LECTURE #1

Econometrics I

INTRODUCTION TO ECONOMETRICS & COURSE INFORMATION & QUICK STATISTICS REVIEW

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

Summer semester 2024, February 20

Outline

Introduction to econometrics

Course information

Quick statistics review
- Probability distributions and their features
- Recap of selected distributions
- LLN and CLT
- Hypothesis testing

Outline

Introduction to econometrics

Course information

Quick statistics review

- Probability distributions and their features
- Recap of selected distributions
- LLN and CLT
- Hypothesis testing

What is econometrics?

- Econometrics has evolved from statistics as a separate discipline:
  - experimental vs. non-experimental (observational) data
  - analysis of non-experimental data comes with additional issues
- Focus on:
  - estimating economic relationships
  - empirical testing of economic theories or hypotheses
  - evaluating economic policies
  - predicting, forecasting economic variables
- Readings:
  - Chapter 1: 1.1, 1.2, 1.4

Econometrics as a separate discipline

- 1930: the first Econometrics Society in Cleveland, OH, USA
  - Joseph A. Schumpeter, Harold Hotelling, Karl Menger, Norbert Wiener, or Irving Fisher (president), among others
- 1932: the Cowles Commission for Research in Economics, Colorado Springs, CO, USA (due to Alfred Cowles)
  - Nobel Prize laureates: e.g., Keneth Arrow, Gérard Debreu, James Tobin, Franco Modigliani, Herbert A. Simon, Joseph E. Stiglitz, or Harry Markowitz
- Many Nobel Prize laureates in economics are, in fact, econometricians
  - e.g., Robert Merton, Myron Scholes, James Heckman, Daniel McFadden, Robert Engle, Clive Granger, Thomas Sargent, Christopher Sims. . . 2021: Joshua Angrist, Guido Imbens

Steps in empirical economic analysis

1. Proposing an economic ‘model in words’
2. Converting it into a formalized econometric model
3. Collecting data
4. Selecting an appropriate software
5. Estimating the model, evaluating its quality
6. Reporting and interpreting the results

Economic vs. econometric models

- Economic models are based on economic theory. Usually, only a general form of relationships is given:

$$y = f(\bullet).$$

- Econometric models have a specific form that can be estimated and further analyzed:

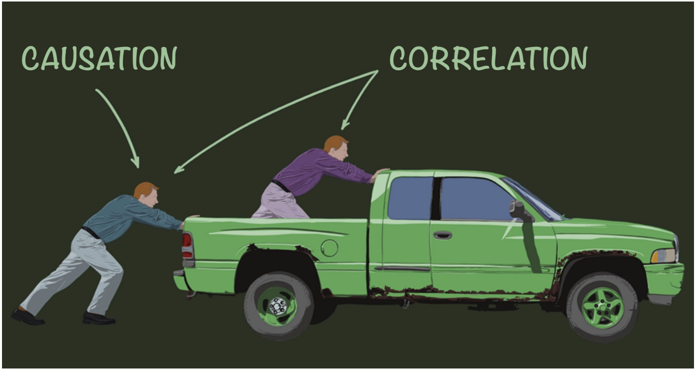
$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \\ y &= \beta_0 x_1^{\beta_1} \ldots x_k^{\beta_k} u, \\ &\vdots \end{aligned}$$

with parameters $\beta_0, \ldots, \beta_k$ and an error term u.

Note on causality and ceteris paribus in econometrics

- In an ideal case, the econometric analysis leads to uncovering a causal effect between variables.
- However, as the saying goes, correlation is not causation.
- The ceteris paribus approach ('holding all other relevant factors fixed') utilized in the econometric analysis could help with the issue $\Rightarrow$ econometric methods can simulate a ceteris paribus experiment.
- Nevertheless, finding a true causal effect is challenging.

