Probability Theory for Econometricians

Sven Otto

March 18, 2025

Table of contents

Welcome

This tutorial provides a concise introduction to the fundamental concepts of probability theory for econometricians and data scientists.

The tutorial keeps the measure-theoretic details of probability theory to a minimum, focusing instead on the practical aspects of probability theory that are most relevant for econometricians.

Current Status: This tutorial is still under development.

The sections presented here originate from the *Statistics for Data Analytics* course taught in Winter Term 2024.

For a quick foundational review, I recommend sections 2 and 3 of Stock and Watson (2019): Textbook Link

1 Probability 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 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
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.

1.3 Events and Probabilities

To quantify the uncertainty in random variables, we need to assign probabilities to different possible outcomes or sets of outcomes. This is where events and probability functions come into play.

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, using our coin toss example where $Y = 1$ represents heads and $Y = 0$ represents tails:

- Elementary events:

$$A_1 = \{Y = 0\} \text{ (coin shows tails)}$$
$$A_2 = \{Y = 1\} \text{ (coin shows heads)}$$
$$A_3 = \{Y = 2.5\} \text{ (impossible outcome)}$$

- Half-open events:

$$A_4 = \{Y \geq 0\} = \{Y \in [0, \infty)\}$$
$$A_5 = \{-1 \leq Y < 1\} = \{Y \in [-1, 1)\}$$

The **probability function** P assigns values between 0 and 1 to events. For a fair coin toss (where $Y = 1$ represents heads and $Y = 0$ represents tails), 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 elementary 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

For our examples:

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

Every event has a **complementary event** (denoted with superscript c), which consists of all outcomes not in the original event. For any pair of events, we can also take the **union** (denoted by $\cup$) and **intersection** (denoted by $\cap$). Let's define further events:

- Complement (all outcomes not in the original event):

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

- Union (outcomes in either event):

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

- Intersection (outcomes in both events):

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

- Combinations of multiple events:

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

- **Certain event** (contains all possible outcomes):

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

- **Empty event** (contains no outcomes):

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

For the coin toss experiment, we can verify the probabilities of all these events:

- $P(A_1) = 0.5$ (probability of tails)
- $P(A_2) = 0.5$ (probability of heads)
- $P(A_3) = 0$ (coin never shows 2.5)
- $P(A_4) = 1$ (coin always shows a non-negative value)
- $P(A_5) = 0.5$ (only tails falls in this interval)
- $P(A_6) = 0$ (coin never shows a negative value)
- $P(A_7) = 0.5$ (same as probability of tails)
- $P(A_8) = 0.5$ (contains only tails)

- $P(A_9) = 1$ (contains all possible coin outcomes)
- $P(A_{10}) = 1$ (the certain event always occurs)
- $P(A_{11}) = 0$ (the empty event never occurs)

To illustrate how events and probabilities apply in other contexts, consider our education level example. If Y represents years of schooling with possible values $\{4, 10, 12, 13, 14, 16, 18, 21\}$, we might define the event $B = \{Y \geq 16\}$ representing "has at least a Bachelor's degree." The probability $P(B)$ would then represent the proportion of the population with at least a Bachelor's degree.

1.4 Probability Function

Now that we have defined events, we need a formal way to assign probabilities to them consistently. The probability function P assigns probabilities to events within the Borel sigma-algebra (denoted as $\mathcal{B}$), which contains all events we would ever need to compute probabilities for in practice. This includes our previously mentioned events $A_1, \ldots, A_{11}$, any interval of the form $\{Y \in (a, b)\}$ with $a, b \in \mathbb{R}$, and all possible unions, intersections, and complements of these events.

Two events A and B are **disjoint** if $A \cap B = \{\}$, meaning they have no common outcomes. For example, $A_1 = \{Y = 0\}$ and $A_2 = \{Y = 1\}$ are disjoint (a coin cannot show both heads and tails simultaneously), while A_1 and $A_4 = \{Y \geq 0\}$ are not disjoint since $A_1 \cap A_4 = \{Y = 0\}$.

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

Basic Rules of Probability

Fundamental Axioms:

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

Implied Properties:

- $P(Y \notin \mathbb{R}) = P(\{\}) = 0$ 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 (monotonicity)
- $P(A^c) = 1 - P(A)$ for the complement event of A
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A, B

The first three properties listed above are known as the axioms of probability, first formalized by Andrey Kolmogorov in 1933. The remaining properties follow as logical consequences of these axioms.

Let's consider a practical example: In our education survey, suppose we know the following probabilities:

- $P(\text{Primary education}) = 0.1$
- $P(\text{Secondary education}) = 0.6$
- $P(\text{Tertiary education}) = 0.3$

These events are disjoint (a person cannot simultaneously have exactly primary and exactly secondary education as their highest level), and they cover all possibilities (everyone has some highest level of education). Using the axioms:

1. Each probability is non-negative (satisfying axiom 1)
2. The sum $0.1 + 0.6 + 0.3 = 1$ (satisfying axiom 2)
3. The probability of having either primary or secondary education is $P(\text{Primary or Secondary}) = P(\text{Primary}) + P(\text{Secondary}) = 0.1 + 0.6 = 0.7$ (using axiom 3 for disjoint events)

From the implied properties, we can also calculate that the probability of not having tertiary education is $P(\text{No tertiary}) = 1 - P(\text{Tertiary}) = 1 - 0.3 = 0.7$.

1.5 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 CDF of the variable *coin* is

$$F(a) = \begin{cases} 0 & a < 0, \\ 0.5 & 0 \leq a < 1, \\ 1 & a \geq 1, \end{cases} \tag{1.1}$$

with the following CDF plot:

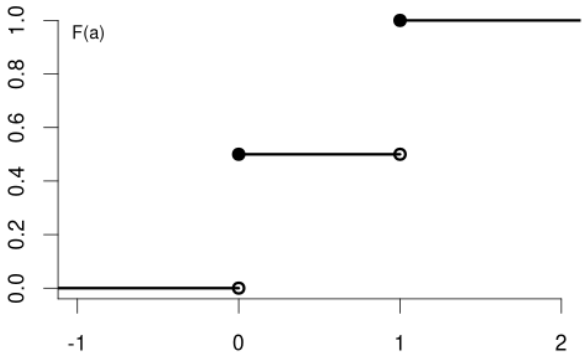


Figure 1.1: CDF of coin (discrete random variable)

The CDF of the variable *education* could be:

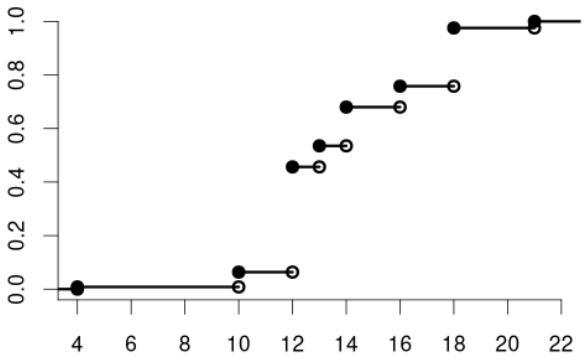


Figure 1.2: CDF of education (discrete random variable)

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

Notice the key difference: the CDF of a **continuous random variable** (like wage) is smooth, while the CDF of a **discrete random variable** (like coin and education) contains jumps and

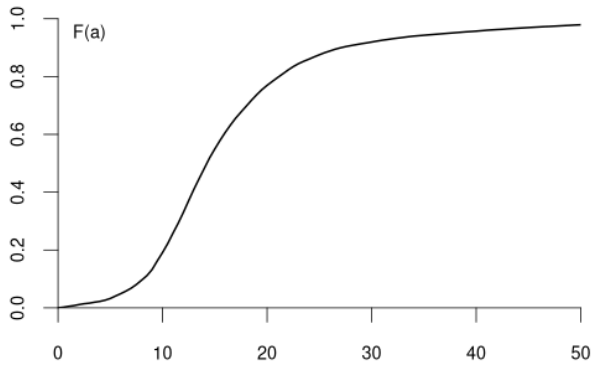


Figure 1.3: CDF of wage (continuous random variable)

is flat between these jumps. The height of each jump corresponds to the probability of that specific value occurring.

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

- **Non-decreasing**: $F(a) \leq F(b)$ for $a \leq b$.
  Reflects the monotonicity of probability when the event $\{Y \leq a\}$ is contained in $\{Y \leq b\}$ for $a < b$.
- **Limits at 0 and 1**: $\lim_{a \to -\infty} F(a) = 0$ and $\lim_{a \to \infty} F(a) = 1$.
  Ensures the total probability equals 1 and impossible events have zero probability.
- **Right-continuity**: $\lim_{\varepsilon \to 0, \varepsilon > 0} F(a + \varepsilon) = F(a)$.
  Ensures $P(Y \leq a)$ includes $P(Y = a)$, which matters especially for discrete variables with jumps in the CDF. This property means that at any point a, the CDF value includes the probability mass exactly at that point, making $F(a) = P(Y \leq a)$ rather than $P(Y < a)$.

By the basic rules of probability, we can compute the probability of any event of interest if we know the CDF $F(a)$. Here are the most common calculations:

Probability Calculations Using the CDF (for $a < b$):

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

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

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

which is the jump height at a. For continuous random variables, point probabilities are always zero, while for discrete random variables, they can be positive.

Here, $\lim_{\varepsilon \to 0, \varepsilon > 0} F(a - \varepsilon)$ denotes the left limit at a while $\lim_{\varepsilon \to 0, \varepsilon > 0} F(a + \varepsilon)$ denotes the right limit at a. When approaching any point from the left, the CDF can have a jump at that point, while when approaching from the right, the CDF cannot jump (due to right-continuity).

