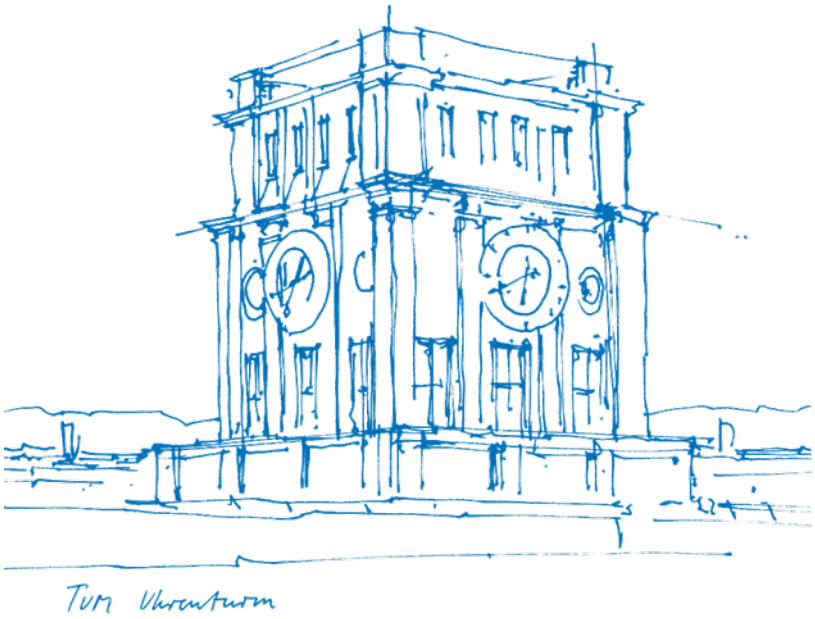
Applied Econometrics
TUM School of Management
Technical University of Munich

Empirical Research Methods - Lecture 1

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, April 17, 2024



Some ads: Student Organizations at TUM

1. **180 Degrees Consulting Munich**
2. Academy Consult
3. **Center for Digital Technology & Management (CDTM)**
4. Society of Sommeliers
5. TUfast | Moto electric
6. TUM Blockchain Club Lecture

Organizational Issues

Organizational Issues

Lecture: Wednesday 13.15-14.45, Room 0980 (Audimax).
Note: On Wednesday, June 5, 2024, the lecture will be in room 2750.

Tutorials: (starting in the second week)

1. Monday, 16.45-18.15, room 1402
2. Tuesday, 13.15-14.45, room 2750 (video recording scheduled)
3. Thursday, 11.30-13.00, romm 2760
4. Friday, 9.45-11.15, room 1402

Tutors: Rebecca Groh and Gabriel Vollert.
Due to the public holiday and the TUM Sustainability Day, the lectures on **May 1** and **June 12** will not take place, however, the tutorials in these weeks will be held as usual. Further, the tutorials on **May 9**, **May 20**, and **May 30** are cancelled due to public holidays. If your class is affected by one of those holidays, visit one of the other three classes in that respective week.

Organizational Issues

Tutorials:

In the tutorials you will be required to use *R studio*, which is an interface for using *R*. Both *R* and *R studio* can be used for free.

You can find an installation video and a brief introduction on our TUM-Moodle page.

Please bring your own computer to the tutorials (with R already installed) and try to solve the problem sets **before** the tutorials.

Organizational Issues

Discussion Forum on Moodle:

Use this forum to ask and answer questions, and discuss topics of the lecture and class with your fellow students.

Our staff is only passively monitoring the forum and might pick up questions that were asked there for discussion in class.

Organizational Issues

Replication Weeks and Grade Bonus:

During the semester, you will be able to participate in 2 replication assignments. The task will be to replicate some results from published articles using the statistical package R.

If you successfully participate in **both** replication weeks, you will obtain a grad bonus of 0.3 in the final exam.

We will offer additional support in either of these two assignments. The dates will be announced soon.

Organizational Issues

Grading: Written exam on Thursday July 25, 2024; 08:30 - 10:30; on campus

Office hour: by arrangement, office.econometrics@mgt.tum.de
Room 2413 (0504.02.413)

Course homepage: via TUM-Moodle

Organizational Issues

We will briefly discuss fundamental statistical concepts in the first two weeks as well as in the first tutorials.

There will be some overlap to your course in introductory statistics.

The focus of this course will be more on identification of causal effects.

You will see that once identification is established, estimation is often very simple.

Anything else? You can also write us an email: office.econometrics@mgt.tum.de

Organizational Issues

About me

Ph.D. in Economics (2012) from University of Munich (LMU)

Head of the Health Econometrics Unit at the Max Planck Society until 2020

Professor of Microeconometrics at University of Mannheim and Professor of Economics at LMU.

Professor for Applied Econometrics at TUM since 2021

Research interests: Econometrics, Data Science and Machine Learning

Applications in Applied Microeconometrics, in particular Health, Labour, Experimental and Insurance Economics

Course topics

1. Empirical Research & Econometrics
2. (Multivariate) Random Variables
3. Linear Regressions
   - 3.1 The Population Model
   - 3.2 Assumptions
   - 3.3 Identification & Estimation of Parameters
4. Interpretation of Parameters
5. Hypothesis Testing
6. Omitted Variables & Panel Data Estimation
7. Program Evaluation & Identification Strategies
8. Probit & Logit Regressions

If time permits, we might look at limited dependent variable models and/or instrumental variable estimation.

