Probability Theory for Econometricians

Sven Otto

February 29, 2024

Table of contents

Welcome

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

1 Probability

1.1 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.

1.2 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**.

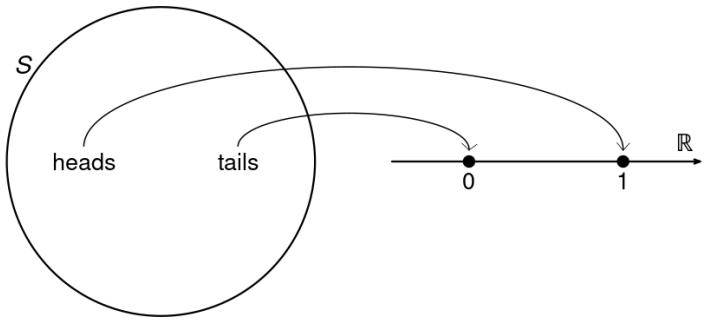


Figure 1.1: Bernoulli random variable

- *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 1.1: ISCED 2011 levels

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

1.3 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

1.4 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.

1.5 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:

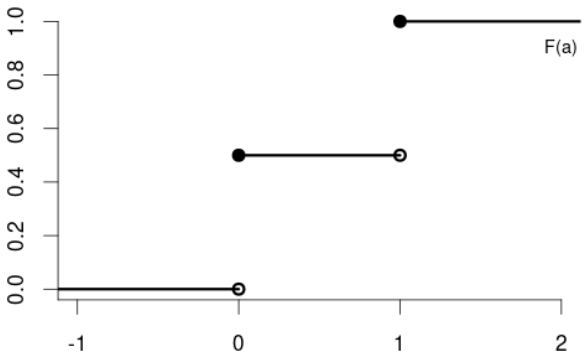


Figure 1.2: CDF of coin

The CDF of the variables *education* is

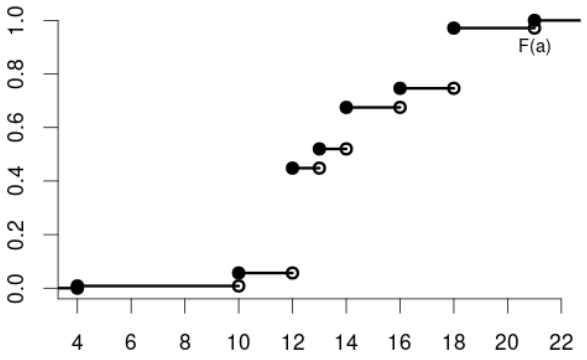


Figure 1.3: CDF of education

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 $a < b$):

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

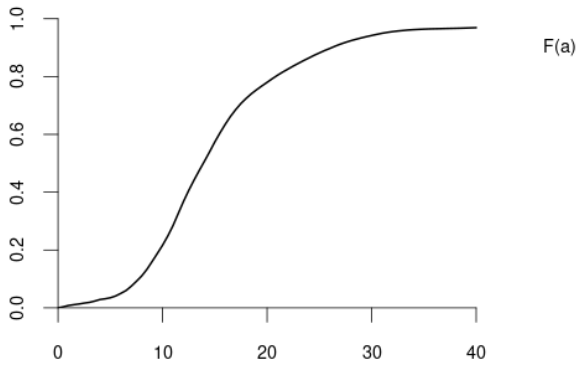


Figure 1.4: CDF of wage

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

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

$$\pi(a) = P(Y = a) = F(a) - \lim_{\epsilon \to 0} F(a - \epsilon) = \text{“jump height at } a\text{”}. \tag{1.1}$$

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 1.1. 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.

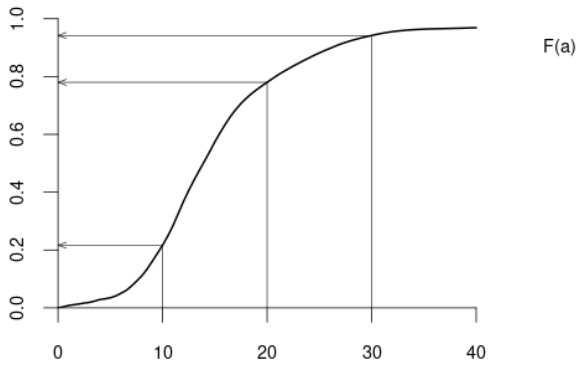


Figure 1.5: CDF of wage evaluated at some points

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.