Let's use our coin toss example to illustrate how to calculate different probabilities using the CDF in Equation 1.1:

1. $P(Y \leq 0.5) = F(0.5) = 0.5$
2. $P(Y > 0.5) = 1 - F(0.5) = 1 - 0.5 = 0.5$
3. $P(Y = 0) = F(0) - \lim_{\varepsilon \to 0, \varepsilon > 0} F(0 - \varepsilon) = 0.5 - 0 = 0.5$
4. $P(-1 < Y \leq 2) = F(2) - F(-1) = 1 - 0 = 1$

1.6 Probability Mass Function

In the previous section, we defined the point probability $P(Y = a)$ as the height of the jump in the CDF at point a. These point probabilities are systematically organized in the probability mass function:

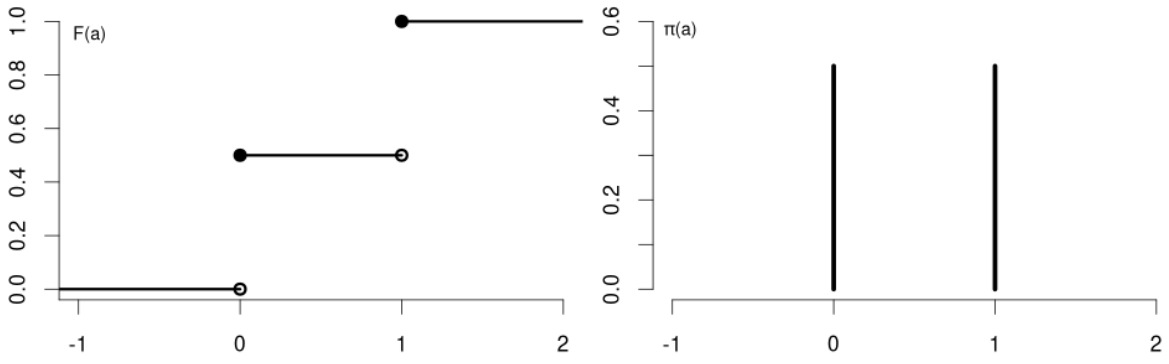
Probability Mass Function (PMF)

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

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



(a) CDF of coin (b) PMF of coin

Figure 1.4: Coin variable: CDF (left) and PMF (right)

The *education* variable has the following PMF:

$$\pi(a) = P(Y = a) = \begin{cases} 0.008 & \text{if } a = 4 \\ 0.055 & \text{if } a = 10 \\ 0.393 & \text{if } a = 12 \\ 0.079 & \text{if } a = 13 \\ 0.145 & \text{if } a = 14 \\ 0.078 & \text{if } a = 16 \\ 0.218 & \text{if } a = 18 \\ 0.024 & \text{if } a = 21 \\ 0 & \text{otherwise} \end{cases}$$

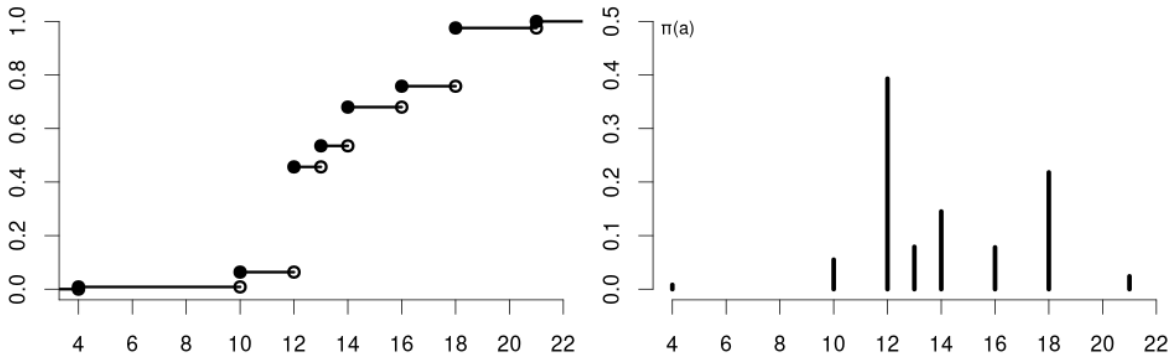
Note that these probability values sum to 1.

The **support** $\mathcal{Y}$ of Y is the set of all values that Y can take with non-zero probability: $\mathcal{Y} = \{a \in \mathbb{R} : \pi(a) > 0\}$.

For the coin variable, the support is $\mathcal{Y} = \{0, 1\}$, while for the education variable, the support is $\mathcal{Y} = \{4, 10, 12, 13, 14, 16, 18, 21\}$.

Any valid PMF must satisfy the following properties:

- **Non-negativity**: $\pi(a) \geq 0$ for all $a \in \mathbb{R}$
- **Sum to one**: $\sum_{a \in \mathcal{Y}} \pi(a) = 1$
- **Relationship to CDF**: $F(b) = \sum_{a \in \mathcal{Y}, a \leq b} \pi(a)$



(a) CDF of education

(b) PMF of education

Figure 1.5: Education variable: CDF (left) and PMF (right)

1.7 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 (PDF)

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

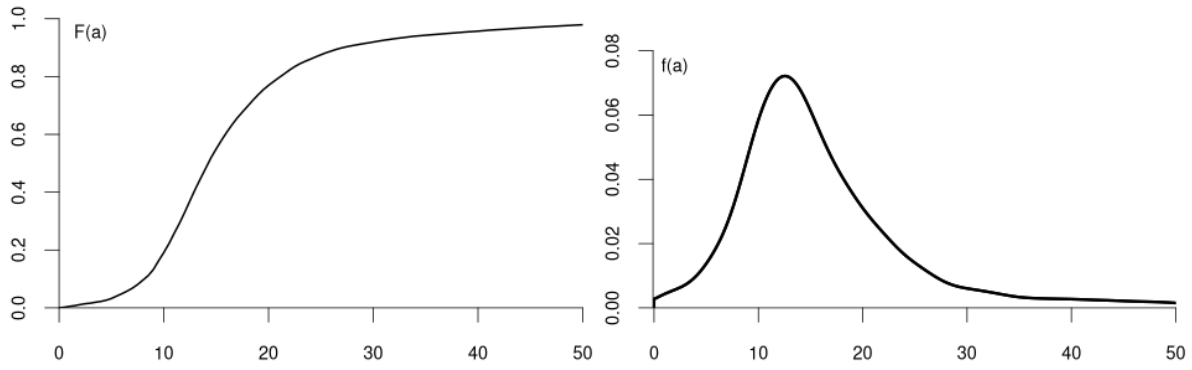
$$f(a) = \frac{d}{da}F(a).$$

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



(a) CDF of wage (b) PDF of wage

Figure 1.6: Wage variable: CDF (left) and PDF (right)

The **support** of a continuous random variable Y with PDF f is the set $\mathcal{Y} = \{a \in \mathbb{R} : f(a) > 0\}$, which contains all values where the density is positive. For instance, the support of the *wage* variable is $\mathcal{Y} = \{a \in \mathbb{R} : a \geq 0\}$, reflecting that wages cannot be negative.

Basic Rules for Continuous Random Variables (with $a \leq b$):

- $P(Y = a) = \int_a^a f(u)\ \mathrm{d}u = 0$
- $P(Y \leq a) = P(Y < a) = F(a) = \int_{-\infty}^a f(u)\ \mathrm{d}u$
- $P(Y > a) = P(Y \geq a) = 1 - F(a) = \int_a^{\infty} f(u)\ \mathrm{d}u$
- $P(a < Y < b) = F(b) - F(a) = \int_a^b f(u)\ \mathrm{d}u$
- $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b) = P(a \leq Y < b)$

Unlike the PMF, which directly gives probabilities, the PDF does not represent probability directly. Instead, the probability is given by the area under the PDF curve over an interval. The PDF value $f(a)$ itself can be greater than 1, as long as the total area under the curve equals 1.

It is important to note that for continuous random variables, the probability of any single point is zero. This is why, as shown in the last rule above, the inequalities (strict or non-strict) don't affect the probability calculations for intervals. This stands in contrast to discrete random variables, where the inclusion of endpoints can change the probability value.

1.8 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** (CCDF),

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

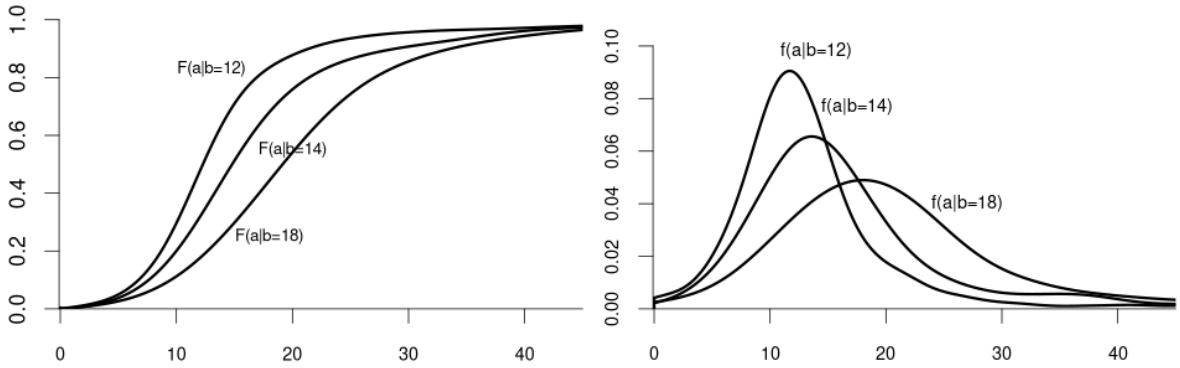
represents the distribution of a random variable Y given that another random variable Z takes a specific value b. It answers the question: "If we know that $Z = b$, what is the distribution of Y?"

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

Figure 1.7: Wage distributions conditional on education level

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**.

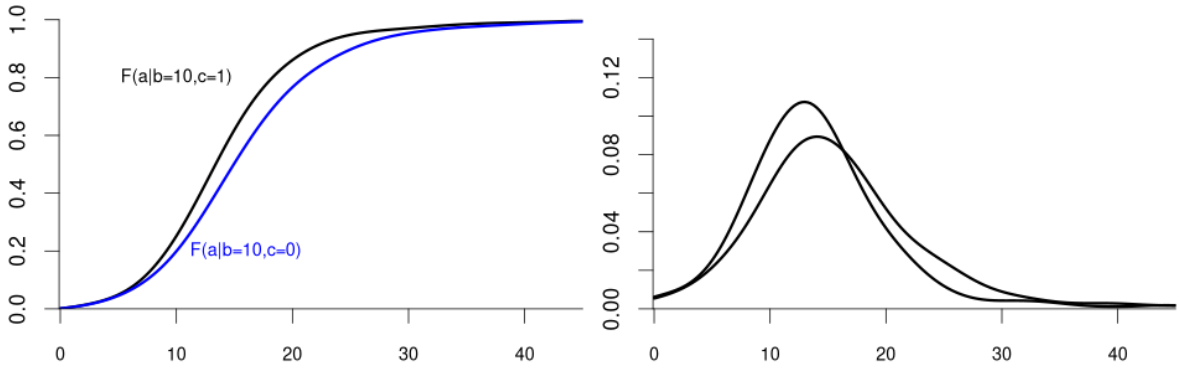
Note that the conditional CDF $F_{Y|Z=b}(a)$ can only be defined for values of b in the support of Z.

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 1.8: Wage distributions conditional on 10 years of 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.