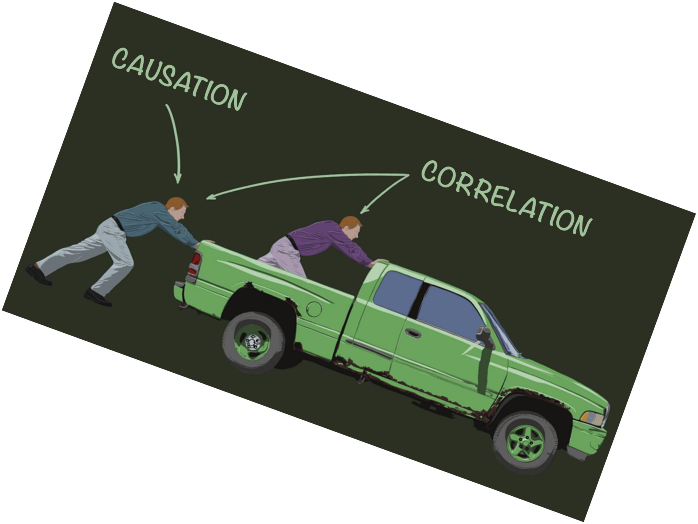
Causality in econometrics



Source: facebook.com/VedatorCZ [2020-10]

▸ Further examples

Ceteris paribus in econometrics



Source: facebook.com/VedatorCZ [2020-10]

Outline

Econometrics I course: People and times

- **Lecturer:** Jiri Kukacka, Ph.D.
  - email: jiri.kukacka@fsv.cuni.cz
  - web: ies.fsv.cuni.cz/en/contacts/people/58305408
- **Code:** JEB109 (link to SIS)
- **Lectures:** Tuesday, 11:00, lecture hall 'Michal Mejstrik' 109
- **Office/consultation hours:** Tuesday, 13:30–15:30, room 406, by appointment via email, or we can schedule a Meet call
- **Seminars:** Tue 12:30 [ML] and 15:30 [AS], Wed 8:00 [MPet], 12:30 [HN], and 15:30 [HN], Thu 17:00 [MPav], PC lab 016
- **Teaching assistants:**
  - Mgr. Martina Luskova: 38341240@fsv.cuni.cz
  - Hieu Nguyen, M.Sc.: HieuThiHoang.Nguyen@cerge-ei.cz
  - Mgr. Margarita Pavlova: Margarita.Pavlova@cerge-ei.cz
  - Mathieu Petit, M.Sc.: 79846277@fsv.cuni.cz
  - Attila Sarkany, M.Sc.: 77127687@fsv.cuni.cz

Econometrics I course: Requirements and grading

- **Pre-requisity:** JEB105 Statistics (& Mathematics I + II)
- **Course requirements:**
  - 2 home assignments (0–20 points: 10 + 10)
  - midterm written test (0–30 points)
  - final written exam (0–50 points, necessary condition to pass: more than 25 points)
  - detailed info in SIS: JEB109
- **Standard FSV grading policy (Directive S_SO_002):**
  - 90+ to 100 points result in 'A' ('Excellent')

    . . .
  - 50 or fewer points result in 'F' (not passed)
- **6 (ETCS) credits**

Econometrics I course: Details and dates

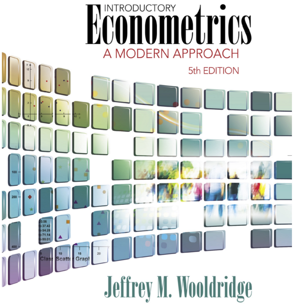
- **Home assignments:**
  - announced via SIS on Thursdays, due in two weeks
  - teams of two, completely via SIS ('Study group roster' app)
  - 'academic integrity', deadlines non-negotiable, solo $\Rightarrow$ 0
  - points and feedback via SIS
- **Midterm test:**
  - 65-minute, written, closed-book, IDs, lecture hall MM 109
  - **April 2**, 11:00, results via SIS
- **Final exam:**
  - 100-minute, written, closed-book, IDs, lecture hall MM 109
  - exam terms (always from 11:00):
    May 28, **June 4**, **June 11**, **June 27**
  - resit term: **September 3** (11:00)
  - registration (from May 6, 8:00 a.m.) and results via SIS
  - specific details (changes) at the beginning of May

Econometrics I course: Materials and software

- **Course materials:**
  - lecture handouts and R codes (Sun)
  - seminar handouts, datasets, and R codes (Sun)
  - a sketch of solutions and answers to exercises (Fri)
  - solutions to home assignments (Fri)
  - a specimen final exam (beginning of May)
- **Software:**
  - R and RStudio (freeware, all platforms, PC lab 016)
  - mostly at seminars (+ 'pen and paper' + discussion)
  - strongly suggested course: Data Analysis in R (if you are new to R, review at least the recorded introductory lectures)
- **Communication, please, in advance:**
  - Q $\longrightarrow$ SIS: JEB109 $\longrightarrow$ TAs $\longrightarrow$ lecturer/consultation $\longrightarrow$ A

Core textbook

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*. Cengage Learning, 5th Ed., .pdf (6th Ed., 7th Ed.)