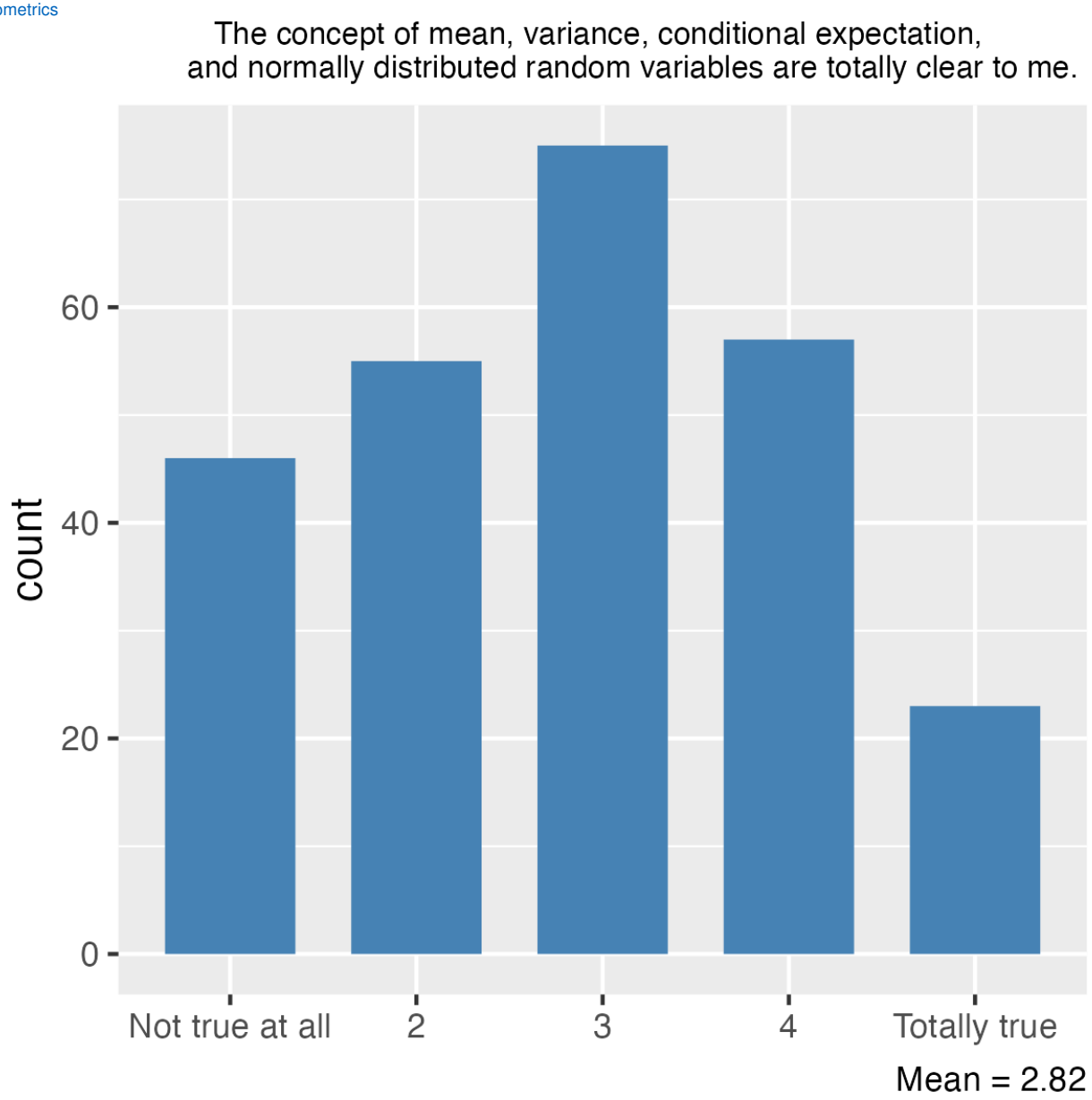
Helpful literature

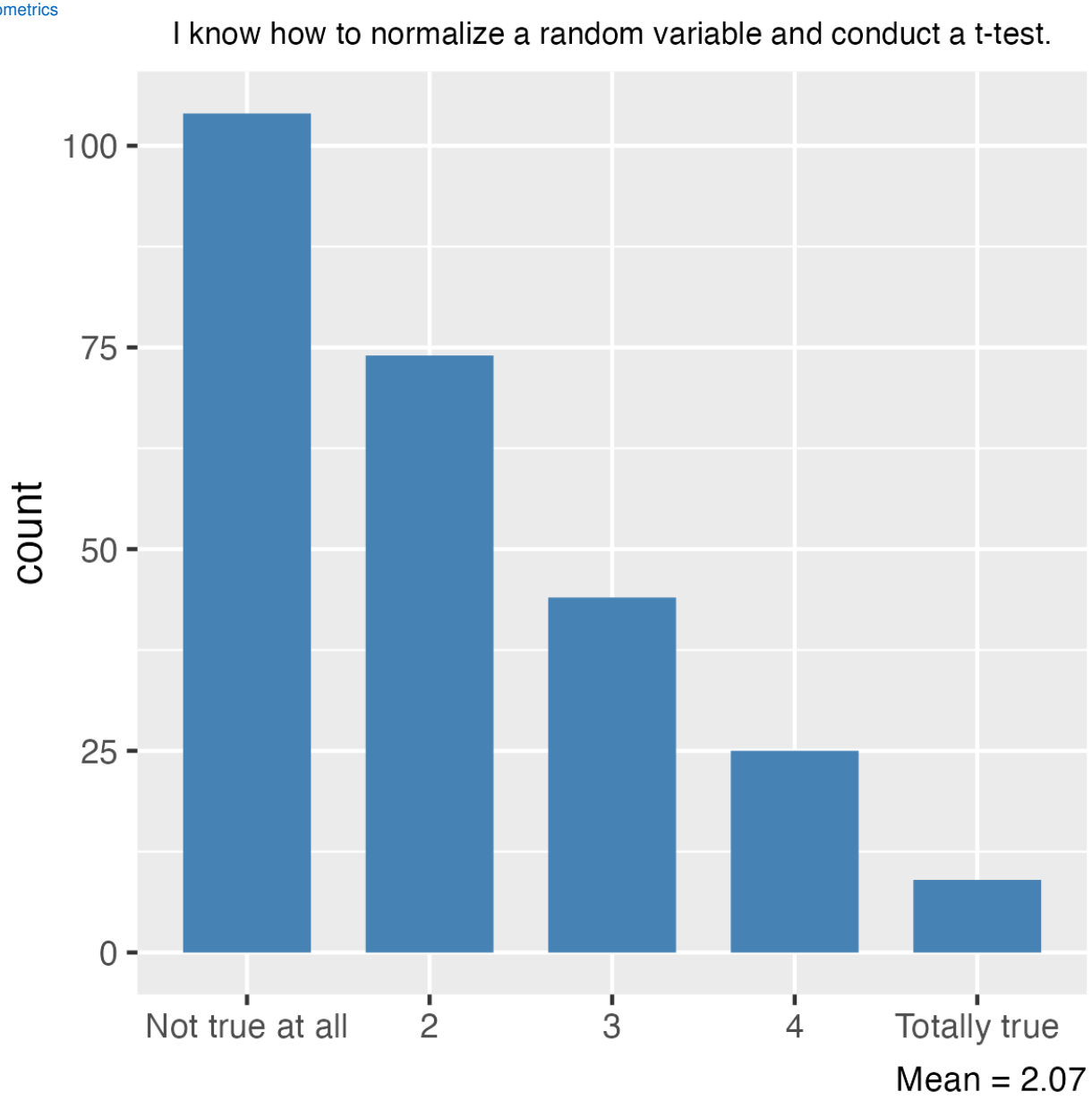
Main reference:

Cunningham: Causal Inference. Available here
`https://mixtape.scunning.com/index.html`

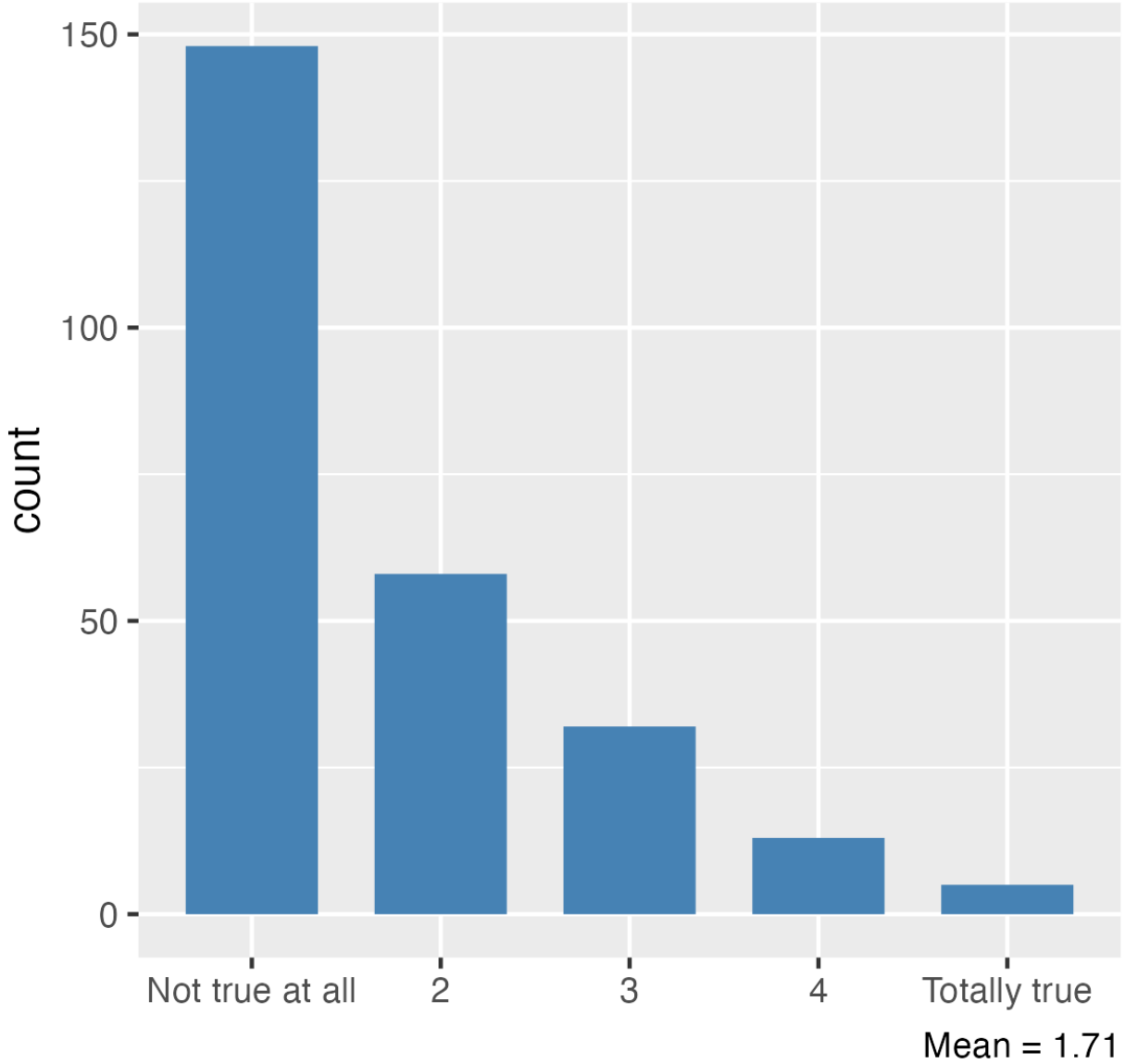
Very advanced:

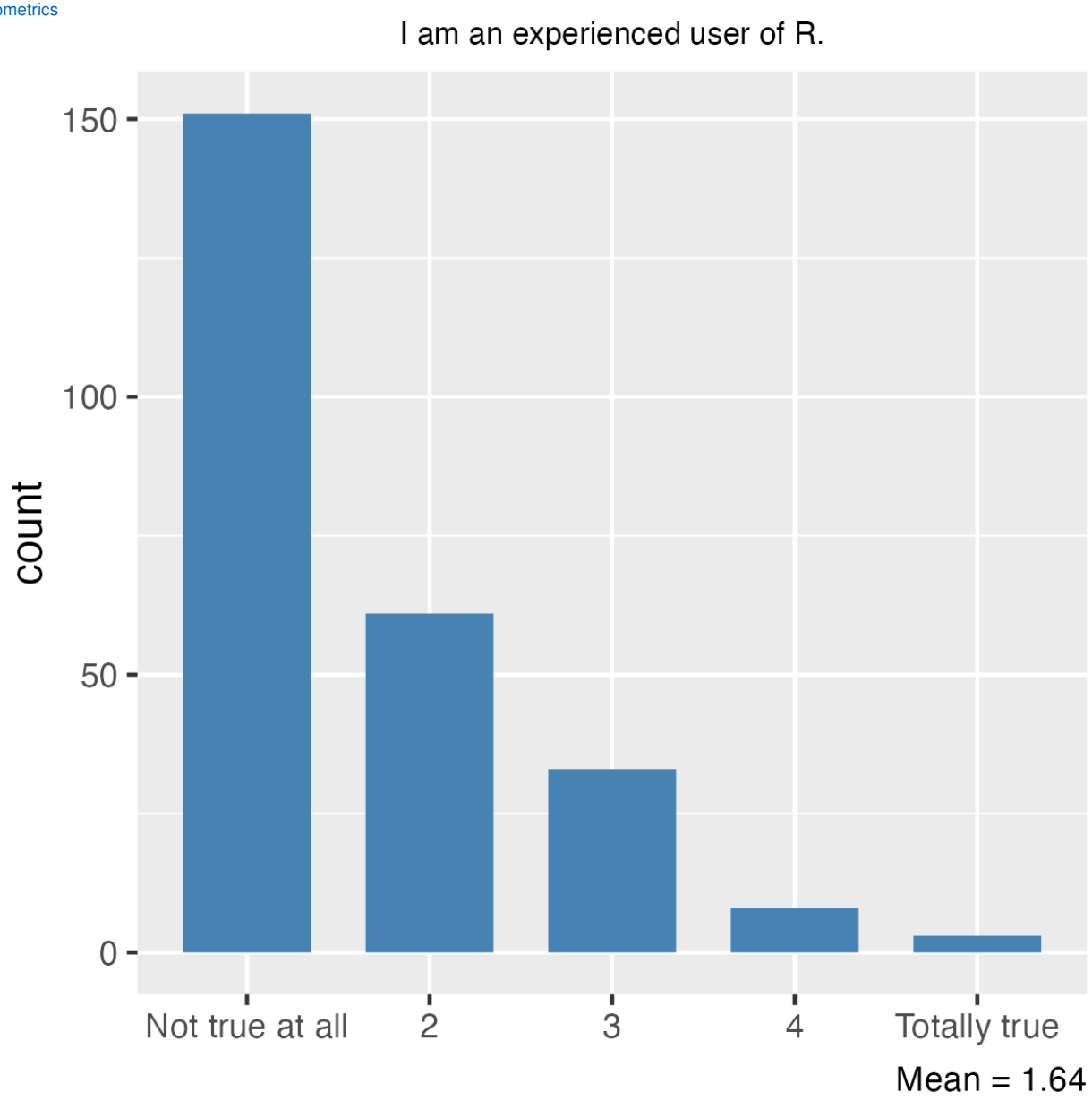
Hansen: Econometrics. Relevant pages available via TUM-Moodle


The concept of mean, variance, conditional expectation,
and normally distributed random variables are totally clear to me.
count
60
40
20
0
Not true at all
2
3
4
Totally true
Mean = 2.82


I know how to normalize a random variable and conduct a t-test.
count
100
75
50
25
0
Not true at all
2
3
4
Totally true
Mean = 2.07

I know how to derive the OLS estimator and calculate standard errors.




I am an experienced user of R.
150
100
50
0
count
Not true at all
2
3
4
Totally true
Mean = 1.64

Outline of today's lecture

Why Empirical Research?
What is Econometrics?
Main Types of Datasets

Why Empirical Research?

Why Empirical Research?

Theory often provides little guidance about sign and/or magnitudes of economic relationships.

A good example is **private returns to education**: say, expected wage gain for one additional year of schooling.

We may agree that this value is positive (w/o using any data) but the exact number we can only learn from data.

Analysis of data plays crucial role in understanding economic behavior (e.g. investments in education).

Almost always, our data is a (random) sample of the population (for example, private returns to education for **people working in Germany**).

Why Empirical Research?

Population: Group or collection of possible entities of interest (e.g. school districts, German population). We will think of populations as infinitely large.

Key idea of sampling: We don't want to collect or cannot collect data on every member of a population.

Therefore, we draw objects at random from the population (with equal probability).

We learn about the population mean (aka expected value) from the sample mean.

Why Empirical Research?

Different aims of sampling

A) Punishment and deterrence device: for example, fare dodging or speeding tickets

In case A, the main aim is to affect the behavior of the observations in the sample.

B) Drawing conclusions about the entire population:
- a) How many items are broken?
- b) Expected wage gain for one additional year of schooling?

In case B, we actually do **not** really care about the **mean wage** in our sample, we just use it to learn the **expected wage** in the population.

To do so, we need econometrics.

What is Econometrics?

What is Econometrics?

Use of statistical methods to analyze economic data.

Estimate relationships between economic variables.

Econometrics involves knowledge about

- economic theory
- your dataset
- statistics (& very recently also data science)

What is Econometrics?

Two main goals in econometrics:

1. **Prediction or forecasting**
   For example, Amazon: “Customers who bought this item also bought ...”
   A forecast is a prediction about the value of a variable in the future,
   like GDP next year or weather tomorrow.

2. **Estimating causal effects**
   Causality means that a certain action or treatment (e.g., smaller class sizes)
   leads to a specific consequence (e.g., more clever students).

We do not need causality to make a good prediction/forecast. Example:
Observing that people use an umbrella allows you to “predict”: It is raining.
But, of course, the act of using an umbrella does not cause it to rain.

What is Econometrics?

In empirical research we are often interested in causal effects. For example,

- by how much do we expect sales to fall if we increase prices by 1%?
- by how much do we expect crime to fall if we rise compulsory schooling by one year (social returns to education)?

Ideally, we would have an **experiment**: manipulate x, measure y.

Causal (average) effect of x on y:

- $\mathrm{E}(y|x = \mathit{hot}) - \mathrm{E}(y|x = \mathit{cold})$ if x is binary
- $\frac{\partial \mathrm{E}(y|x)}{\partial x}$ if x is continuous

What is Econometrics?

In economics we rarely have **experimental data**
(so called, **R**andomized **C**ontrolled **T**rials).

For several reasons (you will learn in my course), experiments are often considered a gold standard in research.

In economics, this discussion goes back to (at least) LaLonde's famous critique (LaLonde, AER 1986).

Analyzing experimental data is (almost) always very **easy**. The main difficulty is setting up the experiment in the right way.

In recent years more and more economists use experimental data.

What is Econometrics?

What is Econometrics?

In economics we often analyze **observational** data. That is we observe (economic) behavior & have no chance to manipulate x.