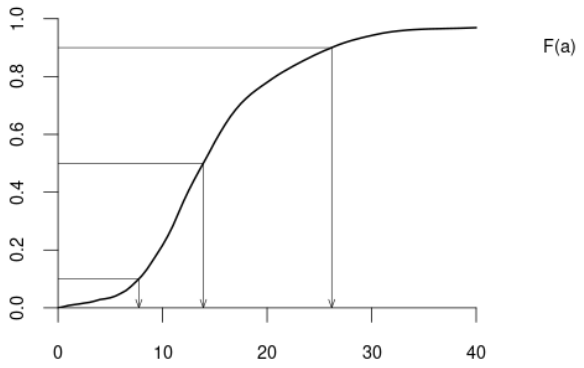


Figure 1.6: Quantiles of variable wage

– 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.

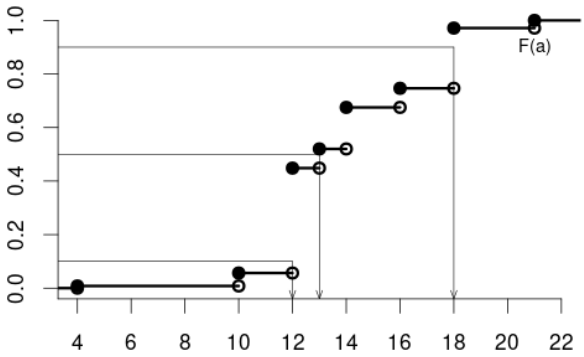


Figure 1.7: Quantiles of variable education

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.

1.6 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

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

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

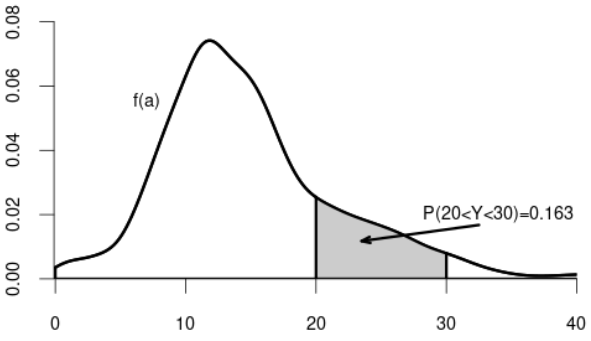


Figure 1.8: PDF of the variable wage

Probability rule for the PDF:

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

1.7 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.

1.7.1 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

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

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

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

1.7.2 Expectation of a continuous random variable

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

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

Using numerical integration for the density of Figure 1.8 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

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

1.7.3 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

$$E[Y] = \int_{-\infty}^{\infty} u \, \mathrm{d}F(u). \tag{1.2}$$

Note that Equation 1.2 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

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

1.7.4 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 1.7.3 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

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

1.8 Descriptive features of a distribution

Table 1.2: 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

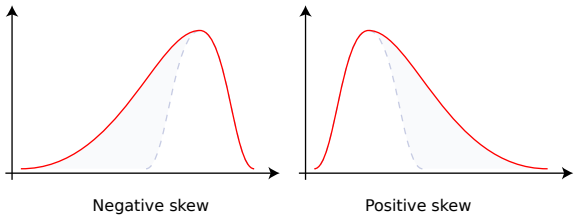


Figure 1.9: Positive and negative skewness

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 1.8) 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:

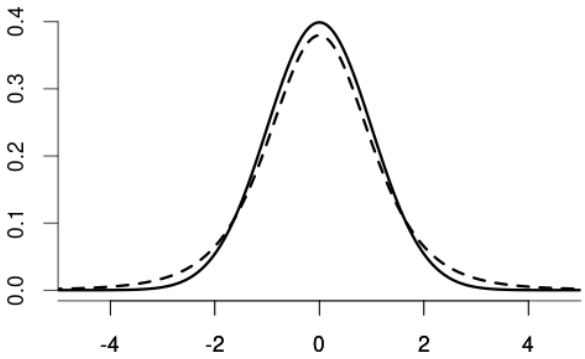


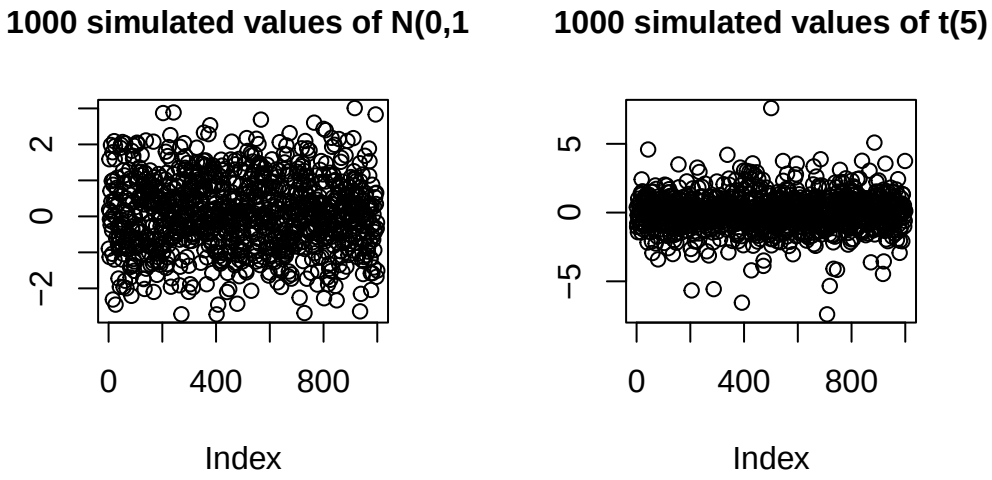
Figure 1.10: PDFs of the standard normal distribution (solid) and the t_5-distribution (dashed)

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 $\mathcal{N}(0,1)$:

```
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 $E[Y^4]$.

1.8.1 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

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

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.

1.9 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

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

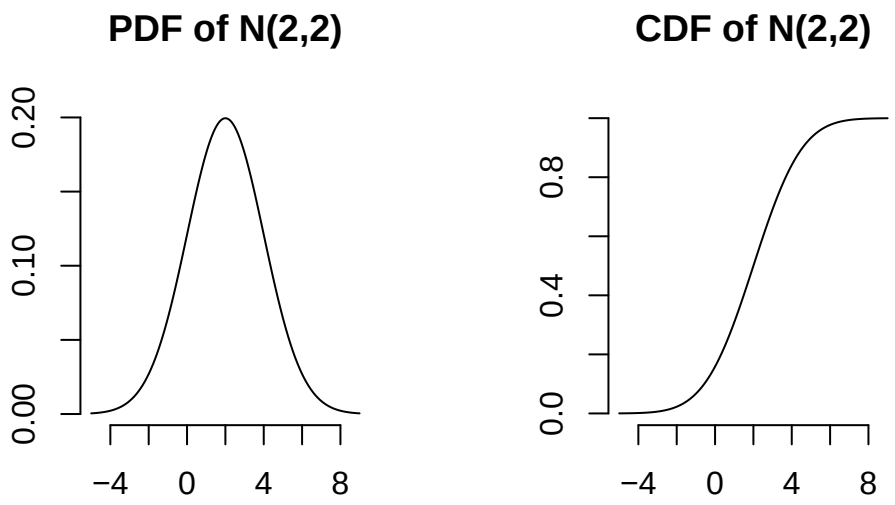
and CDF

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

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

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

```
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.

1.10 Additional reading

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

1.11 R-codes

statistics-sec2.R

2 Dependence

2.1 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.

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

2.2 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 2.1: 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 2.2: 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.