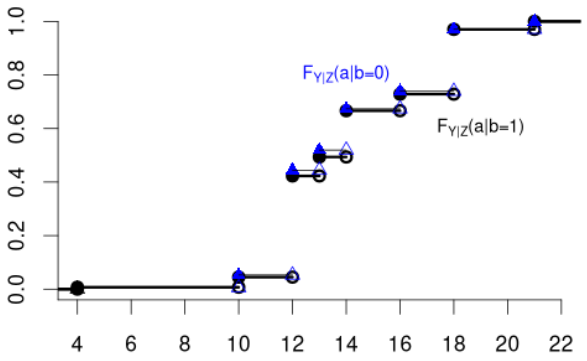
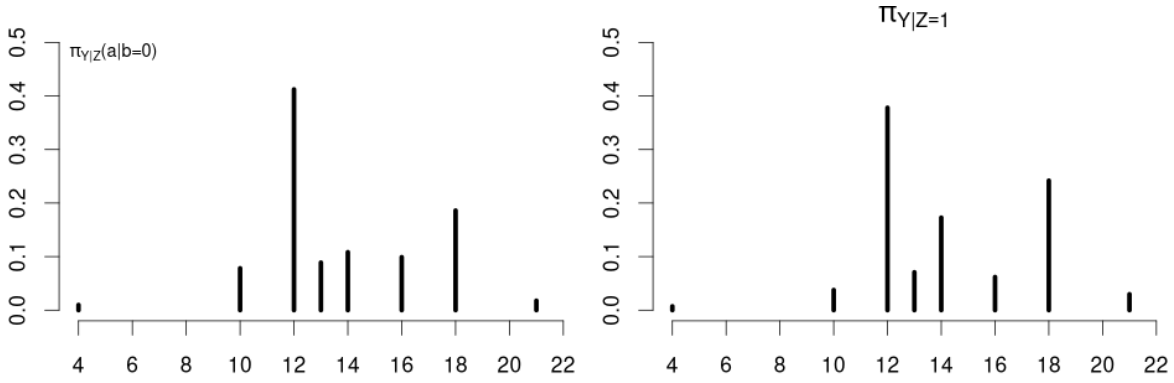


Figure 1.9: Conditional CDFs of education given married



(a) Education given unmarried

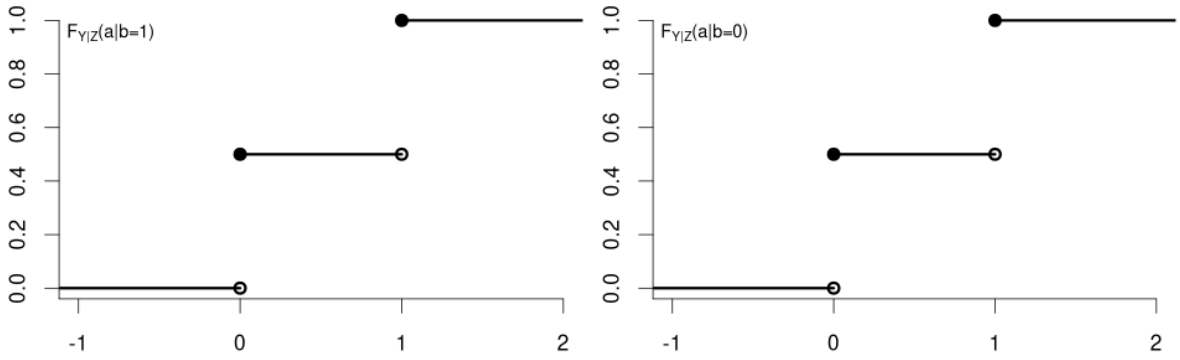
(b) Education given married

Figure 1.10: Conditional PMFs of education for unmarried (left) and married (right) individuals

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.

Clearly, *education* and *married* are dependent random variables. For example, $\pi_{Y|Z=0}(12) > \pi_{Y|Z=1}(12)$ 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 unmarried (b) Coin flip given married

Figure 1.11: Conditional CDFs of a coin flip for unmarried (left) and married (right) individuals

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**.

1.9 Independence of Random Variables

In the previous section, we saw that the distribution of a coin flip remains the same regardless of a person's marital status, illustrating the concept of independence. Let's now formalize this important concept.

Independence

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

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

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

> **Technical Note**: More rigorously, the independence condition should state "for almost every b" rather than "for all b". This means the condition must hold 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". 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.

For discrete random variables, independence can be expressed using PMFs: Y and Z are independent if and only if $\pi_{Y|Z=b}(a) = \pi_Y(a)$ for all a in the support of Y and all b in the support of Z. Similarly, for continuous random variables, independence means the conditional PDF factorizes $f_{Y|Z=b}(a) = f_Y(a)$.

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 all (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 all $(b_1, \ldots, b_n)$.

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:

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

1.10 Independent and Identically Distributed

An important concept in statistics is that of an independent and identically distributed (i.i.d.) sample. This arises naturally when we consider multiple random variables that share the same distribution and do not influence each other.

i.i.d. Sample / Random Sample

A collection of random variables $Y_1, \ldots, Y_n$ is **i.i.d.** (independent and identically distributed) if:

1. They are mutually independent: for each $i = 1, \ldots, n$,

$$F_{Y_i|Y_1=b_1,\ldots,Y_{i-1}=b_{i-1},Y_{i+1}=b_{i+1},\ldots,Y_n=b_n}(a) = F_{Y_i}(a)$$

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

2. They have the same distribution function: $F_{Y_i}(a) = F(a)$ for all $i = 1, \ldots, n$ and all a.

For example, consider n coin flips, where each Y_i represents the outcome of the i-th flip (with $Y_i = 1$ for heads and $Y_i = 0$ for tails). If the coin is fair and the flips are performed independently, then $Y_1, \ldots, Y_n$ form an i.i.d. sample with

$$F(a) = F_{Y_i}(a) = \begin{cases} 0 & a < 0 \\ 0.5 & 0 \leq a < 1 \\ 1 & a \geq 1 \end{cases} \qquad \text{for all } i = 1, \ldots, n.$$

Similarly, if we randomly select n individuals from a large population and measure their wages, the resulting measurements $Y_1, \ldots, Y_n$ can be treated as an i.i.d. sample. Each Y_i follows the same distribution (the wage distribution in the population), and knowledge of one person's wage doesn't affect the distribution of another's. The function F is called the **population distribution** or the **data-generating process (DGP)**.

1.11 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 the vector inequality $\boldsymbol{X}_1 \leq \boldsymbol{a}$ represents the intersection of component-wise inequalities, i.e., $\{X_{11} \leq a_1\} \cap \{X_{12} \leq a_2\} \cap \cdots \cap \{X_{1k} \leq a_k\}$.

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

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

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 all $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$.

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 of Random Vectors

The concept of i.i.d. samples naturally extends to random vectors. A collection of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is **i.i.d.** 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 all $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$.

An **i.i.d. dataset** (or **random sample**) is one where each multivariate observation not only comes from the same population distribution F but is independent of the others.

2 Expected Value

The CDF, PMF, and PDF fully characterize the probability distribution of a random variable but contain too much information for practical interpretation. We usually need summary measures that capture essential characteristics of a distribution. The **expectation** or **expected value** is the most important measure of the central tendency. It gives you the average value you can expect to get if you repeat the random experiment multiple times.

2.1 Discrete Case

As previously defined, a discrete random variable Y is one that can take on a countable number of distinct values. The probability that Y takes a specific value a is given by the probability mass function (PMF) $\pi(a) = P(Y = a)$.

2.1.1 Expectation

Expected Value (Discrete Case)

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

$$E[Y] = \sum_{u \in \mathcal{Y}} u \cdot \pi(u). \tag{2.1}$$

The expected value can be interpreted as the long-run average outcome of the random variable Y if we were to observe it repeatedly in independent experiments. For example, if we flip a fair coin many times, the proportion of heads will approach 0.5, which is the expected value of the coin toss random variable.

Example: Binary Random Variable

A **binary** or **Bernoulli** random variable Y takes on only two possible values: 0 and 1. The support is $\mathcal{Y} = \{0, 1\}$, and the PMF is $\pi(1) = p$ and $\pi(0) = 1 - p$ for some $p \in (0, 1)$. The expected value of Y is:

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

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

Example: Education Variable

Using the variable *education* with its PMF values introduced previously, we can calculate the expected value:

$$\begin{aligned} E[Y] &= 4 \cdot \pi(4) + 10 \cdot \pi(10) + 12 \cdot \pi(12) + 13 \cdot \pi(13) \\ &\quad + 14 \cdot \pi(14) + 16 \cdot \pi(16) + 18 \cdot \pi(18) + 21 \cdot \pi(21) \\ &= 0.032 + 0.55 + 4.716 + 1.027 + 2.03 + 1.248 + 3.924 + 0.504 \\ &= 14.031 \end{aligned}$$

So, the expected value of *education* is 14.031 years, which corresponds roughly to the completion of short-cycle tertiary education (ISCED level 5).

2.1.2 Conditional Expectation

Previously, we introduced conditional probability distributions, which describe the distribution of a random variable given that another random variable takes a specific value. Building on this foundation, we can define the conditional expectation, which measures the expected value of a random variable when we have information about another random variable.

Conditional Expectation Given a Fixed Value

For a discrete random variable Y with conditional PMF $\pi_{Y|Z=b}(a)$, the conditional expectation of Y given $Z = b$ is defined as:

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

This formula closely resembles the unconditional expectation, but uses the conditional PMF instead of the marginal PMF. The conditional expectation $E[Y|Z=b]$ can be interpreted as the average value of Y we expect to observe, given that we know Z has taken the value b.

Example: Education Given Marital Status

Let's examine the conditional PMFs of *education* given *marital status* studied previously.

For unmarried individuals ($Z=0$):

$$\pi_{Y|Z=0}(a) = \begin{cases} 0.01 & \text{if } a=4 \\ 0.07 & \text{if } a=10 \\ 0.43 & \text{if } a=12 \\ 0.09 & \text{if } a=13 \\ 0.10 & \text{if } a=14 \\ 0.09 & \text{if } a=16 \\ 0.19 & \text{if } a=18 \\ 0.02 & \text{if } a=21 \\ 0 & \text{otherwise} \end{cases}$$

For married individuals ($Z=1$):

$$\pi_{Y|Z=1}(a) = \begin{cases} 0.01 & \text{if } a=4 \\ 0.03 & \text{if } a=10 \\ 0.38 & \text{if } a=12 \\ 0.07 & \text{if } a=13 \\ 0.17 & \text{if } a=14 \\ 0.06 & \text{if } a=16 \\ 0.25 & \text{if } a=18 \\ 0.03 & \text{if } a=21 \\ 0 & \text{otherwise} \end{cases}$$

The conditional expectation of *education* for unmarried individuals is:

$$
\begin{aligned}
E[Y|Z=0] &= 4 \cdot 0.01 + 10 \cdot 0.07 + 12 \cdot 0.43 + 13 \cdot 0.09 \\
&\quad + 14 \cdot 0.10 + 16 \cdot 0.09 + 18 \cdot 0.19 + 21 \cdot 0.02 \\
&= 13.75
\end{aligned}
$$

The conditional expectation of *education* for married individuals is:

$$
\begin{aligned}
E[Y|Z=1] &= 4 \cdot 0.01 + 10 \cdot 0.03 + 12 \cdot 0.38 + 13 \cdot 0.07 \\
&\quad + 14 \cdot 0.17 + 16 \cdot 0.06 + 18 \cdot 0.25 + 21 \cdot 0.03 \\
&= 14.28
\end{aligned}
$$

We observe that the expected education level is higher for married individuals (14.28 years) compared to unmarried individuals (13.75 years), which suggests a dependence between *marital status* and *education*.

