, econometrics is a fascinating set of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends."

(Studenmund, 2016, *Using Econometrics: A Practical Guide*)

What is econometrics?

- Econometrics is the 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?"

Why do we need econometrics?

"We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists."

(Studenmund, 2016, *Using Econometrics: A Practical Guide*)

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 (Yd)

- Theoretical functional relationship:

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

- Econometrics allows us to specify:

$$Q = 27.7 - 0.11P + 0.03P_s + 0.23Yd$$

Introductory Econometrics course

- **Lecturer:** Jiri Kukacka, Ph.D.
  - email: jiri.kukacka@fsv.cuni.cz
  - web: ies.fsv.cuni.cz/en/staff/kukacka
- **Lectures:** Wednesday, 11:00, lecture hall 314, Opletalova 26
- **Office hours:** Wednesday, 13:30–15:00, room 406, by appointment via email, please, or we can Meet/Zoom
- **Seminars:** Thursday, 15:30 and 17:00, room 016
- **Teaching assistants:**
  - Jan Sila (jan.sila@fsv.cuni.cz)
  - Periklis Brakatsoulas (peribrak@gmail.com)
- **Hybrid form of teaching:**
  - simultaneous online broadcast via Zoom
  - links and a passcode in SIS

Introductory Econometrics course

- **Course requirements:**
  - 3 homework assignments (0-40 points in total: 13+13+14)
  - Final exam online (Moodle part: 0-30 points; oral part: 0-30 points; to pass each part, a student has to achieve at least 15 points; to pass the exam, a student has to pass both parts)
  - Detailed info in SIS: JEM062
- **Grading policy (Dean's Measure no. 20/2019):**
  - 90+ to 100 points result in 'A' ('Excellent')
  - 80+ to 90 points result in 'B' ('Very good')
  - 70+ to 80 points result in 'C' ('Good')
  - 60+ to 70 points result in 'D' ('Satisfactory')
  - 50+ to 60 points result in 'E' ('Sufficient')
  - 50 or less points result in 'F' (not passed)

Introductory Econometrics course

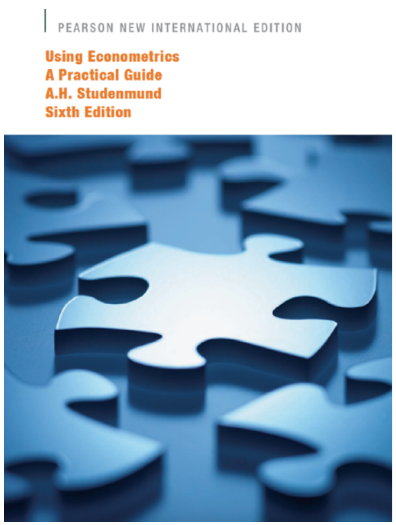
- **Homework assignments:**
  - Announced via SIS on Wednesdays, due in 8 days via SIS
  - Teams of two
  - 'Academic integrity' required
  - Results and feedback via SIS
- **Final exam online:**
  - Moodle part (open-book) $\longrightarrow$ oral part (closed-book)
  - Good and stable internet connection and a webcam required
  - Term 1: January 12 (Wed), 2022, 11:00
  - Term 2: January 19 (Wed), 2022, 18:00
  - Term 3: January 26 (Wed), 2022, 11:00
  - Term 4: February 2 (Wed), 2022, 18:00
  - Enrolment and results via SIS
  - Specific details will be provided at the beginning of December

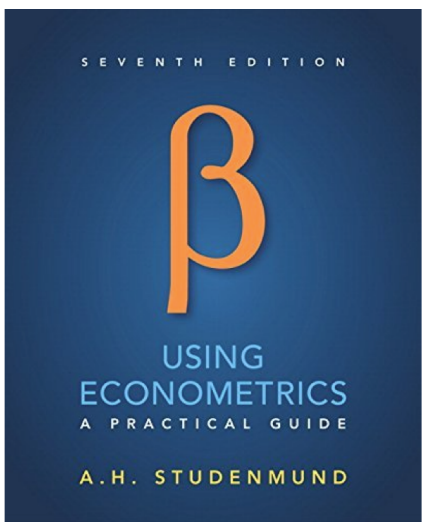
Introductory Econometrics course

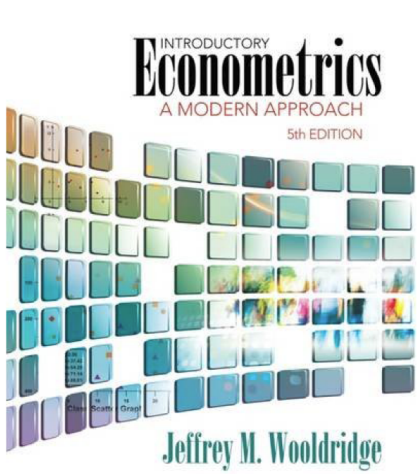
- **Course materials:**
  - Lecture handouts (Mo)
  - Seminar handouts and datasets (Mo)
  - A sketch of solutions and answers to exercises (Thu or Fri)
  - Solutions to home assignments (Fri)
  - A specimen final exam (beginning of December)
- **Software:**
  - Gretl (free, open-source, all platforms)
  - Excel (provided to students by the Faculty)

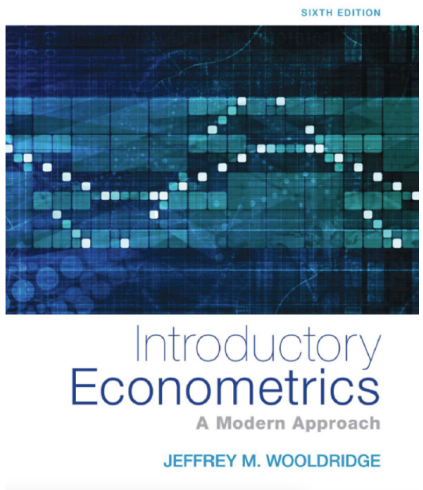
Core textbooks

- Studenmund, A. H. (2016). *Using Econometrics: A Practical Guide*. Pearson Educ., 7th Edition (pdf)
- Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage, 6th Edition (pdf)