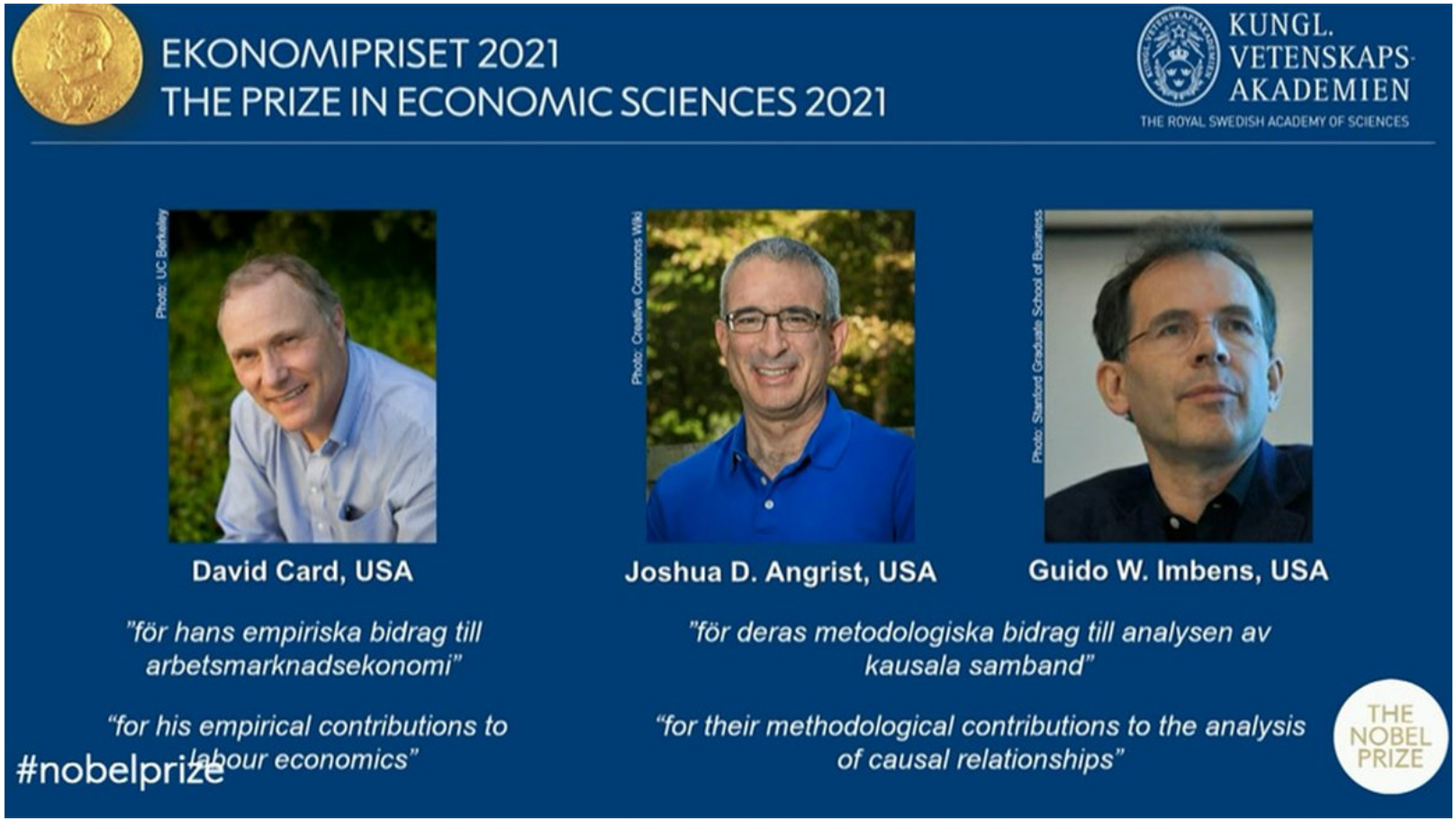
Here, we often resort to *ceteris paribus* analysis: $E(y|x, \mathbf{c})$
Holding all other factors **c** fixed, we estimate the causal effect of x.

Analyzing observational data is (almost) always very **hard**.

Many of the methods we are going to discuss are (primarily) intended to learn **causal effects** from observational data.

There is still a lot of research being done in this field.

What is Econometrics?



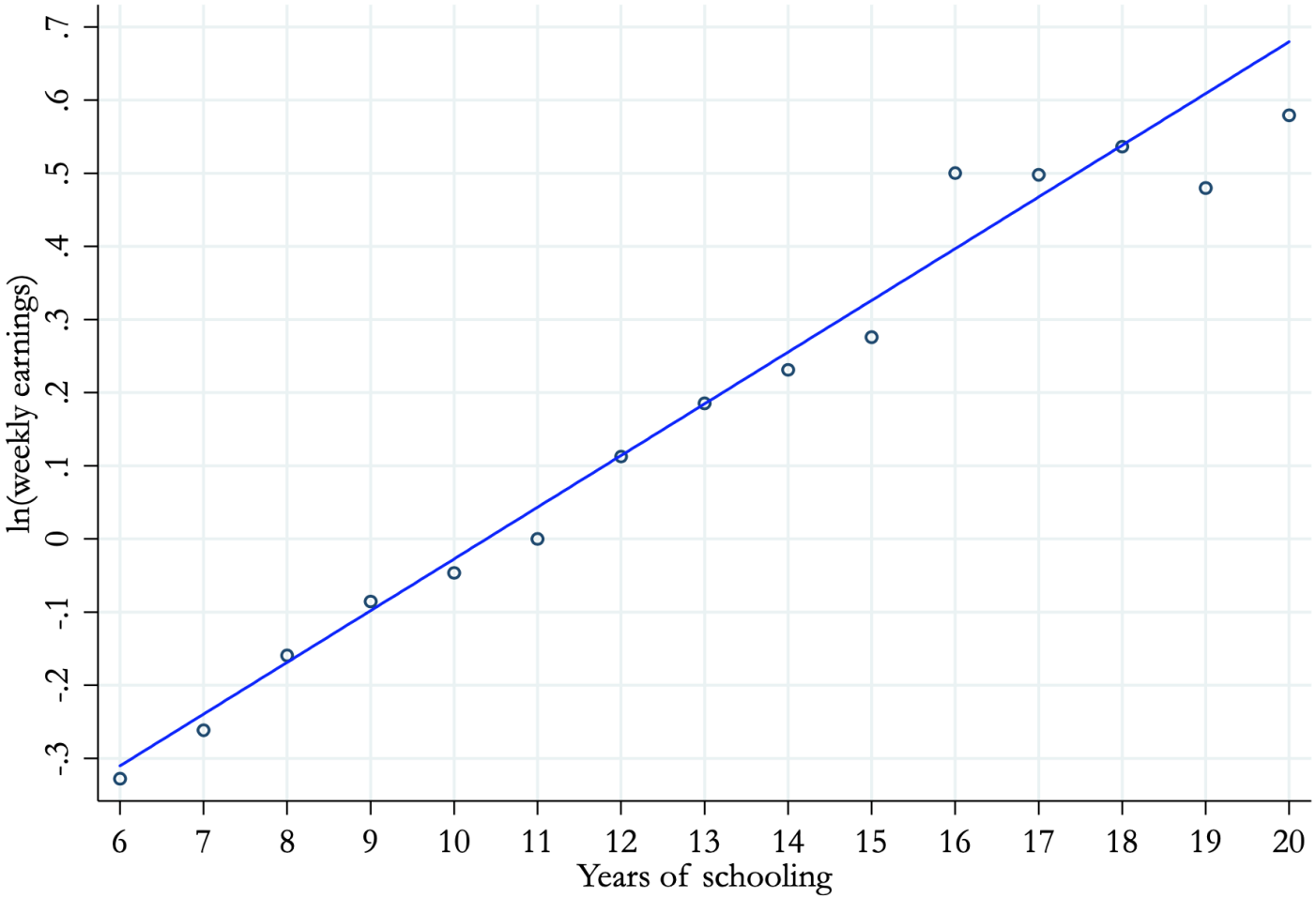
What is Econometrics?

Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

ANSWERING CAUSAL QUESTIONS USING OBSERVATIONAL DATA

The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel

Figure 1: The cross-sectional relationship between earnings and schooling



Notes: The figure is based on the data used by Angrist and Krueger (1991). The data set comes from the 1980 US Census. It covers men born 1930-1939. Log earnings for individuals with 11 years of schooling are normalized to 0.

Suppose we have access to observational data for a large number of individuals on their earnings, Y_i, and whether they have completed high school, $D_i = 1$, or not, $D_i = 0$. The observed average difference in earnings across the two groups, is an estimate of $\Delta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$. It is unlikely that Δ equals the causal effect of interest, however. To see this formally, subtract and add $E[Y_i(0)|D_i = 1]$, a counterfactual term, which yields:

$$\underbrace{\Delta}_{\text{Difference in means}} = \underbrace{E\big[\big(Y_i(1) - Y_i(0)\big)\big|D_i = 1\big]}_{\text{causal effect on high school completers}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

The first term in this expression, $E\big[\big(Y_i(1) - Y_i(0)\big)\big|D_i = 1\big]$, is the causal effect of interest. In our high school example, it provides the answer to the question: What is the effect on wage earnings of completing high school, among those that did so? The second term, $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$, represents selection bias. It measures how different the earnings would have been in the two populations — completers and non-completers — had they not completed high school. For reasons described above, high school completers would likely earn more than non-completers, even without high-school, implying that the selection effect is positive in this example. In such a case, the comparison of means, Δ, provides an upward biased estimate of the causal effect of high school education on earnings.