2.1.3 Conditional Expectation Function (CEF)

So far, we have used $E[Y|Z=b]$ to denote the conditional expectation of Y given a specific value b of Z. This is a fixed number for each value of b. A related concept is the **Conditional Expectation Function**, denoted as $E[Y|Z]$ without specifying a particular value for Z.

Conditional Expectation Function (CEF)

The conditional expectation function $E[Y|Z]$ represents a random variable that depends on the random outcome of Z. It is a function that maps each possible value of Z to the corresponding conditional expectation:

$$
E[Y|Z] = m(Z) \quad \text{where} \quad m(b) = E[Y|Z=b]
$$

Here, $m(\cdot)$ is the function that represents the CEF, mapping each possible value of Z to the corresponding conditional expectation.

The CEF is random precisely because it is a function of the random variable Z. Before we observe the value of Z, we cannot determine the value of $E[Y|Z]$. Once we observe Z, the CEF gives us the expected value of Y corresponding to that specific observation. This makes $E[Y|Z]$ a random variable whose value depends on the random outcome of Z. In contrast, $E[Y|Z=b]$ is a deterministic scalar non-random value.

For our marital status example, the CEF is:

$$E[Y|Z] = m(Z) = \begin{cases} 13.75 & \text{if } Z = 0 \text{ (unmarried)} \\ 14.28 & \text{if } Z = 1 \text{ (married)} \end{cases}$$

In our population, the marginal PMF of *married* is

$$\pi_Z(a) = \begin{cases} 0.4698 & \text{if } a = 0 \text{ (unmarried)} \\ 0.5302 & \text{if } a = 1 \text{ (married)} \\ 0 & \text{otherwise.} \end{cases}$$

Using these values the PMF of $E[Y|Z]$ is:

$$\pi_{E[Y|Z]}(a) = P(E[Y|Z] = a) = \begin{cases} 0.4698 & \text{if } a = 13.75 \\ 0.5302 & \text{if } a = 14.28 \\ 0 & \text{otherwise.} \end{cases}$$

2.1.4 Law of Iterated Expectations (LIE)

Law of Iterated Expectations

For two random variables Y and Z:

$$E[Y] = E[E[Y|Z]]$$

This elegant equation states that the expected value of Y can be found by first calculating the conditional expectation of Y given Z (which gives us the random variable $E[Y|Z]$), and then taking the expected value of this random variable. In other words, we are taking the expectation of the conditional expectation.

The Law of Iterated Expectations is a fundamental tool in econometrics with numerous applications. It is particularly important for understanding the properties of estimators in the presence of conditioning variables like in regression analysis.

To understand why this law holds, let's consider an intuitive argument based on the **law of total probability**. For discrete random variables, the law of total probability tells us that we can find the overall probability of an event $Y = a$ by considering all possible scenarios $Z = b$ that could lead to that event. More precisely, $P(Y = a)$ equals the weighted sum

of conditional probabilities $P(Y = a|Z = b)$ across all possible values b of Z, where each conditional probability is weighted by $P(Z = b)$:

$$\pi_Y(a) = \sum_{b \in \mathcal{Z}} \pi_{Y|Z=b}(a) \cdot \pi_Z(b)$$

The LIE follows a similar logic. We can think of the overall expectation of Y as a weighted average of conditional expectations $E[Y|Z = b]$ across all possible values of Z, with each conditional expectation weighted by the probability of the corresponding Z value:

$$E[Y] = \sum_{u \in \mathcal{Z}} E[Y|Z = u] \cdot \pi_Z(u)$$

The right hand side is precisely what $E[E[Y|Z]]$ means: take the conditional expectation function $E[Y|Z]$ and average it over all possible values $u \in \mathcal{Z}$ of Z, where $\mathcal{Z}$ is the support of Z.

For our *education* and *marital status* example, the LIE gives us:

$$\begin{aligned} E[Y] &= E[E[Y|Z]] \\ &= E[Y|Z = 0] \cdot \pi_Z(0) + E[Y|Z = 1] \cdot \pi_Z(1) \\ &= 13.75 \cdot 0.4698 + 14.28 \cdot 0.5302 \\ &= 6.460 + 7.571 \\ &= 14.031 \end{aligned}$$

This matches exactly with our directly calculated expected value of 14.031 years from the marginal PMF.

2.1.5 Conditioning Theorem (CT)

Conditioning Theorem / Factorization Property

For two random variables Y and Z:

$$E[ZY|Z] = Z \cdot E[Y|Z]$$

To see this, let's first consider the case for a specific value $Z = b$:

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

When we consider this factorization across all possible values of Z rather than a fixed value b, we get the general form of the Conditioning Theorem: $E[ZY|Z] = Z\cdot E[Y|Z]$.

The conditioning theorem states that we can factor out the conditioning variable Z from the conditional expectation. The intuition is that when we condition on Z, we're essentially treating it as if we already know its value, so it behaves like a constant within the conditional expectation. Since summation is linear and constants can be factored out, Z can be factored out of $E[ZY|Z]$.

This theorem is particularly useful in econometric derivations, especially when working with regression models.

For example, in our marital status context, if we want to compute $E[ZY|Z]$ (the conditional expectation of education multiplied by marital status, given marital status), we get:

$$E[ZY|Z] = Z\cdot E[Y|Z] = \begin{cases} 0\cdot 13.75 = 0 & \text{if } Z=0 \text{ (unmarried)} \\ 1\cdot 14.28 = 14.28 & \text{if } Z=1 \text{ (married)} \end{cases}$$

I've evaluated your draft for the "Continuous Case" subsection, and it's a good start. Here's my suggested improved version with better organization, more detailed explanations, and enhanced examples:

2.2 Continuous Case

Now, let's extend our discussion of expected values to continuous random variables, which are characterized by probability density functions (PDFs) rather than probability mass functions (PMFs).

Expected Value (Continuous Case)

The **expectation** or **expected value** of a continuous random variable Y with PDF $f_Y(u)$ and support $\mathcal{Y}$ is defined as

$$E[Y] = \int_{\mathcal{Y}} u f_Y(u)\,du = \int_{-\infty}^{\infty} u f_Y(u)\,du. \tag{2.2}$$

Intuitively, this integral calculates a weighted average of all possible values of Y, where the weight of each value is given by its density. This is analogous to the discrete case, where we

computed a weighted sum. The key difference is that continuous random variables have infinitely many possible values within their support, requiring integration rather than summation.

2.2.1 Conditional Expectation for Continuous Variables

For continuous random variables, the conditional expectation given that $Z = b$ is defined similarly:

Conditional Expectation (Continuous Case)

For a continuous random variable Y with conditional PDF $f_{Y|Z=b}(u)$, the conditional expectation of Y given $Z = b$ is:

$$E[Y|Z = b] = \int_{-\infty}^{\infty} u f_{Y|Z=b}(u)\, du$$

The same principles of conditional expectation functions (CEF) that we discussed for discrete variables apply here as well. The CEF $E[Y|Z] = m(Z)$ is a random variable that depends on the random outcome of Z, where $m(b) = E[Y|Z = b]$ is the function mapping each possible value of Z to the corresponding conditional expectation.

2.2.2 Examples with Continuous Random Variables

Example 1: Uniform Distribution

A random variable Y follows a **uniform distribution** on the interval $[0, 1]$ if its PDF is constant across this interval:

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

The expected value is calculated as:

$$E[Y] = \int_{-\infty}^{\infty} u f(u)\, du = \int_0^1 u \cdot 1\, du = \int_0^1 u\, du = \left.\frac{u^2}{2}\right|_0^1 = \frac{1}{2}$$

So the expected value of a uniform random variable on $[0, 1]$ is exactly $\frac{1}{2}$, the midpoint of the interval. More generally, for a uniform distribution on $[a, b]$, the expected value is $\frac{a+b}{2}$.

Example 2: Wage Distribution

Let's return to our *wage* variable from Part 1. Suppose the wage distribution in our population has the following PDF:

$$f(u) = \begin{cases} \frac{1}{20}e^{-u/20} & \text{if } u \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

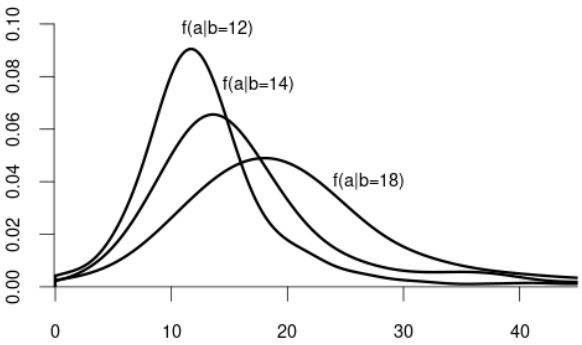
This represents an exponential distribution with parameter $\lambda = 1/20$. The expected value is:

$$E[Y] = \int_0^\infty u \cdot \frac{1}{20}e^{-u/20}\, du = \frac{1}{20}\int_0^\infty u \cdot e^{-u/20}\, du$$

Using integration by parts or leveraging the known mean of an exponential distribution ($1/\lambda$), we get $E[Y] = 20$ EUR per hour.

Example 3: Wage Given Education

From Figure 1.7b, we saw that the conditional distribution of *wage* given *education* varies substantially:



(a) Conditional PDFs of wage given education

These different distributions lead to different conditional expectations:

- $E[Y|Z = 12] = 14.3$ EUR/hour
- $E[Y|Z = 14] = 17.8$ EUR/hour
- $E[Y|Z = 18] = 27.0$ EUR/hour

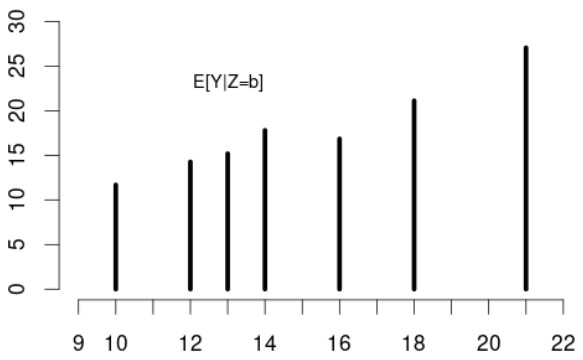


Figure 2.2: CEF of wage given education

The CEF plot is given below:

The increasing conditional expectations reflect the positive relationship between education and wages. This dependency confirms that education and wages are not independent random variables.

Example 4: Wage Given Experience

Now consider the relationship between *wage* (Y) and years of *experience* (Z). Labor economics theory suggests that wages typically increase with experience but at a diminishing rate. Suppose empirical analysis reveals that the conditional expectation of wage given experience level b follows a quadratic form:

$$m(b) = E[Y|Z = b] = 19 + 0.5b - 0.013b^2$$

This functional form captures the initial increase in wages with experience (positive linear term $0.5b$) and the diminishing returns over time (negative quadratic term $-0.013b^2$).