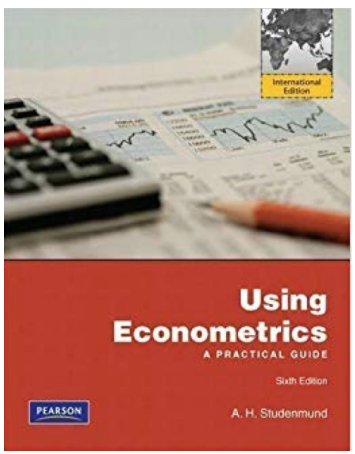


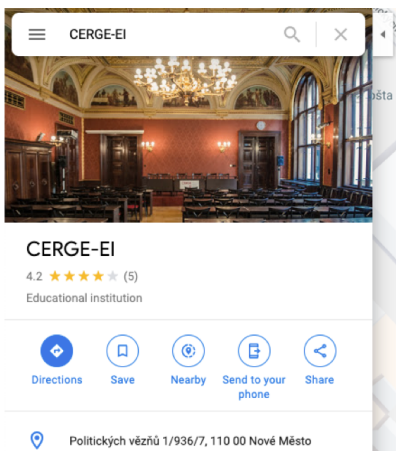
Libraries

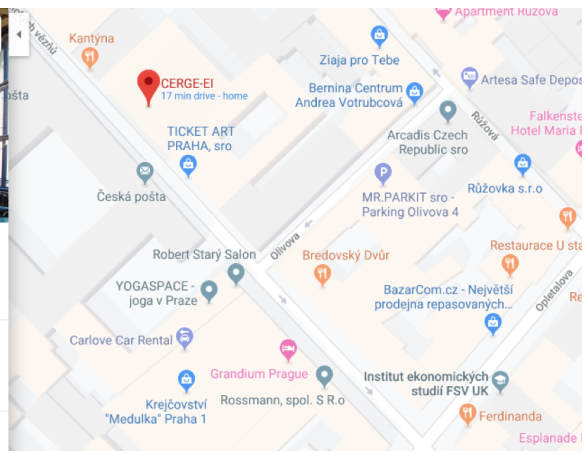
IES library: ies.fsv.cuni.cz/en/node/189

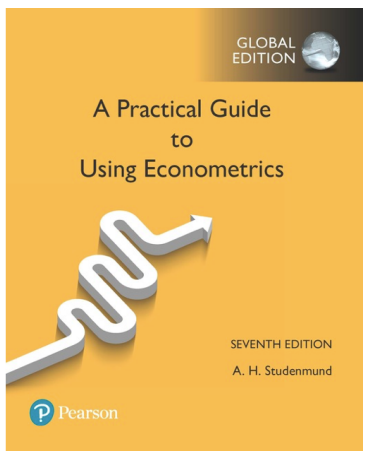
CERGE-EI library: olib.cerge.cuni.cz

↓ e-book (7th Ed. 2017)









Course content

- **Lectures and HAs (tentative):**
  - Lecture 1: Intro & Recap of statistical background
  - Lectures 2–6: Linear regression model & Hypotheses testing, HA#1 & HA#2
    Lecture 5, October 27: ONLINE ONLY
  - Lectures 7–10: Violations of model assumptions, HA#3
    (no lecture on November 17)
  - Lecture 11: Introduction to qualitative dependent variables
  - Lecture 12: Revision, Questions & Answers
  - Detailed info in SIS: JEM062

- **Seminars (exercise sessions, practicals):**
  - Serve to clarify and apply concepts presented in lectures
  - Both 'pen and paper' and software exercises
  - Discussion
  - October 28: ONLINE ONLY; November 18: no seminars

Lecture #1

- **Recap of statistical background**
  - Probability theory
  - Statistical inference
- Readings:
  - Studenmund (2016 & 17, [2014]): Chapters 1.1, [15/17 Statistical Principles]
  - Wooldridge (2016, 2012): Appendix B, C-1–C-3

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
  - A sum of random variables is a random variable too...
- **Discrete random variable:** has a countable number of possible values
  - Example: the number of times that a coin will be flipped before a head is obtained, gender, outcome of rolling dice
- **Continuous random variable:** can take on any value in an interval
  - Example: height, temperature, speed, individual wealth, time until a breakdown of an engine

Discrete random variables

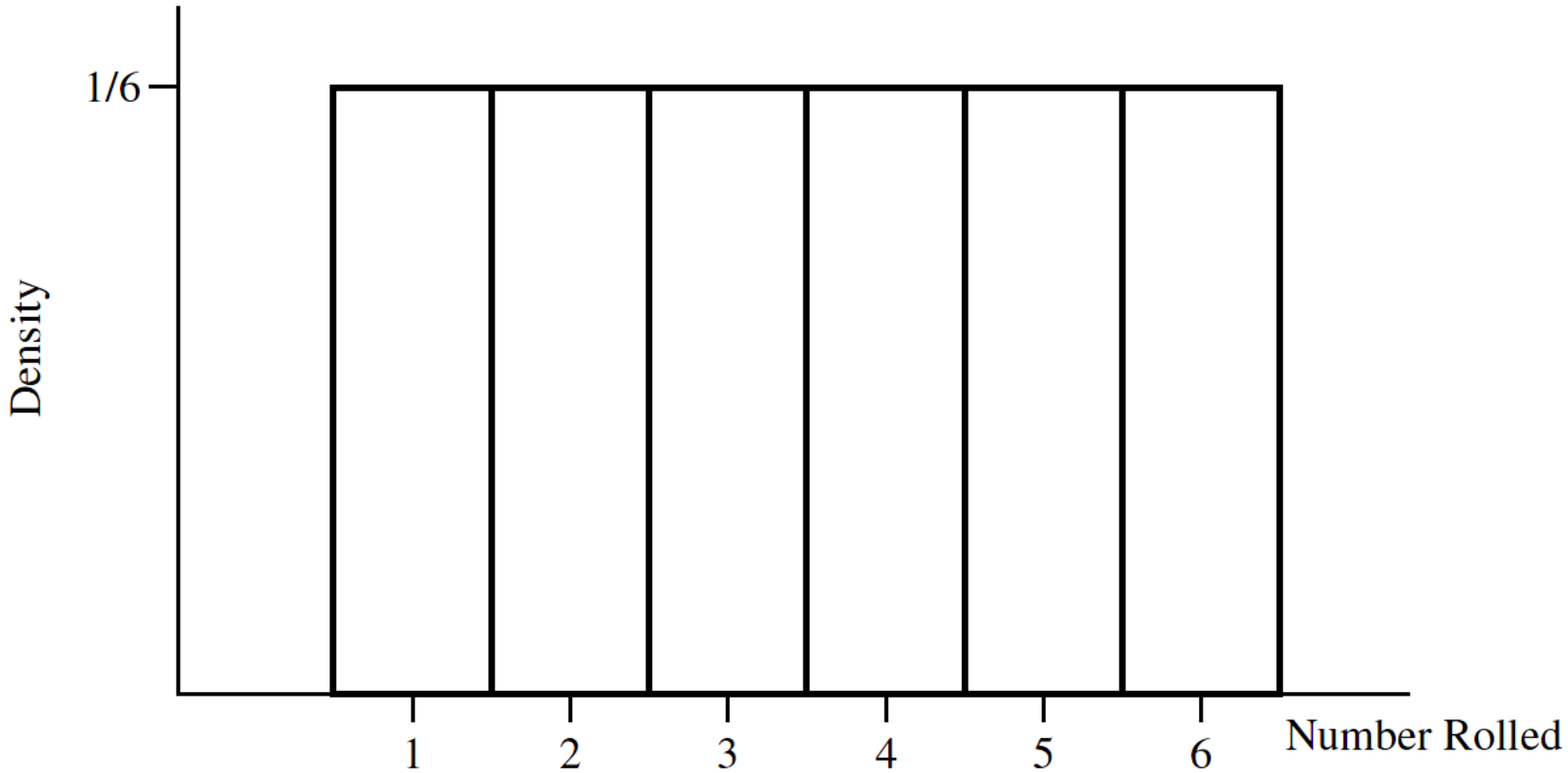
- Described by listing the possible values and the associated probability that it takes on each value
- **Probability distribution** of a variable X that can take values x_1, x_2, x_3, ... :