Alternative (more advanced):

- Baltagi, B. H. (2021). Econometrics. Springer, 6th Ed.
- Greene, W. H. (2018). Econometric Analysis. Pearson Education, 8th Ed.
- Wooldridge, J. M. (2010). Econometric Analysis . . . MIT Press, 2nd Ed.

Course syllabus & schedule

- L #1: Introduction & Information & Quick stats review
  - there are no seminars
- L #2 – #4: Chapters 1 + 2 + 3, HA #1 due on March 28
  - *Simple + Multiple regression analysis: Basics + OLS*
- L #5 and #6: Chapters 4 + 5
  - *Gauss-Markov theorem + Inference + Asymptotics*
- Midterm test instead of lecture: April 2
  - there are no seminars
- L #7 – #9: Chapters 6 + 7, HA #2 due on May 9
  - *Further issues + Selection of variables + Qualitative analysis*
- L #10 and #11: Chapters 8 + 9
  - *Heteroskedasticity + Specification and data issues*
- L #12 (May 21): Revision of key concepts (Chapters 1–9)
  - there are no seminars
- detailed info in SIS: JEB109 (Dean's and Rector's Day, holiday)

Outline

Quick statistics review

- We will now recap the essential concepts students should be familiar with from their statistics courses that will be used throughout our econometrics classes.
- It will be a very brief, quick, and intuitive overview, primarily for information on what students are expected to know.
- Please refer to JEB105 Statistics and the Wooldridge (2012) appendices for details.

- Readings:
  - SIS information: JEB109
  - Appendix B, C.3, C.6

Outline

Distribution functions

- We deal with random variables that form a random sample (*i.i.d* n-tuple, usually with unknown probability distribution) and their realizations (observations $\rightarrow$ observed data set).
- For any random variable X, the cumulative distribution function (CDF) F_X is defined as

$$F_X(x) = P(X \leq x),\ x \in \mathbb{R}.$$

- Some of the properties of F_X:
  - nondecreasing
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to +\infty} F_X(x) = 1$
- For continuous random variables, we have the density function $f_X(x)$ of X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt,\ -\infty < x < +\infty.$$

Selected moments and other features

- Mean and its estimators: sample mean, median
- Variance and standard deviation
- Bivariate association: covariance and correlation

(Population) mean

- Measure of central tendency of a random variable
- Expected value $\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$
- Some of the properties:
  - $\mathbb{E}(X + c) = \mathbb{E}(X) + c$, where c is a constant
  - $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
  - $\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i)$
  - $\mathbb{E}(cX) = c\mathbb{E}(X)$, where c is a constant
- Popular estimators of the expected value:
  - sample mean/average: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - median: 50th percentile

Variance

- Measure of variability/dispersion and uncertainty
- If $\mathbb{E}(X^2) < +\infty$, then

$$\mathrm{Var}(X) = \mathbb{E}\big(X - \mathbb{E}(X)\big)^2 = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$$

  is called variance of X.
- Variance is the second centered moment:

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} \big(x - \mathbb{E}(X)\big)^2 f_X(x)dx.$$

- Standard deviation σ: a positive square root of variance
- Some of the properties:
  - $\mathrm{Var}(X) \geq 0$
  - $\mathrm{Var}(b) = 0$, where b is a constant
  - $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$, where a, b are constants

Covariance

- Measure of linear association between two random variables
- If $\mathbb{E}(X^2) < +\infty$ and $\mathbb{E}(Y^2) < +\infty$, then

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

  is called covariance between X and Y.
- Some of the properties:
  - let $X_1, \ldots, X_n$ be random variables with finite second moms. Then

$$\text{Var}(a_1 X_1 + \ldots + a_n X_n + b) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \text{Cov}(X_i, X_j).$$

