Probability Theory for Econometricians

Sven Otto

February 27, 2025

Table of contents

Welcome

This tutorial gives a short introduction to the most important basic concepts of probability theory and statistics for econometricians.

This tutorial is still under construction. The two sections presented here are the first two sections of my course Statistics for Data Analytics from Winter Term 2023, which contains a review of probability theory.

For a quick review of the basics, I recommend sections 2 and 3 of Stock and Watson (2019): LINK

1 Distribution

1.1 Random Experiment

From an empirical perspective, a dataset is just a fixed array of numbers. Any summary statistic we compute – like a sample mean, sample correlation, or OLS coefficient – is simply a function of these numbers.

These statistics provide a snapshot of the data at hand but do not automatically reveal broader insights about the world. To add deeper meaning to these numbers, identify dependencies, and understand causalities, we must consider how the data were obtained.

A **random experiment** is an experiment whose outcome cannot be predicted with certainty. In statistical theory, any dataset is viewed as the result of such a random experiment. While individual outcomes are unpredictable, patterns emerge when experiments are repeated.

The gender of the next person you meet, daily fluctuations in stock prices, monthly music streams of your favorite artist, or the annual number of pizzas consumed – all involve a certain amount of randomness and emerge from random experiments. Probability theory gives us the tools to analyze this randomness systematically.

1.2 Random experiment

From an empirical perspective, a dataset is just a fixed array of numbers. Any summary statistic we compute – like a sample mean, sample correlation, or OLS coefficient – is simply a function of these numbers.

These statistics provide a snapshot of the data at hand but do not automatically reveal broader insights about the world. To add deeper meaning to these numbers, identify dependencies, and understand causalities, we must consider how the data were obtained.

A **random experiment** is an experiment whose outcome cannot be predicted with certainty. In statistical theory, any dataset is viewed as the result of such a random experiment.

The gender of the next person you meet, daily fluctuations in stock prices, monthly music streams of your favorite artist, or the annual number of pizzas consumed – all involve a certain amount of randomness and emerge from random experiments.

1.3 Random Variables

A **random variable** is a numerical summary of a random experiment. An **outcome** is a specific result of a random experiment. The **sample space** S is the set/collection of all potential outcomes.

Let's consider some examples:

- *Coin toss:* The outcome of a coin toss can be "heads" or "tails". This random experiment has a two-element sample space: $S = \{heads, tails\}$. We can express the experiment as a binary random variable:

$$Y = \begin{cases} 1 & \text{if outcome is heads,} \\ 0 & \text{if outcome is tails.} \end{cases}$$

- *Gender:* If you conduct a survey and interview a random person to ask them about their gender, the answer may be "female", "male", or "diverse". It is a random experiment since the person to be interviewed is selected randomly. The sample space has three elements: $S = \{female, male, diverse\}$. To focus on female vs. non-female, we can define the female dummy variable:

$$Y = \begin{cases} 1 & \text{if the person is female,} \\ 0 & \text{if the person is not female.} \end{cases}$$

  Similarly, dummy variables for *male* and *diverse* can be defined.

- *Education level:* If you ask a random person about their education level according to the ISCED-2011 framework, the outcome may be one of the eight ISCED-2011 levels. We have an eight-element sample space:

$$S = \{Level\ 1, Level\ 2, Level\ 3, Level\ 4, Level\ 5, Level\ 6, Level\ 7, Level\ 8\}.$$

The eight-element sample space of the education-level random experiment provides a natural ordering. We define the random variable *education* as the number of years of schooling of the interviewed person, with values corresponding to typical completion times in the German education system:

$$Y = \text{years of schooling} \in \{4, 10, 12, 13, 14, 16, 18, 21\}.$$

Table 1.1: ISCED 2011 levels

ISCED.level	Education.level	Years.of.schooling
1	Primary	4
2	Lower Secondary	10

Table 1.1: ISCED 2011 levels

ISCED.level	Education.level	Years.of.schooling
3	Upper secondary	12
4	Post-Secondary	13
5	Short-Cycle Tertiary	14
6	Bachelor's	16
7	Master's	18
8	Doctoral	21

- *Wage*: If you ask a random person about their income per working hour in EUR, there are infinitely many potential answers. Any (non-negative) real number may be an outcome. The sample space is a continuum of different wage levels. The wage level of the interviewed person is already numerical. The random variable is

$$Y = \text{income per working hour in EUR.}$$

Random variables share the characteristic that their value is uncertain before conducting a random experiment (e.g., flipping a coin or selecting a random person for an interview). Their value is always a real number and is determined only once the experiment's outcome is known.

2 Probability

2.1 Random experiments

A random experiment is a procedure or situation where the result is uncertain and determined by a probabilistic mechanism. An **outcome** is a specific result of a random experiment. The **sample space** S is the set/collection of all potential outcomes.

Let's consider some examples:

- *Coin toss*: The outcome of a coin toss can be 'heads' or 'tails'. This random experiment has a two-element sample space: $S = \{heads, tails\}$.
- *Gender*: If you conduct a survey and interview a random person to ask them about their gender, the answer may be 'female', 'male', or 'diverse'. It is a random experiment since the person to be interviewed is selected randomly. The sample space has three elements: $S = \{female, male, diverse\}$.
- *Education level*: If you ask a random person about their education level according to the ISCED-2011 framework, the outcome may be one of the eight ISCED-2011 levels. We have an eight-element sample space:

$$S = \{Level\ 1, Level\ 2, Level\ 3, Level\ 4, Level\ 5, Level\ 6, Level\ 7, Level\ 8\}.$$

- *Wage*: If you ask a random person about their income per working hour in EUR, there are infinitely many potential answers. Any (non-negative) real number may be an outcome. The sample space is a continuum of different wage levels.

2.2 Random variables

A **random variable** is a numerical summary of a random experiment. In econometrics and applied statistics, we always express random experiments in terms of random variables. Let's define some random variables based on the random experiments above:

- *Coin*: A two-element sample space random experiment can be transformed to a binary random variable, i.e., a random variable that takes either 0 or 1. We define the *coin* random variable as

$$Y = \begin{cases} 1 & \text{if outcome is heads,} \\ 0 & \text{if outcome is tails.} \end{cases}$$

  A binary random variable is also called **Bernoulli random variable**.

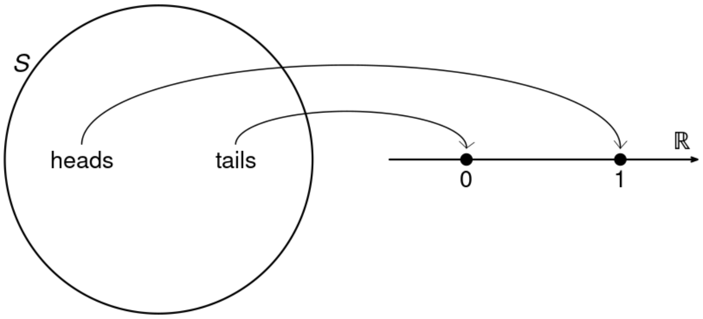


Figure 2.1: Bernoulli random variable

- *Female dummy*: The three-element sample space of the gender random experiment does not provide any natural ordering. A useful way to transform it into random variables are **dummy variables**. The *female* dummy variable is a Bernoulli random variable with

$$Y = \begin{cases} 1 & \text{if the person is female,} \\ 0 & \text{if the person is not female.} \end{cases}$$

  Similarly, dummy variables for *male* and *diverse* can be defined.

- *Education*: The eight-element sample space of the education-level random experiment provides a natural ordering. We define the random variable *education* as the number of years of schooling of the interviewed person:

$$Y = \text{number of years of schooling} \in \{4, 10, 12, 13, 14, 16, 18, 21\}.$$

- *Wage*: The wage level of the interviewed is already numerical. The random variable is

$$Y = \text{income per working hour in EUR.}$$

Table 2.1: ISCED 2011 levels

[H]

ISCED level	Education level	Years of schooling
1	Primary	4
2	Lower Secondary	10
3	Upper secondary	12
4	Post-Secondary	13
5	Short-Cycle Tertiary	14
6	Bachelor's	16
7	Master's	18
8	Doctoral	21

2.3 Probability function

In the case of a fair coin, it is natural to assign the following probabilities to the coin variable: $P(Y = 0) = 0.5$ and $P(Y = 1) = 0.5$. By definition, the coin variable will never take the value 2.5, so the corresponding probability is $P(Y = 2.5) = 0$. We may also consider intervals, e.g., $P(Y \geq 0) = 1$ and $P(-1 \leq Y < 1) = 0.5$

The **probability function** P assigns values between 0 and 1 to **events**. Specific subsets of the real line define events. Any real number defines an event, and any open, half-open, or closed interval represents an event as well, e.g.,

$$A_1 = \{Y = 0\}, \quad A_2 = \{Y = 1\}, \quad A_3 = \{Y = 2.5\}$$

and

$$A_4 = \{Y \geq 0\}, \quad A_5 = \{-1 \leq Y < 1\}.$$

We may take **complements**

$$A_6 := A_4^c = \{Y \geq 0\}^c = \{Y < 0\},$$

as well as **unions** and **intersections**:

$$\begin{aligned} A_7 &:= A_1 \cup A_6 = \{Y = 0\} \cup \{Y < 0\} = \{Y \leq 0\}, \\ A_8 &:= A_4 \cap A_5 = \{Y \geq 0\} \cap \{-1 \leq Y < 1\} = \{0 \leq Y < 1\}. \end{aligned}$$

Unions and intersections can also applied iteratively,

$$A_9 := A_1 \cup A_2 \cup A_3 \cup A_5 \cup A_6 \cup A_7 \cup A_8 = \{Y \in (-\infty, 1] \cup \{2.5\}\},$$

and by taking complements, we obtain the full real line and the empty set:

$$\begin{aligned} A_{10} &:= A_9 \cup A_9^c = \{Y \in \mathbb{R}\}, \\ A_{11} &:= A_{10}^c = \{\}. \end{aligned}$$

You may verify that $P(A_1) = 0.5$, $P(A_2) = 0.5$, $P(A_3) = 0$, $P(A_4) = 1$ $P(A_5) = 0.5$, $P(A_6) = 0$, $P(A_7) = 0.5$, $P(A_8) = 0.5$, $P(A_9) = 1$, $P(A_{10}) = 1$, $P(A_{11}) = 0$. If you take the variables *education* or *wage*, the probabilities of these events may be completely different.

To make probabilities a mathematically sound concept, we have to define to which events probabilities are assigned and how these probabilities are assigned. We consider the concept of a **sigma algebra** to collect all events.

Sigma algebra

A collection $\mathcal{B}$ of sets is called sigma algebra if it satisfies the following three properties:

1. $\{\} \in \mathcal{B}$ (empty set)
2. If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$
3. If $A_1, A_2, \ldots \in \mathcal{B}$, then $A_1 \cup A_2 \cup \ldots \in \mathcal{B}$.

If you take all events of the form $\{Y \in (a, b)\}$, where $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, and if you add all unions, intersections, and complements of these events, and again all unions, intersections, and complements of those events, and so on, you will obtain the so-called **Borel sigma algebra**. The Borel sigma algebra contains all events we assign probabilities to, the **Borel sets**.

Probabilities must follow certain conditions. The following axioms ensure that these conditions are fulfilled:

Probability function

A probability function P is a function $P : \mathcal{B} \to [0, 1]$ that satisfies the Axioms of Probability:

1. $P(A) \geq 0$ for every $A \in \mathcal{B}$
2. $P(Y \in \mathbb{R}) = 1$
3. If $A_1, A_2, A_3 \ldots$ are disjoint then

$$A_1 \cup A_2 \cup A_3 \cup \ldots = P(A_1) + P(A_2) + P(A_3) + \ldots$$

Recall that two events A and B are **disjoint** if they have no outcomes in common, i.e., if $A \cap B = \{\}$. For instance, A_1 and A_2 are $A_1 = \{Y = 0\}$ and $A_2 = \{Y = 1\}$ are disjoint, but A_1 and $A_4 = \{Y \geq 0\}$ are not disjoint, since $A_1 \cap A_4 = \{Y = 0\}$ is nonempty.

Probabilities are a well-defined concept if we use the Borel sigma algebra and the axioms of probability. The mathematical details are developed in the field of measure theory.

The axioms of probability imply the following rules of calculation:

Basic rules of probability

- $0 \leq P(A) \leq 1$ for any event A
- $P(A^c) = 1 - P(A)$ for the complement event of A
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A, B (inclusion-exclusion principle)
- $P(A) \leq P(B)$ if $A \subset B$
- $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint

2.4 Distribution

The **distribution** of a random variable Y is characterized by the probabilities of all events of Y in the Borel sigma algebra. The distribution of the *coin* variable is fully characterized by the probabilities $P(Y = 1) = 0.5$ and $P(Y = 0) = 0.5$. We can compute the probabilities of all other events using the basic rules of probability. The probability mass function summarizes these probabilities:

Probability mass function (PMF)

The probability mass function (PMF) of a random variable Y is

$$\pi(a) := P(Y = a), \quad a \in \mathbb{R}$$

The PMF of the *coin* variable is

$$\pi(a) = P(Y = a) = \begin{cases} 0.5 & \text{if } a \in \{0, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

The *education* variable may have the following PMF:

$$\pi(a) = P(Y = a) = \begin{cases} 0.008 & \text{if } a = 4 \\ 0.048 & \text{if } a = 10 \\ 0.392 & \text{if } a = 12 \\ 0.072 & \text{if } a = 13 \\ 0.155 & \text{if } a = 14 \\ 0.071 & \text{if } a = 16 \\ 0.225 & \text{if } a = 18 \\ 0.029 & \text{if } a = 21 \\ 0 & \text{otherwise} \end{cases}$$

The PMF is useful for distributions where the sum of the PMF values over a discrete (finite or countably infinite) number of domain points equals 1, as in the examples above. These distributions are called **discrete distributions**.

Another example of a discrete distribution is the **Poisson distribution** with parameter $\lambda > 0$, which has the PMF

$$\pi(a) = \begin{cases} \frac{e^{-\lambda}\lambda^a}{a!} & \text{if } a = 0, 1, 2, 3, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

It has a countably infinite number of domain points with nonzero PMF values, and its probabilities sum to 1, i.e., $\sum_{a=0}^{\infty} \pi(a) = e^{-\lambda} \sum_{a=0}^{\infty} \frac{\lambda^a}{a!} = 1$ since the exponential function has the power series representation $e^{\lambda} = \sum_{a=0}^{\infty} \frac{\lambda^a}{a!}$.

Not all random variables are discrete, e.g., the *wage* variable takes values on a continuum. The cumulative distribution function is a unifying concept summarizing the distribution of any random variable.

2.5 Cumulative distribution function

Cumulative distribution function (CDF)

The cumulative distribution function (CDF) of a random variable Y is

$$F(a) := P(Y \leq a), \quad a \in \mathbb{R},$$

The CDF of the variable *coin* is

$$F(a) = \begin{cases} 0 & a < 0, \\ 0.5 & 0 \leq a < 1, \\ 1 & a \geq 1, \end{cases}$$

with the following CDF plot:

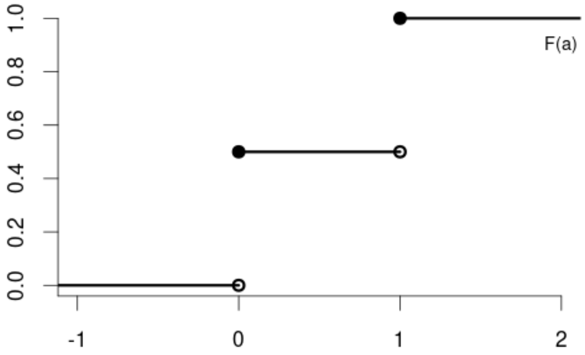


Figure 2.2: CDF of coin

The CDF of the variables *education* is

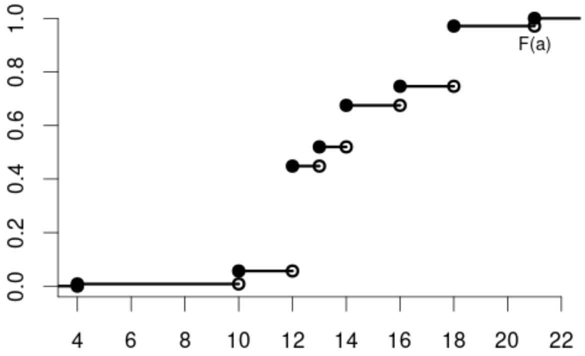


Figure 2.3: CDF of education

and the CDF of the variable *wage* may have the following form:

By the basic rules of probability, we can compute the probability of any event if we know the probabilities of all events of the form $\{Y \leq a\}$.

Some basic rules for the CDF (for $a < b$):

- $P(Y \leq a) = F(a)$
- $P(Y > a) = 1 - F(a)$

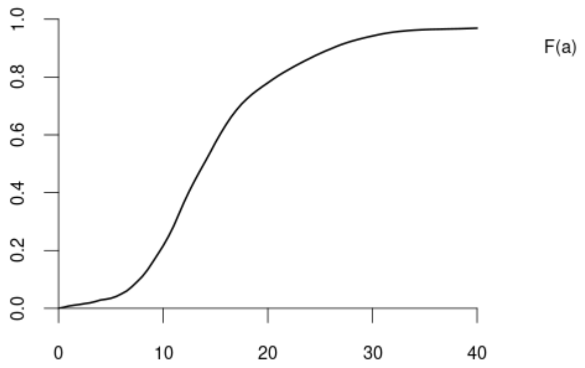


Figure 2.4: CDF of wage

- $P(Y < a) = F(a) - \pi(a)$
- $P(Y \geq a) = 1 - P(Y < a)$
- $P(a < Y \leq b) = F(b) - F(a)$
- $P(a < Y < b) = F(b) - F(a) - \pi(b)$
- $P(a \leq Y \leq b) = F(b) - F(a) + \pi(a)$
- $P(a \leq Y < b) = P(a \leq Y \leq b) - \pi(b)$

Some CDFs have jumps/steps, and some CDFs are smooth/continuous. If F has a jump at domain point a, then the PMF at a is

$$\pi(a) = P(Y = a) = F(a) - \lim_{\epsilon \to 0} F(a - \epsilon) = \text{“jump height at } a\text{”}. \tag{2.1}$$

If F is continuous at domain point a, we have $\lim_{\epsilon \to 0} F(a - \epsilon) = F(a)$, which implies that $\pi(a) = P(Y = a) = 0$.

We call the random variable a **discrete random variable** if the CDF contains jumps and is flat between the jumps. A discrete random variable has only a finite (or countably infinite) number of potential outcomes. The values of the PMF correspond to the jump heights in the CDF as defined in Equation 2.1. The **support** $\mathcal{Y}$ of a discrete random variable Y is the set of all points $a \in \mathbb{R}$ with nonzero probability mass, i.e. $\mathcal{Y} = \{a \in \mathbb{R} : \pi(a) > 0\}$. The probabilities of a discrete random variable sum to 1, i.e., $\sum_{a \in \mathcal{Y}} \pi(a) = 1$.

The Bernoulli variables *coin* and *female* are discrete random variables with support $\mathcal{Y} = \{0, 1\}$. The variable *eduaction* has support $\mathcal{Y} = \{4, 10, 12, 13, 14, 16, 18, 21\}$. A Poisson random variable has thr support $\mathcal{Y} = \mathbb{N} \cup \{0\}$.

We call a random variable a **continuous random variable** if the CDF is continuous at every point $a \in \mathbb{R}$. A continuous random variable has $\pi(a) = P(Y = a) = 0$ for all $a \in \mathbb{R}$. The basic rules for the CDF become simpler in the case of a continuous random variable:

Rules for the CDF of a continuous random variable (for $a < b$):

- $P(Y \leq a) = P(Y < a) = F(a)$
- $P(Y \geq a) = P(Y > a) = 1 - F(a)$
- $P(a < Y \leq b) = P(a \leq Y < b) = F(b) - F(a)$
- $P(a < Y < b) = P(a \leq Y \leq b) = F(b) - F(a)$

Single-outcome events are null sets and occur with probability zero. Therefore, the PMF is not suitable to describe the distribution of a continuous random variable. We use the CDF to compute probabilities of interval events as well as their unions, intersections, and complements.

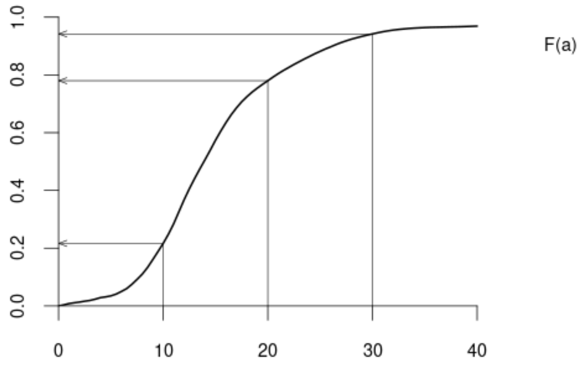


Figure 2.5: CDF of wage evaluated at some points

For instance, $P(Y \leq 30) = 0.942$, $P(Y \leq 20) = 0.779$, $P(Y \leq 10) = 0.217$, and $P(10 \leq Y \leq 20) = 0.779 - 0.217 = 0.562$.