$$\begin{aligned} P(X = x_1) &= p_1 \\ P(X = x_2) &= p_2 \\ P(X = x_3) &= p_3 \\ &\vdots \\ \sum_{i=1} p_i &= 1 \end{aligned}$$

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

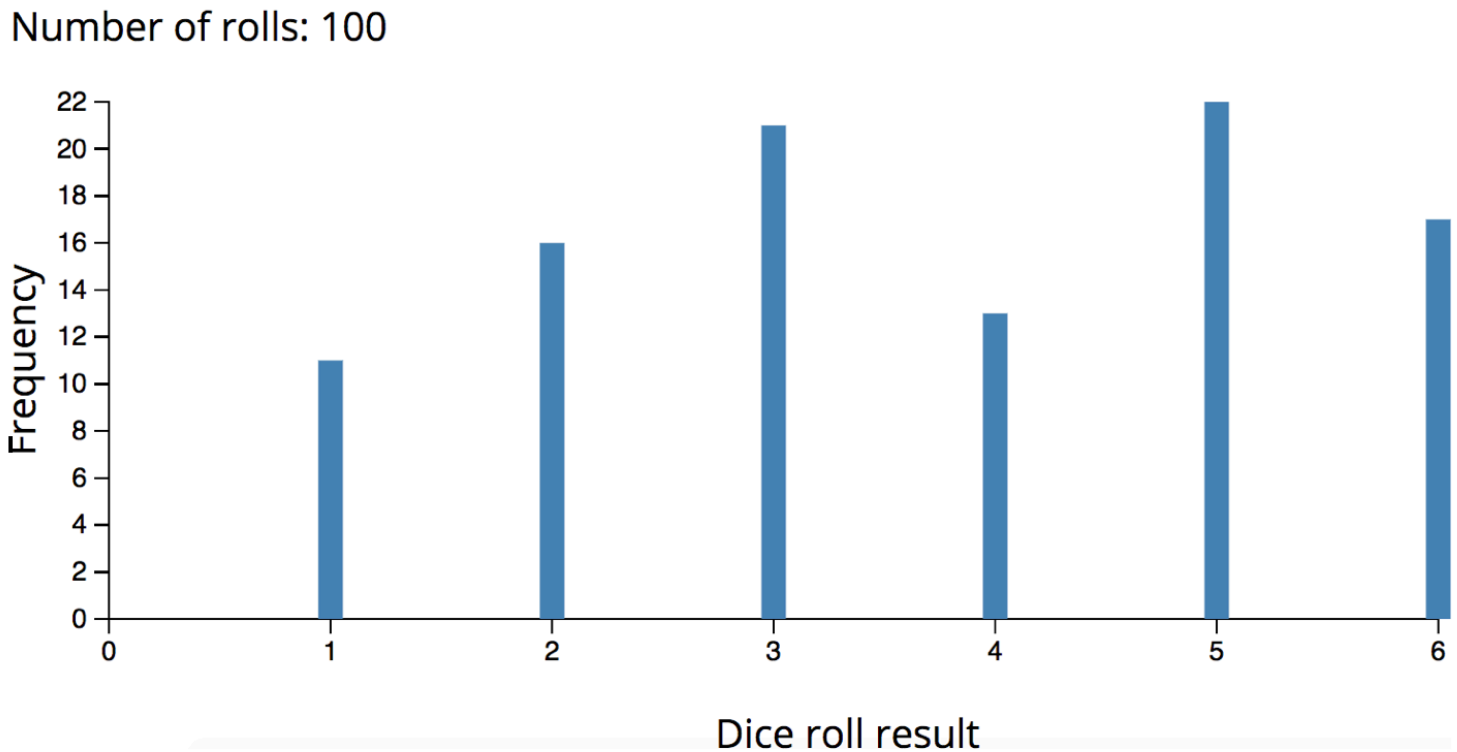
$$F_X(x) = P(X \leq x) = \sum_{i=1, x_i \leq x} P(X = x_i)$$

Six-sided die: probability distribution



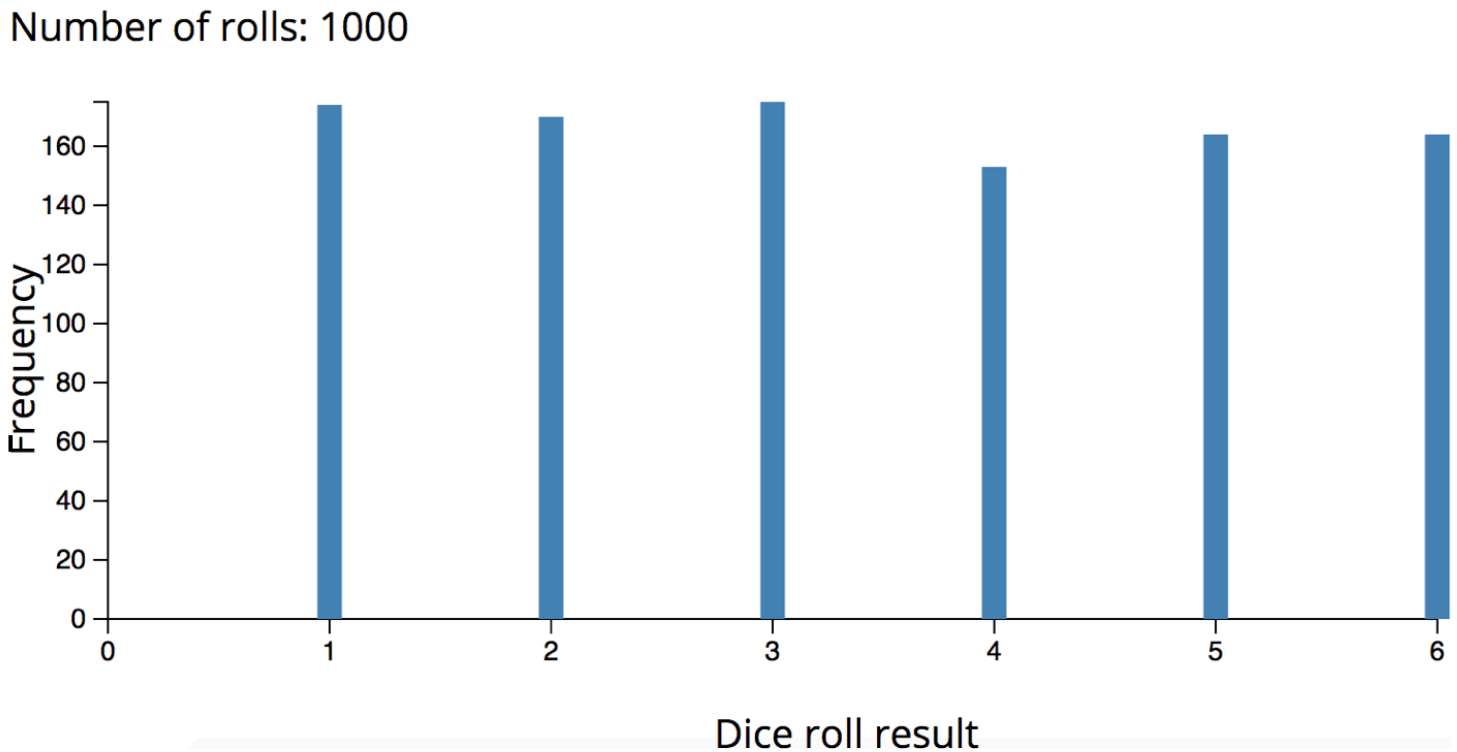
Source: Studenmund (2014, pg. 509)