  - let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two sets of random variables with finite second moments. Then

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i + b, \sum_{j=1}^{m} c_j Y_j + d \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i c_j \text{Cov}(X_i, Y_j).$$

Correlation coefficient

- Standardized measure of linear association
- Let X and Y be two random variables with finite second moments and non-zero variances. Then

$$\mathrm{Corr}(X, Y) = \rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  is called the coefficient of correlation between X and Y.
- It holds that (due to the Cauchy-Schwarz inequality):

$$-1 \leq \rho_{X,Y} \leq 1.$$

Uncorrelated random variables

Having $\rho_{X,Y} = \text{Cov}(X, Y) = 0$ gives some important and practical implications:

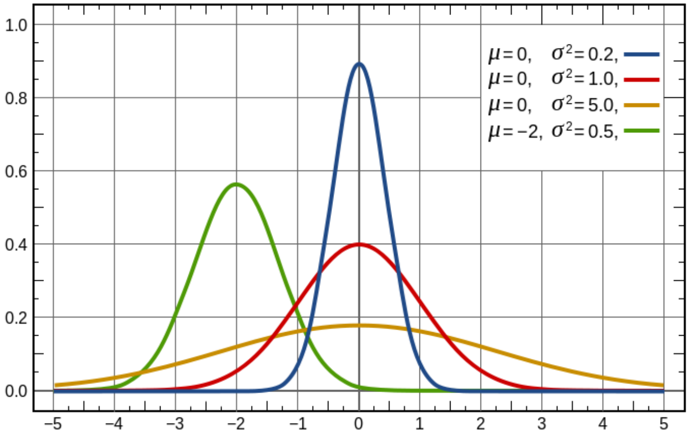
- $\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$
- $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

Outline

Normal distribution

- Usually referred to as the Gaussian distribution (after Carl Friedrich Gauss).
- Normally distributed random variable $X \sim N(\mu, \sigma^2)$ has:
  - $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$
  - $\mathbb{E}(X) = \mu$
  - $\mathrm{Var}(X) = \sigma^2$
- $N(0, 1)$ is the standard normal distribution.
- It is useful to remember the most important quantiles of the standard normal distribution:
  - $\Phi(1.645) = 0.95$
  - $\Phi(1.960) = 0.975$ (≈ 2, 'two-sigma rule')
  - $\Phi(2.326) = 0.99$
  - $\Phi(2.576) = 0.995$

Normal distribution



Source: Wikipedia

Normal distribution

Source: Wikipedia

χ^2 distribution

- If $Z_1, \ldots, Z_k$ are *i.i.d.* random variables with $N(0,1)$, then

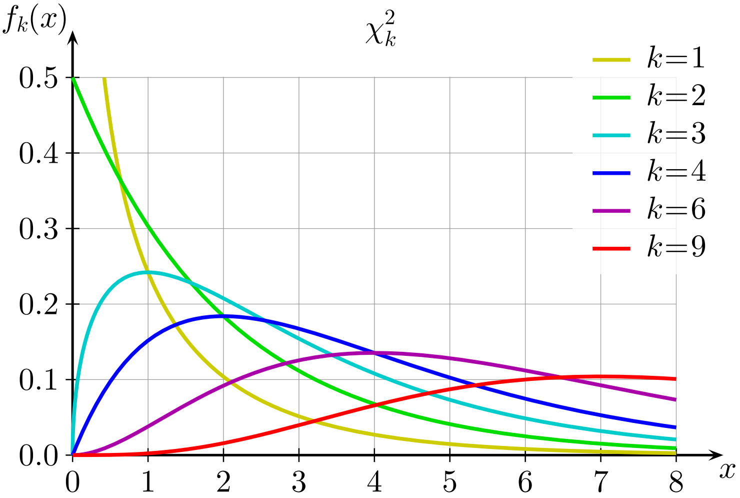
$$X = Z_1^2 + \ldots + Z_n^2 \sim \chi_k^2.$$