For some specific values:

- $m(5) = E[Y|Z = 5] = 21.2$ EUR/hour
- $m(10) = E[Y|Z = 10] = 22.7$ EUR/hour
- $m(20) = E[Y|Z = 20] = 23.8$ EUR/hour
- $m(30) = E[Y|Z = 30] = 22.3$ EUR/hour

The conditional expectation function is maximized at the point where its derivative equals zero:

$$\frac{d}{db}m(b) = 0.5 - 0.026b = 0 \implies b = \frac{0.5}{0.026} \approx 19.2$$

This suggests that, on average, wages peak at around 19.2 years of experience in this population.

As a function of the random variable Z, the CEF is:

$$E[Y|Z] = m(Z) = 19 + 0.5Z - 0.013Z^2$$

This is itself a random variable because its value depends on the random outcome of Z.

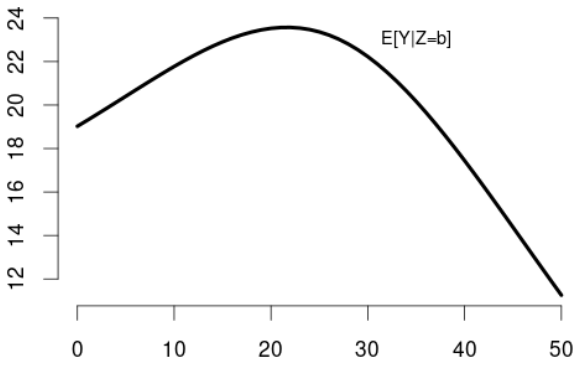


Figure 2.3: CEF of wage given experience

2.3 General Case

We can define the expected value of a random variable Y in a unified way that applies to both discrete and continuous cases using its CDF $F_Y(u)$.

Expected Value (General Definition)

$$E[Y] = \int_{-\infty}^{\infty} u \, \mathrm{d}F_Y(u) \tag{2.3}$$

This formula uses the **Riemann-Stieltjes integral**, which generalizes the familiar Riemann integral. To understand this, recall that the standard **Riemann integral** $\int_a^b g(x)\,dx$ is defined as:

$$\int_a^b g(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} g(x_i^*)\Delta x_i$$

where $[a, b]$ is partitioned into n subintervals $[x_{i-1}, x_i]$ of width $\Delta x_i = x_i - x_{i-1}$, and x_i^* is any point in the i-th subinterval. The limit is taken as the maximum width of all subintervals approaches zero.

In contrast, the **Riemann-Stieltjes integral** $\int_a^b g(x)\, dh(x)$ is defined as:

$$\int_a^b g(x)\, dh(x) = \lim_{n\to\infty} \sum_{i=1}^{n} g(x_i^*)\Delta h_i$$

where $\Delta h_i = h(x_i) - h(x_{i-1})$ represents the increment in the function h over the i-th subinterval.

Intuitively, while the standard Riemann integral weighs the function values by increments in the x-axis, the Riemann-Stieltjes integral weighs them by increments in another function, allowing us to seamlessly handle both continuous distributions (where we integrate against a smooth CDF) and discrete distributions (where the integrator function has jumps).

For infinite intervals, as in $\int_{-\infty}^{\infty} u\, dF_Y(u)$, the integral is defined as:

$$\int_{-\infty}^{\infty} u\, dF_Y(u) = \lim_{a\to-\infty, b\to\infty} \int_a^b u\, dF_Y(u)$$

2.3.1 Special Case: Continuous Random Variables

For a continuous random variable Y with PDF $f_Y(u)$, the CDF $F_Y(u)$ is differentiable with

$$\frac{dF_Y(u)}{du} = f_Y(u)$$

Hence:

$$\mathrm{d}F_Y(u) = f_Y(u)\, \mathrm{d}u$$

Substituting this into our unified definition:

$$E[Y] = \int_{-\infty}^{\infty} u\, \mathrm{d}F_Y(u) = \int_{-\infty}^{\infty} u \cdot f_Y(u)\, \mathrm{d}u$$

This recovers the standard definition for continuous random variables we saw earlier.

2.3.2 Special Case: Discrete Random Variables

For a discrete random variable, the CDF $F_Y(u)$ has jumps at each point $u \in \mathcal{Y}$ where Y can take values. The size of each jump is:

$$\Delta F_Y(u) = F_Y(u) - F_Y(u^-) = P(Y = u) = \pi_Y(u)$$

where $F_Y(u^-) = \lim_{\varepsilon \to 0, \varepsilon > 0} F_Y(u - \varepsilon)$ is the left limit of F_Y at u.

For discrete variables, the Riemann-Stieltjes integral simplifies to a sum over these jumps:

$$E[Y] = \int_{-\infty}^{\infty} u \, \mathrm{d}F_Y(u) = \sum_{u \in \mathcal{Y}} u \cdot \Delta F_Y(u) = \sum_{u \in \mathcal{Y}} u \cdot \pi_Y(u)$$

This matches our earlier definition for discrete random variables.

2.3.3 Why the General Case Matters

The unified approach to expected values offers several important advantages:

1. **Conceptual and practical simplicity**: Using the CDF as the foundation emphasizes that distinguishing between discrete and continuous random variables isn't critically important for defining expectations. In econometric practice, expectations are computed without needing to categorize the distribution first, as the same principles apply regardless of distribution type.
2. **Handling mixed and exotic distributions**: Many real-world variables have distributions that don't fit neatly into pure categories—like wages with both continuous values and "spikes" at round numbers, or insurance claims that are zero with positive probability but continuous for positive values. The general definition also accommodates more exotic theoretical cases like the Cantor distribution.
3. **Unified theoretical development**: For developing econometric theory, having a single definition simplifies proofs and ensures results apply broadly without requiring separate cases for different distribution types, allowing us to focus on understanding the properties and relationships between random variables.

2.3.4 General Definition of Conditional Expectation

Just as we can define the expected value in a unified way, we can also define conditional expectation in a general form that applies to all types of random variables.

Conditional Expectation (General Definition)

For random variables Y and Z, the conditional expectation of Y given $Z = b$ is defined as:

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

where $F_{Y|Z=b}(u)$ is the conditional CDF of Y given $Z = b$.

Similarly, for any random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$, the conditional expectation of Y given $\boldsymbol{Z} = \boldsymbol{b}$ is:

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

This definition uses the Riemann-Stieltjes integral with respect to the conditional CDF. For continuous random variables with conditional PDF $f_{Y|Z=b}(u)$, this becomes:

$$E[Y|Z = b] = \int_{-\infty}^{\infty} u \cdot f_{Y|Z=b}(u) \, \mathrm{d}u$$

For discrete random variables with conditional PMF $\pi_{Y|Z=b}(u)$, it simplifies to:

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

The **conditional expectation function** (CEF) is then defined as:

$$E[Y|Z] = m(Z)$$

where $m(b) = E[Y|Z = b]$ for each possible value b that Z can take. This makes $E[Y|Z]$ a random variable whose value depends on the random outcome of Z.

Similarly, if $\boldsymbol{Z}$ is a vector, we have:

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

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

This can also be extended to conditioning on a $n \times k$ matrix of random variables $\boldsymbol{X}$ (e.g., a regressor matrix), which gives $E[Y|\boldsymbol{X}]$. This extension is particularly important in econometrics for regression analysis.

2.3.5 Conditional Expectation and Independence

When two random variables Y and Z are independent, the conditional distributions simplify considerably. As we saw in the first section, independence means that the conditional distribution of Y given $Z = b$ is the same as the marginal distribution of Y. This fundamental property has important implications for conditional expectations.

Conditional Expectation and Independence

If random variables Y and Z are independent, then:

$$E[Y|Z = b] = E[Y] \quad \text{for all } b$$

And consequently:

$$E[Y|Z] = E[Y]$$

In other words, when Y and Z are independent, knowing the value of Z provides no information about the expected value of Y. The conditional expectation equals the unconditional expectation for every possible value of Z.

To understand why this holds, recall that for independent random variables, the conditional CDF equals the marginal CDF. Using the general definition of conditional expectation:

$$E[Y|Z = b] = \int_{-\infty}^{\infty} u \, \mathrm{d}F_{Y|Z=b}(u) = \int_{-\infty}^{\infty} u \, \mathrm{d}F_Y(u) = E[Y]$$

The middle equality holds because $F_{Y|Z=b}(u) = F_Y(u)$ for all u when Y and Z are independent.

This means that the conditional expectation function $E[Y|Z] = m(Z)$ reduces to a constant function:

$$m(b) = E[Y] \quad \text{for all } b$$

For example, recall our coin flip example from Part 1, where we noted that the distribution of a coin toss remains the same regardless of whether a person is married or unmarried:

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

where Y represents the coin outcome and Z is the marital status. In this case, $E[Y|Z = 0] = E[Y|Z = 1] = E[Y] = 0.5$, reflecting that the expected outcome of a fair coin toss is 0.5 regardless of the marital status of the person tossing it. Hence, $E[Y|Z] = E[Y]$ for this example.

2.3.6 Expected Value of Functions

Often, we are interested not just in the expected value of a random variable Y itself, but in the expected value of some function of Y, such as Y^2 or $\log(Y)$.

Expected Value of a Function

For any function $g(\cdot)$, the expected value of $g(Y)$ is:

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

For conditional expectations:

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

The conditional expectation function is $E[g(Y)|Z] = m(Z)$, where $m(b) = E[g(Y)|Z = b]$.

As discussed above for the different cases, $\mathrm{d}F_Y(u)$ can be replaced by the PMF or the PDF:

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

Example: Transformation of a Binary Variable

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

$$\begin{aligned} E[\log(Y+1)] &= \log(1)\cdot\pi_Y(0) + \log(2)\cdot\pi_Y(1) \\ &= \log(1)\cdot\frac{1}{2} + \log(2)\cdot\frac{1}{2} \\ &= \frac{\log(2)}{2} \approx 0.347 \end{aligned}$$

This approach allows us to compute expectations of arbitrary functions of random variables.