Quantiles

For a continuous random variable Y the α-quantile $q(\alpha)$ is defined as the solution to the equation $\alpha = F(q(\alpha))$, or, equivalently, as the inverse of the distribution function:

$$q(\alpha) = F^{-1}(\alpha)$$

- $q(\cdot)$ is a function from $(0,1)$ to $\mathbb{R}$.
- Some quantiles have special names:
  - The median is the 0.5 quantile.
  - The quartiles are the 0.25, 0.5 and 0.75 quantiles.

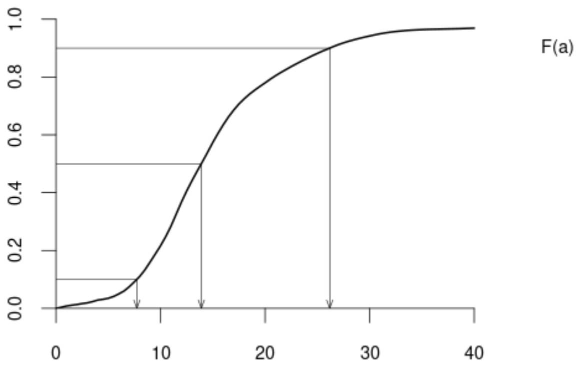


Figure 2.6: Quantiles of variable wage

- The deciles are the 0.1, 0.2,… , 0.9 quantiles.

From the quantile plot, we find that $q(0.1) = 7.73$, $q(0.5) = 13.90$, $q(0.9) = 26.18$. Under this wage distribution, the median wage is 13.90 EUR, the poorest 10% have a wage of less than 7.33 EUR, and the richest 10% have a wage of more than 26.18 EUR.

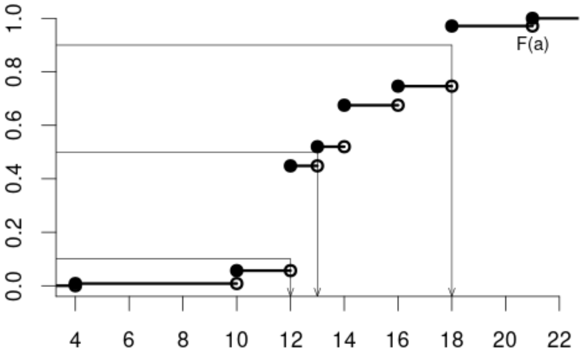


Figure 2.7: Quantiles of variable education

The median of *education* is 13, the 0.1-quantile is 12, and the 0.9-quantile is 18.

A CDF has the following properties:

(i) it is *non-decreasing*,
(ii) it is *right-continuous* (jumps may occur only when the limit point is approached from the left)
(iii) the left limit is zero: $\lim_{a \to -\infty} F(a) = 0$
(iv) the right limit is one: $\lim_{a \to \infty} F(a) = 1$.

Any function F that satisfies these four properties defines a probability distribution. Typically, distributions are divided into discrete and continuous distributions. Still, it may be the case that a distribution does not fall into either of these categories (for instance, if a CDF has jumps on some domain points and is continuously increasing on other domain intervals). In any case, the CDF characterizes the entire distribution of any random variable.

2.6 Probability density function

For discrete random variables, both the PMF and the CDF characterize the distribution. In the case of a continuous random variable, the PMF does not yield any information about the distribution since it is zero. The continuous counterpart of the PMF is the density function:

Probability density function

The probability density function (PDF) or simply density function of a continuous random variable Y is a function $f(a)$ that satisfies

$$F(a) = \int_{-\infty}^{a} f(u) \text{ d}u$$

The density $f(a)$ is the derivative of the CDF $F(a)$ if it is differentiable:

$$f(a) = \frac{d}{da} F(a).$$

Properties of a PDF:

(i) $f(a) \geq 0$ for all $a \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f(u) \text{ d}u = 1$

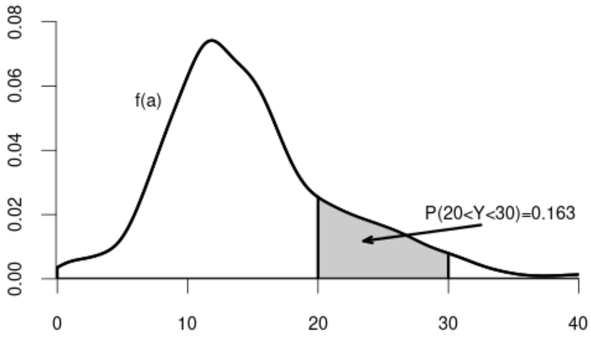


Figure 2.8: PDF of the variable wage

Probability rule for the PDF:

$$P(a < Y < b) = \int_a^b f(u) \,\mathrm{d}u = F(b) - F(a)$$

2.7 Expected value

The **expectation** or **expected value** is the most important measure of the central tendency of a distribution. It gives you the average value you can expect to get if you repeat the random experiment multiple times. We define the expectation first for discrete random variables, then continuous random variables, and finally give a unified definition for all random variables.

2.7.1 Expectation of a discrete random variable

The **expectation** or **expected value** of a discrete random variable Y with PMF $\pi(\cdot)$ and support $\mathcal{Y}$ is defined as

$$E[Y] = \sum_{u \in \mathcal{Y}} u\pi(u).$$

For the *coin* variable, we have $\mathcal{Y} = \{0, 1\}$ and therefore

$$E[Y] = 0 \cdot \pi(0) + 1 \cdot \pi(1) = 0.5.$$

For the variable *education* we get

$$\begin{aligned} E[Y] &= 4 \cdot \pi(4) + 10 \cdot \pi(10) + 12 \cdot \pi(12) \\ &\quad + 13 \cdot \pi(13) + 14 \cdot \pi(14) + 16 \cdot \pi(16) \\ &\quad + 18 \cdot \pi(18) + 21 * \pi(21) = 13.557 \end{aligned}$$

The expectation of a Poisson distributed random variable Y with parameter λ is

$$E[Y] = 0 + \sum_{a=1}^{\infty} a \cdot e^{-\lambda} \frac{\lambda^a}{a!} = e^{-\lambda} \sum_{a=1}^{\infty} \frac{\lambda^a}{(a-1)!} = e^{-\lambda} \sum_{a=0}^{\infty} \frac{\lambda^{a+1}}{a!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

2.7.2 Expectation of a continuous random variable

The **expectation** or **expected value** of a of a continuous random variable Y with PDF $f(\cdot)$ is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \ \mathrm{d}u.$$

Using numerical integration for the density of Figure 6.2 yields the expected value of 16.45 EUR for the wage variable, which is larger than the median value of 13.90 EUR. If the mean is larger than the median, we have a positively skewed distribution, meaning that a few people have high salaries, and many people have medium and low wages.

The uniform distribution on the unit interval $[0,1]$ has the PDF

$$f(u) = \begin{cases} 1 & \text{if } u \in [0,1], \\ 0 & \text{otherwise,} \end{cases}$$

and the expected value of a uniformly distributed random variable Y is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \ \mathrm{d}u = \int_{0}^{1} u \ \mathrm{d}u = \frac{1}{2}.$$

2.7.3 Expectation for general random variables

We can also define the expected value in a unified way for any random variable so we do not have to distinguish between discrete and continuous random variables. Let $F(\cdot)$ be the CDF of the random variable of interest and consider the differential $\mathrm{d}F(u)$, which corresponds to an infinitesimal change in $F(\cdot)$ at u. For a discrete random variable, $F(u)$ changes only if there is a step/jump at u and zero otherwise because it is flat. Thus, for a discrete distribution,

$$\mathrm{d}F(u) = \begin{cases} \pi(u) & \text{if } u \in \mathcal{Y} \\ 0 & \text{if } u \notin \mathcal{Y}. \end{cases}$$

In the case of a continuous random variable with differentiable CDF $F(\cdot)$, we have

$$\mathrm{d}F(u) = f(u) \ \mathrm{d}u,$$

where $f(\cdot)$ is the PDF of the random variable. This gives rise to the following unified definition of the expected value:

The **expectation** or **expected value** of any random variable with CDF $F(\cdot)$ is defined as

$$E[Y] = \int_{-\infty}^{\infty} u \ \mathrm{d}F(u). \tag{2.2}$$

Note that Equation 2.2 is the Riemann-Stieltjes integral of a with respect to the function $F(\cdot)$. Recall that the Riemann integral of u with respect to u over the interval $[-1, 1]$ is

$$\int_{-1}^{1} u \ \mathrm{d}u := \lim_{N\to\infty} \sum_{j=1}^{2N} \Big(\frac{j}{N} - 1\Big)\Big((\tfrac{j}{N} - 1) - (\tfrac{j-1}{N} - 1)\Big) = \lim_{N\to\infty} \sum_{j=1}^{2N} \Big(\frac{j}{N} - 1\Big)\frac{1}{N},$$

for the interval $[-z, z]$ we have

$$\int_{-z}^{z} u \ \mathrm{d}u := \lim_{N\to\infty} \sum_{j=1}^{2N} z\Big(\frac{j}{N} - 1\Big)\frac{z}{N},$$

and we obtain $\int_{-\infty}^{\infty} u \ \mathrm{d}u := \lim_{z\to\infty} \int_{-z}^{z} u \ \mathrm{d}u$ for the integral over the entire real line. Note that $z/N = z(\frac{j}{N} - 1) - z(\frac{j-1}{N} - 1)$ corresponds to a change in u on $[-z, z]$ so we approximate

$$\mathrm{d}u \approx z(\tfrac{j}{N} - 1) - z(\tfrac{j-1}{N} - 1) = \tfrac{z}{N}$$

and let N tend to infinity. In the case of the Riemann-Stieltjes integral, where we integrate with respect to changes in a function $F(\cdot)$, i.e., $\mathrm{d}F(u)$. In an interval $[-z, z]$, we have

$$\mathrm{d}F(u) \approx F\Big(z(\tfrac{j}{N} - 1)\Big) - F\Big(z(\tfrac{j-1}{N} - 1)\Big),$$

and we define

$$\int_{-z}^{z} u \ \mathrm{d}F(u) := \lim_{N\to\infty} \sum_{j=1}^{2N} z\Big(\tfrac{j}{N} - 1\Big) F\Big(z(\tfrac{j}{N} - 1)\Big) - F\Big(z(\tfrac{j-1}{N} - 1)\Big)$$

$$\int_{-\infty}^{\infty} u \ \mathrm{d}F(u) := \lim_{z\to\infty} \int_{-z}^{z} u \ \mathrm{d}F(u)$$

2.7.4 Properties of the expected value

The expected value is a measure of central tendency. It is a **linear** function. For any two random variables Y and Z and any $a, b \in \mathbb{R}$, we have

$$E[aY + bZ] = aE[Y] + bE[Z].$$

The expected value has some optimality properties in terms of prediction. The best predictor of a random variable Y in the mean square error sense is the value g^* that minimizes $E[(Y - g)^2]$ over g. We have

$$E[(Y - g)^2] = E[Y^2] - 2gE[Y] + g^2,$$

and minimizing over g yields

$$\frac{\mathrm{d}E[(Y-g)^2]}{\mathrm{d}g} = -2E[Y] + 2g,$$

which is zero if $g = E[Y]$. The second derivative is positive. Therefore, the expected value is the **best predictor** for a random variable if you do not have any further information available.

We often transform random variables by taking, for instance, squares Y^2 or logs $\log(Y)$. For any transformation function $g(\cdot)$, the expectation of the transformed random variable $g(Y)$ is

$$E[g(Y)] = \int_{-\infty}^{\infty} g(u) \ \mathrm{d}F(u),$$

where $\mathrm{d}F(u)$ can be replaced by the PMF or the PDF as discussed in Section 2.7.3 for the different cases. For instance, if we take the *coin* variable Y and consider the transformed random variable $\log(Y+1)$, the expected value is

$$E[\log(Y+1)] = \log(1) \cdot \frac{1}{2} + \log(2) \cdot \frac{1}{2} = \frac{\log(2)}{2}$$

Moments

The r-th moment of a random variable Y is defined as

$$E[Y^r] = \int_{-\infty}^{\infty} u^r \ \mathrm{d}F(u) = \begin{cases} \sum_{u \in \mathcal{Y}} u^r \pi(u) & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} u^r f(u) \mathrm{d}u & \text{if } Y \text{ is continuous.} \end{cases}$$

2.8 Descriptive features of a distribution

Table 2.2: Some important features of the distribution of Y

$E[Y^r]$	r-th moment of Y
$E[(Y-E[Y])^r]$	r-th central moment of Y
$Var[Y] = E[(Y-E[Y])^2]$	variance of Y
$sd(Y) = \sqrt{Var[Y]}$	standard deviation of Y
$E[((Y-E[Y])/sd(Y))^r]$	r-th standardized moment of Y
$skew = E[((Y-E[Y])/sd(Y))^3]$	skewness of Y
$kurt = E[((Y-E[Y])/sd(Y))^4]$	kurtosis of Y

The mean is a measure of central tendency and equals the expected value. The variance and standard deviation are measures of dispersion. We have

$$Var[Y] = E[(Y-E[Y])^2] = E[Y^2] - E[Y]^2$$

and

$$Var[a + bY] = b^2 Var[Y]$$

for any $a, b \in \mathbb{R}$. The skewness

$$skew = \frac{E[(Y - E[Y])^3]}{sd(Y)^3} = \frac{E[Y^3] - 3E[Y^2]E[Y] + 2E[Y]^3}{(E[Y^2] - E[Y]^2)^{3/2}}$$

is a measure of asymmetry

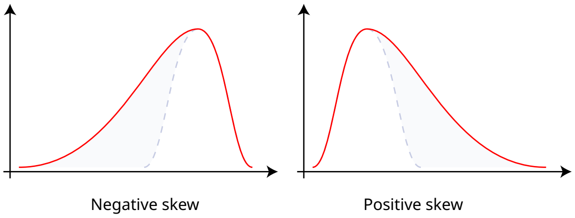


Figure 2.9: Positive and negative skewness

A random variable Y has a **symmetric distribution** about 0 if $F(u) = 1 - F(-u)$. If Y has a density, it is symmetric if $f(x) = f(-x)$. If Y is symmetric about 0, then the skewness is 0. The skewness of the variable *wage* (see Figure 6.2) is positive, i.e., the distribution is positively skewed. The **standard normal distribution** $\mathcal{N}(0, 1)$, which has the density

$$f(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

Below you find a plot of the PDFs of $N(0, 1)$ together with the t_5-distribution, which is the t-distribution with 5 degrees of freedom:

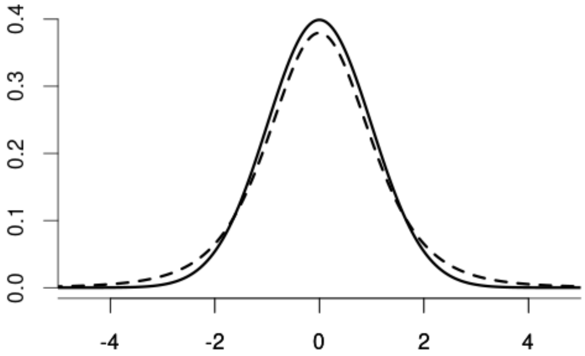


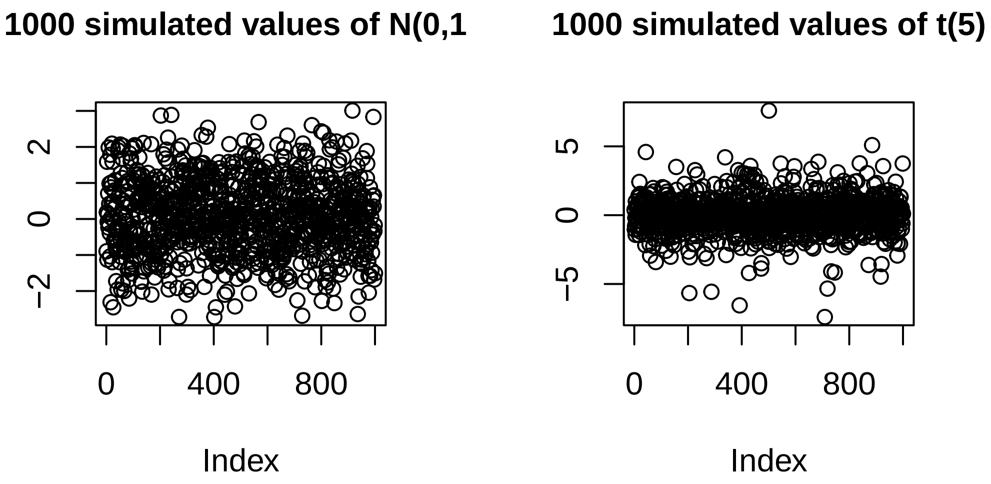
Figure 2.10: PDFs of the standard normal distribution (solid) and the t_5-distribution (dashed)

The standard normal distribution and the t(5) distribution have skewness 0. The kurtosis

$$kurt = \frac{E[(Y - E[Y])^4]}{sd(Y)^4} = \frac{E[Y^4] - 4E[Y^3]E[Y] + 6E[Y^2]E[Y]^2 - 3E[Y]^4}{(E[Y^2] - E[Y]^2)^2}$$

is a measure of how likely extreme outliers are. The standard normal distribution has kurtosis 3 and the t(5) distribution has kurtosis 9 so that outliers in $t(5)$ are more likely than in $\mathcal{N}(0,1)$:

```r
par(mfrow=c(1,2), cex.main=1)
plot(rnorm(1000), main = "1000 simulated values of N(0,1)", ylab = "")
plot(rt(1000,5), main = "1000 simulated values of t(5)", ylab = "")
```



The kurtosis of the variable *wage* is also larger than 3, meaning outliers are much more likely than in the standard normal distribution. In this case, the positive skewness means that more people have a wage less than the average, and the large kurtosis means that there are very few people with exceptionally high salaries (outliers).

All features discussed above are functions of the first four moments $E[Y]$, $E[Y^2]$, $E[Y^3]$ and $E[Y^4]$.

2.8.1 Heavy-tailed distributions

Expectations might be infinity. For instance, the simple Pareto distribution has the PDF

$$f(a) = \begin{cases} \frac{1}{a^2} & \text{if } a > 1, \\ 0 & \text{if } a \leq 1, \end{cases}$$

and the expected value is

$$E[X] = \int_{-\infty}^{\infty} af(a) \text{ da} = \int_{1}^{\infty} \frac{1}{a} \text{ da} = \log(a)|_1^{\infty} = \infty.$$

The game of chance from the St. Petersburg paradox (see https://en.wikipedia.org/wiki/St._Petersburg_paradox) is an example of a discrete random variable with infinite expectation.

There are distributions with finite mean with some higher moments that are infinite. For instance, the first $m-1$ moments of the t_m distribution (Student's-t distribution with m degrees of freedom) are finite, but the m-th moment and all higher order moments are infinite. Random variables with infinite first four moments have a so-called **heavy-tailed distribution** and may produce huge outliers. Many statistical procedures are only valid if the underlying distribution is not heavy-tailed.

