Probability Theory for Econometricians

Sven Otto

March 16, 2025

Table of contents

Welcome

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

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:

- Elementary events:

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

- Half-open events:

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

The **probability function** P assigns values between 0 and 1 to events. For a fair coin toss (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:

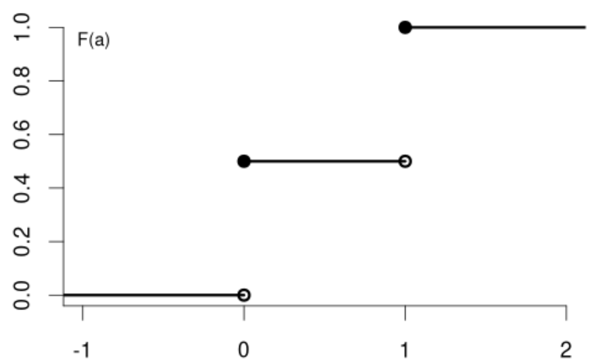


Figure 1.1: CDF of coin (discrete random variable)

The CDF of the variable *education* could be:

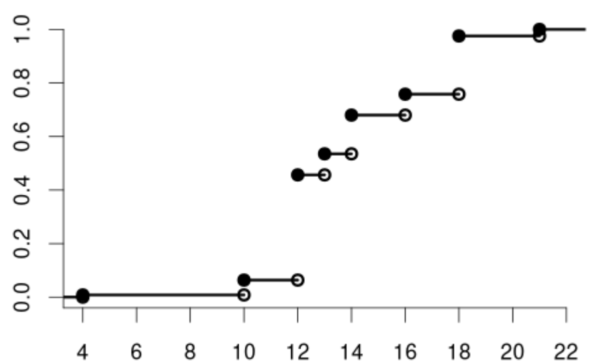


Figure 1.2: CDF of education (discrete random variable)

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

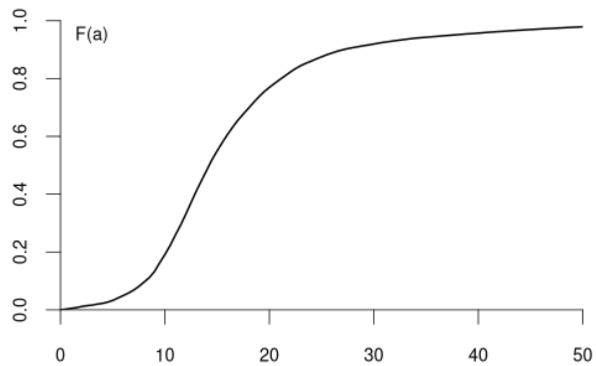


Figure 1.3: CDF of wage (continuous random variable)

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.

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:

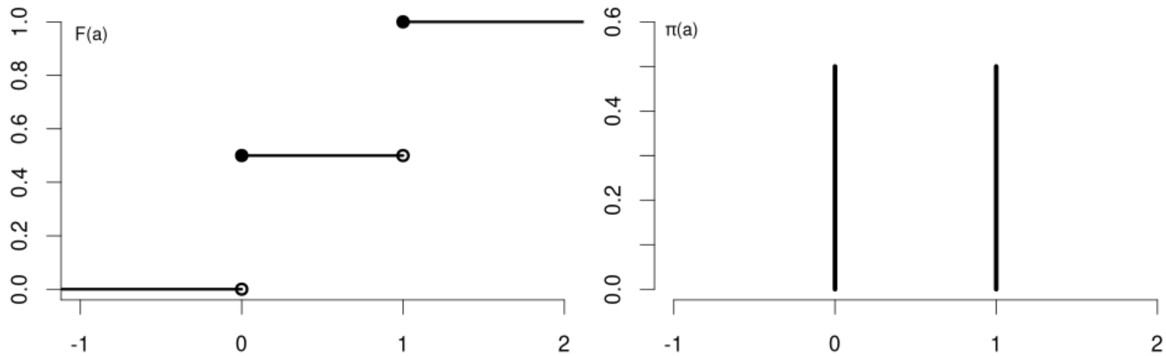
Probability Mass Function (PMF)

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

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

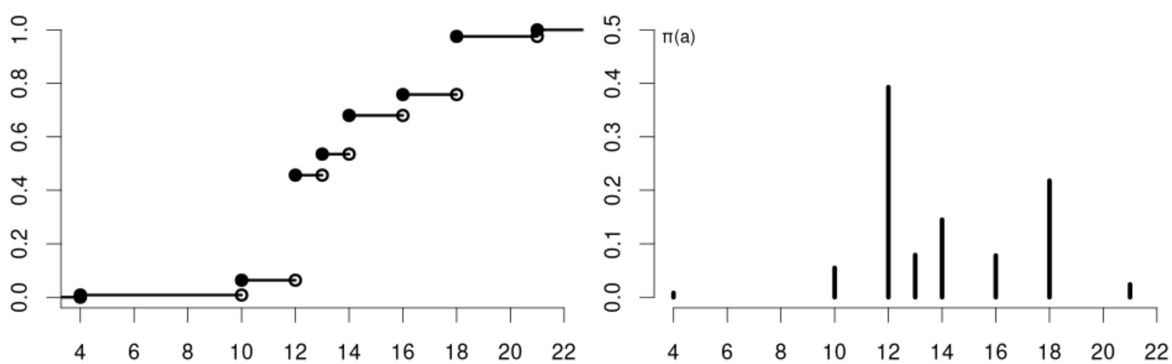


(a) CDF of coin (b) PMF of coin

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

The *education* variable has the following PMF:

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



(a) CDF of education (b) PMF of education

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

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**: $F(b) = \sum_{a \in \mathcal{Y}, a \leq b} \pi(a)$

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:

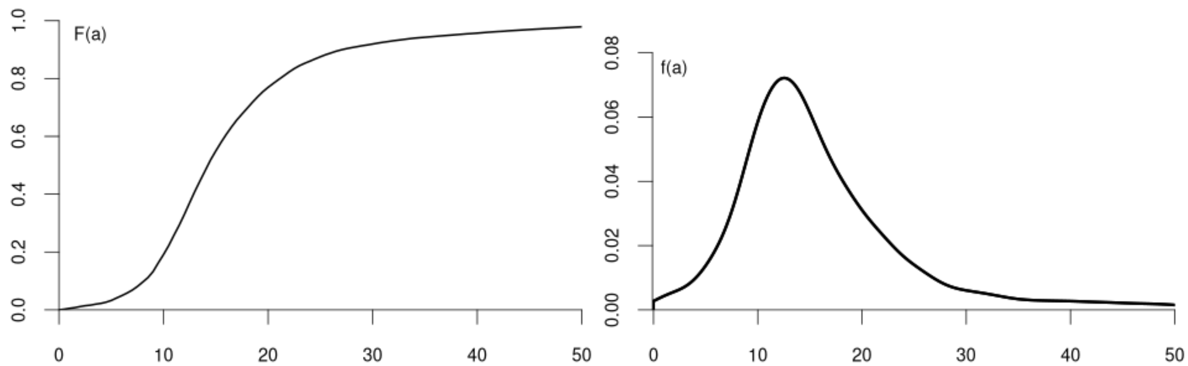
$$F(a) = \int_{-\infty}^{a} f(u) \, \mathrm{d}u$$

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

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 $a \leq b$):



(a) CDF of wage

(b) PDF of wage

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

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

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),

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

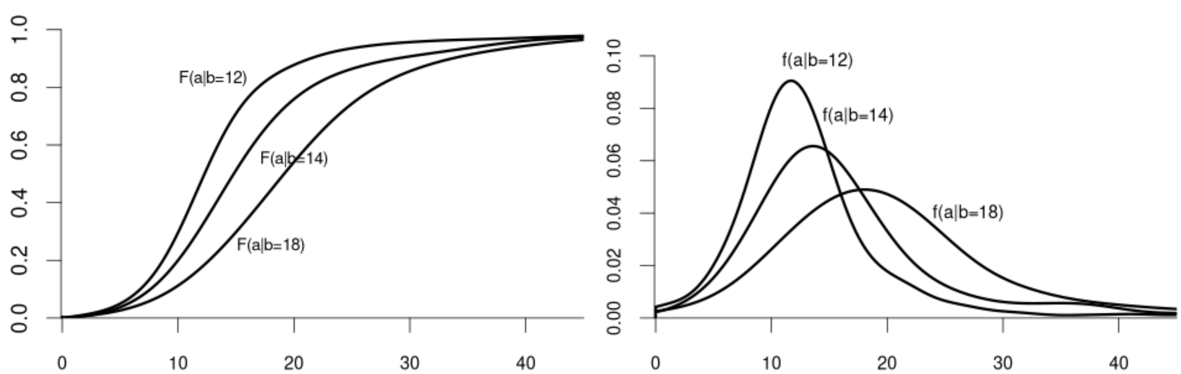
represents the distribution of a random variable Y given that another random variable Z takes a specific value b. It answers the question: "If we know that $Z = b$, what is the distribution of Y?"

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**.