2.3.7 Moments and Related Measures

We can define various moments of a random variable and functions of these moments using our general expectation framework:

Moments and Related Measures

- r**-th moment** of Y:
$$E[Y^r] = \int_{-\infty}^{\infty} u^r \, \mathrm{d}F_Y(u)$$
- r**-th central moment**:
$$E[(Y - E[Y])^r] = \int_{-\infty}^{\infty} (u - E[Y])^r \, \mathrm{d}F_Y(u)$$
- **Variance** (2nd central moment):
$$\mathrm{Var}(Y) = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (u - E[Y])^2 \, \mathrm{d}F_Y(u)$$
- **Standard deviation**:
$$\mathrm{sd}(Y) = \sqrt{\mathrm{Var}(Y)}$$
- r**-th standardized moment**:
$$E\left[\left(\frac{Y - E[Y]}{\mathrm{sd}(Y)}\right)^r\right] = \int_{-\infty}^{\infty} \left(\frac{u - E[Y]}{\mathrm{sd}(Y)}\right)^r \, \mathrm{d}F_Y(u)$$
- **Skewness** (3rd standardized moment):
$$\mathrm{ske}(Y) = E\left[\left(\frac{Y - E[Y]}{\mathrm{sd}(Y)}\right)^3\right]$$
- **Kurtosis** (4th standardized moment):
$$\mathrm{kur}(Y) = E\left[\left(\frac{Y - E[Y]}{\mathrm{sd}(Y)}\right)^4\right]$$

Similarly, conditional versions of these moments can be defined. For example:

- The r-th conditional moment:
$$E[Y^r | Z = b] = \int_{-\infty}^{\infty} u^r \, \mathrm{d}F_{Y|Z=b}(u)$$

- The conditional variance:

$$\begin{aligned}\operatorname{Var}(Y|Z=b) &= E[(Y - E[Y|Z=b])^2|Z=b] \\ &= \int_{-\infty}^{\infty} (u - E[Y|Z=b])^2 \, \mathrm{d}F_{Y|Z=b}(u)\end{aligned}$$

- The conditional variance function:

$$\operatorname{Var}(Y|Z) = v(Z), \quad \text{where } v(b) = \operatorname{Var}(Y|Z=b)$$

2.4 Properties of Expectation

The general expected value operator has several important properties. Here we focus on three fundamental properties: linearity, the Law of Iterated Expectations, and the Conditioning Theorem.

2.4.1 Linearity

Linearity of Expectation

For any constants $a, b \in \mathbb{R}$ and random variable Y:

$$E[a + bY] = a + bE[Y]$$

The same property holds for conditional expectations:

$$E[a + bY|Z] = a + bE[Y|Z]$$

This property tells us that the expectation of a linear transformation of a random variable equals the same linear transformation of the expectation.

To understand why this holds, consider the definition of expectation using the Riemann-Stieltjes integral:

$$E[a + bY] = \int_{-\infty}^{\infty} (a + bu) \, \mathrm{d}F_Y(u)$$

We can separate this into two parts:

$$\int_{-\infty}^{\infty} (a + bu) \, \mathrm{d}F_Y(u) = \int_{-\infty}^{\infty} a \, \mathrm{d}F_Y(u) + \int_{-\infty}^{\infty} bu \, \mathrm{d}F_Y(u)$$

The first integral equals a because $\int_{-\infty}^{\infty} \mathrm{d}F_Y(u) = 1$ for any probability distribution. The second integral equals $b \cdot E[Y]$. Combining these results gives us $a + bE[Y]$.

2.4.2 Law of Iterated Expectations (LIE)

Law of Iterated Expectations

For any random variables Y and Z:

$$E[Y] = E[E[Y|Z]]$$

The LIE states that the expected value of Y can be found by first taking the conditional expectation of Y given Z, and then taking the expectation of this result over the distribution of Z.

This result relies on the law of total probability, which connects marginal and conditional distributions. For discrete random variables, this law states:

$$\pi_Y(u) = \sum_{b \in \mathcal{Z}} \pi_{Y|Z=b}(u) \cdot \pi_Z(b)$$

For continuous random variables with densities, the law takes the form:

$$f_Y(u) = \int_{-\infty}^{\infty} f_{Y|Z=b}(u) \cdot f_Z(b)\, db$$

In the general case using CDFs, the law of total probability is expressed as:

$$\mathrm{d}F_Y(u) = \int_{-\infty}^{\infty} \mathrm{d}F_{Y|Z=b}(u)\, \mathrm{d}F_Z(b)$$

When we evaluate $E[E[Y|Z]]$, we are calculating:

$$E[E[Y|Z]] = \int_{-\infty}^{\infty} E[Y|Z=b]\, \mathrm{d}F_Z(b)$$

Expanding the inner conditional expectation:

$$E[E[Y|Z]] = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} u\, \mathrm{d}F_{Y|Z=b}(u) \right) \mathrm{d}F_Z(b)$$

Applying the general form of the law of total probability, this double integral simplifies to:

$$E[E[Y|Z]] = \int_{-\infty}^{\infty} u\, \mathrm{d}F_Y(u) = E[Y]$$

2.4.3 Conditioning Theorem (CT)

Conditioning Theorem

For any random variables Y and Z, and any measurable function $g(\cdot)$:

$$E[g(Z)Y|Z] = g(Z)E[Y|Z]$$

The Conditioning Theorem states that when we condition on Z, we can factor out any function of Z from the conditional expectation.

To see why this holds, consider a specific value $Z = b$. The conditional expectation becomes:

$$E[g(Z)Y|Z = b] = E[g(b)Y|Z = b] = g(b)E[Y|Z = b]$$

The first equality follows because $g(Z) = g(b)$ when $Z = b$. The second equality uses the linearity property of expectation, treating $g(b)$ as a constant.

Since this relationship holds for every possible value of Z, we have the general result:

$$E[g(Z)Y|Z] = g(Z)E[Y|Z]$$

2.5 Heavy Tails: When Expectations Fail to Exist

Our previous discussions assumed that expected values and covariance matrices exist, but this isn't always guaranteed. Some probability distributions have such slow decay in their tails that moments of certain order may be infinite or undefined. These "heavy-tailed" distributions present special challenges for statistical analysis.

2.5.1 Infinite Expectations

Infinite Expectation

A random variable Y has an infinite expectation if:

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

When this occurs, the expected value $E[Y]$ either diverges to positive infinity, negative infinity, or is not well-defined due to both positive and negative parts of the integral diverging.

In such cases, the sample mean does not converge to any finite value as the sample size increases.

The sample mean of i.i.d. samples from most distributions converges to the population mean as sample size increases (a property known as consistency). However, there are exceptional cases where consistency fails because the population mean itself is infinite.

2.5.2 Examples of Distributions with Infinite Moments

Pareto Distribution

The simple Pareto distribution with parameter $\alpha = 1$ has the PDF:

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

The expected value is:

$$E[Y] = \int_{-\infty}^{\infty} uf(u)\ \mathrm{d}u = \int_{1}^{\infty} \frac{u}{u^2}\ \mathrm{d}u = \int_{1}^{\infty} \frac{1}{u}\ \mathrm{d}u = \log(u)|_{1}^{\infty} = \infty$$

Since the population mean is infinite, the sample mean cannot converge to any finite value and is therefore inconsistent.

St. Petersburg Paradox

The game of chance from the St. Petersburg paradox is a discrete example with infinite expectation. In this game, a fair coin is tossed until a tail appears; if the first tail is on the nth toss, the payoff is 2^n dollars. The expected payoff is:

$$E[Y] = \sum_{n=1}^{\infty} 2^n \cdot \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = \infty$$

This infinity arises from the infinite sum of 1's, reflecting the unbounded potential payoffs in the game.

Cauchy Distribution

The Cauchy distribution (also known as the t-distribution with 1 degree of freedom) has the PDF:

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

The Cauchy distribution presents a fascinating case where the sample mean of n observations has exactly the same distribution as a single observation, regardless of how large n becomes. This means the sample mean does not converge to any value as sample size increases.

t-Distribution with Few Degrees of Freedom

- **Cauchy distribution** (t_1): No finite moments
- t_2 **distribution**: Finite mean but infinite variance
- t_3 **distribution**: Finite variance but infinite skewness
- t_4 **distribution**: Finite skewness but infinite kurtosis

More generally, for a t-distribution with m degrees of freedom:

$$E[Y^k] < \infty \text{ for } k < m$$

$$E[Y^k] = \infty \text{ for } k \geq m$$

2.5.3 Real-World Examples

Heavy-tailed distributions arise in many real-world phenomena where extreme events and large outliers are common:

1. **Financial returns**: Stock market crashes and extreme price movements
2. **Income and wealth distributions**: Extreme wealth concentration
3. **Natural disasters**: Extreme earthquakes, floods, or storms

3 Multiple Random Variables

3.1 Expectations with Multiple Random Variables

In the previous sections, we focused on single random variables and their properties. Now we extend these concepts to scenarios involving multiple random variables, which is essential for analyzing relationships between variables in statistical modeling.

Recall that for any univariate function $g(\cdot)$, the expected value of $g(Y)$ and the conditional expectation given $Z = b$ are:

$$E[g(Y)] = \int_{-\infty}^{\infty} g(u)\ \mathrm{d}F_Y(u), \quad E[g(Y)|Z = b] = \int_{-\infty}^{\infty} g(u)\ \mathrm{d}F_{Y|Z=b}(u)$$

where $F_Y(a)$ is the marginal CDF of Y and $F_{Y|Z=b}(a)$ is the conditional CDF of Y given $Z = b$.

For functions of multiple random variables, we extend this approach using multivariate functions. For a bivariate function $h(Y, Z)$, we can calculate the expected value using the Law of Iterated Expectations (LIE):

Expected Value of a Bivariate Function

For a bivariate function $h(Y, Z)$, the expected value can be calculated as:

$$E[h(Y, Z)] = E[E[h(Y, Z)|Z]]$$

This double expectation can be expressed as:

$$\begin{aligned} E[E[h(Y, Z)|Z]] &= \int_{-\infty}^{\infty} E[h(Y, Z)|Z = b]\ \mathrm{d}F_Z(b) \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(u, b)\ \mathrm{d}F_{Y|Z=b}(u) \right)\ \mathrm{d}F_Z(b) \end{aligned}$$

3.1.1 Important Special Cases

Sum of Random Variables

For the sum of two random variables, $h(Y, Z) = Y + Z$, we have:

$$\begin{aligned} E[Y + Z] &= E[E[Y + Z|Z]] \\ &= E[E[Y|Z] + Z] \quad \text{(linearity of conditional expectation)} \\ &= E[E[Y|Z]] + E[Z] \quad \text{(linearity of expectation)} \\ &= E[Y] + E[Z] \quad \text{(by the LIE)} \end{aligned}$$

Linearity of Expectation for Sums

The expected value of a sum equals the sum of the expected values:

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

This property holds regardless of whether Y and Z are independent or dependent.

Product of Random Variables

For the product of two random variables, $h(Y, Z) = YZ$, we have:

$$\begin{aligned} E[YZ] &= E[E[YZ|Z]] \\ &= E[Z \cdot E[Y|Z]] \quad \text{(by the CT)} \end{aligned}$$

When Y and Z are independent, $E[Y|Z] = E[Y]$, so:

$$E[YZ] = E[Z \cdot E[Y]] = E[Z] \cdot E[Y]$$

Expected Value of a Product

- **General case**: $E[YZ] = E[Z \cdot E[Y|Z]]$
- **Independent case**: If Y and Z are independent, then $E[YZ] = E[Y] \cdot E[Z]$

$E[YZ]$ is also known as the **first cross moment** of Y and Z.

