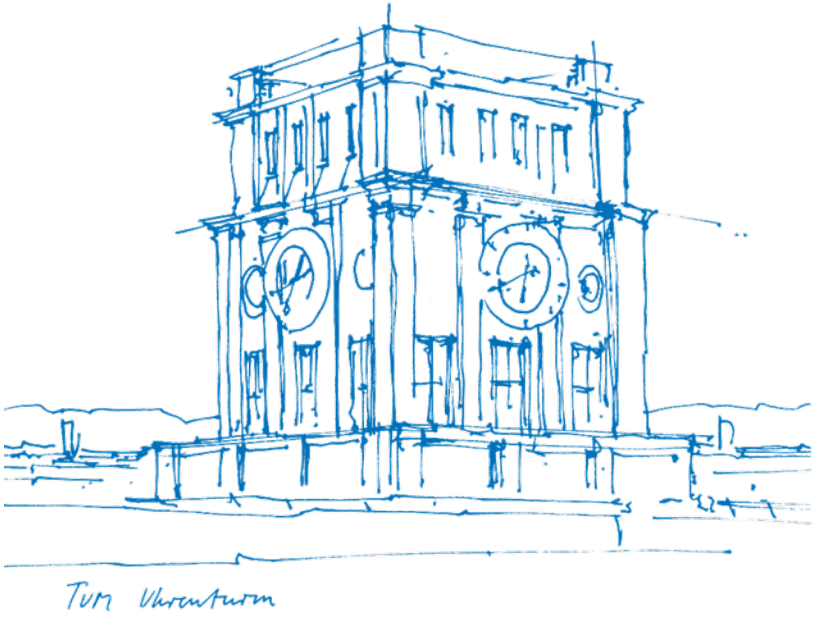
Applied Econometrics
TUM School of Management
Technical University of Munich

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

- Distribution of Random Variables
- Moments of a Distribution

Random sampling

Sample Mean and its Distribution

Hypothesis Tests

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 $0, 1, 2, 3, \dots$, 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.

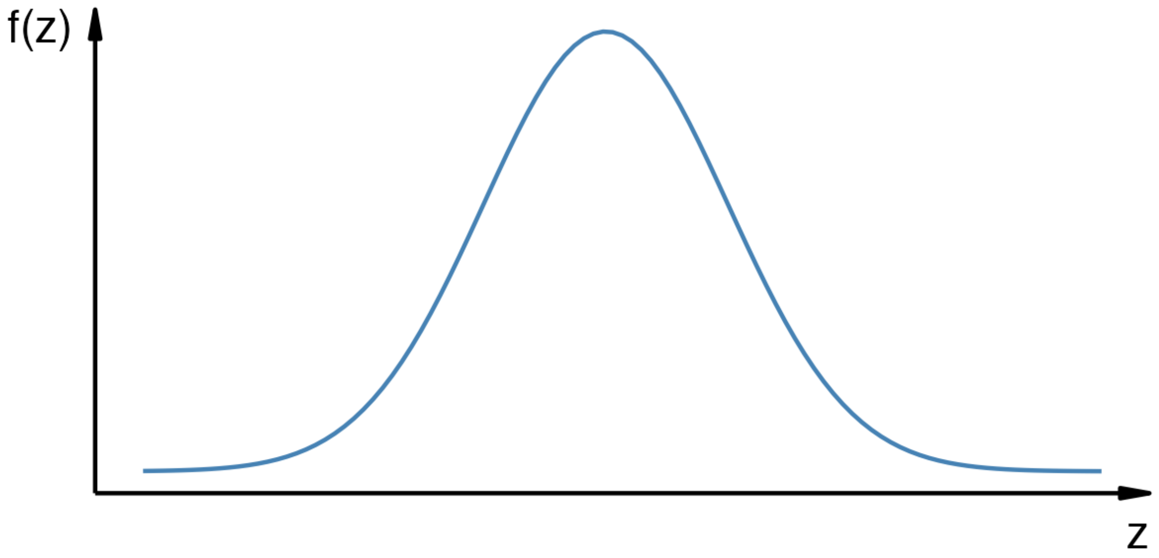
Example: number of students skipping the next lecture (in a course with 30 students)