Main Types of Datasets

Main Types of Datasets

Examples of datasets in Management and Economics:

- Micro Census Data (e.g. from the US or Germany)
- Survey data (e.g. SOEP or Share)
- Administrative data from social security, insurances, banks, etc.
- Process data collected in the internet: Google Analytics, Amazon, Uber, etc.

Main Types of Datasets:

- Cross sections
- Panels
- Repeated (pooled) cross sections
- Time series

Main Types of Datasets

Cross-sectional datasets:

Sample of individuals, households, firms, countries, etc., each observed **only once**.
That is, $N > 1$, $T = 1$, where
$N =$ number of observations and $T =$ number of time periods (e.g., years).

Every row is one observation (e.g., individuals or firms).
Every column is one variable (e.g., characteristics of an individual like age or income).

The order of the rows/observations is arbitrary and unimportant.
Columns we need to distinguish into outcomes and explanatory variables.

Observations are typically considered to be independent of each other.

We can study differences across people or firms during a single time period.

Example of Cross-sectional data

The nlsw88.dta data set contains information on wages, education and job tenure.

idcode NLS id	**age** age in current year	**wage** hourly wage	**grade** current grade completed	**collgrad** college graduate
1	37	11.739125	12	0
2	37	6.400963	12	0
3	42	5.016723	12	0
4	43	9.033813	17	1
6	42	8.083731	12	0
7	39	4.629630	12	0
9	37	10.491142	12	0
12	40	17.206116	18	1
13	40	13.083735	14	0

Main Types of Datasets

Panel datasets:

We observe our cross-sectional units (individuals, firms, etc) over more than one time period (months, years).

That is, $N > 1$, $T > 1$; individuals are observed at multiple points in time.

Usually, the dataset is sorted by individuals and within each individual sorted by time.

This is particularly important for more advanced panel data estimation techniques.

We can study differences across people or firms and over time (additional source of variation).

Example of Panel Data Sets

Key difference to repeated cross section: Same units (people, houses, etc.) are followed over time. We can use more advanced estimation techniques (later more). The wagepan.dta data set contains panel data on wages.

id	year	lwage	educ	hours	married	agric	black	exper
1	1985	1.69989097	14	2864	0	0	0	6
1	1986	-0.72026259	14	2994	0	0	0	7
1	1987	1.66918790	14	2640	0	0	0	8
2	1985	1.60858822	13	2164	0	0	0	9
2	1986	1.57238543	13	2749	0	0	0	10
2	1987	1.82033384	13	2476	0	0	0	11
3	1985	2.26666212	12	2340	1	0	0	9
3	1986	2.06994390	12	2340	1	0	0	10
3	1987	2.87316084	12	2340	1	0	0	11

Main Types of Datasets

Repeated (pooled) cross sections:

Several cross sections drawn independently over time.

That is $N > 1$, $T > 1$ individuals are not observed at multiple points in time.

Crucial difference to panel data is that we cannot follow cross-sectional units (individuals) over time.

Repeated cross sections are usually easier to collect than panel datasets.

Often used to evaluate policy changes.

Example of Pooled Cross Sections

cps78_85.dta contains information on education and wages from 1978 and 1985.

id	year	lwage	educ	age	married	exper	female
1	1978	1.2150	12	25	0	8	0
2	1978	1.6094	12	47	1	30	1
3	1978	2.1401	6	49	1	38	0
4	1978	2.0732	12	36	1	19	0
5	1978	1.6490	12	28	1	11	0
6	1978	1.7148	8	56	1	43	0
7	1978	1.0986	11	18	0	2	0
8	1978	1.8326	15	29	0	9	1
9	1978	0.3567	16	38	0	17	1
10	1978	2.1547	15	43	1	23	0
11	1985	2.1972	10	43	1	27	0
12	1985	1.7047	12	38	1	20	0
13	1985	1.3350	12	22	0	4	1
14	1985	2.3514	12	47	1	29	1
15	1985	2.7080	12	58	1	40	0

Main Types of Datasets

Time series data:

Repeated observation of the same variable (examples: GDP, prices, interest rate).

That is, $N = 1, T > 1$

Order of data is important (every row contains a specific time period).

Observations are typically not independent over time (fuel price yesterday, today and tomorrow are strongly related).

Mainly used to forecast future values (say, the oil price tomorrow).

We won't talk about time-series econometrics in this course.

Example of Time Series Data

The data set phillips.dta includes annual unemployment and inflation rates for the US economy.

year	unem	inf	inf_1	unem_1
1949	5.9	-1.2	8.1	3.8
1950	5.3	1.3	-1.2	5.9
1951	3.3	7.9	1.3	5.3
1952	3.0	1.9	7.9	3.3
1953	2.9	0.8	1.9	3.0
1954	5.5	0.7	0.8	2.9
1955	4.4	-0.4	0.7	5.5
1956	4.1	1.5	-0.4	4.4
1957	4.3	3.3	1.5	4.1
1958	6.8	2.8	3.3	4.3
1959	5.5	0.7	2.8	6.8

Example of Time Series Data

Symptomatic COVID-19 cases per 100,000, by vaccination status and age group.

Meldejahr	Meldewoche	Ungeimpfte 18–59 Jahre	Grundimmunisierte 18–59 Jahre	Ungeimpfte 60+ Jahre	Grundimmunisierte 60+ Jahre
2022	1	195.97	187.38	108.29	55.56
2022	2	185.99	189.28	77.38	47.39
2022	3	224.91	208.47	82.05	53.20
2022	4	254.31	210.56	90.42	53.74
2022	5	268.82	209.18	110.80	56.13
2022	6	258.74	226.40	112.91	65.61
2022	7	224.07	202.98	101.75	65.66
2022	8	201.43	174.59	97.23	60.90
2022	9	167.89	174.26	86.24	60.43
2022	10	164.66	173.43	87.51	55.89
2022	11	127.92	141.52	76.48	50.77

Recommended reading

To repeat this week please read chapter 1 (Introduction).

For next week please read chapter:

2.8 Expected value
2.9 Variance
2.10 Covariance

`https://mixtape.scunning.com/index.html`

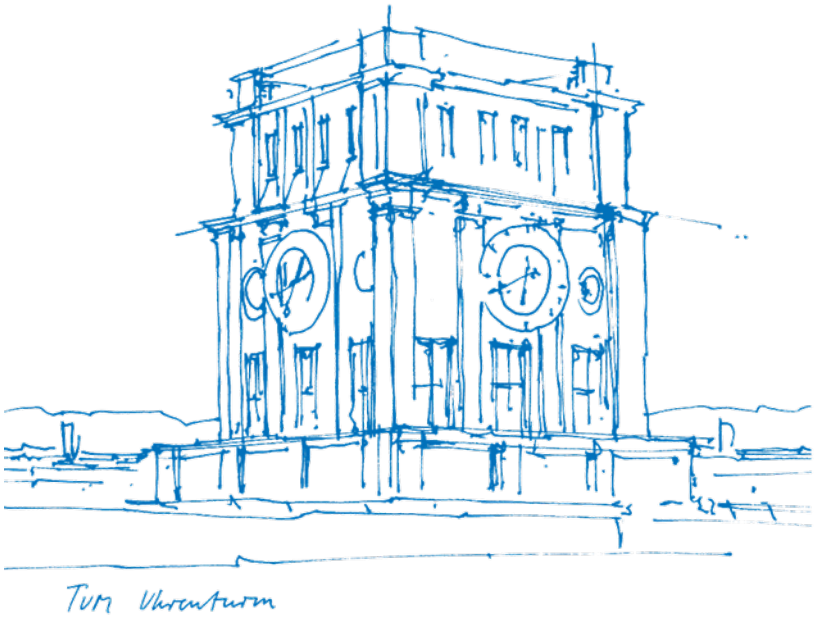
Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de

Applied Econometrics
TUM School of Management
Technical University of Munich

Empirical Research Methods - Lecture 2

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, April 25, 2024



Outline of today's lecture

Random Variables

Concept of Random Variables

Random variables are variables whose realization is uncertain beforehand. Your grade in an exam not only depends on your preparations but also on the difficulty of the exam and many other factors.