2.9 The normal distribution

A random variable X is normally distributed with parameters (μ, σ^2) if it has the density

$$f(a \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(a-\mu)^2}{2\sigma^2}\Big).$$

We write $Y \sim \mathcal{N}(\mu, \sigma^2)$. Mean and variance are

$$E[Y] = \mu, \quad var[Y] = \sigma^2.$$

Special case: standard normal distribution $\mathcal{N}(0,1)$ with density

$$\phi(a) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{a^2}{2}\Big)$$

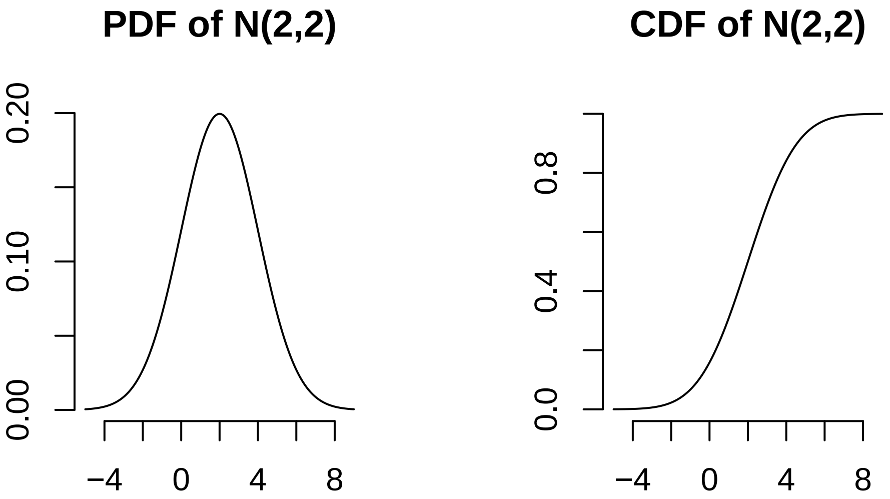
and CDF

$$\Phi(a) = \int_{-\infty}^{a} \phi(u)\mathrm{d}u.$$

$\mathcal{N}(0,1)$ is symmetric around zero:

$$\phi(a) = \phi(-a), \quad \Phi(a) = 1 - \Phi(-a)$$

```r
par(mfrow=c(1,2), bty="n", lwd=1)
x <- seq(-5,9,by=0.01)
plot(x,dnorm(x,2,2),ylab="",xlab="", type="l", main= "PDF of N(2,2)")
plot(x,pnorm(x,2,2),ylab="",xlab="", type="l", main = "CDF of N(2,2)")
```



If $Y_1, \ldots, Y_n$ are normally distributed and $c_1, \ldots, c_n \in \mathbb{R}$, then $\sum_{j=1}^n c_j Y_j$ is normally distributed.

2.10 Additional reading

- Stock and Watson (2019), Section 2
- Hansen (2022a), Section 1-2
- Davidson and MacKinnon (2004), Section 1

2.11 R-codes

statistics-sec2.R

3 Dependence

3.1 Multivariate random variables

In statistics, we typically study multiple random variables simultaneously. We can collect k random variable $X_1, \ldots, X_k$ in a **random vector**

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} = (X_1, \ldots, X_k)'.$$

We also call X a **k-variate random variable**.

Since X is a random vector, its outcome is also vector-valued, e.g. $X = x \in \mathbb{R}^k$ with $x = (x_1, \ldots, x_k)'$. Events of the form $\{X \leq x\}$ mean that each component of the random vector X is smaller than the corresponding values of the vector x, i.e.

$$\{X \leq x\} = \{X_1 \leq x_1, \ldots, X_k \leq x_k\}.$$

3.2 Bivariate random variables

If $k = 2$, we call X a **bivariate random variable**. Consider, for instance, the coin toss Bernoulli variable Y with $P(Y = 1) = 0.5$ and $P(Y = 0) = 0.5$, and let Z be a second coin toss with the same probabilities. $X = (Y, Z)$ is a bivariate random variable where both entries are discrete random variables. Since the two coin tosses are performed separately from each other, it is reasonable to assume that the probability that the first and second coin tosses show 'heads' is 0.25, i.e., $P(\{Y = 1\} \cap \{Z = 1\}) = 0.25$. We would expect the following joint probabilities:

Table 3.1: Joint probabilities of coin tosses

	$Z = 1$	$Z = 0$	any result
$Y = 1$	0.25	0.25	0.5
$Y = 0$	0.25	0.25	0.5
any result	0.5	0.5	1

The probabilities in the above table characterize the **joint distribution** of Y and Z. The table shows the values of the **joint probability mass function**:

$$\pi_{YZ}(a,b) = \begin{cases} 0.25 & \text{if } a \in \{0,1\} \text{ and } b \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}$$

Another example are the random variables Y, a dummy variable for the event that the person has a high wage (more than 25 USD/hour), and Z, a dummy variable for the event that the same person has a university degree. Similarly, $X = (Y, Z)$ is a bivariate random variable consisting of two univariate Bernoulli variables. The joint probabilities might be as follows:

Table 3.2: Joint probabilities of wage and education dummies

	Z=1	Z=0	any education
Y=1	0.19	0.12	0.31
Y=0	0.17	0.52	0.69
any wage	0.36	0.64	1

The joint probability mass function is

$$\pi_{YZ}(a,b) = \begin{cases} 0.19 & \text{if } a=1, b=1, \\ 0.12 & \text{if } a=1, b=0, \\ 0.17 & \text{if } a=0, b=1, \\ 0.52 & \text{if } a=0, b=0, \\ 0 & \text{otherwise.} \end{cases}$$

The **marginal probability mass function** of Y is

$$\begin{aligned} \pi_Y(a) = P(Y=a) &= \pi_{YZ}(a,0) + \pi_{YZ}(a,1) \\ &= \begin{cases} 0.19 + 0.12 = 0.31 & \text{if } a = 1, \\ 0.17 + 0.52 = 0.69 & \text{if } a = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

and the **marginal probability mass function** of Z is

$$\begin{aligned} \pi_Z(b) = P(Z=b) &= \pi_{YZ}(0,b) + \pi_{YZ}(1,b) \\ &= \begin{cases} 0.19 + 0.17 = 0.36 & \text{if } b = 1, \\ 0.12 + 0.52 = 0.64 & \text{if } b = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

An example of a continuous bivariate random variable is $X = (Y, Z)$, where Y is the wage level in EUR/hour and Z is the labor market experience of the same person measured in years.

3.3 Bivariate distributions

Bivariate distribution

The joint distribution function of a bivariate random variable (Y, Z) is

$F_{YZ}(a,b) = P(Y \leq a, Z \leq b) = P(\{Y \leq a\} \cap \{Z \leq b\})$.

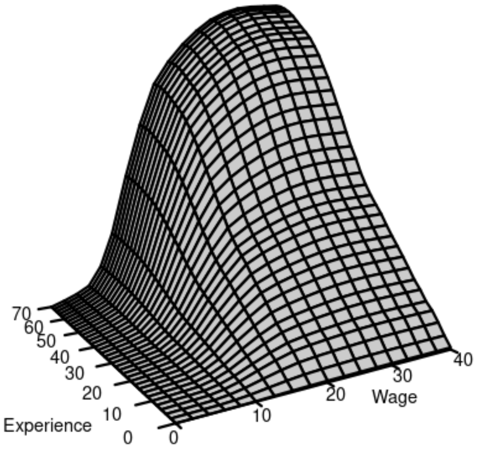


Figure 3.1: Joint CDF of wage and experience

Calculation of probabilities using a bivariate distribution function:

$$P(Y \leq a, Z \leq b) = F_{YZ}(a,b)$$
$$P(a < Y \leq b, c < Z \leq d) = F_{YZ}(b,d) - F_{YZ}(b,c) - F_{YZ}(a,d) + F_{YZ}(a,c)$$

Marginal distributions

The marginal distributions of Y and Z are

$$\begin{aligned} F_Y(a) = P(Y \leq a) = P(Y \leq a, Z < \infty) &= \lim_{b \to \infty} F_{YZ}(a,b), \\ F_Z(b) = P(Z \leq b) = P(Y < \infty, Z \leq b) &= \lim_{a \to \infty} F_{YZ}(a,b) \end{aligned}$$

Bivariate density function

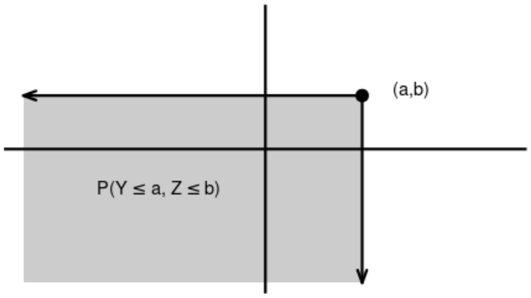


Figure 3.2: Calculate probabilities using the joint CDF

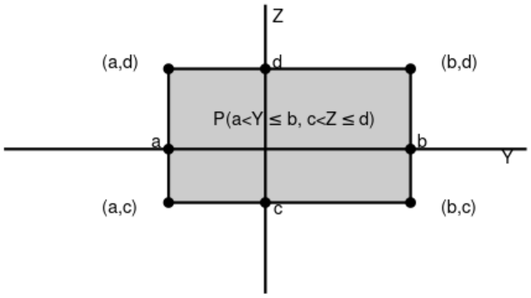


Figure 3.3: Calculate probabilities using the joint CDF

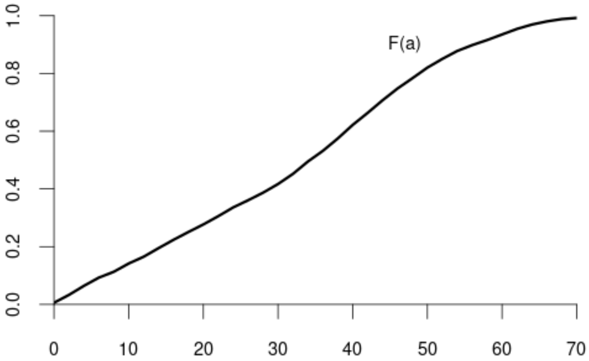


Figure 3.4: Marginal CDF of experience

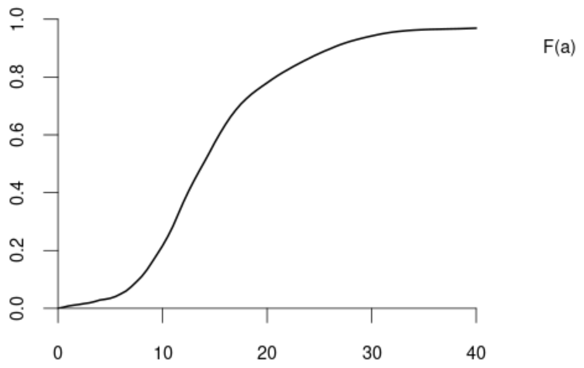


Figure 3.5: Marginal CDF of wage

The joint density function of a bivariate continuous random variable (Y, Z) with differentiable joint CDF $F_{YZ}(a, b)$ equals

$$f_{YZ}(a,b) = \frac{\partial^2}{\partial a \partial b} F_{YZ}(a,b).$$

The marginal densities of Y and Z are

$$f_Y(a) = \frac{d}{da} F_Y(a) = \int_{-\infty}^{\infty} f_{YZ}(a,b) \mathrm{d}b,$$
$$f_Z(b) = \frac{d}{db} F_Z(b) = \int_{-\infty}^{\infty} f_{YZ}(a,b) \mathrm{d}a.$$

3.4 Correlation

Consider the bivariate continuous random variable (Y, Z) with joint density $f_{YZ}(a, b)$. The expected value of $g(Y, Z)$, where $g(\cdot, \cdot)$ is any real-valued function, is given by

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(a,b) f_{YZ}(a,b) \ \mathrm{d}a \ \mathrm{d}b.$$

The first **cross moment** of Y and Z is $E[YZ]$. We have $E[YZ] = E[g(Y, Z)]$ for the function $g(Y, Z) = Y \cdot Z$. Therefore,

$$E[YZ] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab f_{YZ}(a,b) \ \mathrm{d}a \ \mathrm{d}b.$$

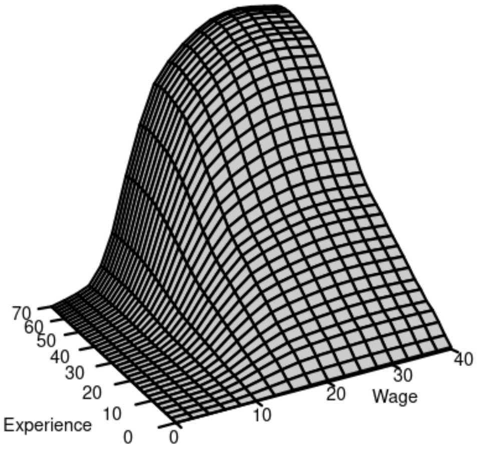


Figure 3.6: Joint CDF of wage and experience

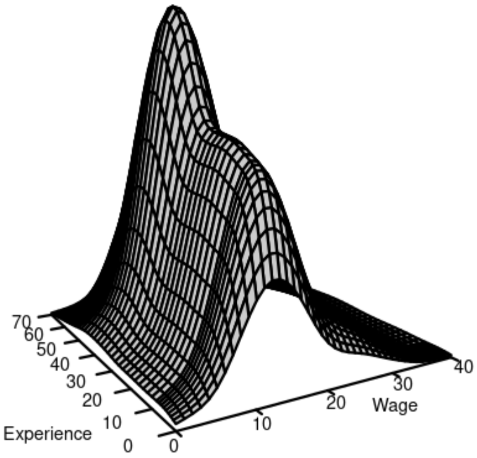


Figure 3.7: Joint PDF of wage and experience

The **covariance** of Y and Z is defined as

$$Cov(Y,Z) = E[(Y - E[Y])(Z - E[Z])] = E[YZ] - E[Y]E[Z].$$

The covariance of Y and Y is the variance:

$$Cov(Y,Y) = Var[Y].$$

The variance of the sum of two random variables depends on the covariance:

$$Var[Y + Z] = Var[Y] + 2Cov(Y,Z) + Var[Z]$$

The **correlation** of Y and Z is

$$Corr(Y,Z) = \frac{Cov(Y,Z)}{sd(Y)sd(Z)}$$

Uncorrelated

Y and Z are **uncorrelated** if $Corr(Y,Z) = 0$, or, equivalently, if $Cov(Y,Z) = 0$.

If Y and Z are uncorrelated, we have

$$\begin{aligned} E[YZ] &= E[Y]E[Z] \\ var[Y + Z] &= var[Y] + var[Z] \end{aligned}$$

3.5 Independence

Two events A and B are independent if

$$P[A \cap B] = P[A]P[B].$$

For instance, in the bivariate random variable of Table 3.1 (two coin tosses), we have

$$P(Y = 1, Z = 1) = 0.25 = 0.5 \cdot 0.5 = P(Y = 1)P(Z = 1).$$

Hence, $\{Y = 1\}$ and $\{Z = 1\}$ are independent events. In the bivariate random variable of Table 3.2 (wage/education), we find

$$P(Y = 1, Z = 1) = 0.19 \neq P(Y = 1)P(Z = 1) = 0.31 \cdot 0.36 = 0.1116.$$

Therefore, the two events are not independent. In this case, the two random variables are dependent.

Independence

Y and Z are **independent** random variables if, for all a and b, the bivariate distribution function is the product of the marginal distribution functions:

$$F_{YZ}(a,b) = F_Y(a)F_Z(b).$$

If this property is not satisfied, we say that X and Y are **dependent**.

The random variables Y and Z of Table 3.1 are independent, and those of Table 3.2 are dependent.

If Y and Z are independent and have finite second moments, then Y and Z are uncorrelated. The reverse is not true!

3.6 Random vectors

The above concepts can be generalized to any k-variate random vector $X = (X_1, \ldots, X_k)$. The joint CDF of X is

$$F_X(x) = P(X_1 \leq x_1, \ldots, X_k \leq x_k).$$

X has independent entries if

$$F_X(x) = \prod_{i=1}^{k} P(X_i \leq x_i) = \prod_{i=1}^{k} F_{X_i}(x_i)$$

If $F_X(x)$ is a continuous CDF, the joint k-dimensional density is

$$f_X(x) = f_X(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_X(x_1, \ldots, x_k).$$

The expectation vector of X is

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_k] \end{pmatrix},$$

and the covariance matrix of X is

$$\begin{aligned} Var[X] &= E[(X - E[X])(X - E[X])'] \\ &= \begin{pmatrix} Var[X_1] & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var[X_2] & \ldots & Cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var[X_k] \end{pmatrix} \end{aligned}$$

For any random vector X, the covariance matrix $Var[X]$ is symmetric and positive semi-definite.

3.7 Conditional distributions

Conditional probability

The conditional probability of an event A given an event B with $P(B) > 0$ is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Let's revisit the wage and schooling example from Table 3.2:

$$P(Y = 1 \mid Z = 1) = \frac{P(\{Y = 1\} \cap \{Z = 1\})}{P(Z = 1)} = \frac{0.19}{0.36} = 0.53$$

$$P(Y = 1 \mid Z = 0) = \frac{P(\{Y = 1\} \cap \{Z = 0\})}{P(Z = 0)} = \frac{0.12}{0.64} = 0.19$$

Note that

$$P(Y = 1 \mid Z = 1) = 0.53 > 0.31 = P(Y = 1)$$

implies

$$P(\{Y = 1\} \cap \{Z = 1\}) > P(Y = 1) \cdot P(Z = 1).$$

If $P(A \mid B) = P(A)$, then the events A and B are independent. If $P(A \mid B) \neq P(A)$, they are dependent.

Conditional distribution of continuous variables

Consider the density $f_{YZ}(a, b)$ of two continuous random variables Y and Z. The **conditional density** of Y given $Z = b$ is

$$f_{Y|Z}(a \mid b) = \frac{f_{YZ}(a, b)}{f_Z(b)}.$$

The **conditional distribution** of Y given $Z = b$ is

$$F_{Y|Z}(a \mid b) = \int_0^a f_{Y|Z}(u \mid b) \, \mathrm{d}u.$$

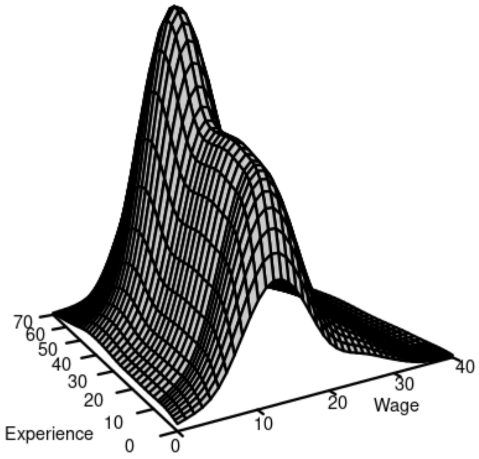


Figure 3.8: Joint PDF of wage and experience

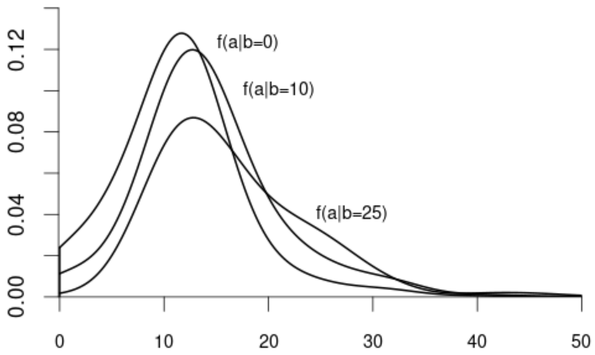


Figure 3.9: Conditional PDFs of wage given experience

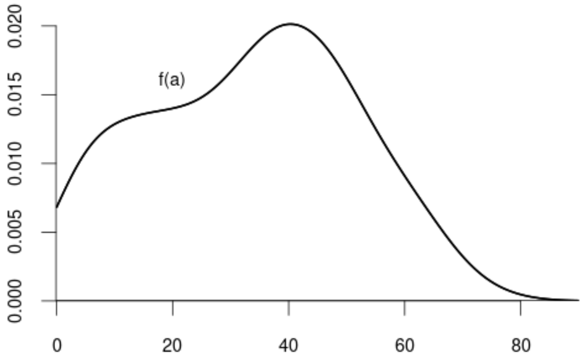


Figure 3.10: PDF of variable experience

If Y is continuous and Z is discrete, the **conditional distribution function** of Y given $\{Z = b\}$ with $P(Z = b) > 0$ is

$$F_{Y|Z}(a \mid b) = P(Y \leq a \mid Z = b) = \frac{P(Y \leq a, Z = b)}{P(Z = b)}.$$

If $F_{Y|Z}(a \mid b)$ is differentiable with respect to b, the **conditional density** of Y given $Z = b$ is

$$f_{Y|Z}(a \mid b) = \frac{\partial}{\partial a} F_{Y|Z}(a \mid b).$$

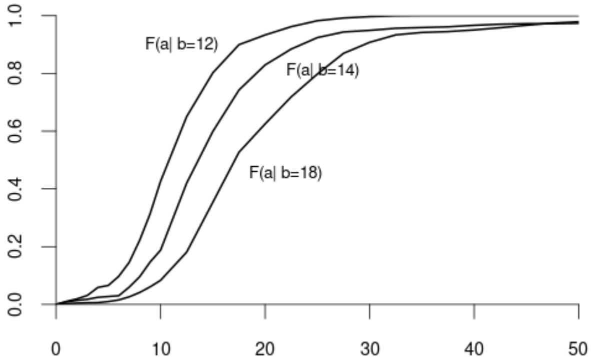


Figure 3.11: Conditional CDFs of wage given education

We often are interested in conditioning on multiple variables, such as the wage given a particular education and experience level. Let $f(y, x) = f(y, x_1, \ldots, x_k)$ be the joint density of the composite random vector $(Y, X_1, \ldots, X_k)$ with $X = (X_1, \ldots, X_k)$. The conditional density of a random variable Y given $X = x = (x_1, \ldots, x_k)'$ is

$$f_{Y|X}(y \mid x) = f(y \mid x_1, \ldots, x_k) = \frac{f(y, x_1, \ldots, x_k)}{f_X(x_1, \ldots, x_k)} = \frac{f(y, x)}{f_X(x)}$$

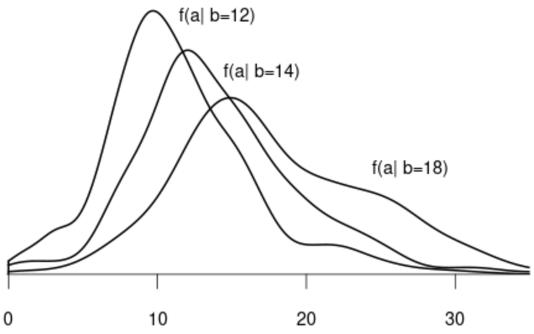


Figure 3.12: Conditional PDFs of wage given education

The conditional distribution of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \int_0^y f(u \mid x) \, \mathrm{d}u.$$

3.8 Conditional expectation

Conditional expectation function

The **conditional expectation** of Y given $X = x$ is the expected value of the distribution $F_{Y|X}(y \mid x)$. For continuous Y with conditional density $f_{Y|X}(y \mid x)$, the conditional expectation is

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) \, \mathrm{d}y.$$

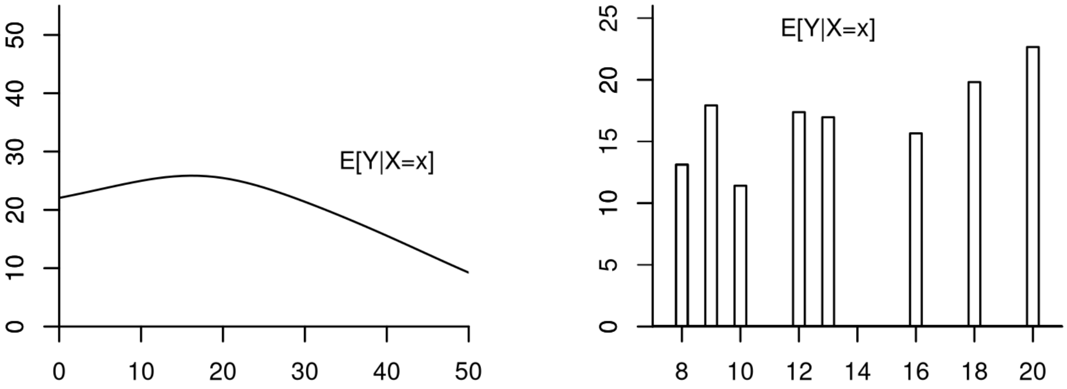
Consider again the wage and experience example. Suppose that the conditional expectation has the functional form

$$E[wage \mid experience = x] = m(x) = 14.5 + 0.9x - 0.017x^2.$$

E.g., for $x = 10$ we have $E[wage \mid experience = 10] = m(10) = 21.8$.