Example: Product of Education and Wage

Consider the random variables *education* (Z) and *wage* (Y) from our earlier examples. These variables are dependent, with $E[Y|Z = b]$ following the pattern shown in the CEF plot from the previous section.

If $E[Y|Z = b] = 2 + 1.2b$ (a simplified linear relationship), then:

$$\begin{aligned} E[YZ] &= E[Z \cdot E[Y|Z]] \\ &= E[Z \cdot (2 + 1.2Z)] \\ &= E[2Z + 1.2Z^2] \\ &= 2E[Z] + 1.2E[Z^2] \end{aligned}$$

If $E[Z] = 14$ years (mean education) and $E[Z^2] = 210$ (second moment of education), then:

$$E[YZ] = 2 \cdot 14 + 1.2 \cdot 210 = 28 + 252 = 280$$

3.1.2 Extending to Three or More Variables

For functions of three or more random variables, we can extend this approach by nesting conditional expectations. For $h(X, Y, Z)$:

$$E[h(X, Y, Z)] = E[E[E[h(X, Y, Z)|X, Y]|X]]$$

This formula allows us to decompose the expectation iteratively, conditioning on one variable at a time.

3.2 Covariance and Correlation

Having explored expectations involving multiple random variables, we now introduce measures that quantify the relationship between random variables: covariance and correlation.

3.2.1 Covariance

Covariance

The **covariance** between two random variables Y and Z is defined as:

$$\text{Cov}(Y, Z) = E[(Y - E[Y])(Z - E[Z])]$$

An equivalent and often more practical definition is:

$$\text{Cov}(Y, Z) = E[YZ] - E[Y]E[Z]$$

The equivalence of these definitions can be shown by expanding the first expression:

$$\begin{aligned}
\text{Cov}(Y, Z) &= E[(Y - E[Y])(Z - E[Z])] \\
&= E[YZ - Y \cdot E[Z] - Z \cdot E[Y] + E[Y]E[Z]] \\
&= E[YZ] - E[Y]E[Z] - E[Y]E[Z] + E[Y]E[Z] \\
&= E[YZ] - E[Y]E[Z]
\end{aligned}$$

Covariance measures the direction of the linear relationship between two variables:

- $\text{Cov}(Y, Z) > 0$: Y and Z tend to move in the same direction (positive relationship)
- $\text{Cov}(Y, Z) < 0$: Y and Z tend to move in opposite directions (negative relationship)
- $\text{Cov}(Y, Z) = 0$: Y and Z have no linear relationship

Using our previous results for $E[YZ]$, we can express covariance in terms of conditional expectations:

$$\begin{aligned}
\text{Cov}(Y, Z) &= E[YZ] - E[Y]E[Z] \\
&= E[Z \cdot E[Y|Z]] - E[Y]E[Z]
\end{aligned}$$

This formula is particularly useful when we know the conditional expectation $E[Y|Z]$.

Properties of Covariance

1. **Symmetry**: $\text{Cov}(Y, Z) = \text{Cov}(Z, Y)$

2. **Linearity in each argument**: For constants a, b and random variables Y, Z, W, V:

$$\begin{aligned}&\text{Cov}(aY + Z, bW + V)\\&= ab\text{Cov}(Y, W) + a\text{Cov}(Y, V) + b\text{Cov}(Z, W) + \text{Cov}(Z, V)\end{aligned}$$

3. **Variance as a special case**: $\text{Var}(Y) = \text{Cov}(Y, Y) = E[Y^2] - E[Y]^2$

4. **Independence implies zero covariance**:
   If Y and Z are independent, then $\text{Cov}(Y, Z) = 0$

3.2.2 Covariance and Independence

It's important to note that while independence implies zero covariance, the converse is not generally true: zero covariance does not imply independence.

A classic example of variables that have zero covariance yet are dependent is when Y is a standard normal random variable and $Z = Y^2$. These variables are clearly dependent (knowledge of Y completely determines Z), but:

$$\text{Cov}(Y, Y^2) = E[Y \cdot Y^2] - E[Y]E[Y^2] = E[Y^3] - E[Y]E[Y^2]$$

For a standard normal variable, $E[Y] = 0$ and $E[Y^3] = 0$ (due to symmetry around 0), so:

$$\text{Cov}(Y, Y^2) = 0 - 0 \cdot E[Y^2] = 0$$

This highlights that covariance only captures linear relationships between variables, not all forms of dependence.

3.2.3 Correlation

While covariance measures the direction of association between variables, its magnitude depends on the scales of the variables. The **correlation coefficient** standardizes this measure to be scale-invariant:

Correlation Coefficient

The correlation coefficient between random variables Y and Z is defined as:

$$\text{Corr}(Y, Z) = \rho_{Y,Z} = \frac{\text{Cov}(Y, Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}$$

The correlation coefficient has the following properties:

1. $-1 \leq \rho_{Y,Z} \leq 1$
2. $\rho_{Y,Z} = 1$ implies a perfect positive linear relationship
3. $\rho_{Y,Z} = -1$ implies a perfect negative linear relationship
4. $\rho_{Y,Z} = 0$ implies no linear relationship

3.3 Expected Value Vector and Covariance Matrix

When working with multivariate data, we often need to analyze several random variables simultaneously. Let's consider a k-dimensional random vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)'$, where the prime denotes vector transposition.

Expected Value Vector

The **expected value vector** (or mean vector) of a random vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)'$ is defined as:

$$\boldsymbol{\mu}_Z = E[\boldsymbol{Z}] = (E[Z_1], E[Z_2], \ldots, E[Z_k])'$$

Each component $E[Z_i]$ is calculated according to:

$$E[Z_i] = \int_{-\infty}^{\infty} z_i \, \mathrm{d}F_{Z_i}(z_i)$$

where F_{Z_i} represents the marginal CDF of Z_i.

The expected value vector provides the central location of the multivariate distribution, serving as a natural extension of the univariate expected value.

Covariance Matrix

The **covariance matrix** of a random vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)'$, denoted by $\boldsymbol{\Sigma}_Z$, is defined as:

$$\boldsymbol{\Sigma}_Z = E[(\boldsymbol{Z} - \boldsymbol{\mu}_Z)(\boldsymbol{Z} - \boldsymbol{\mu}_Z)']$$

Expanding this definition, we get a $k \times k$ matrix:

$$\boldsymbol{\Sigma}_Z = \begin{pmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \cdots & \text{Cov}(Z_1, Z_k) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \cdots & \text{Cov}(Z_2, Z_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Z_k, Z_1) & \text{Cov}(Z_k, Z_2) & \cdots & \text{Var}(Z_k) \end{pmatrix}$$

In this matrix:

- Diagonal elements $\Sigma_{ii} = \text{Var}(Z_i)$ represent the variance of each component
- Off-diagonal elements $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$ represent the covariance between components

Properties of the Covariance Matrix

1. **Symmetry**: $\boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Z'$ since $\text{Cov}(Z_i, Z_j) = \text{Cov}(Z_j, Z_i)$
2. **Positive Semi-Definiteness**: For any non-zero vector $\boldsymbol{a} \in \mathbb{R}^k$, $\boldsymbol{a}'\boldsymbol{\Sigma}_Z\boldsymbol{a} \geq 0$
3. **Linear Transformations**: For a matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b}$, then:
   - $E[\boldsymbol{Y}] = \boldsymbol{A}E[\boldsymbol{Z}] + \boldsymbol{b}$
   - $\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}_Z\boldsymbol{A}'$

The positive semi-definiteness of the covariance matrix follows because $\boldsymbol{a}'\boldsymbol{\Sigma}_Z\boldsymbol{a} = \text{Var}(\boldsymbol{a}'\boldsymbol{Z})$, which is the variance of a linear combination of the components of $\boldsymbol{Z}$, and variance is always non-negative.

The **correlation matrix** standardizes the covariance matrix by dividing each covariance by the product of the corresponding standard deviations:

Correlation Matrix

The correlation matrix of a random vector $\boldsymbol{Z}$, denoted by $\boldsymbol{R}_Z$, is defined as:

$$\boldsymbol{R}_Z = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}$$

where $\rho_{ij} = \frac{\text{Cov}(Z_i, Z_j)}{\sqrt{\text{Var}(Z_i)\text{Var}(Z_j)}}$ is the correlation coefficient between Z_i and Z_j.

Mathematically, if $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sqrt{\text{Var}(Z_i)}$, then:

$$\boldsymbol{R}_Z = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}_Z\boldsymbol{D}^{-1}$$

4 Stochastic Convergence

Building on the concepts of the previous sections, we now turn to stochastic convergence which helps us understand the behavior of estimators as sample sizes increase. Stochastic convergence provides the theoretical framework for understanding when and how our estimates approach the true population parameters, which is essential for conducting valid statistical inference in econometric analysis.

4.1 Estimation

4.1.1 Parameters and Estimators

A **parameter** θ is a characteristic or feature of a population distribution. Parameters are typically fixed but unknown quantities that we aim to learn about through sampling and estimation. Examples of parameters include:

- The mean (expected value) μ of a population distribution
- The variance σ^2 of a population distribution
- The coefficients $\boldsymbol{\beta}$ in a regression model
- The correlation ρ between two random variables

An **estimator** $\widehat{\theta}$ is a function of sample data intended to approximate the unknown parameter θ. Since an estimator is a function of random variables (the sample), it is itself a random variable. When we actually compute the estimator from a specific realized sample, we call the resulting value an **estimate**.

For example, the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an estimator for the population mean $\mu = E[Y]$.

4.1.2 Sequences of Random Variables

When we consider the properties of estimators, we often examine what happens as the sample size increases. This leads us to study sequences of random variables.

A **sequence of random variables** $\{W_n\}_{n=1}^{\infty}$ is an ordered collection of random variables indexed by sample size n. For estimators, we are interested in how the sequence $\{\widehat{\theta}_n\}_{n=1}^{\infty}$ behaves as n increases, where $\widehat{\theta}_n$ represents the estimator based on a sample of size n.

The behavior of such sequences as $n \to \infty$ is the focus of asymptotic theory in econometrics. Understanding this behavior allows us to evaluate the properties of estimators in large samples, even when their exact finite-sample distributions are intractable.

4.2 Convergence in Probability

4.2.1 Definition for General Sequences

Convergence in Probability

A sequence of random variables $\{W_n\}_{n=1}^{\infty}$ **converges in probability** to a constant c if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|W_n - c| > \epsilon) = 0$$

Equivalently, this can be expressed as:

$$\lim_{n\to\infty} P(|W_n - c| \leq \epsilon) = 1$$

This is denoted as $W_n \xrightarrow{p} c$.

Intuitively, convergence in probability means that as the sample size n increases, the probability that W_n deviates from c by more than any fixed positive amount ϵ becomes arbitrarily small.

For example, if $W_n \xrightarrow{p} c$, then for any small $\epsilon > 0$ (say, $\epsilon = 0.01$), we can make $P(|W_n - c| > 0.01)$ as small as we want by choosing a sufficiently large sample size n.