Discrete random variables take on a discrete set of values, like $0, 1, 2, 3, \ldots$, e.g. years of education.

Continuous random variables take on a continuum of possible values, e.g. wages.

The (cumulative) probability distribution characterizes a random variable.

Random Variables

Distribution of Random Variables

Distributions of Discrete Random Variables

In case of *discrete* random variables, the **probability distribution** is just a list or table that links any possible value the random variable can take on to the probability of this value occuring. This is called the **probability mass function (pmf)**. The probabilities of all the values sum up to 1.

The **cumulative distribution function (cdf)** is the probability that the random variable takes on values at most as large as a particular value.

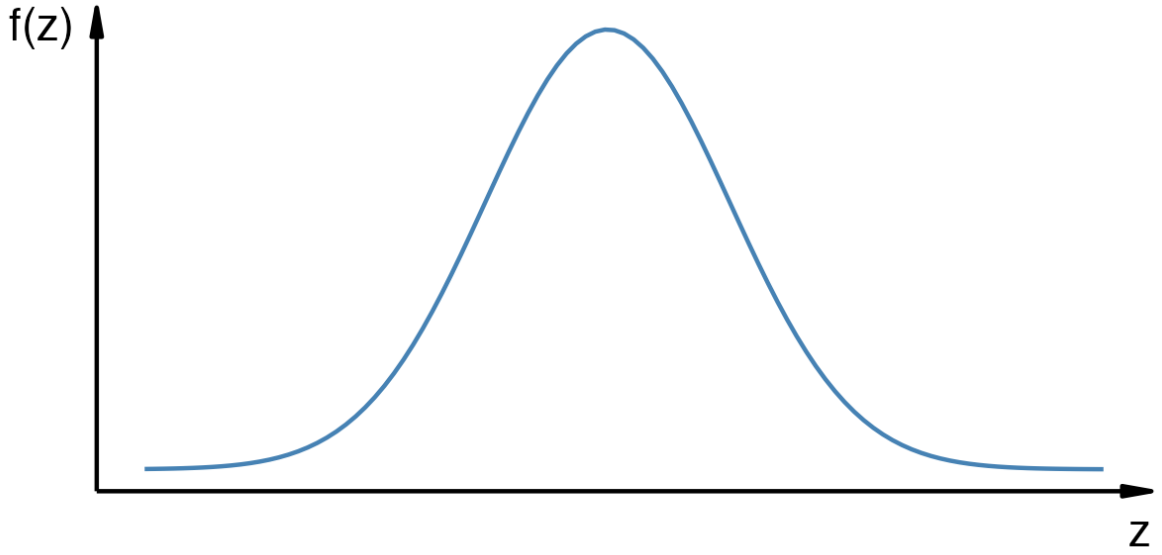
Example: number of students skipping the next lecture (in a course with 30 students)

# of students	pmf	cdf
0	0.20	0.20
1	0.05	0.25
2	0.10	0.35
...	...	...
30	0.01	1

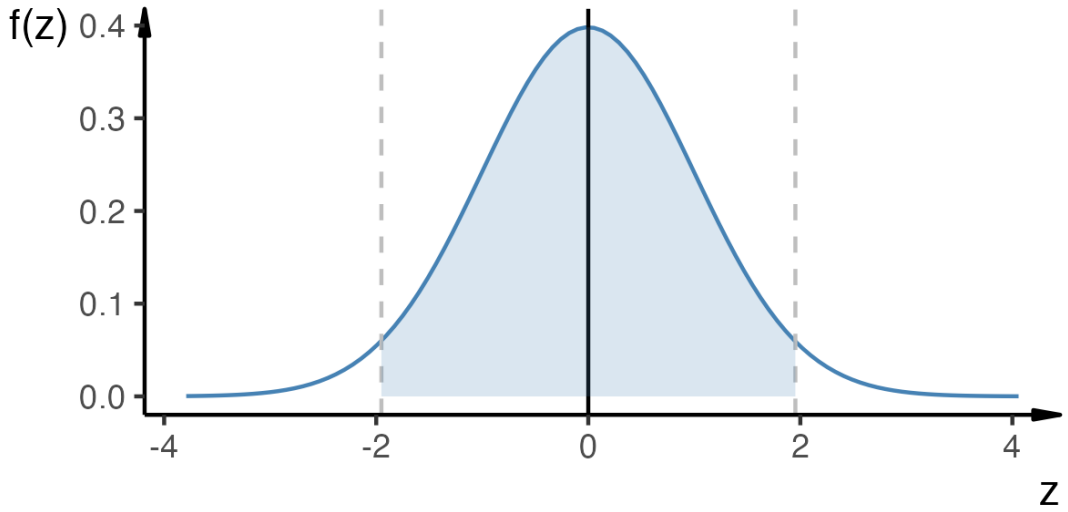
Distribution for Continuous Random Variables

For *continuous* random variables, one cannot produce a list with probabilities for each outcome.

Instead we use the **probability density function, or pdf**, denoted as $f(z)$



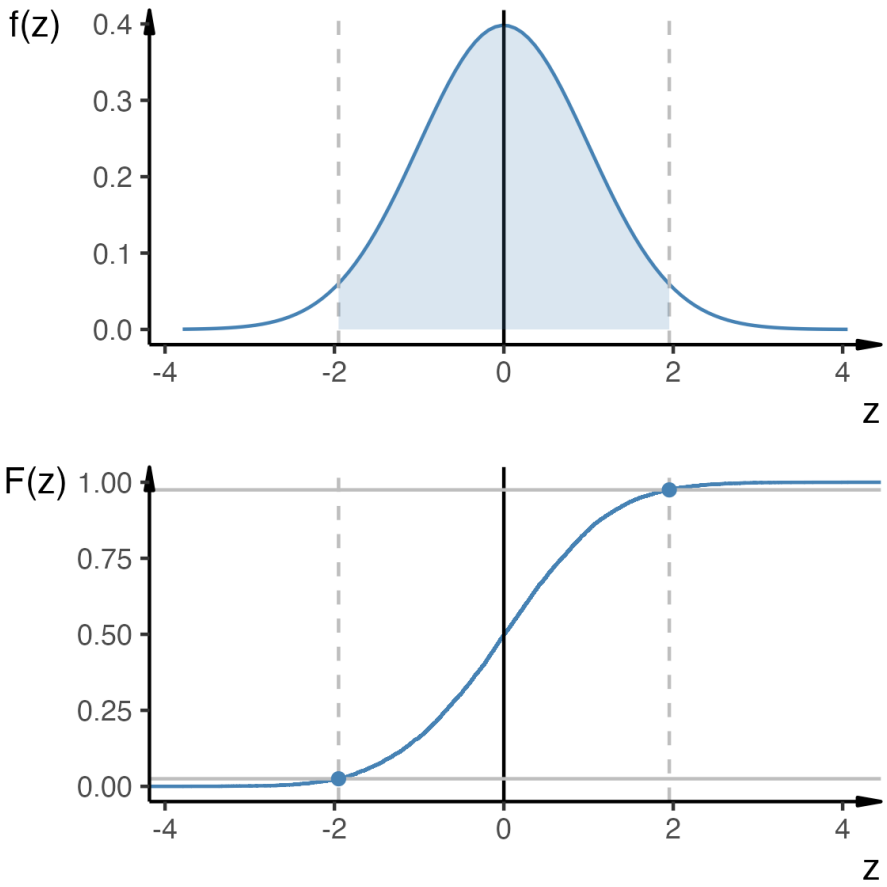
Distribution for Continuous Random Variables