Note that $m(x) = E[wage \mid experience = x]$ is not random. It is a feature of the joint distribution.

Sometimes, it is useful not to fix the experience level to a certain value but to treat it as random:



(a) CEF wage given experience

(b) CEF wage given education

Figure 3.13: Conditional expectation functions

$$
\begin{aligned}
E[wage \mid experience] &= m(experience) \\
&= 14.5 + 0.9 experience - 0.017 experience^2
\end{aligned}
$$

$m(experience) = E[wage \mid experience]$ is a function of the random variable experience and, therefore, itself a random variable.

The conditional expectation function (CEF) of Y given the specific event $\{X = x\}$ is

$$m(x) = E[Y \mid X = x].$$

$m(x)$ is deterministic (non-random) and a feature of the joint distribution.

The conditional expectation function (CEF) of Y given the random vector X is

$$m(X) = E[Y \mid X].$$

$m(X)$ is a function of the random vector X and therefore itself a random variable.

3.9 Law of iterated expectations

Rules of calculation for the conditional expectation

Let Y be a random variable and X a random vector.

(i) Law of the iterated expectations (LIE):

$$E[E[Y \mid X]] = E[Y].$$

A more general LIE: For any two random vectors X and $\widetilde{X}$,

$$E[E[Y \mid X, \widetilde{X}] \mid X] = E[Y \mid X].$$

(ii) Conditioning theorem (CT): For any function $g(\cdot)$,

$$E[g(X)Y \mid X] = g(X)E[Y \mid X].$$

(iii) If Y and X are independent then $E[Y \mid X] = E[Y]$.

3.10 Conditional variance

Conditional variance

If $E[Y^2] < \infty$, the **conditional variance** of Y given the event $\{X = x\}$ is

$$Var[Y \mid X = x] = E[(Y - E[Y \mid X = x])^2 \mid X = x].$$

The conditional variance of Y given the random vector X is

$$Var[Y \mid X] = E[(Y - E[Y \mid X])^2 \mid X].$$

3.11 Best predictor

A typical application is to find a good prediction for the outcome of a random variable Y. Recall that the expected value $E[Y]$ is the best predictor for Y in the sense that $g^* = E[Y]$ minimizes $E[(Y - g)^2]$.

With the knowledge of an additional random vector X, we can use the joint distribution of Y and X to improve the prediction of Y.

It turns out that the CEF $m(X) = E[Y \mid X]$ is the best predictor for Y given the information contained in the random vector X:

Best predictor

If $E[Y^2] < \infty$, then the CEF $m(X) = E[Y \mid X]$ minimizes the expected squared error $E[(Y - g(X))^2]$ among all predictor functions $g(X)$.

Let us find the function $g(\cdot)$ that minimizes $E[(Y-g(X))^2]$:

$$\begin{aligned}
&E[(Y-g(X))^2] = E[(Y-m(X)+m(X)-g(X))^2] \\
&= \underbrace{E[(Y-m(X))^2]}_{=(i)} + 2\underbrace{E[(Y-m(X))(m(X)-g(X))]}_{=(ii)} + \underbrace{E[(m(X)-g(X))^2]}_{(iii)}
\end{aligned}$$

The first term (i) does not depend on $g(\cdot)$ and is finite if $E[Y^2]<\infty$.

For the second term (ii), we use the LIE and CT:

$$\begin{aligned}
&E[(Y-m(X))(m(X)-g(X))] \\
&= E[E[(Y-m(X))(m(X)-g(X)) \mid X]] \\
&= E[E[Y-m(X) \mid X](m(X)-g(X))] \\
&= E[(\underbrace{E[Y \mid X]}_{=m(X)} - m(X))(m(X)-g(X))] = 0
\end{aligned}$$

The third term (iii) $E[(m(X)-g(X))^2]$ is minimal if $m(\cdot)=g(\cdot)$

Therefore, $m(X)=E[Y \mid X]$ minimizes $E[(Y-g(X))^2]$.

The best predictor for Y given X is $m(X)=E[Y \mid X]$, but Y can typically only partially be predicted. We have a prediction error (CEF error)

$$e = Y - E[Y \mid X].$$

The conditional expectation of the CEF error does not depend on X and is zero:

$$\begin{aligned}
E[e \mid X] &= E[(Y-m(X)) \mid X] \\
&= E[Y \mid X] - E[m(X) \mid X] \\
&= m(X) - m(X) = 0
\end{aligned}$$

We say that Y is **conditional mean independent** of Z if $E[Y \mid Z]$ does not depend on Z.

If Y and Z are independent, they are also conditional mean independent, but not necessarily vice versa. If Y and Z are conditional mean independent, they are also uncorrelated, but not necessarily vice versa.

Since the CEF is the best predictor of Y, it is of great interest to study the CEF in practice. Much of the statistical and econometric research deals with methods to approximate and estimate the CEF. This field of statistics is called **regression analysis**.

Consider the following model for Y and X:

$$Y = m(X) + e, \quad E[e \mid X] = 0. \tag{3.1}$$

We call $m(\cdot)$ **regression function** and e **error term**.

From equation Equation 3.1 it follows that

$$E[Y \mid X] = E[m(X) + e \mid X] = E[m(X) \mid X] + E[e \mid X] = m(X).$$

I.e., the nonparametric regression model is a model for the CEF.

If $m(\cdot)$ is a linear function, then Equation 3.1 is a **linear regression model**. We will study this model in detail in the next sections.

3.12 Combining normal variables

Some of the distributions commonly encountered in econometrics are combinations of univariate normal distributions, such as the multivariate normal, chi-squared, Student t, and F distributions.

3.12.1 χ^2-distribution

Let $Z_1, \ldots, Z_m$ be independent $\mathcal{N}(0,1)$ random variables. Then, the random variable

$$Y = \sum_{i=1}^m Z_i^2$$

is **chi-square distributed** with parameter m, written $Y \sim \chi^2_m$.

The parameter m is called the degrees of freedom.

Expectation and variance:

$$E[Y] = m, \quad var[Y] = 2m$$

3.12.2 F-distribution

If $Q_1 \sim \chi^2_m$ and $Q_2 \sim \chi^2_r$, and if Q_1 and Q_2 are independent, then

$$Y = \frac{Q_1/m}{Q_2/r}$$

is ***F*-distributed** with parameters m and r, written $Y \sim F_{m,r}$.

The parameter m is called the degrees of freedom in the numerator; r is the degree of freedom in the denominator.

If $r \to \infty$ then the distribution of mY approaches χ^2_m

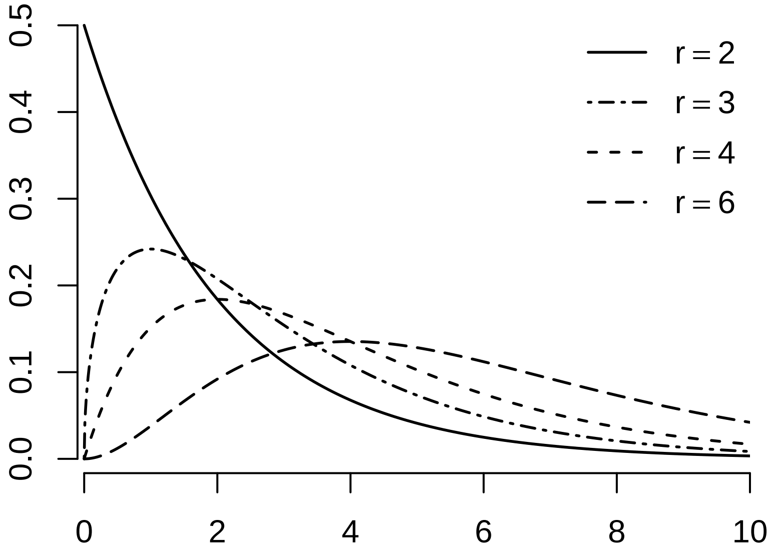


Figure 3.14: χ^2 -distribution

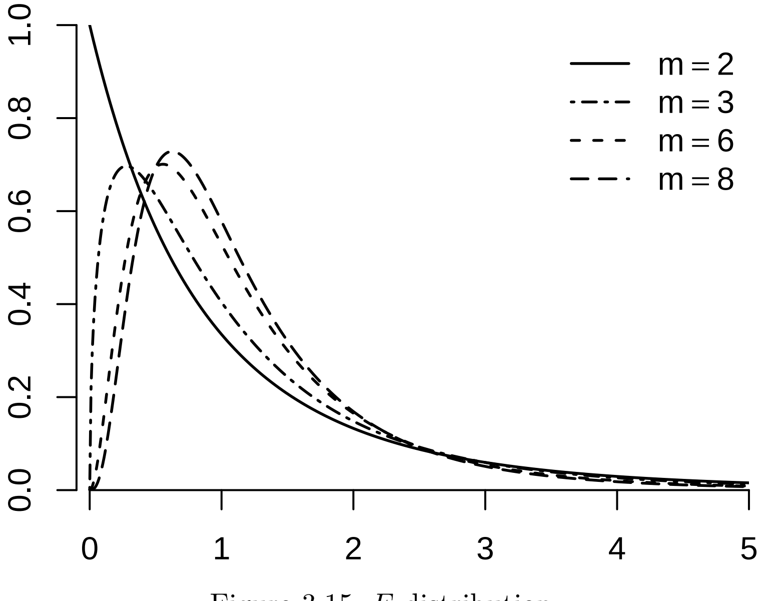


Figure 3.15: F-distribution

3.12.3 Student t-distribution

If $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_m$, and Z and Q are independent, then

$$Y = \frac{Z}{\sqrt{Q/m}}$$

is t**-distributed** with parameter m degrees of freedom, written $Y \sim t_m$.

Expectation, variance, and moments:

$$E[Y] = 0 \quad (\text{if } m \geq 2),$$

$$var[Y] = \frac{m}{m-2} \quad (\text{if } m \geq 3)$$

The first $m-1$ moments are finite: $E[|Y|^r] < \infty$ for $r \leq m-1$ and $E[|Y|^r] = \infty$ for $r \geq m$.

The t-distribution with $m = 1$ is also called **Cauchy distribution**. The t-distributions with 1, 2, 3, and 4 degrees of freedom are heavy-tailed distributions. If $m \to \infty$ then $t_m \to \mathcal{N}(0,1)$

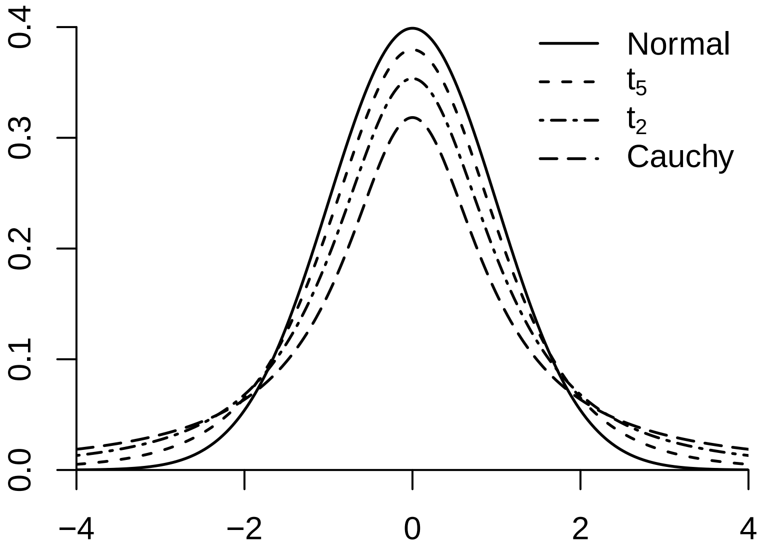


Figure 3.16: Student t-distribution

3.12.4 Multivariate normal distribution

Let $X_1, \ldots, X_k$ be independent $\mathcal{N}(0,1)$ random variables. Then, the k-vector $X = (X_1, \ldots, X_k)'$ has the **multivariate standard normal distribution**, written $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$. Its joint density is

$$f(x) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{x'x}{2}\right).$$

If $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$ and $\widetilde{X} = \mu + \boldsymbol{B}X$ for a $q \times 1$ vector μ and a $q \times k$ matrix $\boldsymbol{B}$, then $\widetilde{X}$ has a **multivariate normal distribution** with parameters μ and $\Sigma = \boldsymbol{B}\boldsymbol{B}'$, written $\widetilde{X} \sim \mathcal{N}(\mu, \Sigma)$. Its joint density is

$$f(x) = \frac{1}{(2\pi)^{k/2}(\det(\Sigma))^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big).$$

The expectation vector and covariance matrix are

$$E[\widetilde{X}] = \mu, \quad var[\widetilde{X}] = \Sigma.$$

3.12.5 R-commands for parametric distributions

	get CDF	quantile function	generate n independent
	$F(a)$	$q(p)$	random numbers
$\mathcal{N}(0,1)$	`pnorm(a)`	`qnorm(p)`	`rnorm(n)`
χ^2_r	`pchisq(a,r)`	`qchisq(p,r)`	`rchisq(n,r)`
t_r	`pt(a,r)`	`qt(p,r)`	`rt(n,r)`
$F_{r,k}$	`pf(a,r,k)`	`qf(p,r,k)`	`rf(n,r,k)`

3.13 Additional reading

- Stock and Watson (2019), Section 2
- Hansen (2022a), Section 4
- Hansen (2022b), Section 2
- Davidson and MacKinnon (2004), Section 1

4 Data

4.1 Datasets

A **univariate dataset** is a sequence of observations $Y_1, \ldots, Y_n$. These n observations can be organized into the **data vector** $\boldsymbol{Y}$, represented as $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. For example, if you conduct a survey and ask five individuals about their hourly earnings, your data vector might look like

$$\boldsymbol{Y} = \begin{pmatrix} 18.22 \\ 23.85 \\ 10.00 \\ 6.39 \\ 7.42 \end{pmatrix}.$$

Typically we have data on more than one variable, such as years of education and the gender. Categorical variables are often encoded as **dummy variables**, which are binary variables. The female dummy variable is defined as 1 if the gender of the person is female and 0 otherwise.

person	wage	education	female
1	18.22	16	1
2	23.85	18	0
3	10.00	16	1
4	6.39	13	0
5	7.42	14	0

A k**-variate dataset** (or multivariate dataset) is a collection of n vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ containing data on k variables. The i-th vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$ contains the data on all k variables for individual i. Thus, X_{ij} represents the value for the j-th variable of individual i.

The full k-variate dataset is structured in the $n \times k$ **data matrix** $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{pmatrix} X_{11} & \ldots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{nk} \end{pmatrix}$$

The i-th row in $\boldsymbol{X}$ corresponds to the values from $\boldsymbol{X}_i$. Since $\boldsymbol{X}_i$ is a column vector, we use the transpose notation $\boldsymbol{X}_i'$, which is a row vector.

The data matrix for our example is

$$\boldsymbol{X} = \begin{pmatrix} 18.22 & 16 & 1 \\ 23.85 & 18 & 0 \\ 10.00 & 16 & 1 \\ 6.39 & 13 & 0 \\ 7.42 & 14 & 0 \end{pmatrix}$$

with data vectors

$$\boldsymbol{X}_1 = \begin{pmatrix} 18.22 \\ 16 \\ 1 \end{pmatrix}, \ \boldsymbol{X}_2 = \begin{pmatrix} 23.85 \\ 18 \\ 0 \end{pmatrix}, \ \ldots \ .$$

Vector and matrix algebra provide a compact mathematical representation of multivariate data and an efficient framework for analyzing and implementing statistical methods. We will use matrix algebra frequently throughout this course.

To refresh or enhance your knowledge of matrix algebra, please consult the following resources:

Crash Course on Matrix Algebra:

matrix.svenotto.com
Section 19.1 of the Stock and Watson textbook also provides a brief overview of matrix algebra concepts.

4.2 R programming language

The best way to learn statistical methods is to program and apply them yourself. Throughout this course, we will use the `R` programming language for implementing empirical methods and analyzing real-world datasets.

If you are just starting with `R`, it is crucial to familiarize yourself with its basics. Here's an introductory tutorial, which contains a lot of valuable resources:

Getting Started with R:

rintro.svenotto.com

For those new to R, I also recommend the interactive R package SWIRL, which offers an excellent way to learn directly within the R environment. Additionally, a highly recommended online book to learn R programming is Hands-On Programming with R.

One of the best features of R is its extensive ecosystem of packages contributed by the statistical community. You find R packages for almost any statistical method out there and many statisticians provide R packages to accompany their research.

One of the most frequently used packages in applied econometrics is the `AER` package ("Applied Econometrics with R"), which provides a comprehensive collection of inferential methods for linear models. You can install the package with the command `install.packages("AER")` and you can load it with

```
library(AER)
```

at the beginning of your code. We will explore several additional packages in the course of the lecture.

4.3 Datasets in R

R includes many built-in datasets and packages of datasets that can be loaded directly into your R environment. For illustration, we consider the `CASchools` dataset available in the `AER` package. This dataset is used in the Stock and Watson textbook *Introduction to Econometrics* in Sections 4–8. It contains information on various characteristics of schools in California, such as test scores, teacher salaries, and student demographics. The data were collected in 1998.