2.3 Bivariate distributions

Bivariate distribution

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

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

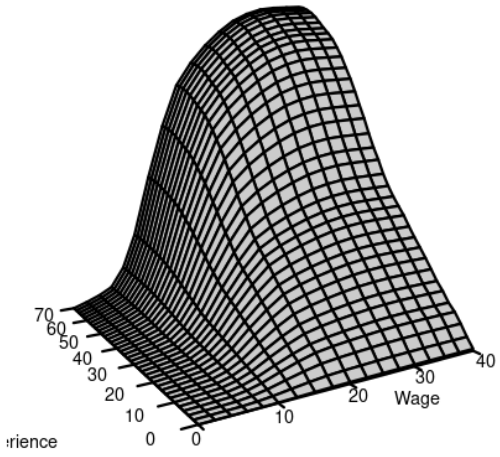


Figure 2.1: Joint CDF of wage and experience

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

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

Bivariate density function

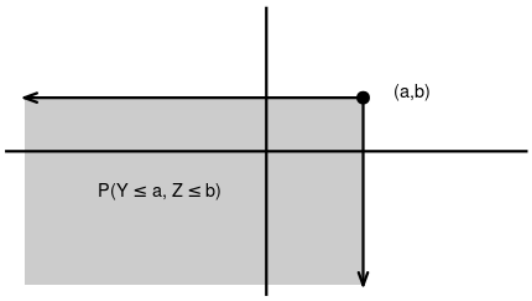


Figure 2.2: Calculate probabilities using the joint CDF

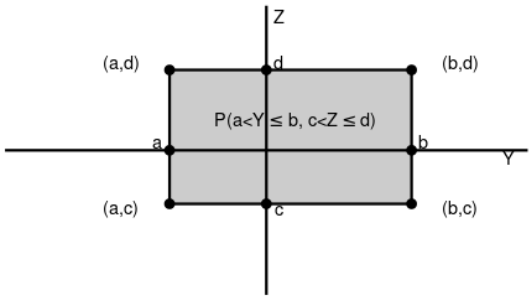


Figure 2.3: Calculate probabilities using the joint CDF

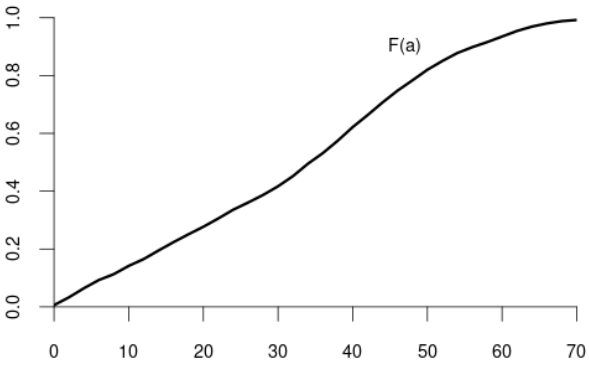


Figure 2.4: Marginal CDF of experience

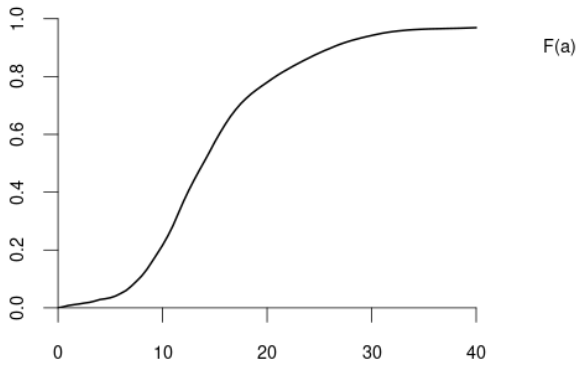


Figure 2.5: Marginal CDF of wage

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

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

2.4 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,

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

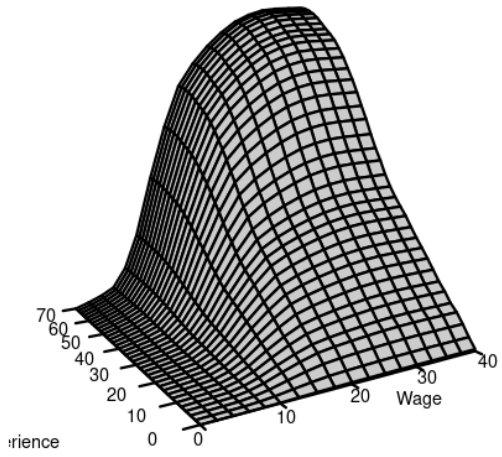


Figure 2.6: Joint CDF of wage and experience

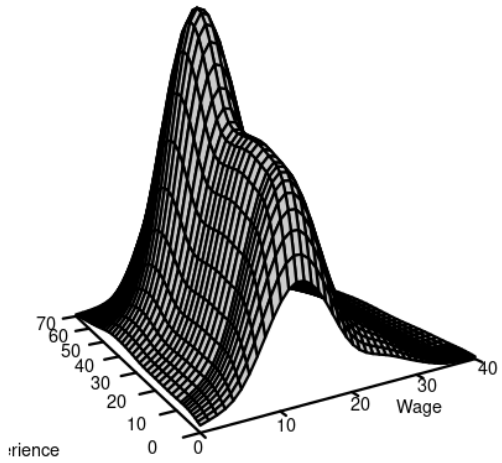


Figure 2.7: Joint PDF of wage and experience

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

$$E[YZ] = E[Y]E[Z]$$
$$var[Y+Z] = var[Y] + var[Z]$$

2.5 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 2.1 (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 2.2 (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 2.1 are independent, and those of Table 2.2 are dependent.

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

2.6 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.

2.7 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 2.2:

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

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

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

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

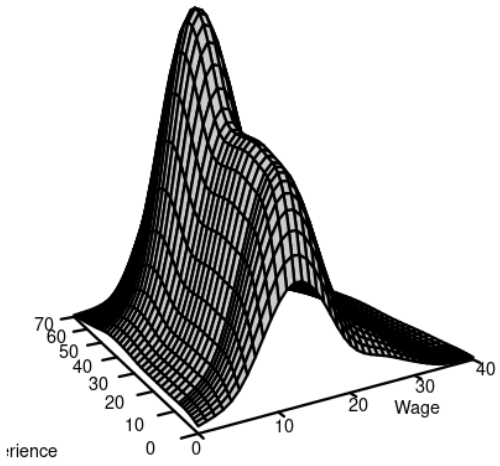


Figure 2.8: Joint PDF of wage and experience

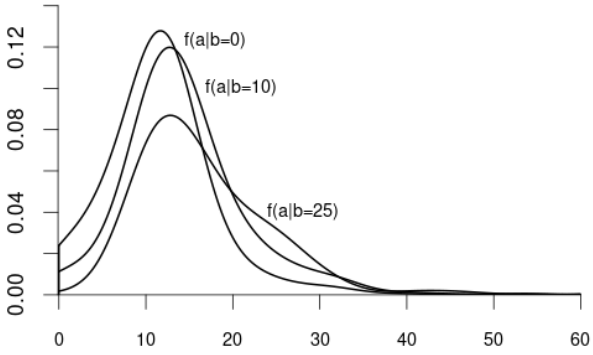


Figure 2.9: Conditional PDFs of wage given experience

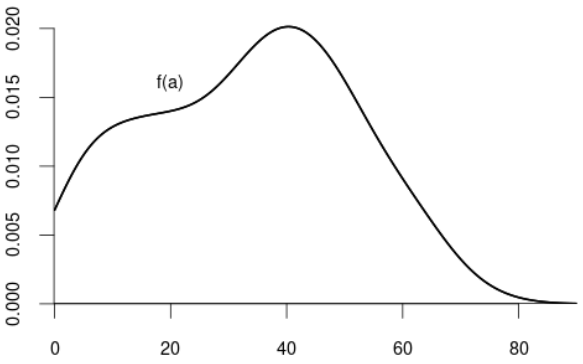


Figure 2.10: PDF of variable experience

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

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

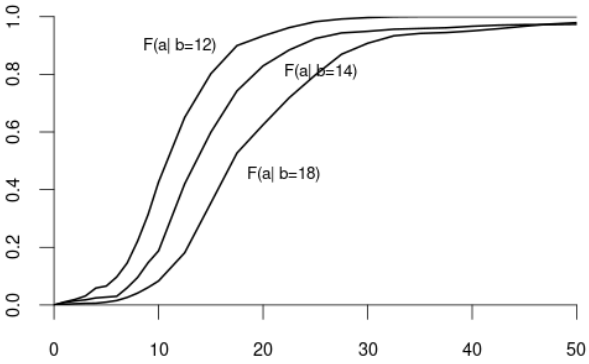


Figure 2.11: Conditional CDFs of wage given education

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

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

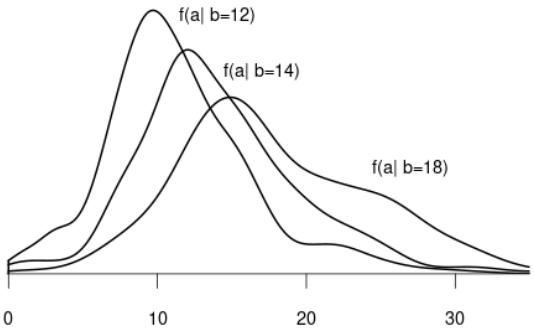


Figure 2.12: Conditional PDFs of wage given education

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

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

2.8 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