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



Source: academo.org/demos/dice-roll-statistics

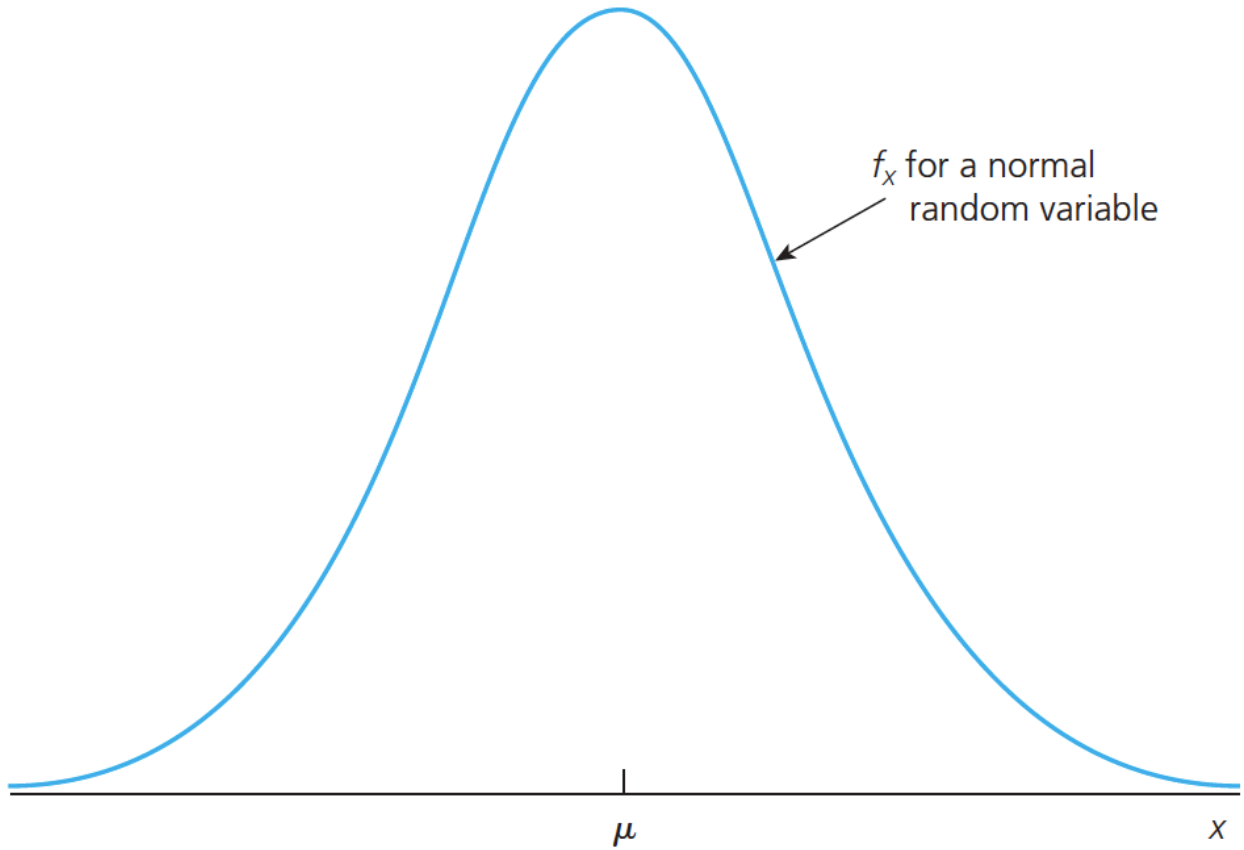
Six-sided die: histogram of data (1000 rolls)



Source: academo.org/demos/dice-roll-statistics

Continuous random variables

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

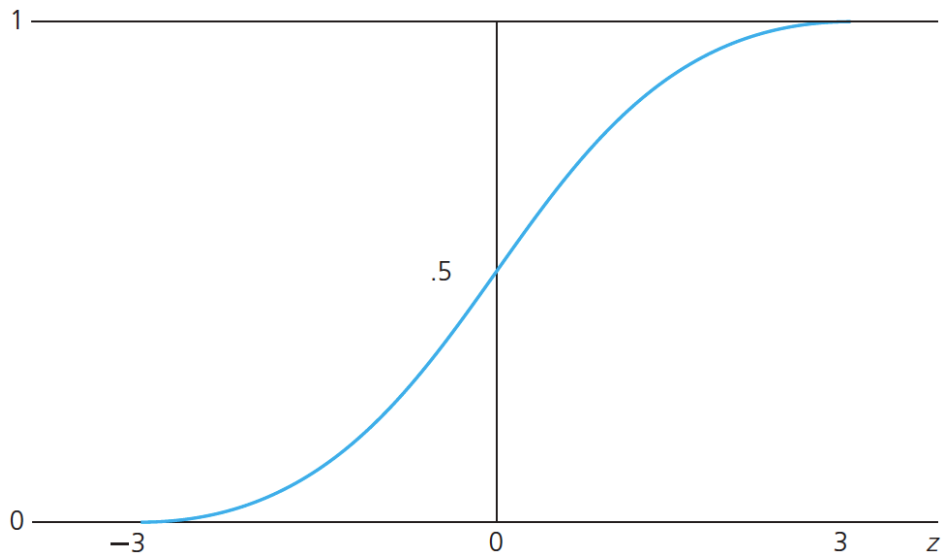


Source: Wooldridge (2016, pg. 666)