The dataset contains the following variables:

Variable	Description
district	School district ID
school	School name
county	County name
grades	Grade span: K-6 or K-8
students	Student count
teachers	Teacher count
calworks	% of CalWorks students
lunch	% receiving free lunch
computer	Number of computers
expenditure	Expenditure per student
income	District average income (thousands $)
english	% of English learners
read	Average reading score
math	Average math score

To load this dataset into your R session, simply use:

```
data(CASchools, package = "AER")
```

The Environment pane in RStudio's top-right corner displays all objects currently in your workspace, including the `CASchools` dataset. You can click on `CASchools` to open a table viewer and explore its contents. To get a description of the dataset, use the `?CASchools` command. The `head()` function displays its first few rows:

```
head(CASchools)
```

```
  district                          school  county grades students teachers
1    75119              Sunol Glen Unified Alameda  KK-08      195    10.90
2    61499            Manzanita Elementary   Butte  KK-08      240    11.15
3    61549     Thermalito Union Elementary   Butte  KK-08     1550    82.90
4    61457 Golden Feather Union Elementary   Butte  KK-08      243    14.00
5    61523        Palermo Union Elementary   Butte  KK-08     1335    71.50
6    62042         Burrel Union Elementary  Fresno  KK-08      137     6.40
  calworks   lunch computer expenditure    income   english  read  math
1   0.5102  2.0408       67    6384.911 22.690001  0.000000 691.6 690.0
2  15.4167 47.9167      101    5099.381  9.824000  4.583333 660.5 661.9
3  55.0323 76.3226      169    5501.955  8.978000 30.000002 636.3 650.9
4  36.4754 77.0492       85    7101.831  8.978000  0.000000 651.9 643.5
5  33.1086 78.4270      171    5235.988  9.080333 13.857677 641.8 639.9
6  12.3188 86.9565       25    5580.147 10.415000 12.408759 605.7 605.4
```

The `CASchools` dataset is stored as a `data.frame`, R's most common data storage class for tabular data as in the data matrix $\boldsymbol{X}$. It organizes data in the form of a table, with variables as columns and observations as rows.

```
class(CASchools)
```

```
[1] "data.frame"
```

To inspect the structure of your dataset, you can use `str()`:

```
str(CASchools)
```

```
'data.frame':   420 obs. of  14 variables:
 $ district   : chr  "75119" "61499" "61549" "61457" ...
 $ school     : chr  "Sunol Glen Unified" "Manzanita Elementary" "Thermalito Union Elementary
 $ county     : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6 25 ...
 $ grades     : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
 $ students   : num  195 240 1550 243 1335 ...
 $ teachers   : num  10.9 11.1 82.9 14 71.5 ...
 $ calworks   : num  0.51 15.42 55.03 36.48 33.11 ...
 $ lunch      : num  2.04 47.92 76.32 77.05 78.43 ...
 $ computer   : num  67 101 169 85 171 25 28 66 35 0 ...
 $ expenditure: num  6385 5099 5502 7102 5236 ...
 $ income     : num  22.69 9.82 8.98 8.98 9.08 ...
 $ english    : num  0 4.58 30 0 13.86 ...
 $ read       : num  692 660 636 652 642 ...
 $ math       : num  690 662 651 644 640 ...
```

The dataset contains variables of different types: `chr` for character/text data, `Factor` for categorical data, and `num` for numeric data.

The variable `students` contains the total number of students enrolled in a school. It is the fifth variable in the data set. To access the variable as a vector, you can type `CASchools[,5]` (the fifth column in your data matrix), or `CASchools[,"students"]`, or simply `CASchool$students`.

If you want to select the variables `students` and `teachers`, you can type `CASchools[,c("students", "teachers")]`. We can define our own dataframe `mydata` that contains a selection of variables:

```
mydata = CASchools[,c("students", "teachers", "english", "income", "math", "read")]
head(mydata)
```

```
  students teachers   english    income  math  read
1      195    10.90  0.000000 22.690001 690.0 691.6
2      240    11.15  4.583333  9.824000 661.9 660.5
3     1550    82.90 30.000002  8.978000 650.9 636.3
4      243    14.00  0.000000  8.978000 643.5 651.9
5     1335    71.50 13.857677  9.080333 639.9 641.8
6      137     6.40 12.408759 10.415000 605.4 605.7
```

The pipe operator `|>` efficiently chains commands. It passes the output of one function as the input to another. For example, `mydata |> head()` gives the same output as `head(mydata)`.

A convenient alternative to select a subset of variables of your dataframe is the `select()` function from the `dplyr` package. Let's chain the `select()` and `head()` function:

```r
library(dplyr)
CASchools |> select(students, teachers, english, income, math, read) |> head()
```

```
  students teachers   english    income  math  read
1      195    10.90  0.000000 22.690001 690.0 691.6
2      240    11.15  4.583333  9.824000 661.9 660.5
3     1550    82.90 30.000002  8.978000 650.9 636.3
4      243    14.00  0.000000  8.978000 643.5 651.9
5     1335    71.50 13.857677  9.080333 639.9 641.8
6      137     6.40 12.408759 10.415000 605.4 605.7
```

Piping in R makes code more readable by allowing you to read operations from left to right in a natural order, rather than nesting functions inside each other from the inside out.

We can easily add new variables to our dataframe, for instance, the student-teacher ratio (the total number of students per teacher) and the average test score (average of the math and reading scores):

```r
# compute student-teacher ratio and append it to mydata
mydata$STR = mydata$students/mydata$teachers
# compute test score and append it to mydata
mydata$score = (mydata$read+mydata$math)/2
```

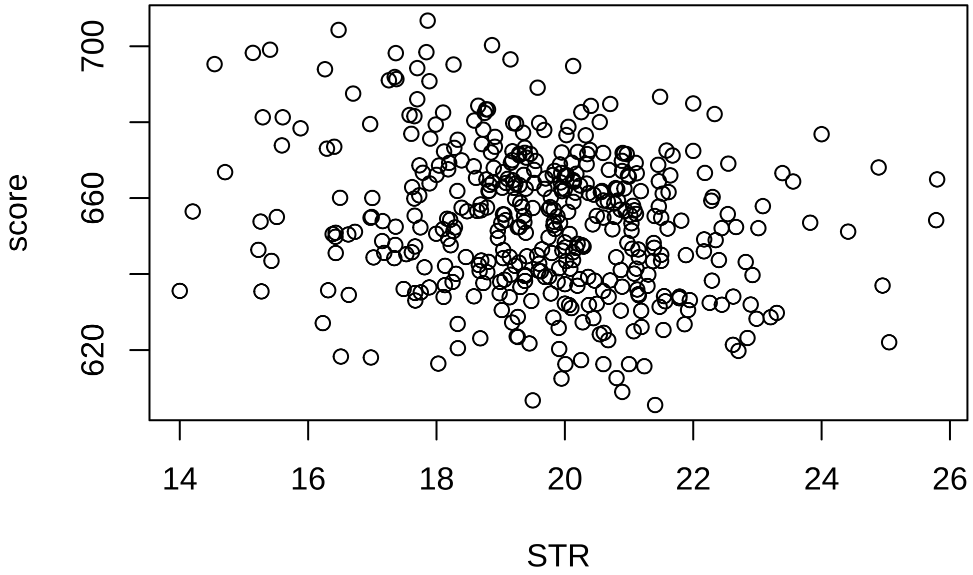
The variable `english` indicates the proportion of students whose first language is not English and who may need additional support. We might be interested in the dummy variable `HiEL`, which indicates whether the proportion of English learners is above 10 percent or not:

```r
# append HiEL to mydata
mydata$HiEL = (mydata$english >= 10) |> as.numeric()
```

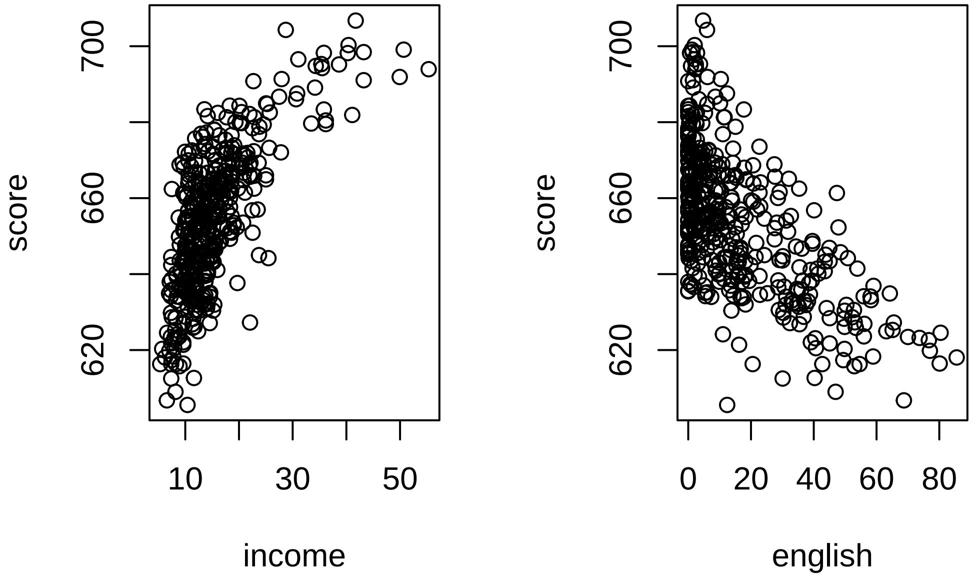
Note that `mydata$english >= 10` is a logical expression with either `TRUE` or `FALSE` values. The command `as.numeric()` creates a dummy variable by translating `TRUE` to `1` and `FALSE` to `0`.

Scatterplots provide further insights:

```r
plot(score~STR, data = mydata)
```



```
par(mfrow = c(1,2))
plot(score~income, data = mydata)
plot(score~english, data = mydata)
```



The option `par(mfrow = c(1,2))` allows to display multiple plots side by side. Try what happens if you replace `c(1,2)` with `c(2,1)`.

4.4 R-codes

statistics-sec01.R

5 Probability

5.1 Random Sampling

From an empirical perspective, a dataset $Y_1, \ldots, Y_n$ or $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is just a fixed array of numbers. Any summary statistic we compute – like a sample mean, sample correlation, or OLS coefficient – is simply a function of these numbers.

These statistics provide a snapshot of the data at hand but do not automatically reveal broader insights about the world. To add deeper meaning to these numbers, identify dependencies, and understand causalities, we must consider how the data were obtained.

A **random experiment** is an experiment whose outcome cannot be predicted with certainty. In statistical theory, any dataset is viewed as the result of such a random experiment.

The gender of the next person you meet, daily fluctuations in stock prices, monthly music streams of your favorite artist, or the annual number of pizzas consumed – all involve a certain amount of randomness and emerge from random experiments.

Sampling is the process of drawing observations from a population. Hence, a dataset is also called a **sample**. Each summary statistic, such as a sample mean or OLS coefficient, is one possible outcome of the random experiment. Repeating the experiment produces a new sample and new statistics.

In statistical theory, the **population** from which we draw observations is treated as infinite. It serves as a theoretical construct that includes not only existing members of a physical population, but all possible future or hypothetical individuals. In coin flip studies, for example, the infinite population represents not just all coin flips ever performed, but all possible coin flips that could theoretically occur in any context at any time.

The goal of **statistical inference** is to learn about the world from the observed sample. This requires assumptions about how the data were collected.

The simplest ideal assumption is **random sampling**, where each observation is drawn independently from the population – like drawing balls from an urn or randomly selecting survey participants. This principle is often called **i.i.d. sampling** (independent and identically distributed sampling). To define these concepts rigorously, we rely on **probability theory**.

5.2 Random variables

A **random variable** is a numerical summary of a random experiment. An **outcome** is a specific result of a random experiment. The **sample space** S is the set/collection of all potential outcomes.

Let's consider some examples:

- *Coin toss:* The outcome of a coin toss can be "heads" or "tails". This random experiment has a two-element sample space: $S = \{heads, tails\}$. We can express the experiment as a binary random variable:
$$Y = \begin{cases} 1 & \text{if outcome is heads,} \\ 0 & \text{if outcome is tails.} \end{cases}$$
- *Gender:* If you conduct a survey and interview a random person to ask them about their gender, the answer may be "female", "male", or "diverse". It is a random experiment since the person to be interviewed is selected randomly. The sample space has three elements: $S = \{female, male, diverse\}$. To focus on female vs. non-female, we can define the female dummy variable:
$$Y = \begin{cases} 1 & \text{if the person is female,} \\ 0 & \text{if the person is not female.} \end{cases}$$
Similarly, dummy variables for *male* and *diverse* can be defined.
- *Education level:* If you ask a random person about their education level according to the ISCED-2011 framework, the outcome may be one of the eight ISCED-2011 levels. We have an eight-element sample space:
$$S = \{Level\ 1, Level\ 2, Level\ 3, Level\ 4, Level\ 5, Level\ 6, Level\ 7, Level\ 8\}.$$
The eight-element sample space of the education-level random experiment provides a natural ordering. We define the random variable *education* as the number of years of schooling of the interviewed person:
$$Y = \text{years of schooling} \in \{4, 10, 12, 13, 14, 16, 18, 21\}.$$
- *Wage*: If you ask a random person about their income per working hour in EUR, there are infinitely many potential answers. Any (non-negative) real number may be an outcome. The sample space is a continuum of different wage levels. The wage level of the interviewed is already numerical. The random variable is
$$Y = \text{income per working hour in EUR.}$$

Random variables share the characteristic that their value is uncertain before conducting a random experiment (e.g., flipping a coin or selecting a random person for an interview). Their value is always a real number and is determined only once the experiment's outcome is known.

Table 5.1: ISCED 2011 levels

[H]

ISCED level	Education level	Years of schooling
1	Primary	4
2	Lower Secondary	10
3	Upper secondary	12
4	Post-Secondary	13
5	Short-Cycle Tertiary	14
6	Bachelor's	16
7	Master's	18
8	Doctoral	21

5.3 Events and probabilities

An **event** of a random variable Y is a specific subset of the real line. Any real number defines an event (elementary event), and any open, half-open, or closed interval represents an event as well.

Let's define some specific events:

- Elementary events:

$$A_1 = \{Y = 0\}, \quad A_2 = \{Y = 1\}, \quad A_3 = \{Y = 2.5\}$$

- Half-open events:

$$\begin{aligned} A_4 &= \{Y \geq 0\} = \{Y \in [0, \infty)\} \\ A_5 &= \{-1 \leq Y < 1\} = \{Y \in [-1, 1)\}. \end{aligned}$$

The **probability function** P assigns values between 0 and 1 to events. For a fair coin toss it is natural to assign the following probabilities:

$$P(A_1) = P(Y = 0) = 0.5, \quad P(A_2) = P(Y = 1) = 0.5$$

By definition, the coin variable will never take the value 2.5, so we assign

$$P(A_3) = P(Y = 2.5) = 0.$$

To assign probabilities to interval events, we check whether the events $\{Y = 0\}$ and/or $\{Y = 1\}$ are subsets of the event of interest.

If both $\{Y = 0\}$ and $\{Y = 1\}$ are contained in the event of interest, the probability is 1. If only one of them is contained, the probability is 0.5. If neither is contained, the probability is 0.

$$P(A_4) = P(Y \geq 0) = 1, \quad P(A_5) = P(-1 \leq Y < 1) = 0.5.$$

Every event has a **complementary event**, and for any pair of events we can take the **union** and **intersection**. Let's define further events:

- Complements:

$$A_6 = A_4^c = \{Y \geq 0\}^c = \{Y < 0\} = \{Y \in (-\infty, 0)\},$$

- Unions:

$$A_7 = A_1 \cup A_6 = \{Y = 0\} \cup \{Y < 0\} = \{Y \leq 0\}$$

- Intersections:

$$A_8 = A_4 \cap A_5 = \{Y \geq 0\} \cap \{-1 \leq Y < 1\} = \{0 \leq Y < 1\}$$

- Iterations of it:

$$A_9 = A_1 \cup A_2 \cup A_3 \cup A_5 \cup A_6 \cup A_7 \cup A_8 = \{Y \in (-\infty, 1] \cup \{2.5\}\},$$

- Certain event:

$$A_{10} = A_9 \cup A_9^c = \{Y \in (-\infty, \infty)\} = \{Y \in \mathbb{R}\}$$

- Empty event:

$$A_{11} = A_{10}^c = \{Y \notin \mathbb{R}\} = \{\}$$

You may verify that $P(A_1) = 0.5$, $P(A_2) = 0.5$, $P(A_3) = 0$, $P(A_4) = 1$ $P(A_5) = 0.5$, $P(A_6) = 0$, $P(A_7) = 0.5$, $P(A_8) = 0.5$, $P(A_9) = 1$, $P(A_{10}) = 1$, $P(A_{11}) = 0$ for the coin toss experiment.

5.4 Probability function

The probability function P assigns probabilities to events. The set of all events for which probabilities can be assigned is called the Borel sigma-algebra, denoted as $\mathcal{B}$.

The previously mentioned events $A_1, \ldots, A_{11}$ are elements of $\mathcal{B}$. Any event of the form $\{Y \in (a, b)\}$ with $a, b \in \mathbb{R}$ is also in $\mathcal{B}$. Moreover, $\mathcal{B}$ includes all possible unions, intersections, and complements of these events. Essentially, it represents the complete collection of events for which we would ever compute probabilities in practice.

A probability function P must satisfy certain fundamental rules (axioms) to ensure a well-defined probability framework:

Basic rules of probability

- $P(A) \geq 0$ for any event A
- $P(Y \in \mathbb{R}) = 1$ for the certain event
- $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint
- $P(Y \notin \mathbb{R}) = 0$ for the empty event

- $0 \leq P(A) \leq 1$ for any event A
- $P(A) \leq P(B)$ if A is a subset of B
- $P(A^c) = 1 - P(A)$ for the complement event of A
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A, B

Two events A and B are **disjoint** if $A \cap B = \{\}$, meaning they have no common outcomes. For instance, $A_1 = \{Y = 0\}$ and $A_2 = \{Y = 1\}$ are disjoint. However, A_1 and $A_4 = \{Y \geq 0\}$ are not disjoint because their intersection, $A_1 \cap A_4 = \{Y = 0\}$, is nonempty.

The first three properties listed above are known as the axioms of probability. The remaining properties follow as logical consequences of these axioms.

5.5 Distribution function

Assigning probabilities to events is straightforward for binary variables, like coin tosses. For instance, knowing that $P(Y = 1) = 0.5$ allows us to derive the probabilities for all events in $\mathcal{B}$. However, for more complex variables, such as *education* or *wage*, defining probabilities for all possible events becomes more challenging due to the vast number of potential set operations involved.

Fortunately, it turns out that knowing the probabilities of events of the form $\{Y \leq a\}$ is enough to determine the probabilities of all other events. These probabilities are summarized in the cumulative distribution function.

Cumulative distribution function (CDF)

The cumulative distribution function (CDF) of a random variable Y is

$$F(a) := P(Y \leq a), \quad a \in \mathbb{R}.$$

The CDF is sometimes simply referred to as the **distribution function**, or the **distribution**.

The cumulative distribution function (CDF) of the variable *coin* is

$$F(a) = \begin{cases} 0 & a < 0, \\ 0.5 & 0 \leq a < 1, \\ 1 & a \geq 1, \end{cases}$$

with the following CDF plot:

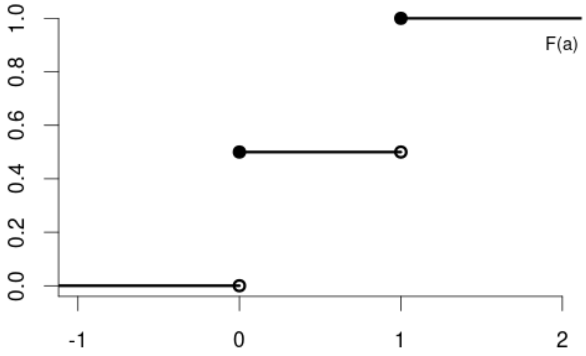


Figure 5.1: CDF of coin

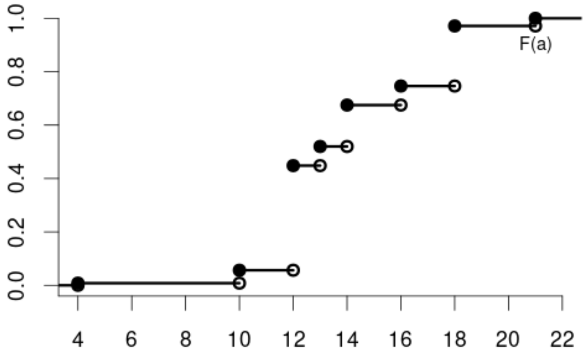


Figure 5.2: CDF of education

The CDF of the variable *education* could be

and the CDF of the variable *wage* may have the following form:

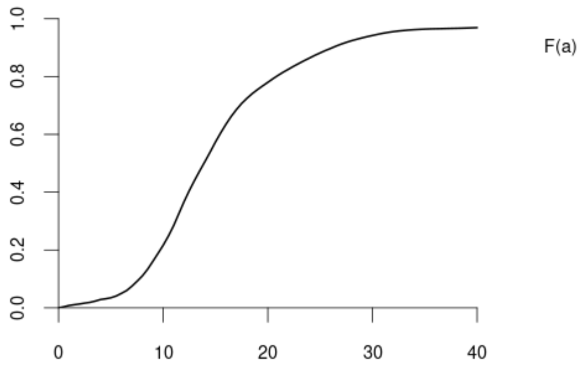


Figure 5.3: CDF of wage

The CDF of a **continuous random variable** is smooth, while the CDF of a **discrete random variable** contains jumps and is flat between jumps. For example, variables like *coin* and *education* are discrete, whereas *wage* is continuous.