- If $X_1, \ldots, X_n$ are independent random variables with $\chi_{k_1}^2, \ldots, \chi_{k_n}^2$, then

$$Y = X_1 + \ldots + X_n \sim \chi_k^2,$$

$$k = \sum_{i=1}^{n} k_i.$$

χ^2 distribution

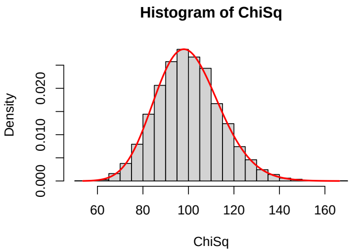


Source: Wikipedia

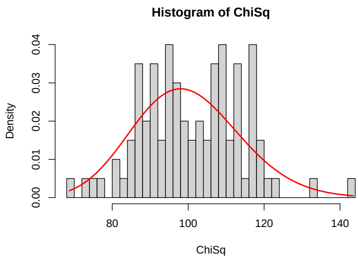
χ^2 distribution: R example

- How the distribution's shape changes with the *df*.
- How close it is to its theoretical density for different numbers of observations/draws.

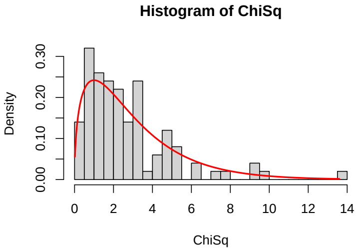
(a) $df = 100, \; n = 10,000$



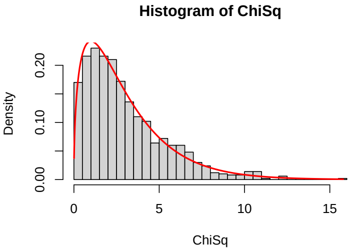
(b) $df = 100, \; n = 100$



(c) $df = 3, \; n = 100$



(d) $df = 3, \; n = 1,000$



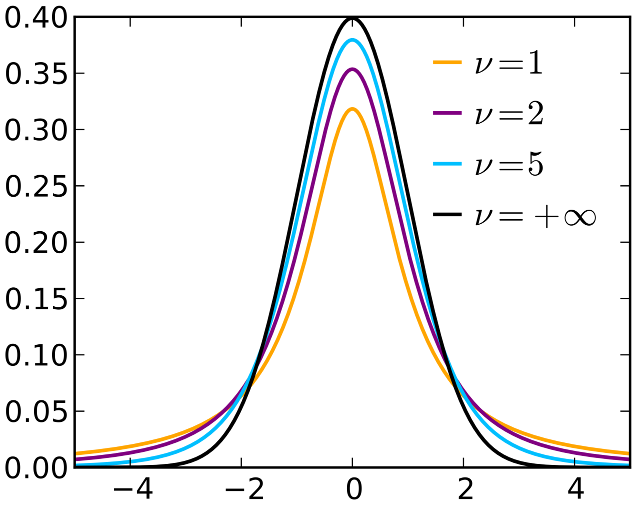
t distribution

- A.k.a. Student's t distribution (due to William Sealy Gosset's pseudonym).
- If $Z \sim N(0,1)$, $V \sim \chi^2_\nu$, Z and V are independent, then

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t_\nu.$$

- For $\nu \to +\infty$, $T \overset{as}{\sim} N(0,1)$.

t distribution

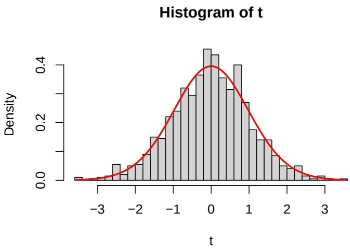


Source: Wikipedia

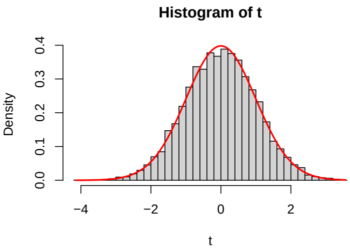
t distribution: R example

- How the distribution's shape changes with the *df*.
- How close it is to its theoretical density for different numbers of observations/draws.