Continuous random variables

Cumulative distribution function (CDF):

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



Source: The standard normal CDF, Wooldridge (2016, pg. 667)

Computational rule: $P(X > x) = 1 - P(X \leq x)$

Expected value vs. median

- **Expected value (mean):**
  - Mean is the (long-run) average value of a random variable
  - It is a **weighted** average of all its possible values
  - The weights are determined by the PDF

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 (seminar #1)

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

17 rolls D6 (ordered): 1 1 1 1 2 2 3 3 **3** 3 4 4 5 5 5 6 6

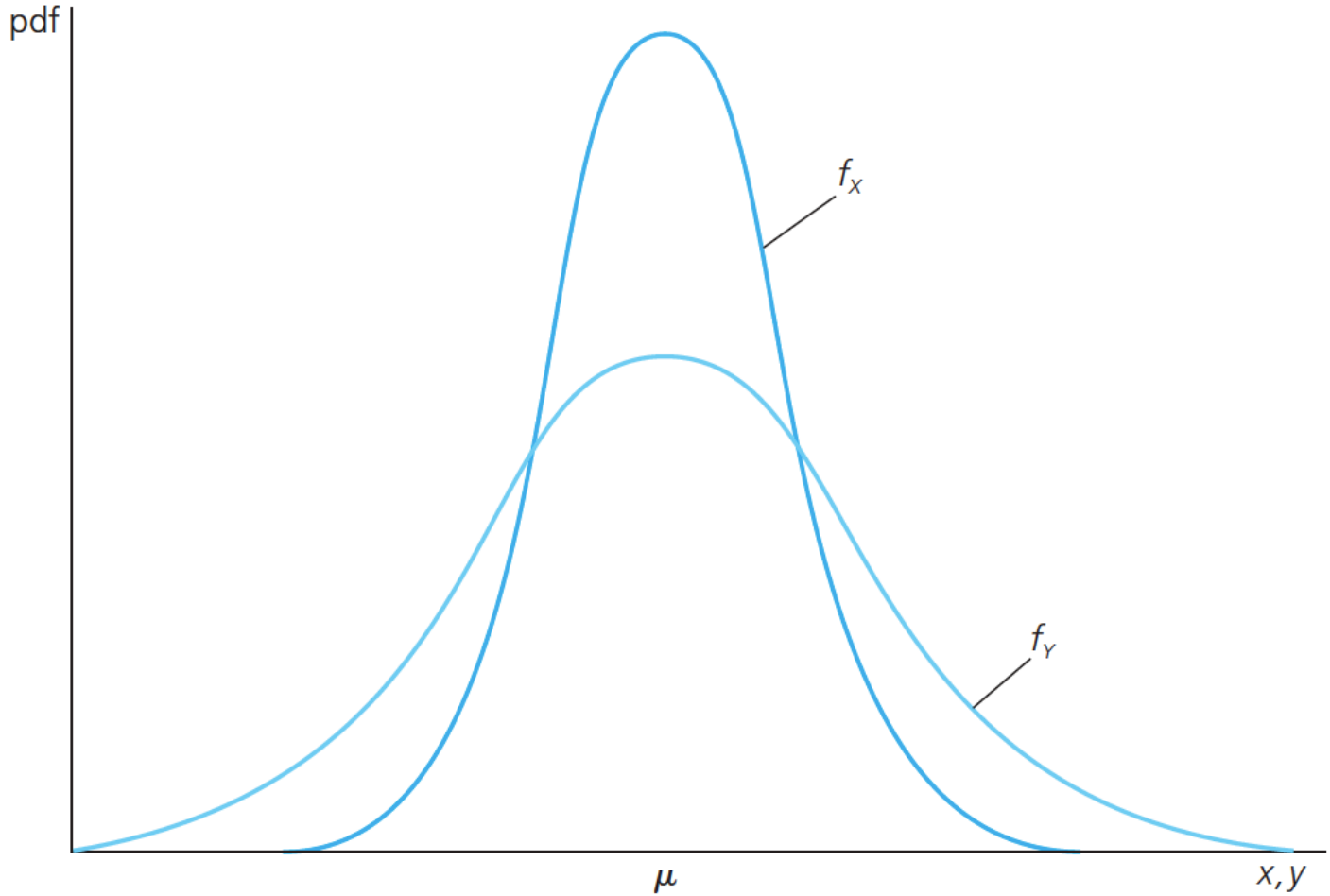
Variance and standard deviation

- **Variance:**
  - Measures the extent to which the values of a random variable are dispersed from the its expected value (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:** $\sigma_X = \sqrt{Var[X]}$

RVs with the same mean but different variances



Source: Wooldridge (2016, pg. 656)

Covariance and correlation

- **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 RVs

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

- **Correlation:**
  - Similar concept to covariance, but easier to interpret
  - It has values between -1 and 1
  - Does not depend on the units of measurement
  - $Corr(X, Y) = 0 \Rightarrow$ no linear relationship between X and Y

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

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)

Sample moments

- Counterparts of theoretical moments of the distribution of X, computed based on observations $X_1, \ldots, X_n$ drawn from this distribution

- **Sample mean:**

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- **Sample variance and standard error:**

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

- **Sample covariance:**

$$Cov_n(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$$

Computational rules

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

where X, Y, Z are random variables and a, b are scalars (constants)

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:

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

- Comp. rule: $Var[\mathbf{Ax}] = \mathbf{A}\,Var[\mathbf{x}]\mathbf{A}'$, $\mathbf{A}$ is a nonrandom matrix

Selected properties of matrix operations (transpose, inverse; Wooldridge (2016): Appendix D)