Any function $F(a)$ with the following properties defines a valid probability distribution:

- Non-decreasing: $F(a) \leq F(b)$ for $a \leq b$;
- Limits at 0 and 1: $\lim_{a \to -\infty} F(a) = 0$ and $\lim_{a \to \infty} F(a) = 1$
- Right-continuity: $\lim_{\varepsilon \to 0, \varepsilon \geq 0} F(a + \varepsilon) = F(a)$

Right-continuity ensures that cumulative probabilities include the probability at each point, which is especially important for discrete variables with their jump points.

The right-continuity property means that the CDF includes the probability mass at each point a, ensuring $P(Y \leq a)$ includes $P(Y = a)$. This property is particularly important for discrete random variables where there are jumps in the CDF.

By the basic rules of probability, we can compute the probability of any event of interest if we know the probabilities of all events of the forms $\{Y \leq a\}$ and $\{Y = a\}$.

Some basic rules for the CDF (for $a < b$):

- $P(Y \leq a) = F(a)$
- $P(Y > a) = 1 - F(a)$
- $P(Y < a) = F(a) - P(Y = a)$
- $P(Y \geq a) = 1 - P(Y < a)$
- $P(a < Y \leq b) = F(b) - F(a)$
- $P(a < Y < b) = F(b) - F(a) - P(Y = b)$
- $P(a \leq Y \leq b) = F(b) - F(a) + P(Y = a)$

- $P(a \leq Y < b) = P(a \leq Y \leq b) - P(Y = b)$

A probability of the form $P(Y = a)$, which involves only an elementary event, is called a **point probability**.

5.6 Probability mass function

The **point probability** $P(Y = a)$ represents the size of the jump at $a \in \mathbb{R}$ in the CDF $F(a)$:

$$P(Y = a) = F(a) - \lim_{\epsilon \to 0, \varepsilon \geq 0} F(a - \varepsilon),$$

which is the jump height at a. We summarize the CDF jump heights or point probabilities in the probability mass function:

Probability mass function (PMF)

The probability mass function (PMF) of a random variable Y is

$$\pi(a) := P(Y = a), \quad a \in \mathbb{R}$$

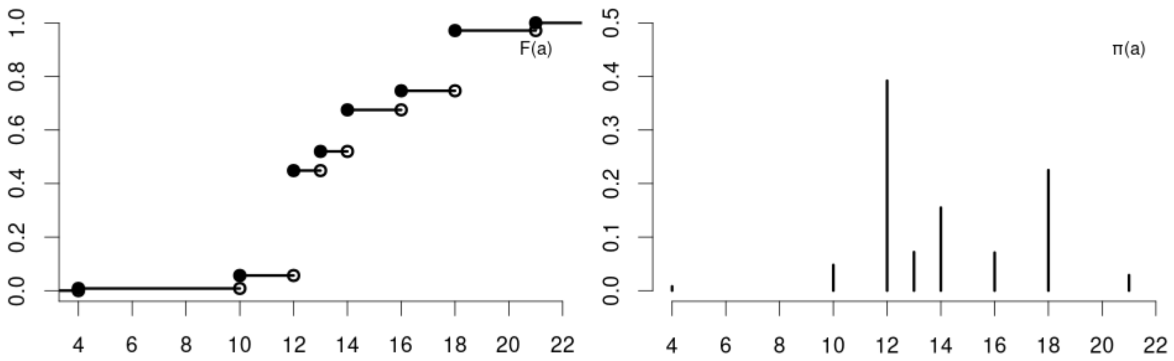
The PMF of the *coin* variable is

$$\pi(a) = P(Y = a) = \begin{cases} 0.5 & \text{if } a \in \{0, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

The *education* variable may have the following PMF:

$$\pi(a) = P(Y = a) = \begin{cases} 0.008 & \text{if } a = 4 \\ 0.048 & \text{if } a = 10 \\ 0.392 & \text{if } a = 12 \\ 0.072 & \text{if } a = 13 \\ 0.155 & \text{if } a = 14 \\ 0.071 & \text{if } a = 16 \\ 0.225 & \text{if } a = 18 \\ 0.029 & \text{if } a = 21 \\ 0 & \text{otherwise} \end{cases}$$

Because continuous variables have no jumps in their CDF, the PMF concept makes only sense for discrete random variables.



(a) CDF of education

(b) PMF of education

5.7 Probability density function

For continuous random variables, the CDF has no jumps, meaning the probability of any specific value is zero, and probability is distributed continuously over intervals. Unlike discrete random variables, which are characterized by both the PMF and the CDF, continuous variables do not have a positive PMF. Instead, they are described by the probability density function (PDF), which serves as the continuous analogue. If the CDF is differentiable, the PDF is given by its derivative:

Probability density function

The **probability density function (PDF)** or simply **density function** of a continuous random variable Y is the derivative of its CDF:

$$f(a) = \frac{d}{da}F(a).$$

Conversely, the CDF can be obtained from the PDF by integration:

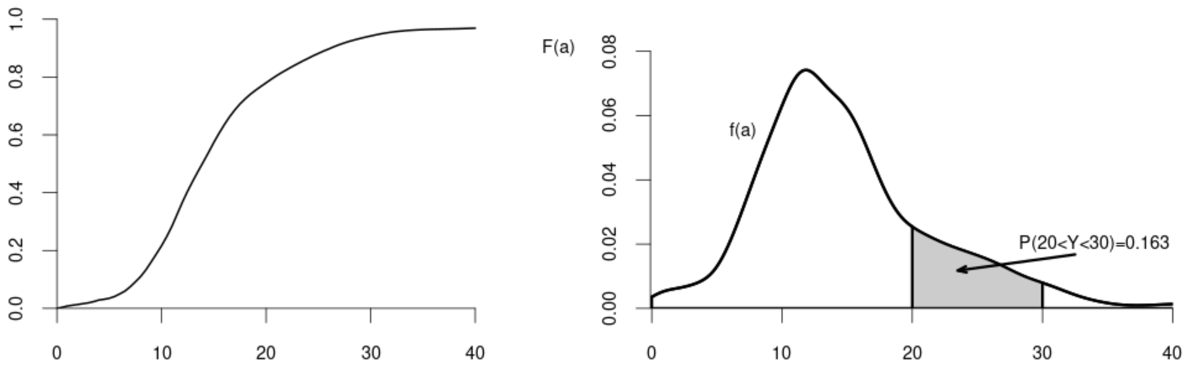
$$F(a) = \int_{-\infty}^{a} f(u)\ \mathrm{d}u$$

Any function $f(a)$ with the following properties defines a valid probability density function:

- Non-negativity: $f(a) \geq 0$ for all $a \in \mathbb{R}$;
- Normalization: $\int_{-\infty}^{\infty} f(u)\ \mathrm{d}u = 1$.

Basic rules for **continuous random variables** (with $a \leq b$):

- $P(Y = a) = \displaystyle\int_{a}^{a} f(u)\ \mathrm{d}u = 0$



(a) CDF of wage

(b) PDF of wage

- $P(Y \leq a) = P(Y < a) = F(a) = \int_{-\infty}^{a} f(u) \, \mathrm{d}u$
- $P(Y > a) = P(Y \geq a) = 1 - F(a) = \int_{a}^{\infty} f(u) \, \mathrm{d}u$
- $P(a < Y < b) = F(b) - F(a) = \int_{a}^{b} f(u) \, \mathrm{d}u$
- $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b) = P(a \leq Y < b)$

5.8 Conditional distribution

The distribution of *wage* may differ between men and women. Similarly, the distribution of *education* may vary between married and unmarried individuals. In contrast, the distribution of a *coin flip* should remain the same regardless of whether the person tossing the coin earns 15 or 20 EUR per hour.

The **conditional cumulative distribution function** (conditional CDF),

$$F_{Y|Z=b}(a) = F_{Y|Z}(a|b) = P(Y \leq a|Z = b),$$

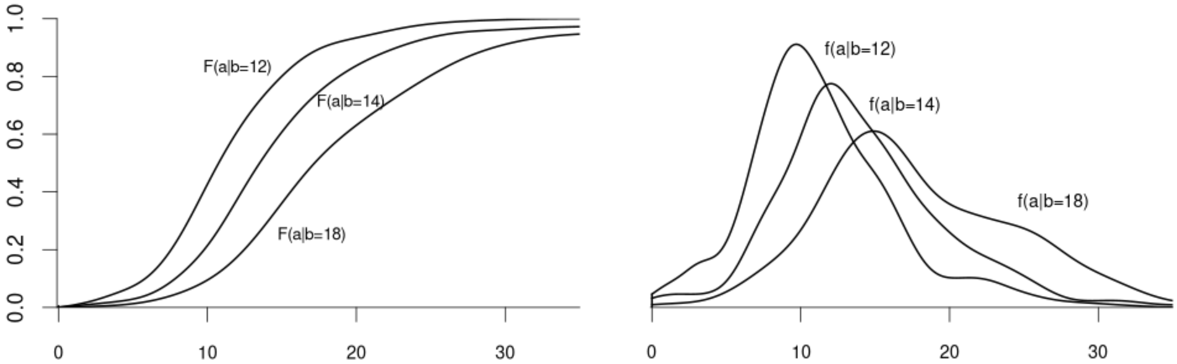
represents the distribution of a random variable Y given that another random variable Z takes a specific value b. It answers the question: "If we know that $Z = b$, what is the distribution of Y?"

For example, suppose that Y represents *wage* and Z represents *education*

- $F_{Y|Z=12}(a)$ is the CDF of wages among individuals with 12 years of education.
- $F_{Y|Z=14}(a)$ is the CDF of wages among individuals with 14 years of education.
- $F_{Y|Z=18}(a)$ is the CDF of wages among individuals with 18 years of education.

Since *wage* is a continuous variable, its conditional distribution given any specific value of another variable is also continuous. The conditional density of Y given $Z = b$ is defined as the derivative of the conditional CDF:

$$f_{Y|Z=b}(a) = f_{Y|Z}(a|b) = \frac{d}{da} F_{Y|Z=b}(a).$$



(a) Conditional CDFs of wage given education (b) Conditional PDFs of wage given education

We observe that the distribution of wage varies across different levels of education. For example, individuals with fewer years of education are more likely to earn less than 20 EUR per hour:

$$P(Y \leq 20|Z = 12) = F_{Y|Z=12}(20) > F_{Y|Z=18}(20) = P(Y \leq 20|Z = 18).$$

Because the conditional distribution of Y given $Z = b$ depends on the value of $Z = b$ we say that the random variables Y and Z are **dependent random variables**.

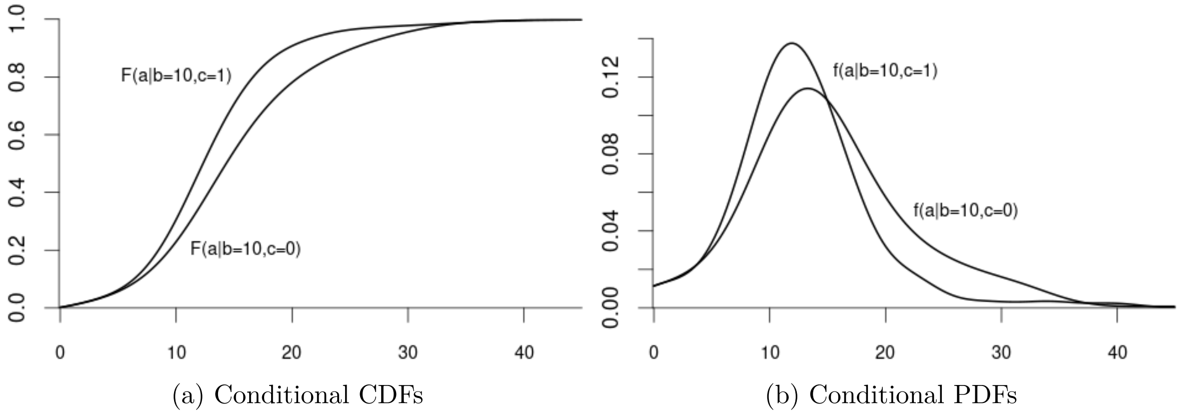
Note that the conditional CDF $F_{Y|Z=b}(a)$ can only be defined for events $Z = b$ that are possible, i.e. b must be in the support of Z. Formally, the support consists of all $b \in \mathbb{R}$ where the cumulative distribution function $F_Z(b)$ is not flat – meaning it either increases continuously or has a jump. For instance, the support of the variable *education* is $\{4, 10, 12, 13, 14, 16, 18, 21\}$ and the support of the variable *wage* is $\{a \in \mathbb{R} : a \geq 0\}$.

We can also condition on more than one variable. Let Z_1 represent the labor market *experience* in years and Z_2 be the *female* dummy variable. The conditional CDF of Y given $Z_1 = b$ and $Z_2 = c$ is:

$$F_{Y|Z_1=b,Z_2=c}(a) = F_{Y|Z_1,Z_2}(a|b, c) = P(Y \leq a|Z_1 = b, Z_2 = c).$$

For example:

- $F_{Y|Z_1=10,Z_2=1}(a)$ is the CDF of wages among women with 10 years of experience.
- $F_{Y|Z_1=10,Z_2=0}(a)$ is the CDF of wages among men with 10 years of experience.



(a) Conditional CDFs

(b) Conditional PDFs

Figure 5.7: Conditional CDFs and PDFs of wage given experience and gender

Clearly the random variable Y and the random vector (Z_1, Z_2) are dependent.

More generally, we can condition on the event that a k-variate random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$ takes the value $\{\boldsymbol{Z} = \boldsymbol{b}\}$, i.e. $\{Z_1 = b_1, \ldots, Z_k = b_k\}$. The conditional CDF of Y given $\{\boldsymbol{Z} = \boldsymbol{b}\}$ is

$$F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) = F_{Y|Z_1=b_1,\ldots,Z_k=b_k}(a).$$

The variable of interest, Y, can also be discrete. Then, any conditional CDF of Y is also discrete. Below is the conditional CDF of *education* given the *married* dummy variable:

- $F_{Y|Z=0}(a)$ is the CDF of education among unmarried individuals.
- $F_{Y|Z=1}(a)$ is the CDF of education among married individuals.

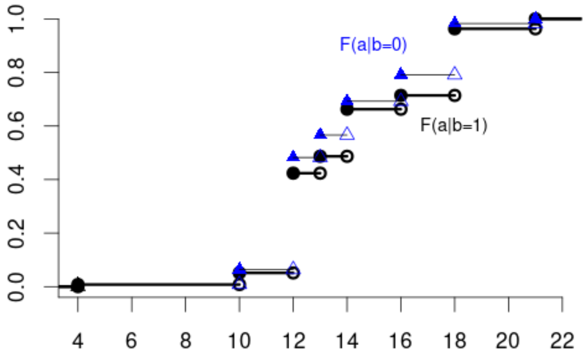


Figure 5.8: Conditional CDFs of education given married

The conditional PMFs $\pi_{Y|Z=0}(a) = P(Y = a|Z = 0)$ and $\pi_{Y|Z=1}(a) = P(Y = a|Z = 1)$ indicate the jump heights of $F_{Y|Z=0}(a)$ and $F_{Y|Z=1}(a)$ at a.

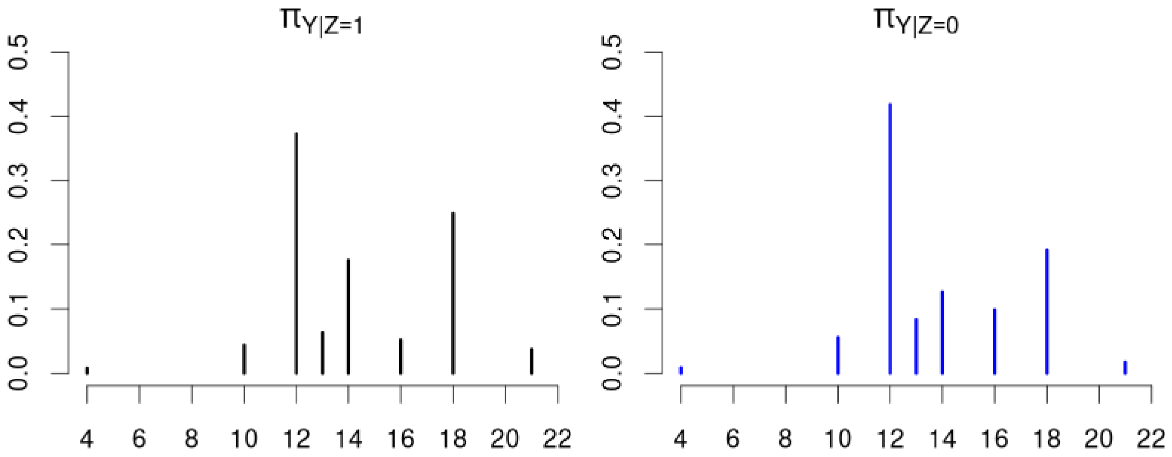
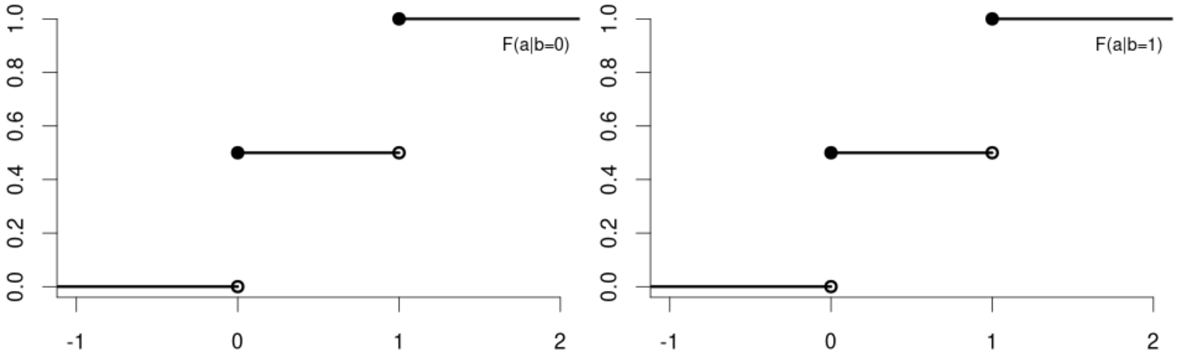


Figure 5.9: Conditional PMFs of education given married

Clearly, *education* and *married* are dependent random variables. E.g., $\pi_{Y|Z=0}(12) > \pi_{Y|Z=1}(12)$ and $\pi_{Y|Z=0}(18) < \pi_{Y|Z=1}(18)$.

In contrast, consider Y = coin flip and Z = married dummy variable. The CDF of a coin flip should be the same for married or unmarried individuals:



(a) Coin flip given married

(b) Coin flip given unmarried

Figure 5.10: Conditional CDFs of a coin flip of a married (left) and unmarried (right) individual

Because

$$F_Y(a) = F_{Y|Z=0}(a) = F_{Y|Z=1}(a) \quad \text{for all } a$$

we say that Y and Z are **independent random variables**.

5.9 Independence of random variables

Independence

Y and Z are **independent** if and only if

$$F_{Y|Z=b}(a) = F_Y(a) \quad \text{for all } a \quad \text{and for almost every } b.$$

Note that if $F_{Y|Z=b}(a) = F_Y(a)$ for all b, then automatically $F_{Z|Y=a}(b) = F_Y(b)$ for all a. Due to this symmetry we can equivalently define independence through the property $F_{Z|Y=a}(b) = F_Z(b)$.

Here, "for almost every b" means for every b in the support of Z, apart from a set of values that has probability 0 under Z. Put differently, the condition must hold for all b-values that Z can actually take, with exceptions allowed only on a set whose probability is 0. Think of it as "for all practical purposes". The condition must hold for all values b that could realistically occur. For instance, we only need independence to hold for non-negative wages. We don't need to check independence for negative wages since they can't occur.

The definition naturally generalizes to Z_1, Z_2, Z_3. They are **mutually independent** if, for each $i \in \{1, 2, 3\}$, the conditional distribution of Z_i given the other two equals its marginal distribution. In CDF form, this means:

(i) $F_{Z_1|Z_2=b_2,Z_3=b_3}(a) = F_{Z_1}(a)$
(ii) $F_{Z_2|Z_1=b_1,Z_3=b_3}(a) = F_{Z_2}(a)$
(iii) $F_{Z_3|Z_1=b_1,Z_2=b_2}(a) = F_{Z_3}(a)$

for all a and for almost every (b_1, b_2, b_3). Here, we need all three conditions.