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

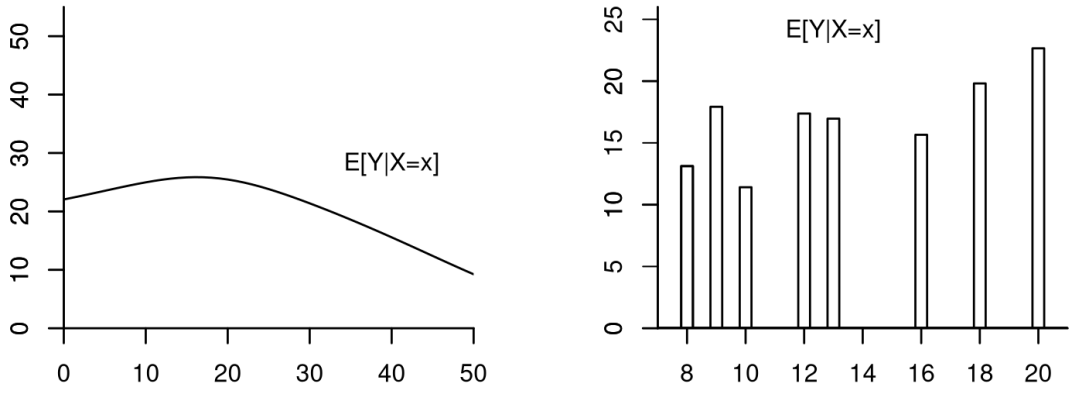
Consider again the wage and experience example. Suppose that the conditional expectation has the functional form

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

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 2.13: 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.

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

2.10 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

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

2.11 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:

$$Y = m(X) + e, \quad E[e \mid X] = 0. \tag{2.1}$$

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

From equation Equation 2.1 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 2.1 is a **linear regression model**. We will study this model in detail in the next sections.

2.12 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.

2.12.1 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:

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

2.12.2 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

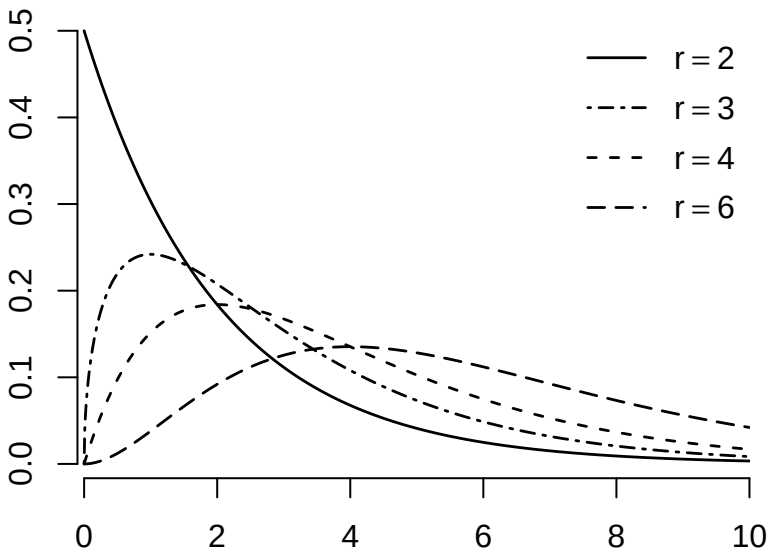


Figure 2.14: χ^2 -distribution

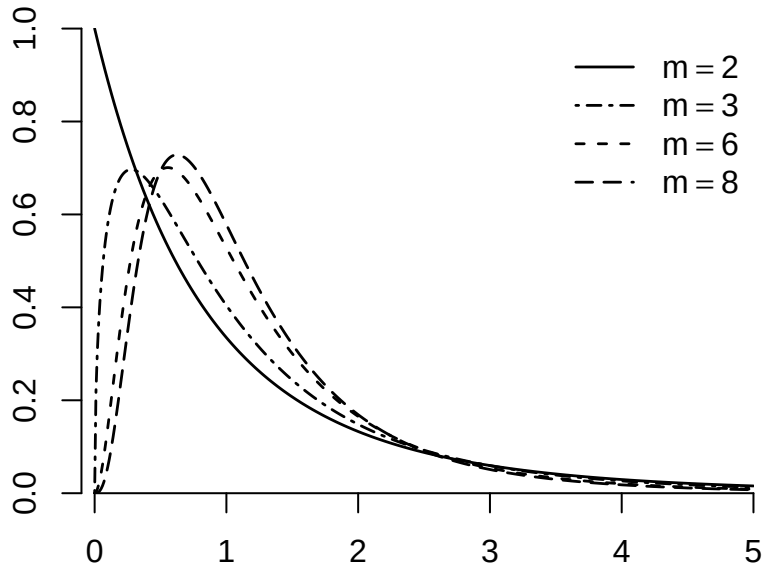


Figure 2.15: F-distribution

2.12.3 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 $t_m \to \mathcal{N}(0,1)$

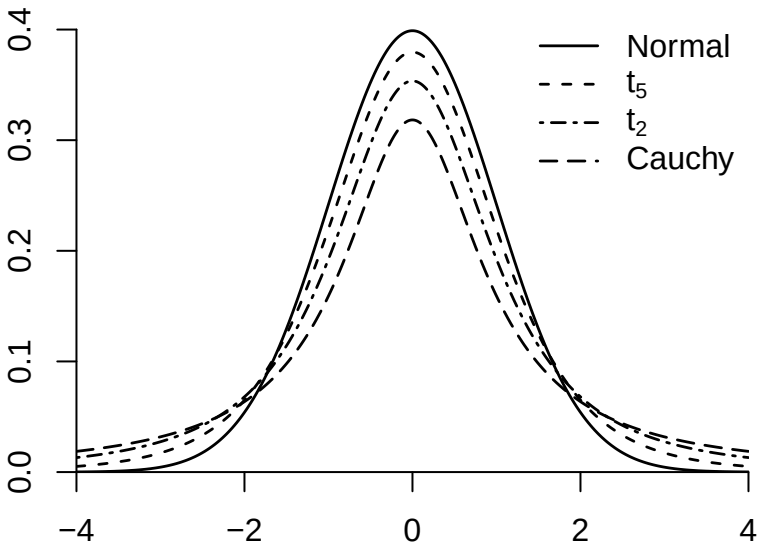


Figure 2.16: Student t-distribution

2.12.4 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

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

2.12.5 R-commands for parametric distributions

	get CDF $F(a)$	quantile function $q(p)$	generate n independent 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)`

2.13 Additional reading

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