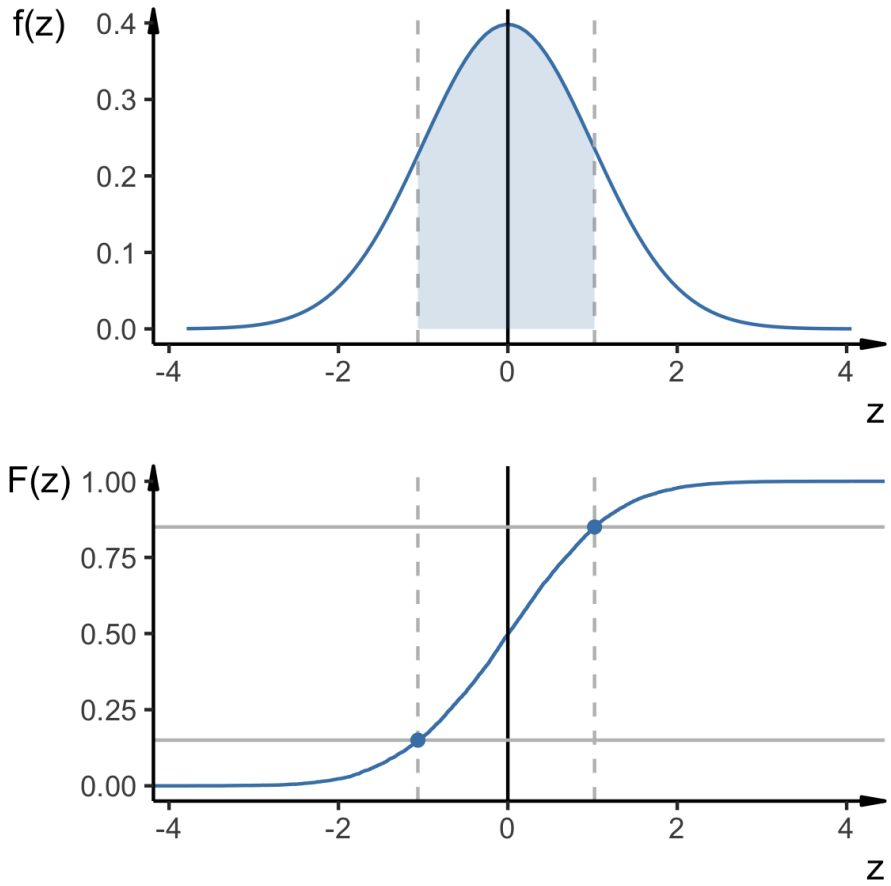
For any two points, a and b, the probability that the random variable Z falls in the interval between a and b is equal to the integral of the pdf f(z) from a to b .

$P(a < Z < b) = \int_a^b f(z)\mathrm{d}z = F(b) - F(a)$, where $F(z) = P(Z < z)$ denotes the **cumulative distribution function** (cdf).

Distribution for Continuous Random Variables



Distribution for Continuous Random Variables



Random Variables

Moments of a Distribution

Moments of a Distribution

The distribution contains all information about a random variable.

However, it is often cumbersome to present.

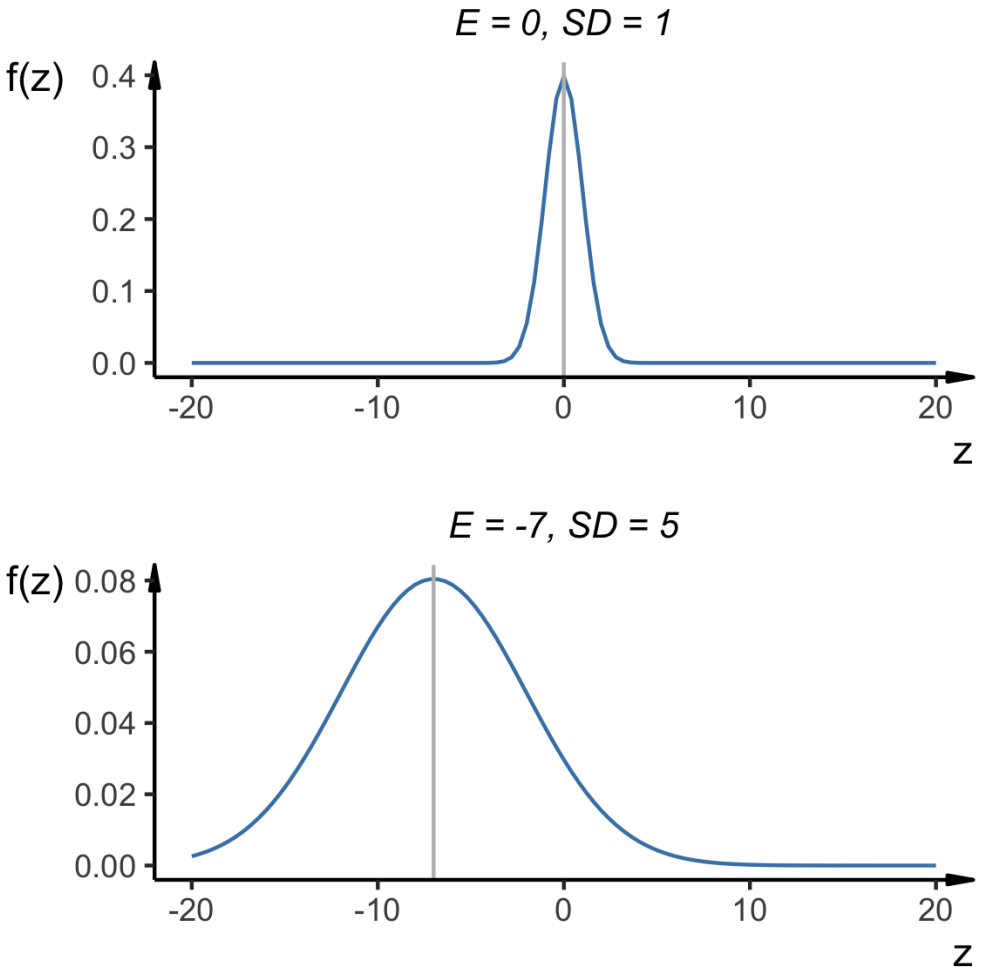
Part of the information can be summarised in functions (aka **moments**) of a random variable.

The most important moments are the **exptected value** and the **variance**.

The expected value (E) measures the **location** of the distribution.

The variance (Var) and standard deviation (SD) measure the **dispersion** of the distribution around its expected value.

Moments & Distributions



Expected Value

The expected value (sometimes also called population mean) is the long-run average of a random variable if you repeat a random process (infinitely) often.

Suppose a **discrete** random variable Y takes on k possible values and p_j denotes the probability that Y takes on the value y_j:

$$\begin{aligned} \mathrm{E}(Y) &= y_1 p_1 + y_2 p_2 + \ldots y_k p_k \\ &= \sum_{j=1}^{k} y_j p_j \end{aligned}$$

Example: Expected value when tossing a coin (with $H = 1$ for head and $H = 0$ for tail):

$$\begin{aligned} \mathrm{E}(H) &= \mathrm{E}(H|H=1)P(H=1) + \mathrm{E}(H|H=0)P(H=0) \\ &= 1 \times p + 0 \times (1-p) = p \end{aligned}$$

Expected Value

The expected value of a **continuous** random variable also is a probability-weighted (long-run) average.

Summations have to be replaced by integrals:

$$\mathrm{E}(Y) = \int y\, f(y)\, \mathrm{d}y$$

where $f(y)$ denotes the probability density function of Y.

Variance and Standard Deviation

Let the expected value be $\mathrm{E}(Y) = \mu_Y$.

The variance of a random variable Y is defined as

$$\mathrm{Var}(Y) = \mathrm{E}\left[(Y - \mu_Y)^2\right] = \ldots = \mathrm{E}(Y^2) - \underbrace{\mathrm{E}(Y)^2}_{\mu_Y^2}$$

and often denoted by σ_Y^2.

The standard deviation then is $\mathrm{SD} = \sqrt{\sigma_Y^2} = \sigma_Y$.

Moments & Distributions

So far, we've talked about population values. That is,

- we have not used any data
- and we did not estimate anything so far.

Population values are generally *unobserved* but they do exist.

The main aim of empirical research is to use data and appropriate assumptions to estimate population values.

Example: average income in Germany.

Random sampling

Population and Random Samples

Population: Group or collection of possible entities of interest (e.g. school districts, German population). We will think of populations as infinitely large.

Key idea: We don't want to collect data on every member of the population. Therefore, we draw *n* objects at random from the population (with equal probability).

We can learn the population mean (aka expected value) from the sample mean.

Population and Random Samples

Suppose you draw at random from the German population and ask about income.

Every resident has the same probability of being drawn $\frac{1}{82{,}000{,}000}$

Let y_i denote the income of the i-th drawn person i.

Because all observations in the sample are selected at random, the values of $y_1, \ldots, y_n$ are themselves random.

Example for i.i.d. Sampling

In this lecture, we restrict ourselves to independently and identically distributed, or "i.i.d", draws from the population.

Identically distributed:
$y_1, \ldots, y_n$ are randomly drawn from the same population.