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

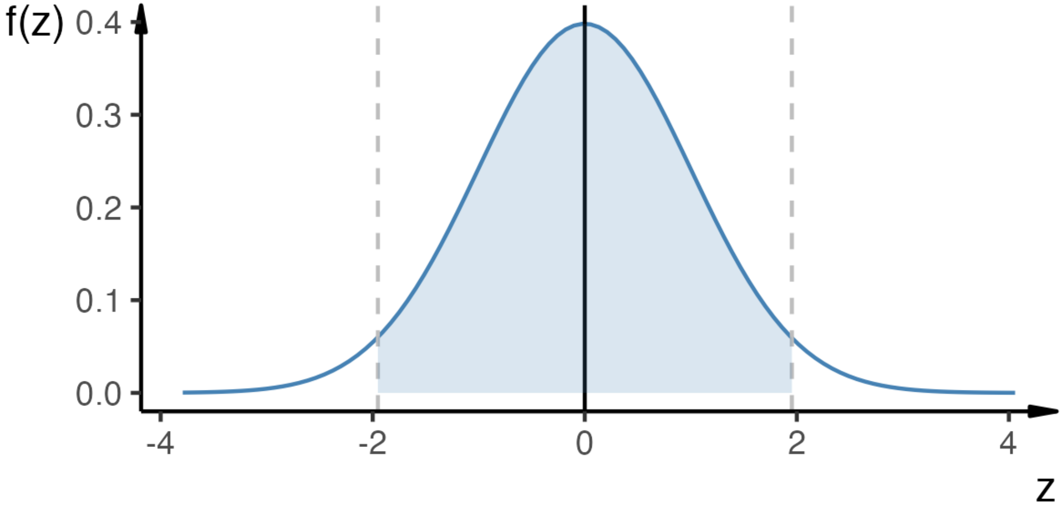
Distribution for Continuous Random Variables

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 $f(z)$



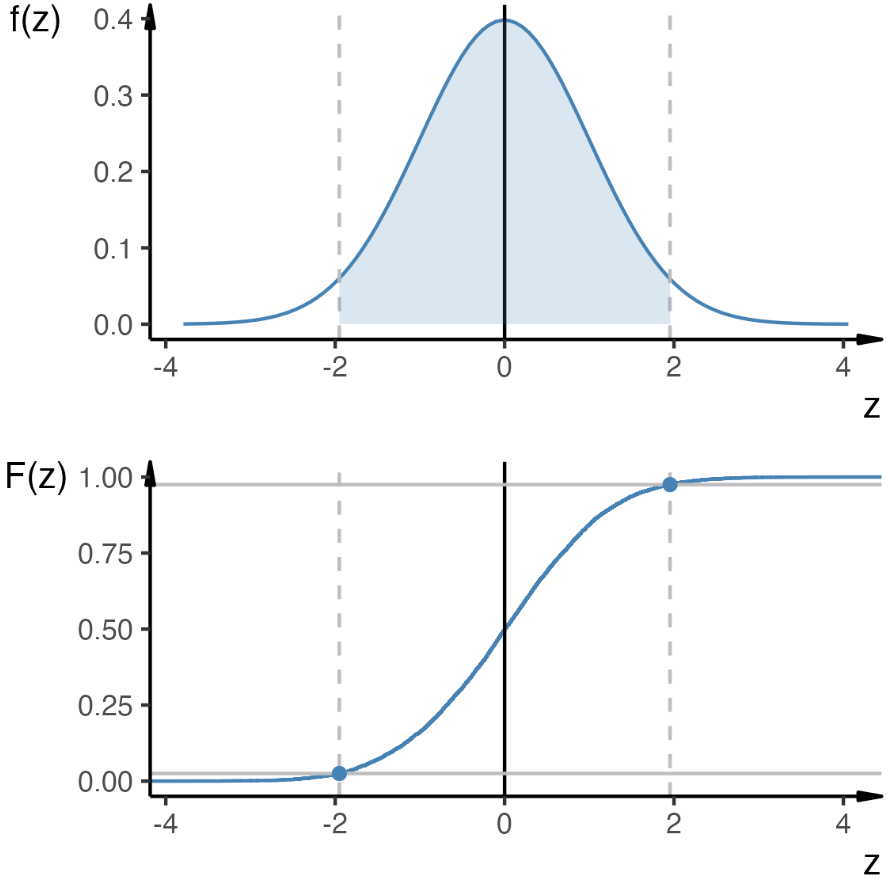
Distribution for Continuous Random Variables