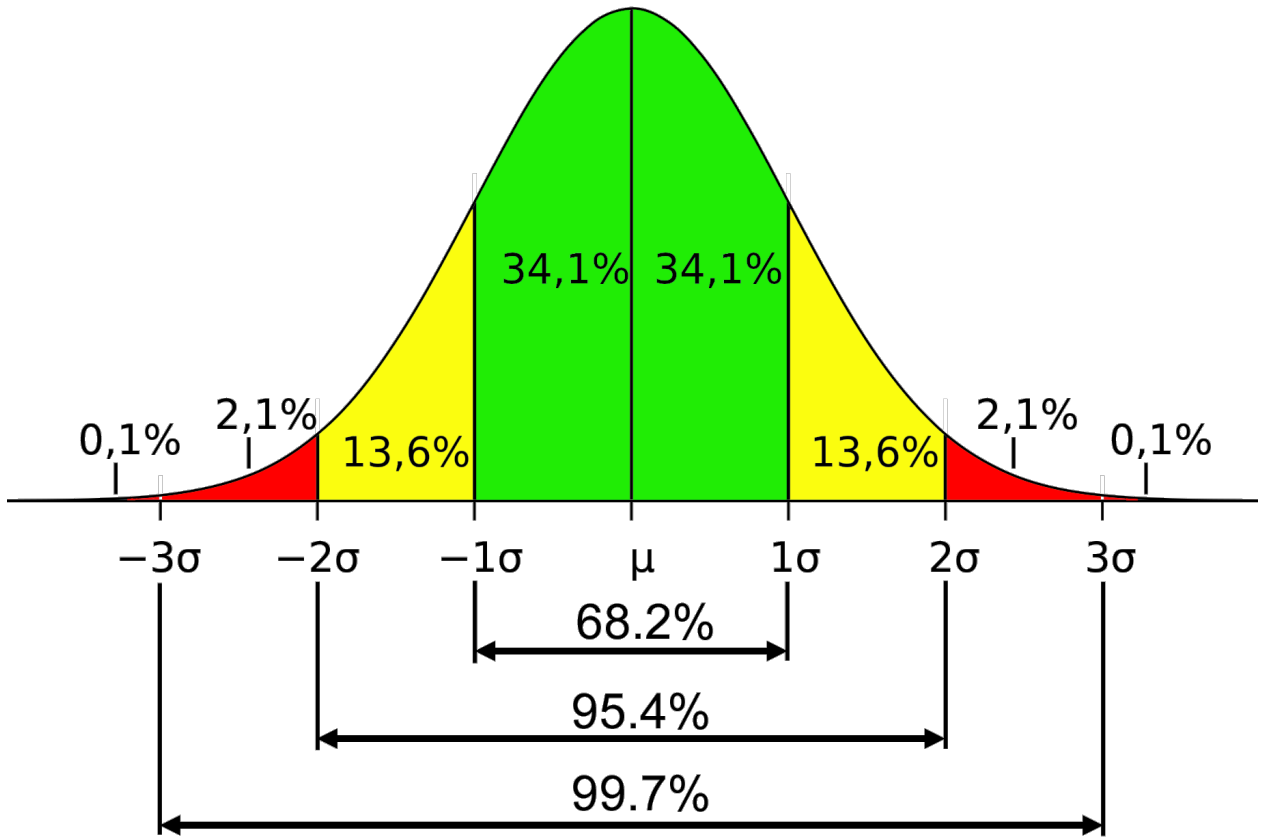
$$
\begin{aligned}
a' &= a \\
(\mathbf{A}')' &= \mathbf{A}^{\star} \\
(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}' \\
(\mathbf{AB})' &= \mathbf{B}'\mathbf{A}'^{\star\star} \\
(\mathbf{ABC})' &= \mathbf{C}'\mathbf{B}'\mathbf{A}' \\
\text{Def.: } \mathbf{D}' &= \mathbf{D} \quad \Leftrightarrow \quad \mathbf{D} \text{ symetric} \\
\mathbf{A}'\mathbf{A} &\ldots \text{ symetric matrix: } (\mathbf{A}'\mathbf{A})' \overset{\star\star}{=} \mathbf{A}'(\mathbf{A}')' \overset{\star}{=} \mathbf{A}'\mathbf{A}
\end{aligned}
$$

$$
\begin{aligned}
(\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\
(\mathbf{A}')^{-1} &= (\mathbf{A}^{-1})'
\end{aligned}
$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are matrices and a is a scalar (constant)

Normal (Gaussian) distribution

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



Source: www.muelaner.com

Normal (Gaussian) distribution

- The most widely used distribution in statistics and econometrics
- Certain random variables appear to roughly follow a normal distribution: human heights and weights, test scores (IQ, grades), leaves of trees, country unemployment rates...
- **Central Limit Theorem**: the sum (or the mean) of a number of independent, identically distributed random variables will tend to be normally distributed, **regardless of their distribution**, if their number is large enough

Standardized random variable

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

$$Z = \frac{X - E[X]}{\sigma_X}$$

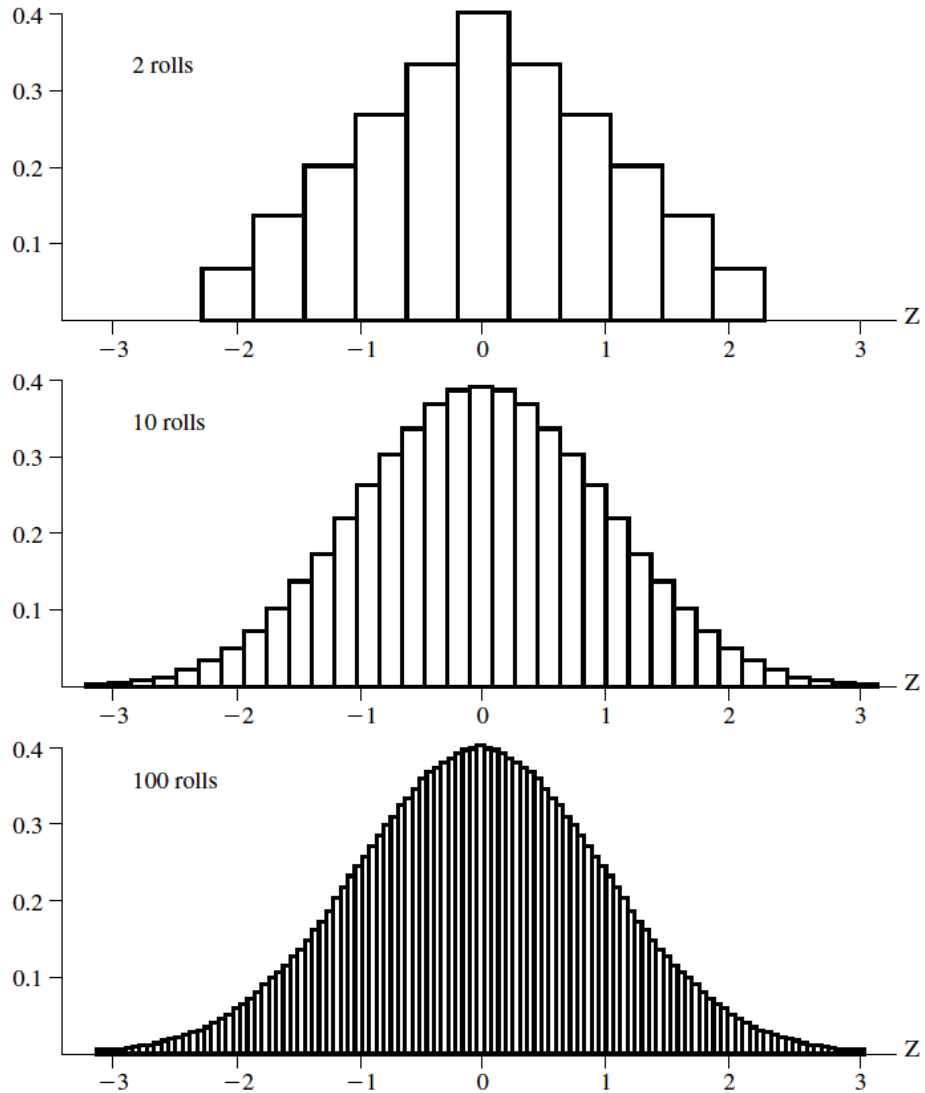
- No matter what are the expected value and variance of X, it always holds that