Independently distributed:
Knowing the value of y_1 has no informational content for the value of y_2.
Income of Heinz Schmidt living in Berlin tells us nothing about the income of Helmut Farbmacher living in Munich.

i.i.d. sampling simplifies many theoretical discussions. Keep in mind, most datasets are indeed not iid. Often, the methods we will discuss work nevertheless (the theoretical proofs are just less intuitive).

Sample Mean and its Distribution

Sample Mean and its Distribution

The **sample mean** is calculated as:

$$\begin{aligned} \bar{y} &= \frac{1}{n}(y_1 + y_2 + \ldots + y_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}$$

Because of random sampling, $\bar{y}$ is itself a random variable.

The value of $\bar{y}$ differs from one randomly drawn sample to the next (e.g., mean income across several random samples).

Because $\bar{y}$ is a random variable, it has itself a probability distribution (called sampling distribution) and we can also think about the expected value and the variance of $\bar{y}$.

Expected value of $\bar{y}$

Suppose that y_i are i.i.d. draws with expectation μ and variance σ^2.

What is the expected value of $\bar{y}$?

$$\mathrm{E}(\bar{y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

So, $\bar{y}$ is an **unbiased** estimator.

Remember the rules:
$\mathrm{E}(X+Y+Z) = E(X)+E(Y)+E(Z)$
$\mathrm{E}(bX) = b\mathrm{E}(X)$ with b being a constant.

Variance of $\bar{y}$

What is the variance of $\bar{y}$?

$$\mathrm{Var}(\bar{y}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(y_i) = \frac{\sigma^2}{n}$$

Why is it important that y_i's are independently distributed in this step?

Remember the rules:
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$
$\mathrm{Var}(bX) = b^2\,\mathrm{Var}(X)$ with b being a constant.

The **standard error** of $\bar{y}$ is the square root of $\mathrm{Var}(\bar{y})$:

$$\mathrm{SE} = \sigma/\sqrt{n}$$

Variance & standard error of the sample mean thus decrease with the sample size.

Law of Large Numbers and Consistency

The **law of large numbers** states that if

1. iid observations $y_1, y_2, \ldots, y_n$
2. $\mathrm{E}(y) = \mu$
3. $\mathrm{Var}(y) = \sigma^2 < \infty$,

Under these conditions, the **sample mean will be really close to the expected value/population mean** if the number of observations (n) in the sample **is large**.

We say, $\bar{y}$ converges in probability to μ or we simply say $\bar{y}$ is **consistent**, and write

$$\bar{y} \xrightarrow{p} \mu$$

Law of Large Numbers and Consistency

Consistency:

$$\bar{y} \xrightarrow{p} \mu$$

Intuitively, the variance of the sample mean (σ^2/n) goes to zero if the numbers of observation n increases, and $\bar{y}$ will very likely fall close to its expected value μ.

Unbiasedness:

$$\mathrm{E}(\bar{y}) = \mu$$

Intuitively, if we calculate the mean for several random samples of the population, then the average of these means will be very close to its expected value μ.

Finite-Sample vs Asymptotic Distribution

So far, we just considered the central tendency and the variance of $\bar{y}$.
But what can we say about the distribution of $\bar{y}$?
Here we need to distinguish two cases.

Case 1: If y_i is normally distributed, ...

The sum of normally distributed variables is again normally distributed.

This means we know the **exact** or **finite-sample distribution** of the sample mean $\bar{y}$ in this case because $\bar{y}$ is a (scaled) sum of the y_i's.

$\bar{y}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$, we write

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Finite-Sample vs Asymptotic Distribution

Case 2: If y_i is **not** normally distributed, ...

The exact sampling distribution of $\bar{y}$ may be very complicated.

Theoretical results, however, allow us to find an **asymptotic distribution**.

Asymptotic theory applies if the number of observations goes to infinity (in practice it means we have a large sample).

Case 2 is far more relevant in applied economics.

Central Limit Theorem

The (Lindeberg-Levy) **central limit theorem** states that if

1. iid observations $y_1, y_2, \ldots, y_n$
2. $\mathrm{E}(y) = \mu$
3. $0 < \mathrm{Var}(y) = \sigma^2 < \infty$,

Under these conditions, the distribution of the sample mean will be arbitrarily well approximated by a normal distribution with mean μ and variance σ^2/n if the number of observations (n) in the sample **is large**. We write

$$\bar{y} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad \bar{y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right).$$

We get the **asymptotic distribution** of $\bar{y}$ irrespective of the distribution of y_i.

Central Limit Theorem

We can standardise the sample mean such that it converges in distribution to a standard normal distribution $N(0,1)$:

$$\sqrt{n}\left(\bar{y}-\mu\right) \xrightarrow{d} N\left(0,\sigma^2\right)$$

or

$$\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N\left(0,1\right)$$

This result is very convenient since it allows us to perform hypothesis tests.

Remember the rules:
$\mathrm{E}(\bar{y}-\mu)=\mu-\mu=0$
$\mathrm{Var}(\sqrt{n}\bar{y}/\sigma)=n\,\mathrm{Var}(\bar{y})/\sigma^2=1$

Hypothesis Tests

Hypothesis Tests

Suppose we want to test the null hypothesis $H_0 : \mu = \mu_0$ vs the alternative hypothesis $H_1 : \mu \neq \mu_0$ for a specific value μ_0.

If the null hypothesis is true (i.e., μ is indeed equal to μ_0), then the t-statistic has to be (asymptotically) standard normally distributed:

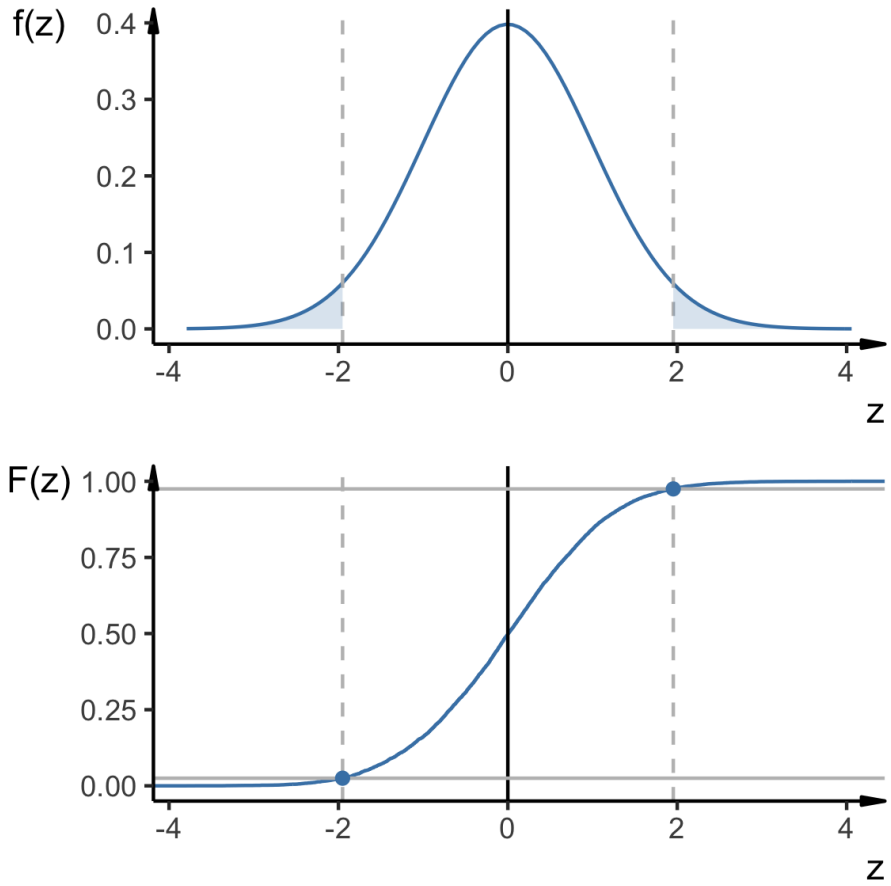
$$t = \frac{\bar{y} - \mu_0}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Note that we need to replace the unknown variance σ by a consistent estimate $\hat{\sigma}^2$. We can obtain such a consistent estimate by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

If $|t| > 2.58$ (or $|t| > 1.96$ or $|t| > 1.64$), we can reject the null hypothesis H_0 at the 1% (or 5% or 10%) level.

Hypothesis Tests



Recommended reading

For next week please read chapter:

2.11 Population model

2.12 Mean independence

in `https://mixtape.scunning.com/index.html`

and

2.1 Introduction

2.2 The Distribution of Wages

in `https://www.ssc.wisc.edu/~bhansen/econometrics/`

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de