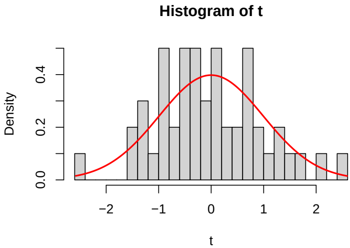
(a) $df = 33,\ n = 1,000$



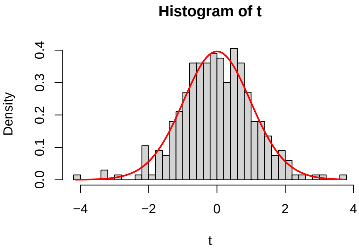
(b) $df = 100,\ n = 10,000$



(c) $df = 100,\ n = 50$



(d) $df = 33,\ n = 333$

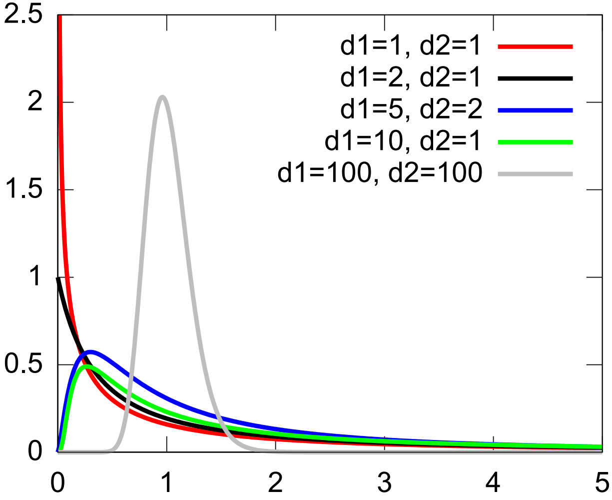


F distribution

- A.k.a. Fisher-Snedecor distribution
- If $V_1 \sim \chi^2_{d_1}$, $V_2 \sim \chi^2_{d_2}$, and V_1 and V_2 are independent, then

$$F = \frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$$

F distribution

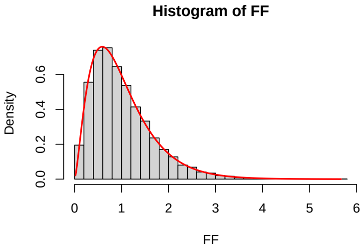


Source: Wikipedia

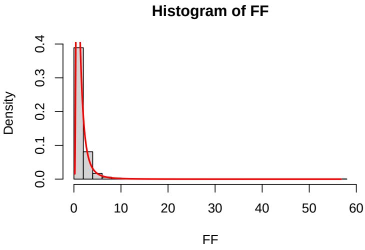
F distribution: R example

- How the distribution's shape changes with sets of *df*, asymmetry.
- How close it is to its theoretical density for different numbers of observations/draws.

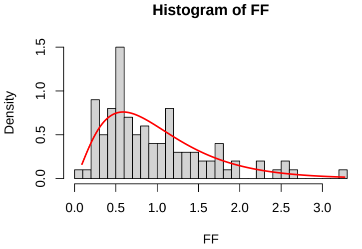
(a) $df1 = 5,\ df2 = 100,\ n = 10,000$



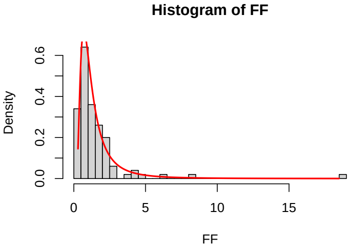
(b) $df1 = 100,\ df2 = 5,\ n = 10,000$



(c) $df1 = 5,\ df2 = 100,\ n = 100$



(d) $df1 = 100,\ df2 = 5,\ n = 100$



Outline

(Weak) law of large numbers (LLN)

- Let $X_1, \ldots, X_n$ be a sequence of *i.i.d.* random variables with mean μ and finite non-zero variance. Then,

$$\operatorname*{plim}_{n \to +\infty} \left(\bar{X}_n \right) = \mu.$$

- Alternatively, then for all $\varepsilon > 0$:

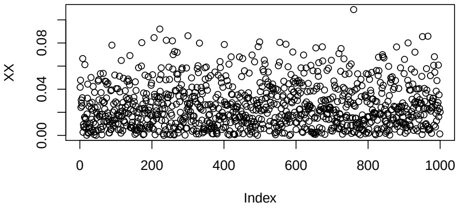
$$\lim_{n \to +\infty} P\left(\left| \bar{X}_n - \mu \right| > \varepsilon \right) = 0.$$

- Intuitively: if we are interested in estimating the population mean μ, we can get arbitrarily close to μ by choosing a sufficiently large sample.