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

- Standard normal distribution:

$$X \sim N(\mu, \sigma^2) \quad \longrightarrow \quad Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

PDF of sums of six-sided dice rolls (standardized)

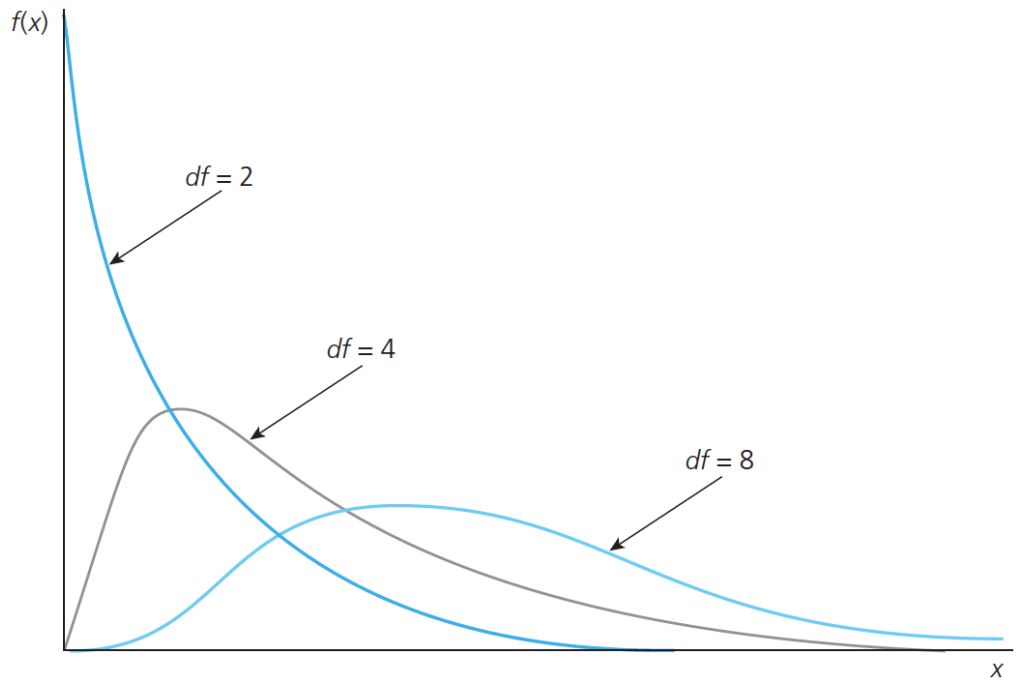


Source: Studenmund (2014, pg. 516)

Chi-squared distribution

- **Chi-squared distribution** with m degrees of freedom: χ^2_m
- Let $Z_i \sim N(0,1)$ for each i and independent, then

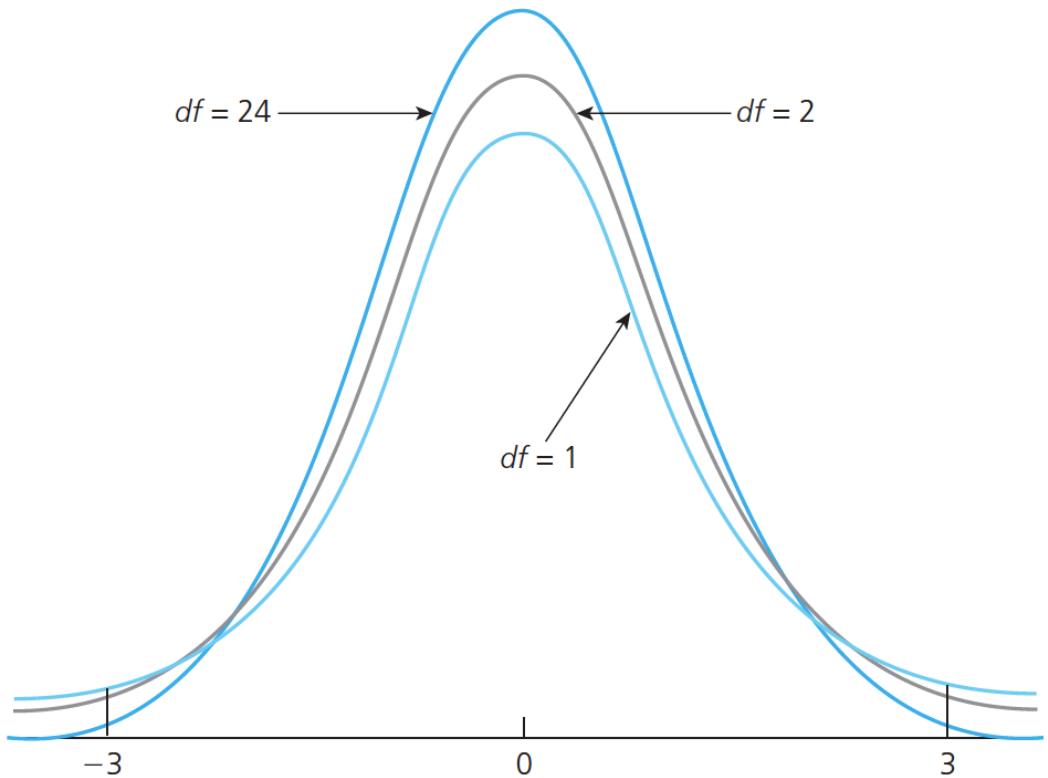
$$X = \sum_{i=1}^{m} Z_i^2 \quad \longrightarrow \quad X \sim \chi^2_m$$