4.2.2 Consistency for a Parameter

Applying the concept of convergence in probability to estimators leads to the important property of **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.

Consistency

An estimator $\widehat{\theta}_n$ is **consistent** for the parameter θ if:

$$\widehat{\theta}_n \xrightarrow{p} \theta \quad \text{as} \quad n \to \infty$$

That is, if for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) = 0$$

Consistency is a minimal requirement for a good estimator. It ensures that with a large enough sample, the estimator will be arbitrarily close to the true parameter with high probability. However, consistency alone doesn't tell us how quickly the estimator approaches the parameter as the sample size increases.

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.

Here's the improved subsection on sufficient condition and MSE decomposition:

I've analyzed your updated version in paste.txt, and I can make the proof more concise by directly substituting the variance and bias. Here's a revised version that maintains comprehensiveness while being more concise:

4.2.3 Sufficient Condition for Consistency and the MSE Decomposition

A powerful approach to establishing consistency relies on examining the mean squared error (MSE) of an estimator and applying Markov's inequality.

Markov's Inequality

For any non-negative random variable X and any positive constant a:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Markov's inequality provides an upper bound on the probability that a non-negative random variable exceeds any positive threshold. While this bound may not be tight, it is extremely useful for proving convergence results.

To apply Markov's inequality to establish consistency, we consider the squared deviation between the estimator and the parameter:

$$\begin{aligned} P(|\widehat{\theta}_n - \theta| > \epsilon) &= P((\widehat{\theta}_n - \theta)^2 > \epsilon^2) \\ &\leq \frac{E[(\widehat{\theta}_n - \theta)^2]}{\epsilon^2} \\ &= \frac{MSE(\widehat{\theta}_n)}{\epsilon^2} \end{aligned}$$

This derivation leads directly to a sufficient condition for consistency:

Sufficient Condition for Consistency

Let $\widehat{\theta}_n$ be an estimator for parameter θ. If:

$$\lim_{n\to\infty} MSE(\widehat{\theta}_n) = \lim_{n\to\infty} E[(\widehat{\theta}_n - \theta)^2] = 0$$

Then $\widehat{\theta}_n$ is consistent for θ.

To analyze when this condition holds, we need to understand the components of the MSE. The **bias** of an estimator is defined as:

$$Bias(\widehat{\theta}_n) = E[\widehat{\theta}_n] - \theta$$

The bias measures the systematic deviation of the estimator from the true parameter. An estimator is **unbiased** if $Bias(\widehat{\theta}_n) = 0$ for all sample sizes, and **asymptotically unbiased** if $\lim_{n\to\infty} Bias(\widehat{\theta}_n) = 0$.

The MSE can be decomposed into the sum of the variance and the squared bias. Here's a concise proof:

$$\begin{aligned}
MSE(\widehat{\theta}_n) &= E[(\widehat{\theta}_n - \theta)^2] \\
&= E[(\widehat{\theta}_n - E[\widehat{\theta}_n] + E[\widehat{\theta}_n] - \theta)^2] \\
&= E[(\widehat{\theta}_n - E[\widehat{\theta}_n])^2 + 2(\widehat{\theta}_n - E[\widehat{\theta}_n])(E[\widehat{\theta}_n] - \theta) + (E[\widehat{\theta}_n] - \theta)^2] \\
&= Var(\widehat{\theta}_n) + 2 \cdot 0 \cdot Bias(\widehat{\theta}_n) + [Bias(\widehat{\theta}_n)]^2 \\
&= Var(\widehat{\theta}_n) + [Bias(\widehat{\theta}_n)]^2
\end{aligned}$$

The middle term vanishes because $E[\widehat{\theta}_n - E[\widehat{\theta}_n]] = 0$ by definition.

This fundamental decomposition reveals that estimation error comes from two sources: the variability of the estimator around its expected value (variance) and the systematic deviation of the expected value from the true parameter (bias).

This leads to a practical sufficient condition for consistency:

Practical Sufficient Condition for Consistency

An estimator $\widehat{\theta}_n$ is consistent for θ if both of the following conditions hold as $n \to \infty$:

1. $Bias(\widehat{\theta}_n) \to 0$ (asymptotically unbiased)
2. $Var(\widehat{\theta}_n) \to 0$ (variance approaches zero)

4.2.4 Law of Large Numbers

The **Law of Large Numbers (LLN)** is one of the fundamental results in probability theory and provides a key tool for establishing consistency of many estimators.

Weak Law of Large Numbers (WLLN)

Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with $E[Y_i] = \mu$ and $Var(Y_i) = \sigma^2 < \infty$. Then the sample mean converges in probability to the population mean:

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty$$

The WLLN essentially states that if we take a large enough sample from a population with finite mean, the sample mean will be close to the population mean with high probability.

For the sample mean, we can directly verify the conditions for consistency:

1. $E[\overline{Y}_n] = \mu$ (unbiased for all n)
2. $Var(\overline{Y}_n) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

The LLN extends to functions of sample means as well. If $g(\cdot)$ is a continuous function and $\overline{Y}_n \xrightarrow{p} \mu$, then:

$$g(\overline{Y}_n) \xrightarrow{p} g(\mu)$$

This result, known as the **continuous mapping theorem**, allows us to establish consistency for a wide range of estimators that can be expressed as functions of sample means.

4.2.5 Rate of Convergence

While consistency tells us that an estimator eventually converges to the true parameter, it doesn't indicate how quickly this convergence occurs. The **rate of convergence** provides this information.

Rate of Convergence

A consistent estimator $\widehat{\theta}_n$ converges at rate r_n if:

$$r_n(\widehat{\theta}_n - \theta) = O_p(1)$$

where $O_p(1)$ indicates that the sequence is "bounded in probability."

For many standard estimators, including the sample mean of i.i.d. random variables, the rate of convergence is $r_n = \sqrt{n}$. This means:

$$\sqrt{n}(\widehat{\theta}_n - \theta) = O_p(1)$$

The rate of convergence $r_n = \sqrt{n}$ is often called the "root-n" rate or "parametric" rate of convergence. Estimators with this rate have the property that to halve the average estimation error, we need to quadruple the sample size.

The root mean squared error (RMSE) captures this relationship:

$$RMSE(\widehat{\theta}_n) = \sqrt{E[(\widehat{\theta}_n - \theta)^2]} \approx \frac{C}{\sqrt{n}}$$

where C is a constant that depends on the specific distribution and estimator.

4.3 Convergence in Distribution

While convergence in probability describes how an estimator concentrates around the true parameter, **convergence in distribution** describes the limiting shape of the distribution of the estimator (or a transformation of it).

4.3.1 Limiting Distribution Definition for General Sequences

Convergence in Distribution

A sequence of random variables $\{W_n\}_{n=1}^{\infty}$ **converges in distribution** to a random variable W if:

$$\lim_{n\to\infty} F_{W_n}(x) = F_W(x)$$

for all points x where $F_W(x)$ is continuous. Here, F_{W_n} and F_W are the cumulative distribution functions of W_n and W, respectively.

This is denoted as $W_n \xrightarrow{d} W$.

Unlike convergence in probability, which relates a sequence of random variables to a fixed constant, convergence in distribution relates the sequence to another random variable with a specific distribution.

It's important to note that $W_n \xrightarrow{d} W$ does not mean that the random variables W_n approach W in a pointwise sense. Rather, the distribution function of W_n approaches the distribution function of W.

4.3.2 Consistent Estimator Has Degenerate Limiting Distribution

If an estimator $\widehat{\theta}_n$ is consistent for θ, then:

$$\widehat{\theta}_n \xrightarrow{p} \theta$$

This implies that $\widehat{\theta}_n$ also converges in distribution to the constant θ:

$$\widehat{\theta}_n \xrightarrow{d} \theta$$

However, this limiting distribution is **degenerate** — it places all probability mass at the single point θ. While this confirms consistency, it doesn't provide information about the shape of the sampling distribution for finite n.

4.3.3 Asymptotic Distribution of an Estimator

To obtain a non-degenerate limiting distribution that provides useful information about the sampling variability of a consistent estimator, we typically examine a standardized version of the estimator.

If $\widehat{\theta}_n$ converges to θ at rate r_n, then we study:

$$r_n(\widehat{\theta}_n - \theta)$$

For many estimators with $r_n = \sqrt{n}$, this standardized quantity converges in distribution to a normal random variable:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$$

where V is the asymptotic variance.

The distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ is called the **asymptotic distribution** of the estimator $\widehat{\theta}_n$. For large n, we can approximate:

$$\widehat{\theta}_n \approx N\left(\theta, \frac{V}{n}\right)$$

This approximation is the basis for constructing confidence intervals and conducting hypothesis tests in large samples.

4.3.4 Central Limit Theorem

The **Central Limit Theorem (CLT)** is the key result that establishes the asymptotic normality of many estimators.

Central Limit Theorem (CLT)

Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then:

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Or equivalently:

$$\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

The CLT tells us that the standardized sample mean follows a standard normal distribution in large samples, regardless of the underlying distribution of the individual observations (as long as the variance is finite).

This remarkable result means that we can construct approximate confidence intervals and conduct hypothesis tests for the mean using the normal distribution, even when the population distribution is non-normal, provided the sample size is sufficiently large.

The CLT extends to more complex estimators as well. If an estimator can be expressed as a function of sample means, the **delta method** allows us to derive its asymptotic distribution.

Delta Method

If $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$ and $g(\cdot)$ is a differentiable function with $g'(\theta) \neq 0$, then:

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2 V)$$

4.3.5 The Normal Distribution and Its Properties

Given the central role of the normal distribution in asymptotic theory, it's worthwhile to review its key properties.

Normal Distribution

A random variable X follows a normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if its probability density function (PDF) is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The standard normal distribution, denoted $Z \sim N(0,1)$, has mean 0 and variance 1, with PDF:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

and cumulative distribution function (CDF) $\Phi(z)$.

Key properties of the normal distribution relevant to econometrics include:

1. **Linear combinations**: If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then:

$$aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

2. **Standardization**: If $X \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

3. **Symmetry**: The standard normal distribution is symmetric around zero:

$$\phi(z) = \phi(-z)$$

$$\Phi(z) = 1 - \Phi(-z)$$

4. **Quantiles**: The p-quantile of the standard normal distribution, denoted z_p, satisfies $\Phi(z_p) = p$. Some important quantiles include:

   - $z_{0.975} = 1.96$ (used for 95% confidence intervals)

- $z_{0.995} = 2.58$ (used for 99% confidence intervals)
- $z_{0.9} = 1.28$ (used for 90% confidence intervals)

5. **Multivariate normal distribution**: A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ follows a multivariate normal distribution, denoted $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if every linear combination of its components follows a univariate normal distribution.

The pervasiveness of the normal distribution in asymptotic theory stems from the CLT and related results. Even when the exact finite-sample distribution of an estimator is complex or unknown, its asymptotic normal distribution often provides a good approximation in large samples.