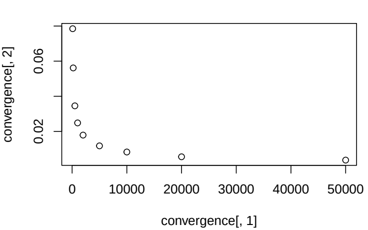
LLN: R example

- Convergence in probability (with increasing numbers of observations n) of the absolute difference between the sample mean and population mean of *i.i.d.* random variables to zero.
- Check the 'rate of convergence'.

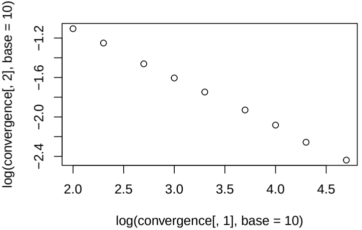
(a) simulated data: $n = 1,000$, $rep = 1,000$



(b) convergence: $\nearrow n$



(c) log-convergence



Central limit theorem (CLT)

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample with mean μ and finite non-zero variance σ^2. Then,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

  has an asymptotic standard normal distribution $N(0, 1)$.

- Intuitively: regardless of the population distribution of X, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution.

  Remarkably, the entire distribution of Z_n also gets arbitrarily close to the shape of the standard normal distribution $N(0, 1)$ as n gets large.

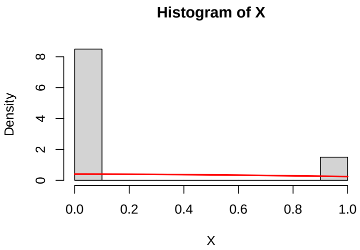
CLT: R example

- Demonstrate CLT on an extreme case: a binomial distribution.
- One trial only, probability of success $p = 0.2$:

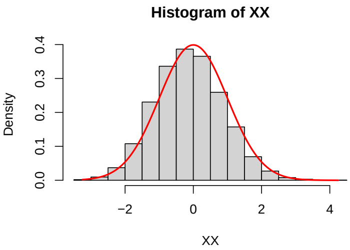
$$\begin{aligned} \mu &= p = 0.2, \\ \sigma^2 &= p(1-p) = 0.16. \end{aligned}$$

- Show 'convergence' for different n (an intuitive way of what a 'finite approximation of infinity' could be).

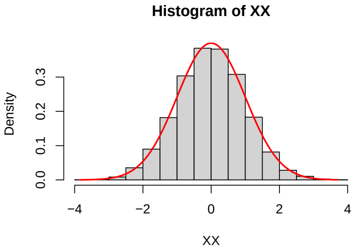
(a) $n = 100, \; rep = 1$



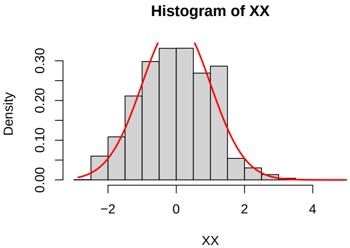
(b) $n = 100, \; rep = 10,000$



(c) $n = 10,000, \; rep = 10,000$



(d) $n = 33, \; rep = 10,000$



Outline

Hypothesis testing

- The null (H_0) and alternative (H_1) hypotheses
- Significance level α:
  - the risk/probability of incorrectly rejecting a true H_0 (Type I error)
- Test statistic:
  - function of the random sample, computed for a particular outcome
- Critical value(s):
  - quantile(s) of the distribution of the test statistic under H_0 signifying the range where H_0 is rejected for a given α
- *p*-value:
  - 'the smallest α at which H_0 is still rejected' for a given test statistic
- Type I error:
  - we reject a true H_0
  - 'conviction of the innocent' (if H_0 : innocent)
- Type II error:
  - we do not reject a false H_0 (usually referred to as β in statistics)
  - 'letting the criminal go' (if H_0 : innocent)

Hypothesis testing

Equivalent approaches (leading to the same results):