Source: Wooldridge (2016, pg. 670)

t distribution

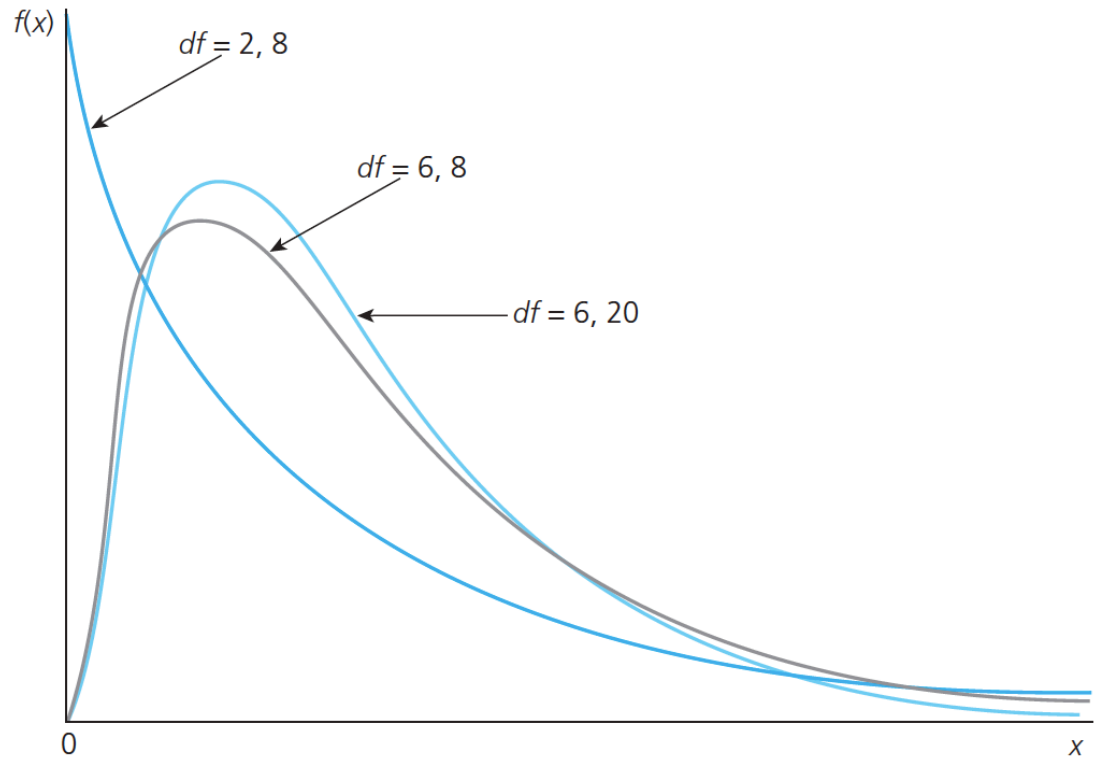
- (Student's) **t distribution** with m degrees of freedom: t_m
- Let $Z \sim N(0,1)$, $X \sim \chi^2_m$, independent: $T = \dfrac{Z}{\sqrt{X/m}} \sim t_m$
- Note also that as m grows, t_m distribution approaches $N(0,1)$



Source: Wooldridge (2016, pg. 671)

F distribution

- (Fisher-Snedecor) **F distribution** with m and o degrees of freedom: $F_{m,o}$
- Let $X \sim \chi^2_m$, $Y \sim \chi^2_o$, independent: $F = \frac{X/m}{Y/o} \sim F_{m,o}$
- Why all important: hypotheses testing, confidence intervals



Source: Wooldridge (2016, pg. 671)

Some terminology

- **Population:** the entire group of items of our research interest
- **Sample:** a part of the population that we actually observe
- **Statistical inference:** use of a sample to draw conclusion about the characteristics of the population from which the sample came
- Examples: opinion polls, medical experiments

Random sampling

- Statistical inference can be performed correctly only on a **random sample**, i.e. a sample that reflects the true distribution of the population
- Each member of the population is equally likely to be included in a random sample
- Each observation in a random sample is an **independent** random variable drawn from the same population

- **Biased sample:** any sample that differs systematically from the population that it is intended to represent

Selection biases

- **Selection bias:** occurs when the selection of the sample systematically excludes or under represents certain groups
  - Example: opinion poll about tuition payments among undergraduate students vs all citizens
- **Self-selection bias:** occurs when we examine data for a group of people who have chosen to be in that group
  - Example: accident statistics of people who buy collision insurance
- **Survivor bias:** occurs when we choose a sample from a current population (survivors) in order to draw inferences about a past population
  - Example: S&P500 companies, medical records of old people
- **Nonresponse bias:** the systematic refusal of some groups to participate in an experiment or to respond to a poll

Some more terminology

- **Parameter:** a true characteristic of the distribution of a variable, whose value is unknown, but can be estimated
  - Example: population mean $E[X]$
- **Estimator:** a sample statistic that is used to estimate the value of the parameter
  - Example: sample mean $\overline{X}_n$
  - Note that the estimator is a random variable (it has a probability distribution, mean, variance,...)
- **Estimate:** the specific value of the estimator that is obtained using an estimation technique and a particular sample

Linearity and linear combination

- In mathematics, a linear function $f(x)$ is a function that satisfies the following two properties:
  - Additivity: $f(x + y) = f(x) + f(y)$
  - Homogeneity of degree 1: $f(\alpha x) = \alpha f(x) \quad \forall\ \alpha$

- **Linear combination** is an expression constructed from a set of terms by multiplying each term by a constant and adding the results
- E.g. a linear combination of x and y would be any expression of the form $ax + by$, where a and b are constants

Source: Wikipedia here (linearity) and here [accessed 2018-09-18].

Properties of an estimator

- An estimator is **unbiased** if the mean of its distribution is equal to the true value of the parameter it is estimating
- An estimator is **consistent** if it converges to the true value of the parameter as the sample size increases
- An estimator is **efficient** if the variance of its sampling distribution is the smallest possible

Properties of an estimator: Example

- Let X_i be observations sampled from a distribution with mean μ and variance σ^2
- Let us consider the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of μ
- It can be shown that:
  1. $E[\overline{X}_n] = \mu$
  2. $\overline{X}_n \to \mu$ as n increases
  3. $\overline{X}_n$ has the smallest variance of all possible estimators of μ
- Hence, the sample mean $\overline{X}_n$ is an unbiased, consistent, and efficient estimator of μ

Summary

- Today, we have revised some important concepts from statistics that we will use throughout our econometrics classes
- It was a very brief and quick overview, serving only for information what students are expected to know already
- The focus was on distributions and their moments, on sampling and estimation terminology

Seminars and the next lecture #2

- In the upcoming **seminars**, we will practice some of the concepts mentioned today:
  - basic statistical concepts (mean, median, variance)
  - work with the standard normal distribution (statistical tables)
  - properties of the sample mean
- In the next **lecture**, we will start with regression analysis and introduce the Ordinary Least Squares (OLS) estimator
- Readings for lecture #2:
  - Studenmund (2016 & 17, [2014]): Chapters 1.2–1.5, 2.1–2.3, 3.1–3.2 [3], 7.1
  - Wooldridge (2016, 2012): Chapters 1, 2-1–2-2, 2-3a, 2-4c