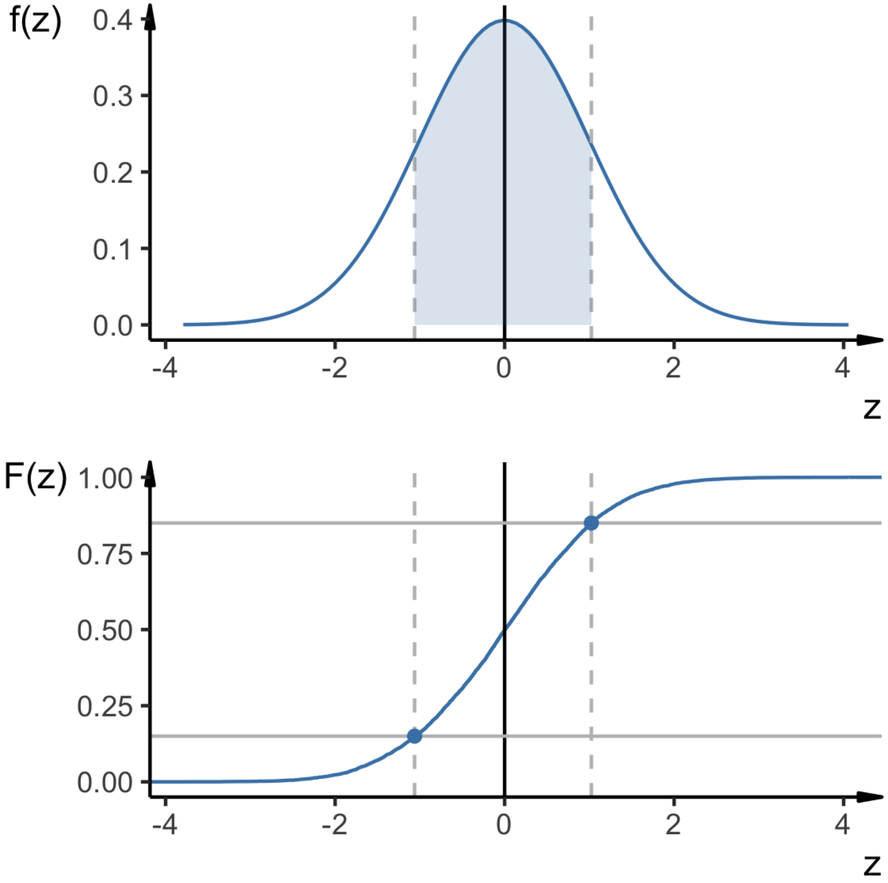
For any two points, *a* and *b*, the probability that the random variable *Z* falls in the interval between *a* and *b* is equal to the integral of the pdf f(z) from *a* to *b* .

$P(a < Z < b) = \int_a^b f(z)\mathrm{d}z = F(b) - F(a)$, 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.

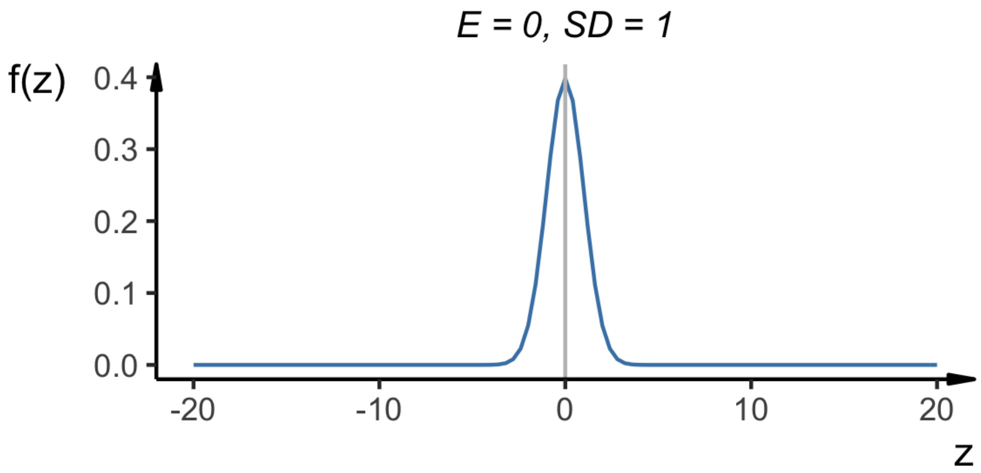
Part of the information can be summarised in functions (aka **moments**) of a random variable.