- define H_0 and H_1
- set the significance level α, e.g., 5%
- obtain a test statistic

a) through the critical value(s):
   - get the critical value(s) of the test statistic under H_0
   - if the test statistic is outside the quantile range set by the critical value(s), reject H_0

b) through the *p*-value:
   - find the complementary CDF probability $P(X > x)$ associated to the quantile equal to the test statistic (*p*-value)
   - if the *p*-value is lower than α, reject H_0

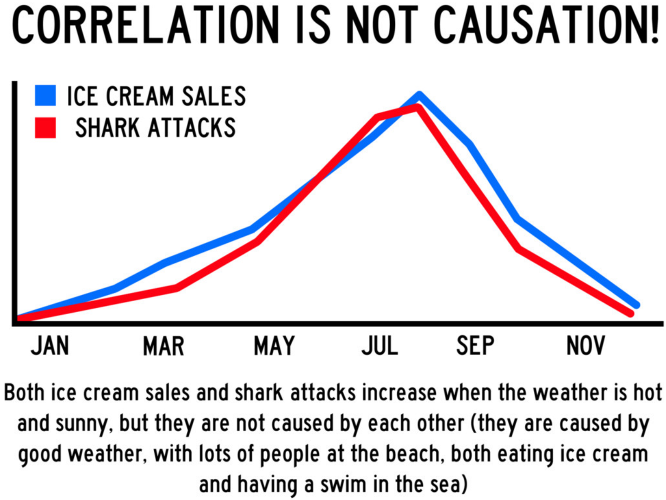
c) through the confidence interval (usually for two-sided H_1):
   - construct a CI for a respective parameter
   - if the hypothesized value under H_0 is outside the CI, reject H_0

- mind the difference between one-sided and two-sided tests

Seminars and the next lecture

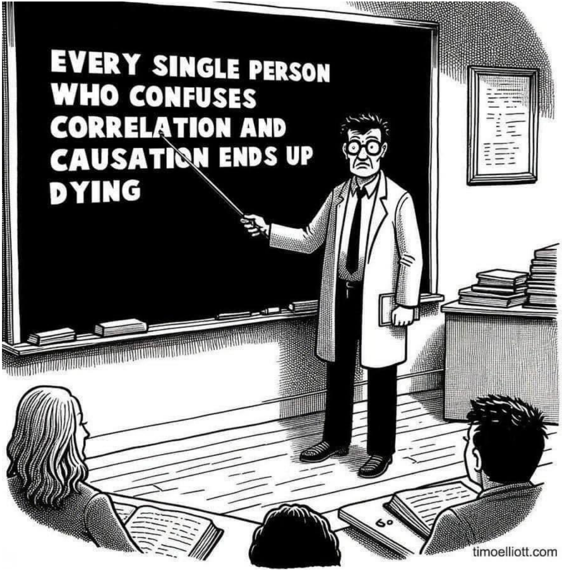
- Seminars:
  - there are no seminars this week
- Next lecture #2:
  - data structures
  - simple regression model
  - deriving the OLS estimator
- Readings for lecture #2:
  - Chapter 1: 1.3, Chapter 2: 2.1, 2.2, 2.4

Appendix I: Causality in econometrics



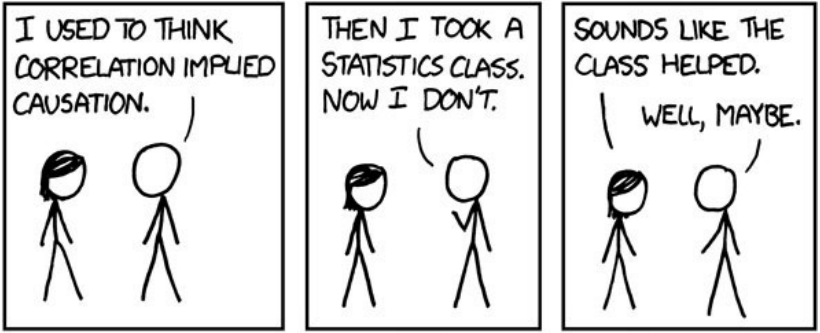
Source: idatassist.com [2021-10-11]

Appendix I: Causality in econometrics



Source: X.com [2024-02-18]

Appendix I: Causality in econometrics



Source: xkcd.com/552 [2021-10-11]

Back