Mutual independence

The random variables $Z_1, \ldots, Z_n$ are **mutually independent** if and only if, for each $i = 1, \ldots, n$,

$$F_{Z_i|Z_1=b_1,\ldots,Z_{i-1}=b_{i-1},\, Z_{i+1}=b_{i+1},\ldots,Z_n=b_n}(a) = F_{Z_i}(a).$$

for all a and almost every $(b_1, \ldots, b_n)$.

An equivalent viewpoint uses the **joint CDF** of the vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$, which is defined as:

$$F_{\boldsymbol{Z}}(\boldsymbol{a}) = F_{Z_1,\ldots,Z_n}(a_1, \ldots, a_n) = P(Z_1 \leq a_1, \ldots, Z_n \leq a_n) = P(\boldsymbol{Z} \leq \boldsymbol{a}),$$

where

$$P(Z_1 \leq a_1, \ldots, Z_n \leq a_n) = P(\{Z_1 \leq a_1\} \cap \ldots \cap \{Z_n \leq a_n\}).$$

Then $Z_1, \ldots, Z_n$ are mutually independent if and only if the joint CDF is the product of the marginal CDFs:

$$F_{\boldsymbol{Z}}(\boldsymbol{a}) = F_{Z_1}(a_1) \cdots F_{Z_n}(a_n) \quad \text{for all } a_1, \ldots, a_n.$$

5.10 Independence of random vectors

Often in practice, we work with multiple variables recorded for different individuals or time points. For example, consider two random vectors:

$$\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1k})', \quad \boldsymbol{X}_2 = (X_{21}, \ldots, X_{2k})'.$$

The conditional distribution function of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2$ takes the value $\boldsymbol{b} = (b_1, \ldots, b_k)'$ is

$$F_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{b}}(\boldsymbol{a}) = P(\boldsymbol{X}_1 \leq \boldsymbol{a}|\boldsymbol{X}_2 = \boldsymbol{b}),$$

where $\boldsymbol{X}_1 \leq \boldsymbol{a}$ means $X_{1j} \leq a_j$ for each coordinate $j = 1, \ldots, k$.

For instance, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ represent the survey answers of two different, randomly chosen people, then $F_{\boldsymbol{X}_2|\boldsymbol{X}_1=\boldsymbol{b}}(\boldsymbol{a})$ describes the distribution of the second person's answers, given that the first person's answers are $\boldsymbol{b}$. If the two people are truly randomly selected and unrelated to one another, we would not expect $\boldsymbol{X}_2$ to depend on whether $\boldsymbol{X}_1$ equals $\boldsymbol{b}$ or some other value $\boldsymbol{c}$. In other words, knowing $\boldsymbol{X}_1$ provides no information that changes the distribution of $\boldsymbol{X}_2$.

Independence of random vectors

Two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are **independent** if and only if

$$F_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{b}}(\boldsymbol{a}) = F_{\boldsymbol{X}_1}(\boldsymbol{a}) \quad \text{for all } \boldsymbol{a} \quad \text{and for almost every } \boldsymbol{b}.$$

This definition extends naturally to mutual independence of n random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$. They are called **mutually independent** if, for each $i = 1, \ldots, n$,

$$F_{\boldsymbol{X}_i|\boldsymbol{X}_1=\boldsymbol{b}_1,\ldots,\boldsymbol{X}_{i-1}=\boldsymbol{b}_{i-1},\boldsymbol{X}_{i+1}=\boldsymbol{b}_{i+1},\ldots,\boldsymbol{X}_n=\boldsymbol{b}_n}(\boldsymbol{a}) = F_{\boldsymbol{X}_i}(\boldsymbol{a})$$

for all $\boldsymbol{a}$ and almost every $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$.

Hence, in an independent sample, what the i-th randomly chosen person answers does not depend on anyone else's answers.

i.i.d. sample / random sample

A collection of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is **i.i.d.** (independent and identically distributed) if they are mutually independent and have the same distribution function F. Formally,

$$F_{\boldsymbol{X}_i|\boldsymbol{X}_1=\boldsymbol{b}_1,\ldots,\boldsymbol{X}_{i-1}=\boldsymbol{b}_{i-1},\boldsymbol{X}_{i+1}=\boldsymbol{b}_{i+1},\ldots,\boldsymbol{X}_n=\boldsymbol{b}_n}(\boldsymbol{a}) = F(\boldsymbol{a})$$

for all $i = 1, \ldots, n$, for all $\boldsymbol{a}$, and almost all $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$.

An **i.i.d. dataset** (or **random sample**) is one where each observation not only comes from the same population distribution F but is independent of the others. The function F is called the **population distribution** or the **data-generating process (DGP)**.

The CPS data are **cross-sectional** data: n individuals are randomly selected from the U.S. population and independently interviewed on k variables. Consequently, these n observations form an i.i.d. sample.

If $Y_1, \ldots, Y_n$ are i.i.d., then $\log(Y_1), \ldots, \log(Y_n)$ are also i.i.d. In fact, any identical transformation of each observation preserves the independence and identical distribution. More formally, if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is i.i.d., then $g(\boldsymbol{X}_1), \ldots, g(\boldsymbol{X}_n)$ is i.i.d. as well, for any function $g(\cdot)$. For instance, if the wages of n interviewed individuals are i.i.d., then their log-wages are also i.i.d.

Sampling methods of obtaining economic datasets that may be considered as random sampling are:

- **Survey sampling**
  Examples: representative survey of randomly selected households from a list of residential addresses; online questionnaire to a random sample of recent customers
- **Administrative records**
  Examples: data from a government agency database, Statistisches Bundesamt, ECB, etc.
- **Direct observation**
  Collected data without experimental control and interactions with the subject. Example: monitoring customer behavior in a retail store
- **Web scraping**
  Examples: collected house prices on real estate sites or hotel/electronics prices on booking.com/amazon, etc.
- **Field experiment**
  To study the impact of a treatment or intervention on a treatment group compared with a control group. Example: testing the effectiveness of a new teaching method by implementing it in a selected group of schools and comparing results to other schools with traditional methods
- **Laboratory experiment**
  Example: a controlled medical trial for a new drug

Examples of cross-sectional data sampling that may produce some dependence across observations are:

- **Stratified sampling**
  The population is first divided into homogenous subpopulations (strata), and a random sample is obtained from each stratum independently. Examples: divide companies into industry strata (manufacturing, technology, agriculture, etc.) and sample from each stratum; divide the population into income strata (low-income, middle-income, high-income).
  The sample is independent within each stratum, but it is not between different strata. The strata are defined based on specific characteristics that may be correlated with the variables collected in the sample.
- **Clustered sampling**
  Entire subpopulations are drawn. Example: new teaching methods are compared to traditional ones on the student level, where only certain classrooms are randomly selected, and all students in the selected classes are evaluated.
  Within each cluster (classroom), the sample is dependent because of the shared environment and teacher's performance, but between classrooms, it is independent.

Other types of data we often encounter in econometrics are time series data, panel data, or spatial data:

- **Time series data** consists of observations collected at different points in time, such as stock prices, daily temperature measurements, or GDP figures. These observations are ordered and typically show temporal trends, seasonality, and autocorrelation.
- **Panel data** involves observations collected on multiple entities (e.g., individuals, firms, countries) over multiple time periods. Every entity thus forms a cluster, within which there is a time series of observations. In this sense, panel data is a specific form of clustered sampling.
- **Spatial data** includes observations taken at different geographic locations, where values at nearby locations are often correlated.

Time series, panel, and spatial data cannot be considered a random sample given their temporal or geographic dependence.

5.11 R-codes

statistics-sec04.R

6 Expectated value

The **expectation** or **expected value** is the most important measure of the central tendency of a distribution. It gives you the average value you can expect to get if you repeat the random experiment multiple times. We define the expectation first for discrete random variables, then continuous random variables, and finally give a unified definition for all random variables.

6.1 Discrete random variables

Recall that a discrete random variable Y is a variable that can take on a countable number of distinct values. Each possible value a has an associated probability $\pi(a) = P(Y = a)$, known as the probability mass function (PMF).

The support $\mathcal{Y}$ of Y is the set of all values that Y can take with non-zero probability:

$$\mathcal{Y} = \{a \in \mathbb{R} : \pi(a) > 0\}.$$

The total probability sums to 1: $\sum_{a \in \mathcal{Y}} \pi(a) = 1$.

The **expectation** or **expected value** of a discrete random variable Y with PMF $\pi(\cdot)$ and support $\mathcal{Y}$ is defined as

$$E[Y] = \sum_{u \in \mathcal{Y}} u\pi(u). \tag{6.1}$$

The expected value of the variable *education* from the previous section is calculated by summing over all possible values:

$$\begin{aligned} E[Y] &= 4 \cdot \pi(4) + 10 \cdot \pi(10) + 12 \cdot \pi(12) \\ &\quad + 13 \cdot \pi(13) + 14 \cdot \pi(14) + 16 \cdot \pi(16) \\ &\quad + 18 \cdot \pi(18) + 21 \cdot \pi(21) = 14.117 \end{aligned}$$

A **binary** or **Bernoulli** random variable Y takes on only two possible values: 0 and 1. The support is $\mathcal{Y} = \{0, 1\}$. The probabilities are

- $\pi(1) = P(Y = 1) = p$
- $\pi(0) = P(Y = 0) = 1 - p$

for some $p \in (0,1)$. The expected value of Y is:

$$\begin{aligned} E[Y] &= 0 \cdot \pi(0) + 1 \cdot \pi(1) \\ &= 0 \cdot (1-p) + 1 \cdot p \\ &= p. \end{aligned}$$

For the variable *coin*, the probability of heads is $p = 0.5$ and the expected value is $E[Y] = p = 0.5$.

6.2 Continuous random variables

For discrete random variables, both the PMF and the CDF characterize the distribution. For continuous random variables, the PMF concept does not apply because the probability of any specific point is zero. The continuous counterpart of the PMF is the density function:

Probability density function

The **probability density function (PDF)** or simply **density function** of a continuous random variable Y with CDF $F(a)$ is a function $f(a)$ that satisfies

$$F(a) = \int_{-\infty}^{a} f(u) \ \mathrm{d}u$$

If the CDF is differentiable, the density $f(a)$ is its derivative:

$$f(a) = \frac{d}{da} F(a).$$

Properties of a PDF:

(i) $f(a) \geq 0$ for all $a \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f(u) \ \mathrm{d}u = 1$

Probability rule for the PDF:

$$P(a < Y < b) = \int_{a}^{b} f(u) \ \mathrm{d}u = F(b) - F(a)$$

The **expectation** or **expected value** of a continuous random variable Y with PDF $f(\cdot)$ is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \ \mathrm{d}u. \tag{6.2}$$

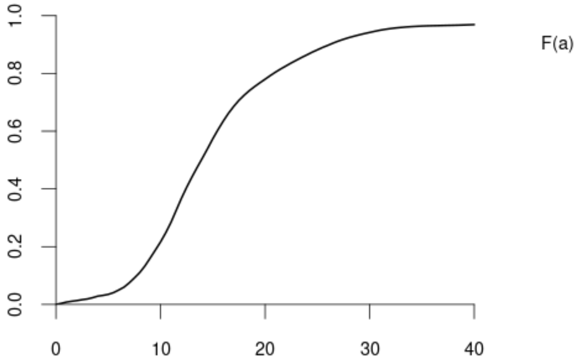


Figure 6.1: CDF of wage

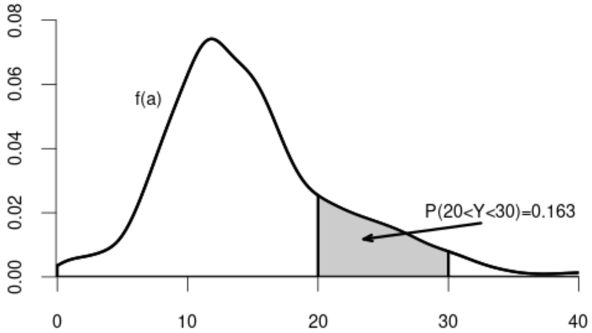


Figure 6.2: PDF of wage

The uniform distribution on the unit interval $[0, 1]$ has the PDF

$$f(u) = \begin{cases} 1 & \text{if } u \in [0, 1], \\ 0 & \text{otherwise}, \end{cases} \tag{6.3}$$

and the expected value of a uniformly distributed random variable Y is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u = \int_0^1 u \, \mathrm{d}u = \frac{1}{2} u^2 \Big|_0^1 = \frac{1}{2}.$$

6.3 Unified definition of the expected value

The expected value of a random variable Y can be defined in a unified way that applies to both discrete and continuous cases by using its CDF $F(u)$:

$$E[Y] = \int_{-\infty}^{\infty} u \, \mathrm{d}F(u). \tag{6.4}$$

This integral, known as the **Riemann-Stieltjes integral**, generalizes the concept of integration to include functions that may not be smooth or differentiable everywhere.

For a continuous random variable with PDF $f(u)$, the CDF $F(u)$ is smooth and differentiable. The relationship between the CDF and the PDF is:

$$\mathrm{d}F(u) = f(u) \, \mathrm{d}u.$$

Substituting this into our unified definition gives:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} u \, \mathrm{d}F(u) \\ &= \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u, \end{aligned}$$

which matches the standard definition of the expected value for continuous random variables as in Equation 6.2.

For a discrete random variable, the CDF $F(u)$ is a step function that increases in jumps at the possible values $u \in \mathcal{Y}$ that Y can take. The "change" or jump in the CDF at each $u \in \mathcal{Y}$ is:

$$\Delta F(u) = F(u) - F(u^-) = P(Y = u) = \pi(u),$$

where $F(u^-)$ is the value of $F(u)$ just before u, and $\pi(u)$ is the PMF of Y.

Integrating with respect to $F(u)$ simplifies to summing over these jumps:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} u \ \mathrm{d}F(u) \\ &= \sum_{u \in \mathcal{Y}} u \ \Delta F(u) \\ &= \sum_{u \in \mathcal{Y}} u\pi(u), \end{aligned}$$

which aligns with the standard definition of the expected value for discrete random variables as in Equation 6.1.

The unified definition $E[Y] = \int_{-\infty}^{\infty} u \ \mathrm{d}F(u)$ allows us to treat all types of random variables consistently, whether the variable is discrete, continuous, or a mixture of both. It can also handle non-standard cases such as distributions with CDFs that are not differentiable everywhere.

6.4 Transformed variables

We often transform random variables by taking, for instance, squares Y^2 or logs $\log(Y)$. For any transformation function $g(\cdot)$, the expectation of the transformed random variable $g(Y)$ is

$$E[g(Y)] = \int_{-\infty}^{\infty} g(u) \ \mathrm{d}F(u),$$

where $F(u)$ is the CDF of Y. As discussed in Section 6.3 for the different cases, $\mathrm{d}F(u)$ can be replaced by the PMF or the PDF, i.e.,

$$\int_{-\infty}^{\infty} g(u) \ \mathrm{d}F(u) = \begin{cases} \sum_{u \in \mathcal{Y}} g(u)\pi(u) & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(u)f(u)\mathrm{d}u & \text{if } Y \text{ is continuous.} \end{cases}$$

For instance, if we take the *coin* variable Y and consider the transformed random variable $\log(Y+1)$, the expected value is

$$E[\log(Y+1)] = \log(1) \cdot \frac{1}{2} + \log(2) \cdot \frac{1}{2} = \frac{\log(2)}{2}$$

We can define the population counterparts of the sample moments and their centralized and standardized versions:

- **r-th moment** of Y:
$$E[Y^r] = \int_{-\infty}^{\infty} u^r \ \mathrm{d}F(u)$$
- **r-th central moment**:
$$E[(Y - E[Y])^r] = \int_{-\infty}^{\infty} (u - E[Y])^r \ \mathrm{d}F(u)$$

- **Variance** (2nd central moment):

$$Var[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (u - E[Y])^2 \, \mathrm{d}F(u)$$

- **Standard deviation**:

$$sd(Y) = \sqrt{Var[Y]}$$

- **r-th standardized moment**:

$$E\Big[\Big(\frac{Y - E[Y]}{sd(Y)}\Big)^r\Big] = \int_{-\infty}^{\infty} \Big(\frac{u - E[Y]}{sd(Y)}\Big)^r \, \mathrm{d}F(u)$$

- **Skewness** (3rd standardized moment):

$$skew(Y) = E\Big[\Big(\frac{Y - E[Y]}{sd(Y)}\Big)^3\Big]$$

- **Kurtosis** (4th standardized moment):

$$kurt(Y) = E\Big[\Big(\frac{Y - E[Y]}{sd(Y)}\Big)^4\Big]$$

6.5 Linearity of the expected value

The expected value is a **linear** function. For any $a, b \in \mathbb{R}$, we have

$$E[aY + b] = aE[Y] + b.$$

For the variance, the following rule applies:

$$Var[aY + b] = a^2 Var[Y].$$

For any two random variables Y and Z, we have

$$E[aY + bZ] = aE[Y] + bE[Z].$$

A similar result for the variance does not hold in general. However, if Y and Z are independent random variables, we have

$$Var[aY + bZ] = a^2 Var[Y] + b^2 Var[Z]. \tag{6.5}$$

6.6 Parameters and estimators

A **parameter** θ is a feature (function) of the population distribution F of some random variable Y. The expectation, variance, skewness, and kurtosis are parameters.

A **statistic** is a function of a sample $Y_1, \ldots, Y_n$. An **estimator** $\widehat{\theta}$ for θ is a statistic intended as a guess about θ. It is a function of the random variables $Y_1, \ldots, Y_n$ and, therefore, a random variable as well. The sample mean, sample variance, sample skewness and sample kurtosis are estimators. When an estimator $\widehat{\theta}$ is calculated in a specific realized sample, we call $\widehat{\theta}$ an **estimate**.

6.7 Estimation of the mean

The expected value $E[Y]$ is also called **population mean** because it is the population counterpart of the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, where the sample $Y_1, \ldots, Y_n$ is identically distributed and has the same distribution as Y. In particular, we have:

$$E[Y_1] = \ldots = E[Y_n] = E[Y].$$

The true population mean $E[Y]$ is unknown in practice, but we can use the sample mean $\overline{Y}$ to estimate it. The sample mean is an unbiased estimator for the population mean because

$$E[\overline{Y}] = \frac{1}{n}\sum_{i=1}^n E[Y_i] = \frac{1}{n}\sum_{i=1}^n E[Y] = E[Y].$$

The **bias** of an estimator is the expected value of the estimator minus the parameter to be estimated. The bias of the sample mean is zero:

$$Bias[\overline{Y}] = E[\overline{Y}] - E[Y] = E[Y] - E[Y] = 0.$$

When repeating random experiments and computing sample means, we can expect the sample means to be distributed around the true population mean, with the population mean at the center of this distribution.

To assess how large the spread around the true population mean is, we can compute the variance:

$$Var[\overline{Y}] = \frac{1}{n^2} Var\Big[\sum_{i=1}^n Y_i\Big]$$

To simplify this term further, let's assume that the sample is i.i.d. (independent and identically distributed), i.e. the observations are randomly sampled from the population. Then, we can apply Equation 6.5:

$$Var\Big[\sum_{i=1}^n Y_i\Big] = \sum_{i=1}^n Var[Y_i].$$

By the identical distribution of the sample, we have

$$Var[Y_1] = \ldots = Var[Y_n] = Var[Y].$$

Therefore, the variance of the sample mean becomes:

$$Var[\overline{Y}] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Y] = \frac{Var[Y]}{n}.$$

The spread of sample means around the true mean becomes smaller, the larger the sample size n is. The more observations we have, the more precisely the sample mean can estimate the true population mean.

6.8 Consistency

Good estimators get closer and closer to the true parameter being estimated as the sample size n increases, eventually returning the true parameter value in a hypothetically infinitely large sample. This property is called **consistency**.

Consistency

An estimator $\widehat{\theta}$ is **consistent** for a true parameter θ if, for any $\epsilon > 0$,

$$P(|\widehat{\theta} - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Equivalently, consistency can be defined by the complementary event:

$$P(|\widehat{\theta} - \theta| \leq \epsilon) \to 1 \qquad \text{as } n \to \infty.$$

If $\widehat{\theta}$ is consistent, we say it **converges in probability** to θ, denoted by

$$\widehat{\theta} \xrightarrow{p} \theta \qquad \text{as } n \to \infty.$$

If an estimator $\widehat{\theta}$ is a continuous random variable, it will almost never reach exactly the true parameter value because point probabilities are zero: $P(\widehat{\theta} = \theta) = 0$.