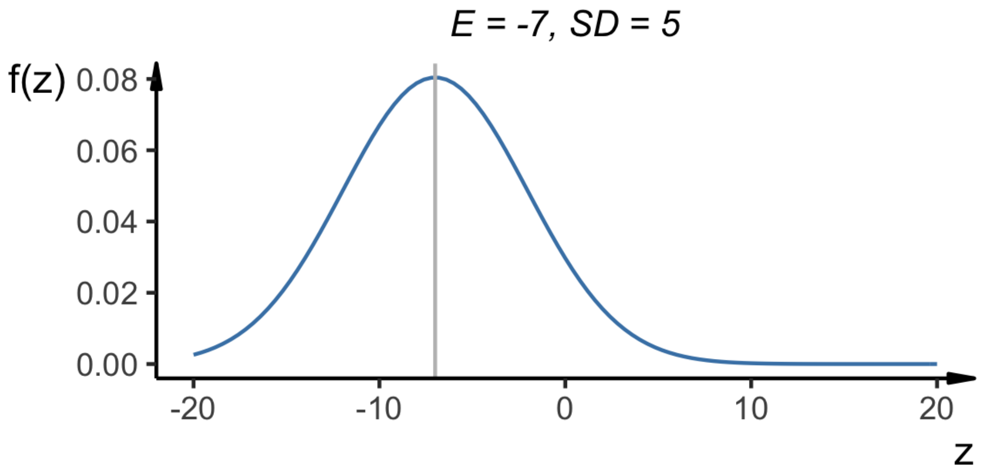
The most important moments are the **exptected value** and the **variance**.

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:

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

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:

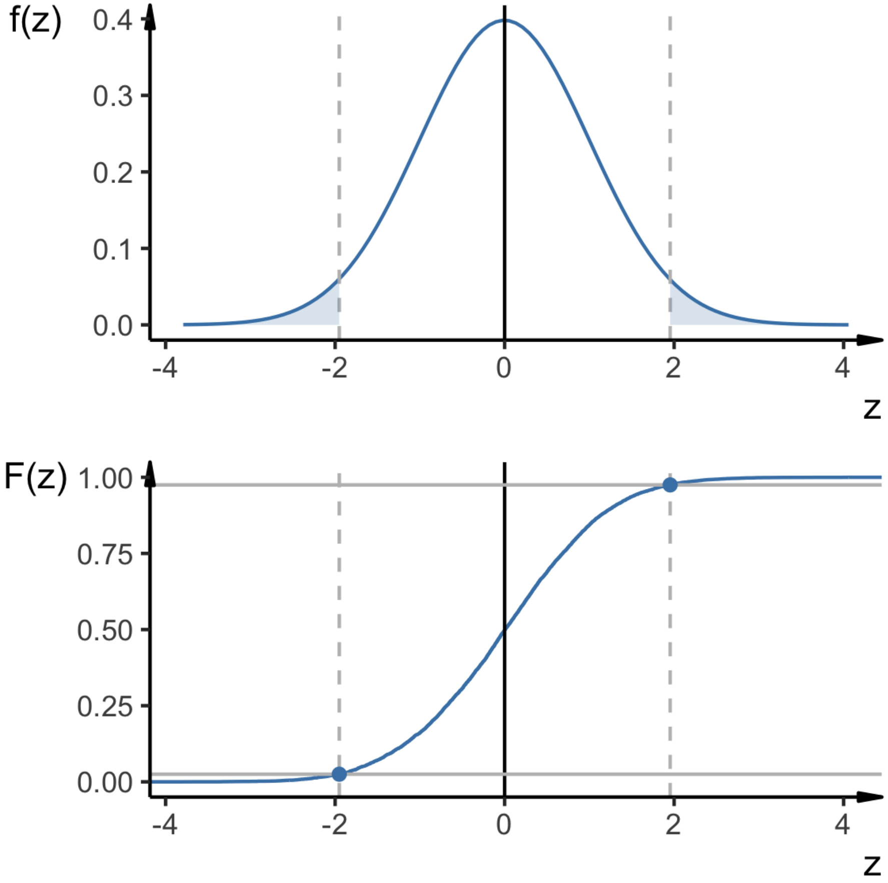
$$t = \frac{\bar{y} - \mu_{\mathbf{0}}}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N\big(0, 1\big).$$

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

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

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