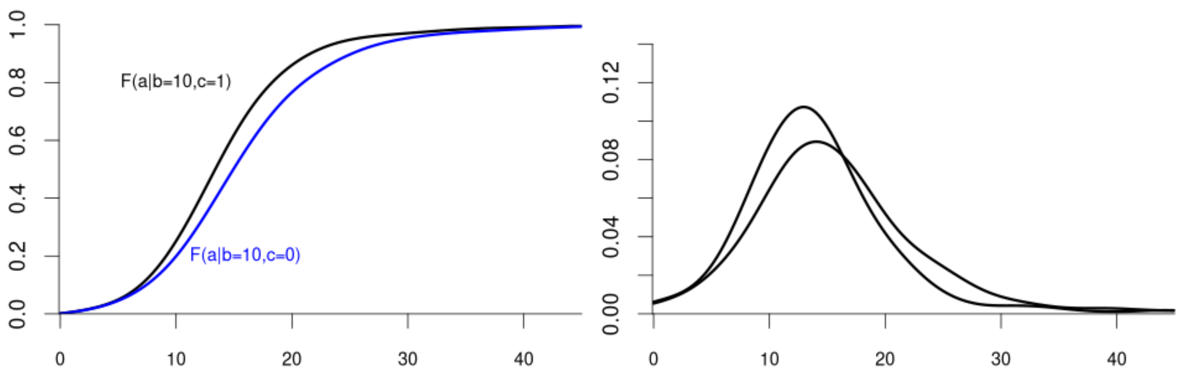
Note that the conditional CDF $F_{Y|Z=b}(a)$ can only be defined for values of b in the support of Z.

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.

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:

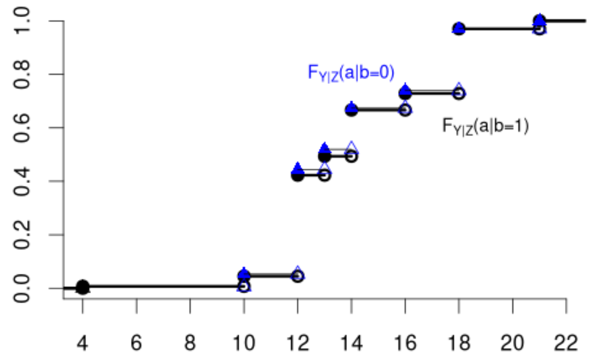
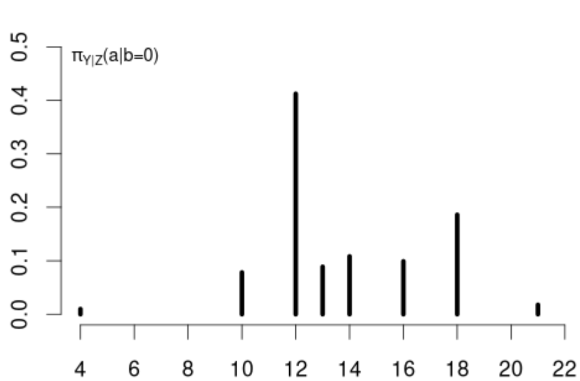
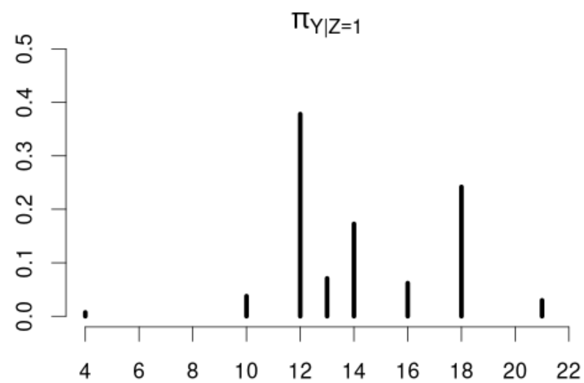


Figure 1.9: Conditional CDFs of education given married

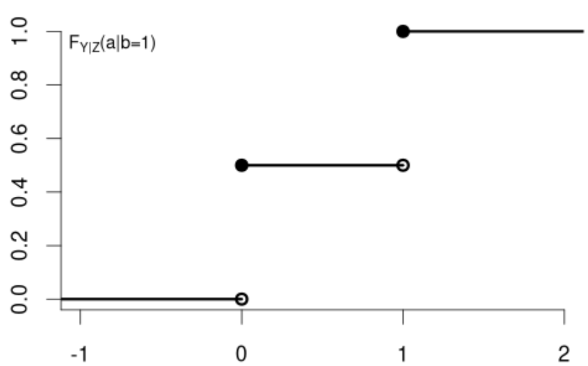


(a) Education given unmarried

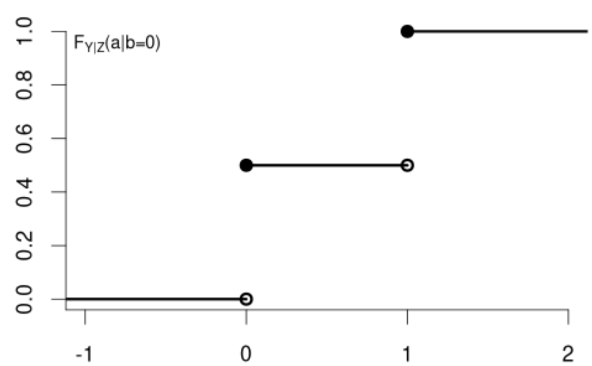


(b) Education given married

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



(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:

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