However, the larger the sample size, the higher should be the probability that $\widehat{\theta}$ is close to the true value θ. Consistency means that, if we fix some small precision value $\epsilon > 0$, then,

$$P(|\widehat{\theta} - \theta| \leq \epsilon) = P(\theta - \epsilon \leq \widehat{\theta} \leq \theta + \epsilon)$$

should increase in the sample size n and eventually reach 1.

An estimator is called **inconsistent** if it is not consistent. An inconsistent estimator is practically useless and leads to false inference. Therefore, it is important to verify that your estimator is consistent.

To show whether an estimator is consistent, we can check the sufficient condition for consistency:

Sufficient condition for consistency

Let $\widehat{\theta}$ be an estimator for some parameter θ. The **bias** of $\widehat{\theta}$ is

$$Bias[\widehat{\theta}] = E[\widehat{\theta}] - \theta.$$

If the **bias** and the **variance** of $\widehat{\theta}$ tends to zero for large sample sizes, i.e., if

i) $Bias[\widehat{\theta}] \to 0 \quad (\text{as } n \to \infty)$,
ii) $Var[\widehat{\theta}] \to 0 \quad (\text{as } n \to \infty)$,

then $\widehat{\theta}$ is consistent for θ.

The reason for this sufficient condition is the fact that

$$P(|\widehat{\theta} - \theta| > \epsilon) \leq Var[\widehat{\theta}] + Bias[\widehat{\theta}]^2,$$

which follows from Markov's inequality.

6.9 Law of large numbers

The sample mean $\overline{Y}$ of an i.i.d. sample is consistent for the population mean $E[Y]$ because

i) $Bias[\overline{Y}] = 0$ for all n;
ii) $Var[\overline{Y}] = Var[Y]/n \to 0$, as $n \to \infty$, provided $Var[Y] < \infty$.

The consistency result of the sample mean is also known as the **law of large numbers (LLN)**:

$$\overline{Y} \xrightarrow{p} E[Y] \qquad \text{as } n \to \infty.$$

Below is an interactive Shiny app to visualize the law of large numbers using simulated data for different sample sizes and different distributions.

SHINY APP: LLN

6.10 Heavy tails

The sample mean of i.i.d. samples from most distributions is consistent. However, there are some exceptional cases where consistency fails. For instance, the simple Pareto distribution has the PDF

$$f(u) = \begin{cases} \frac{1}{u^2} & \text{if } u > 1, \\ 0 & \text{if } u \leq 1, \end{cases}$$

and the expected value is

$$E[X] = \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u = \int_{1}^{\infty} \frac{1}{u} \, \mathrm{d}u = \log(u)|_1^\infty = \infty.$$

The population mean is infinity, so the sample mean cannot converge and is inconsistent. The game of chance from the St. Petersburg paradox (see https://en.wikipedia.org/wiki/St._Petersburg_paradox) is an example of a discrete random variable with infinite expectation.

Another example is the t-distribution with 1 degree of freedom, also denoted as t_1 or Cauchy distribution, which has the PDF

$$f(u) = \frac{1}{\pi(1+u^2)}.$$

The lack of consistency of the sample mean from a t_1 distribution is visualized in the shiny application above.

The Pareto, St. Petersburg, and Cauchy distributions have infinite population mean, and the sample mean of observations from these distributions is inconsistent. These are distributions that produce huge outliers.

There are other distributions that have a finite mean but an infinite variance, skewness, or kurtosis.

For instance, the t_2 distribution has a finite mean but an infinite variance. The t_3 distribution has a finite variance but an infinite skewness. The t_4 distribution has a finite skewness but an infinite kurtosis.

If Y is t_m-distributed (t-distribution with m degrees of freedom), then

$$E[Y], E[Y^2], \ldots, E[Y^{m-1}] < \infty$$

but

$$E[Y^m] = E[Y^{m+1}] = \ldots = \infty.$$

Random variables with infinite first four moments have a so-called **heavy-tailed distribution** and may produce huge outliers. Many statistical procedures are only valid if the underlying distribution is not heavy-tailed.

6.11 Estimation of the variance

Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from some population distribution with population mean $\mu = E[Y]$ and population variance $\sigma^2 = Var[Y] < \infty$.

We introduced two sample cointerparts of σ^2: the sample variance

$$\widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2,$$

and the adjusted sample variance

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2 = \frac{n}{n-1}\widehat{\sigma}_Y^2.$$

The sample variance can be decomposed as

$$\begin{aligned}
\widehat{\sigma}_Y^2 &= \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu + \mu - \overline{Y})^2 \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2 + \frac{2}{n}\sum_{i=1}^n (Y_i - \mu)(\mu - \overline{Y}) + \frac{1}{n}\sum_{i=1}^n (\mu - \overline{Y})^2 \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2 - 2(\overline{Y} - \mu)^2 + (\overline{Y} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2 - (\overline{Y} - \mu)^2
\end{aligned}$$

The mean of $\widehat{\sigma}_Y^2$ is

$$\begin{aligned}
E[\widehat{\sigma}_Y^2] &= \frac{1}{n}\sum_{i=1}^n E[(Y_i - \mu)^2] - E[(\overline{Y} - \mu)^2] = \frac{1}{n}\sum_{i=1}^n Var[Y_i] - Var[\overline{Y}] \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,
\end{aligned}$$

where we used the fact that $Var[\overline{Y}] = \sigma^2/n$.

The sample variance is **downward biased**:

$$Bias[\widehat{\sigma}_Y^2] = E[\widehat{\sigma}_Y^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

On the other hand, the adjusted sample variance is **unbiased**:

$$Bias[s_Y^2] = E[s_Y^2] - \sigma^2 = \frac{n}{n-1}E[\widehat{\sigma}_Y^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

The variance of the sample variance can be computed as

$$Var[\widehat{\sigma}_Y^2] = \frac{\sigma^4}{n}\Big(kurt - \frac{n-3}{n-1}\Big)\frac{(n-1)^2}{n^2},$$

while the variance of the adjusted sample variance is

$$Var[s_Y^2] = \frac{\sigma^4}{n}\Big(kurt - \frac{n-3}{n-1}\Big).$$

As long as the kurtosis of the underlying distribution is finite, the sufficient conditions for consistency are satisfied as the bias and variance tend to zero as $n \to \infty$. The adjusted sample variance is unbiased for any n. The sample variance is biased for fixed n but **asymptotically unbiased** as the bias tends to zero for large n. The sample variance and the adjusted sample variance are consistent for the variance if the sample is i.i.d. and the distribution is not heavy-tailed.

6.12 Bias-variance tradeoff

From a bias perspective, adjusted sample variance s_Y^2 is preferred over $\widehat{\sigma}_Y^2$ because s_Y^2 is unbiased. However, from a variance perspective, $\widehat{\sigma}_Y^2$ is preferred due to its smaller variance. Traditionally, the emphasis on unbiasedness has led to a preference for $\widehat{\sigma}_Y^2$, even at the cost of a higher variance.

A more modern approach balances bias and variance, known as the **bias-variance tradeoff**, by selecting an estimator that minimizes the **mean squared error (MSE)**:

$$MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = Var[\widehat{\theta}] + Bias[\widehat{\theta}]^2.$$

For the variance estimators, the MSEs are

$$MSE[\widehat{\sigma}_Y^2] = Var[\widehat{\sigma}_Y^2] + Bias[\widehat{\sigma}_Y^2]^2 = \frac{\sigma^4}{n}\bigg[\Big(kurt - \frac{n-3}{n-1}\Big)\frac{(n-1)^2}{n^2} + \frac{1}{n}\bigg]$$

and

$$MSE[s_Y^2] = Var[s_Y^2] = \frac{\sigma^4}{n}\Big(kurt - \frac{n-3}{n-1}\Big).$$

Since s_Y^2 is unbiased, its MSE equals its variance.

It is not possible to universally determine which estimator has a lower MSE because this depends on the population kurtosis ($kurt$) of the underlying distribution. However, it can be shown that for all distributions with $kurt \geq 1.5$, the relation $MSE[s_Y^2] > MSE[\widehat{\sigma}_Y^2]$ holds, which implies that $\widehat{\sigma}_Y^2$ is preferred based on the bias-variance tradeoff for all moderately tailed distributions.

To give an indication of typical kurtosis values:

- Symmetric Bernoulli distribution with $P(Y = 0) = P(Y = 1) = 0.5$: kurtosis of 1 (light-tailed).
- Uniform distribution (see Equation 6.3): kurtosis of 1.8 (moderately light-tailed).
- Normal distribution: kurtosis of 3 (moderately tailed).
- t_5 distribution: kurtosis of 9 (moderately heavy-tailed).
- t_4 distribution: infinite kurtosis (heavy-tailed).

Therefore, according to the bias-variance tradeoff, the adjusted sample variance s_Y^2 is preferred only for extremely light-tailed distributions, while $\widehat{\sigma}_Y^2$ is preferred in cases with moderate or higher kurtosis.

In practice, especially with larger samples, the difference between s_Y^2 and $\widehat{\sigma}_Y^2$ becomes negligible, and either estimator is generally acceptable. Therefore, the discussion about a better variance estimator is a bit nitpicky and not of much practical relevance.

However, for instance in high-dimensional regression problems with near multicollinearity ($k \approx n$), the bias-variance tradeoff is crucial. In such cases, biased but low-variance estimators like ridge or lasso (shrinkage estimators) are often preferred over ordinary least squares (OLS).

6.13 R-codes

statistics-sec05.R

7 Conditional expectation

7.1 Conditional mean

Conditional expectation

The **conditional expectation** or **conditional mean** of Y given $\boldsymbol{Z} = \boldsymbol{b}$ is the expected value of the distribution $F_{Y|\boldsymbol{Z}=\boldsymbol{b}}$:

$$E[Y|\boldsymbol{Z} = \boldsymbol{b}] = \int_{-\infty}^{\infty} a \ \mathrm{d}F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a).$$

For continuous Y with conditional density $f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)$, we have $\mathrm{d}F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) = f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) \ \mathrm{d}a$, and the conditional expectation is

$$E[Y|Z = \boldsymbol{b}] = \int_{-\infty}^{\infty} a f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) \ \mathrm{d}a.$$

Similarly, for discrete Y with support $\mathcal{Y}$ and conditional PMF $\pi_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)$, we have

$$E[Y|Z = \boldsymbol{b}] = \sum_{u \in \mathcal{Y}} u \pi_{Y|\boldsymbol{Z}=\boldsymbol{b}}(u).$$

The conditional expectation is a function of $\boldsymbol{b}$, which is a specific value of $\boldsymbol{Z}$ that we condition on. Therefore, we call it the **conditional expectation function**:

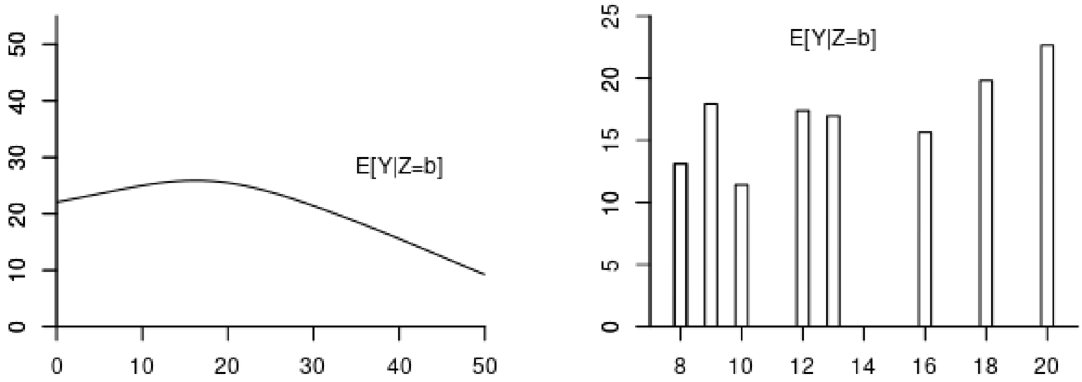
$$m(\boldsymbol{b}) = E[Y|Z = \boldsymbol{b}].$$

Suppose the conditional expectation of wage given experience level b is:

$$m(b) = E[wage|exper = b] = 14.5 + 0.9b - 0.017b^2.$$

For example, with 10 years of experience:

$$m(10) = E[wage|exper = 10] = 21.8.$$



(a) CEF wage given experience

(b) CEF wage given education

Figure 7.1: Conditional expectation functions. The x-axis represents b.

Here, $m(b)$ assigns a specific real number to each fixed value of b; it is a deterministic function derived from the joint distribution of wage and experience.

However, if we treat experience as a random variable, the conditional expectation becomes:

$$m(exper) = E[wage|exper] = 14.5 + 0.9exper - 0.017exper^2.$$

Now, $m(exper)$ is a function of the random variable experexper and is itself a random variable.

In general:

- The conditional expectation given a specific value b is:

  $$m(\boldsymbol{b}) = E[Y|\boldsymbol{Z} = \boldsymbol{b}],$$

  which is deterministic.
- The conditional expectation given the random variable Z is:

  $$m(\boldsymbol{Z}) = E[Y|\boldsymbol{Z}],$$

  which is a random variable because it depends on the random vector $\boldsymbol{Z}$.

This distinction highlights that the conditional expectation can be either a specific number, i.e. $E[Y|\boldsymbol{Z} = \boldsymbol{b}]$, or a random variable, i.e., $E[Y|\boldsymbol{Z}]$, depending on whether the condition is fixed or random.

7.2 Rules of calculation

Rules of Calculation for Conditional Expectation

Let Y be a random variable and $\boldsymbol{Z}$ a random vector. The rules of calculation rules below are fundamental tools for working with conditional expectations:

(i) Law of Iterated Expectations (LIE):

$$E[E[Y|\boldsymbol{Z}]] = E[Y].$$

Intuition: The LIE tells us that if we first compute the expected value of Y given each possible outcome of $\boldsymbol{Z}$, and then average those expected values over all possible values of $\boldsymbol{Z}$, we end up with the overall expected value of Y. It's like calculating the average outcome across all scenarios by considering each scenario's average separately.

More generally, for any two random vectors $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$:

$$E[E[Y|\boldsymbol{Z}, \boldsymbol{Z}^*]|\boldsymbol{Z}] = E[Y|\boldsymbol{Z}].$$

Intuition: Even if we condition on additional information $\boldsymbol{Z}^*$, averaging over $\boldsymbol{Z}^*$ while keeping $\boldsymbol{Z}$ fixed brings us back to the conditional expectation given $\boldsymbol{Z}$ alone.

(ii) Conditioning Theorem (CT):

For any function $g(\boldsymbol{Z})$:

$$E[g(\boldsymbol{Z})\, Y|\boldsymbol{Z}] = g(\boldsymbol{Z})\, E[Y|\boldsymbol{Z}].$$

Intuition: Once we know $\boldsymbol{Z}$, the function $g(\boldsymbol{Z})$ becomes a known quantity. Therefore, when computing the conditional expectation given $\boldsymbol{Z}$, we can treat $g(\boldsymbol{Z})$ as a constant and factor it out.

(iii) Independence Rule (IR):

If Y and $\boldsymbol{Z}$ are independent, then:

$$E[Y|\boldsymbol{Z}] = E[Y].$$

Intuition: Independence means that Y and $\boldsymbol{Z}$ do not influence each other. Knowing the value of $\boldsymbol{Z}$ gives us no additional information about Y. Therefore, the expected value of Y remains the same regardless of the value of $\boldsymbol{Z}$, so the conditional expectation equals the unconditional expectation.

Another way to see this is the fact that, if Y and Z are independent, then

$$F_{Y|Z=b}(a) = F_Y(a) \quad \text{for all } a \text{ and } b.$$

7.3 Expectation of bivariate random variables

We often are interested in expected values of functions involving two random variables, such as the **cross-moment** $E[YZ]$ for variables Y and Z.

If $F(a,b)$ is the joint CDF of (Y,Z), then the cross-moment is defined as:

$$E[YZ] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ab \ \mathrm{d}F(a,b). \tag{7.1}$$

If Y and Z are continuous and $F(a,b)$ is differentiable, the joint probability density function (PDF) of (Y,Z):

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b).$$

This allows us to write the differential of the CDF as

$$\mathrm{d}F(a,b) = f(a,b) \ \mathrm{d}a \ \mathrm{d}b,$$

and the cross-moment becomes:

$$E[YZ] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ab \ \mathrm{d}F(a,b) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} abf(a,b) \ \mathrm{d}a \ \mathrm{d}b.$$

In the *wage* and *experience* example, we have the following joint CDF and joint PDF:

If Y and Z are discrete with joint PMF $\pi(a,b)$ and support $\mathcal{Y}$, the cross moment is

$$E[YZ] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ab \ \mathrm{d}F(a,b) = \sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}} ab \ \pi(a,b).$$

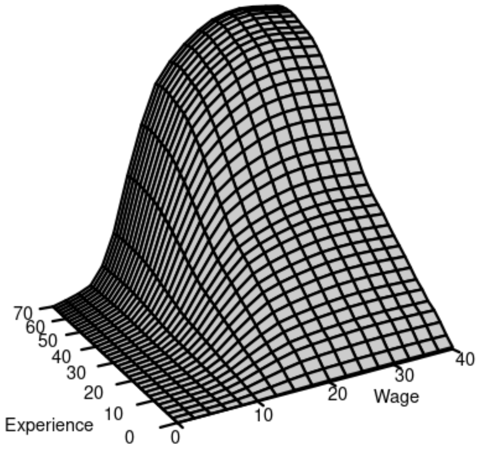


Figure 7.2: Joint CDF of wage and experience

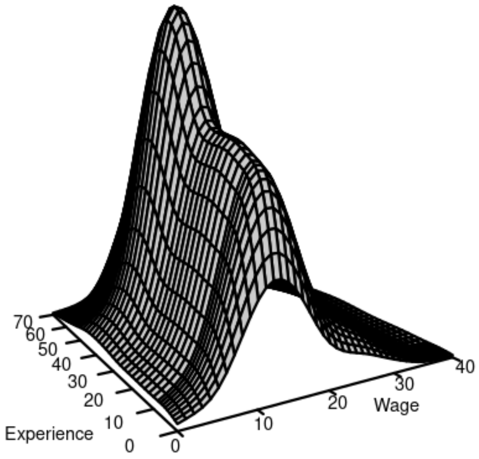


Figure 7.3: Joint PDF of wage and experience

If one variable is discrete and the other is continuous, the expectation involves a mixture of summation and integration.

In general, the expected value of any real valued function $g(Y, Z)$ is given by

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(a,b) \; \mathrm{d}F(a,b).$$

7.4 Covariance and correlation

The **covariance** of Y and Z is defined as:

$$Cov(Y,Z) = E[(Y - E[Y])(Z - E[Z])] = E[YZ] - E[Y]E[Z].$$

The covariance of Y with itself is the variance:

$$Cov(Y,Y) = Var[Y].$$

The variance of the sum of two random variables depends on the covariance:

$$Var[Y+Z] = Var[Y] + 2Cov(Y,Z) + Var[Z]$$

The **correlation** of Y and Z is

$$Corr(Y,Z) = \frac{Cov(Y,Z)}{sd(Y)sd(Z)}$$

where $sd(Y)$ and $sd(Z)$ are the standard deviations of Y and Z, respectively.

Uncorrelated

Y and Z are **uncorrelated** if $Corr(Y,Z) = 0$, or, equivalently, if $Cov(Y,Z) = 0$.

If Y and Z are uncorrelated, then:

$$\begin{aligned} E[YZ] &= E[Y]E[Z] \\ Var[Y+Z] &= Var[Y] + Var[Z] \end{aligned}$$

If Y and Z are independent and have finite second moments, they are uncorrelated. However, the reverse is not necessarily true; uncorrelated variables are not always independent.

7.5 Expectations for random vectors

These concepts generalize to any k-dimensional random vector $\boldsymbol{Z} = (Z_1, \dots, Z_k)$.

The expectation vector of $\boldsymbol{Z}$ is:

$$E[\boldsymbol{Z}] = \begin{pmatrix} E[Z_1] \\ \vdots \\ E[Z_k] \end{pmatrix}.$$

The covariance matrix of $\boldsymbol{Z}$ is:

$$\begin{aligned} Var[\boldsymbol{Z}] &= E[(\boldsymbol{Z} - E[\boldsymbol{Z}])(\boldsymbol{Z} - E[\boldsymbol{Z}])'] \\ &= \begin{pmatrix} Var[Z_1] & Cov(Z_1, Z_2) & \dots & Cov(X_1, Z_k) \\ Cov(Z_2, Z_1) & Var[Z_2] & \dots & Cov(Z_2, Z_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Z_k, Z_1) & Cov(Z_k, Z_2) & \dots & Var[Z_k] \end{pmatrix} \end{aligned}$$

For any random vector $\boldsymbol{Z}$, the covariance matrix $Var[\boldsymbol{Z}]$ is symmetric and positive semi-definite.

7.6 R-codes

statistics-sec07.R