Conditional probabilities and expectation

Daniel Ahlberg, Stockholm University *April 20, 2022*

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let A and B be two events. When $\mathbb{P}(B) > 0$ we define the *conditional probability* of A given B as

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

We interpret the conditional probability of A given B as the probability that A occurs, given the additional information that B occurs. This interpretation can be justified via the interpretation of probabilities as the relative frequency by which an event occurs when an experiment is repeated many times independently of one another. More precisely, if N_B denotes the number of occurrences of the event B in n independent trials, and N_{AB} denotes the number of occurrences of both A and B in the same trials, then, for large values of n,

$$\mathbb{P}(A|B) = \mathbb{P}(A \cap B) \cdot \frac{1}{\mathbb{P}(B)} \approx \frac{N_{AB}}{n} \cdot \frac{n}{N_B} = \frac{N_{AB}}{N_B}.$$

That is, the conditional probability is roughly equal to the relative frequency at which also A occurs among all trials that result in B.

Random variables are central objects in probability theory, and being able to condition on the outcome of a random variable is therefore desirable. Suppose that X and Y are random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.[1] If X is discrete, then it is straightforward to condition on its outcome. Then, if $\{X = x\}$ is an event of positive probability, simply apply (1).

Example 1. Roll two fair dice. Let X denote the sum of the dice and let Y denote the lowest of their values. Then, using (1),

$$\mathbb{P}(Y = 2|X = 7) = \frac{\mathbb{P}(X = 7, Y = 2)}{\mathbb{P}(X = 7)} = \frac{2/36}{6/36} = 1/3,$$

since there are six outcomes for which the sum is 7, of which two correspond to the lowest value being 2. □

On the other hand, if X is continuous, then conditioning with respect to $\{X = x\}$ via (1) does not make sense, as it would involve division by zero. As we shall see, conditioning on a discrete variable is mostly straightforward, but conditioning on a continuous variable will require a great deal of care. While a full treatment of conditioning would require elements from measure theory, we shall in this text satisfy ourselves with a glimpse of the theory, which should be suficient to justify the definitions that we make.[2]

[1] It is important that X and Y are defined on the *same* probability space as we want to address their joint occurrence.

[2] A rigorous treatment of conditioning on continuous random variables was in fact one of the reasons that led Kolmogorov to define probability theory on the basis of measure theory.

Conditional probabilities as probability measures

Given pair $(\Omega, \mathcal{F})$, consisting of a sample space Ω and a collection of events $\mathcal{F}$, there are many possible choices of a probability measure $\mathbb{P} : \mathcal{F} \to [0,1]$ to complete the pair into a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Example 2. Consider the act of tossing a coin. The possible outcomes of this experiment can be described by the sample space $\Omega = \{H, T\}$, and a suitable choice for $\mathcal{F}$ is the collection of all subsets of Ω, i.e. $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. For every $p \in (0,1)$ we obtain a probability measure $\mathbb{P}_p$ on $(\Omega, \mathcal{F})$ where p corresponds to the probability of the coin coming up 'heads'. □

The following result shows that, given a probability measure on $(\Omega, \mathcal{F})$ and an event of positive probability, the conditional probability with respect to this event is nothing but another probability measure on $(\Omega, \mathcal{F})$.

Proposition 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and B an event with $\mathbb{P}(B) > 0$. Then the mapping $\mathbb{P}(\,\cdot\,|B) : \mathcal{F} \to [0,1]$ defined by $A \mapsto \mathbb{P}(A|B)$ is a probability measure.*

Proof. We need to verify that $\mathbb{P}(\,\cdot\,|B)$ satisfies the three probability axioms

(i) $\mathbb{P}(A|B) \geq 0$ for all $A \in \mathcal{F}$;
(ii) $\mathbb{P}(\Omega|B) = 1$;
(iii) $\mathbb{P}\big(\bigcup_{k\geq 1} A_k \big| B\big) = \sum_{k\geq 1} \mathbb{P}(A_k|B)$ for pairwise disjoint events $A_1, A_2, \ldots$.

These are straightforward consequences of (1). (Please check!) □

Recall that a random variable by definition is a function $X : \Omega \to \mathbb{R}$, which for each outcome of the sample space takes a real value. Note that a random variable is defined without reference to the probability measure associated to the pair $(\Omega, \mathcal{F})$.[3] However, the *distribution* of the random variable is determined by the probability measure $\mathbb{P}$ we choose for $(\Omega, \mathcal{F})$. For instance, in the coin tossing example above, let X be the random variable that equals 1 if the coin turns up heads and 0 otherwise. Then X is Bernoulli distributed, but its parameter depends on the bias p of the coin.

Example 3. Roll three fair dice. Let Y denote the number of sixes, and let B be the event that the first die turns up six. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the corresponding probability space. According to the proposition, also $\mathbb{P}(\,\cdot\,|B)$ is a probability measure, so that $(\Omega, \mathcal{F}, \mathbb{P}(\,\cdot\,|B))$ is again a probability space.

The distribution of the random variable Y with respect to $\mathbb{P}$ is binomial with parameters 3 and 1/6. The distribution of $Y - 1$ with respect to $\mathbb{P}(\,\cdot\,|B)$ is binomial with parameters 2 and 1/6, since on the event B we know that Y is at least 1. □

[3] Formally, $\mathcal{F}$ is assumed to be a so-called *σ-algebra* and, to be a random variable, X is required to satisfy $\{X \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$ in order for us to be allowed to measure the probability of every event of this form.

Conditional distributions: discrete vectors

We saw above that conditioning gives rise to new probability measures, and thus changes the distribution of a random variable. In particular, we are interested in conditioning on the outcome of some other random variable. In the discrete setting, the theory is mostly straightforward, so we treat the discrete case in detail first.

Definition 2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and (X, Y) a discrete random vector. For every $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$, the function $\mathbb{P}(\cdot\,|X = x)$ defines a probability measure. The distribution of Y with respect to this measure is referred to as the *conditional distribution* of Y given $X = x$.

As in the unconditional case, we refer to the following functions as the *conditional distribution function* and *conditional mass function* of Y given $X = x$: For all $y \in \mathbb{R}$,

$$F_{Y|X=x}(y) := \mathbb{P}(Y \leq y|X = x) \quad \text{and} \quad p_{Y|X=x}(y) := \mathbb{P}(Y = y|X = x).$$

We note, by (1), that the conditional mass function can be expressed as

$$p_{Y|X=x}(y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{X,Y(x,y)}}{p_X(x)}, \tag{2}$$

so that in the case (as assumed here) that (X, Y) is a discrete random vector, we obtain

$$\sum_{y\in\mathbb{R}} p_{Y|X=x}(y) = \frac{\sum_{y\in\mathbb{R}} p_{X,Y}(x, y)}{p_X(x)} = \frac{p_X(x)}{p_X(x)} = 1.$$

That is, also the conditional distribution of Y given $X = x$ is discrete, and the distribution function can be expressed in terms of the mass function:

$$F_{Y|X=x}(y) = \sum_{z\leq y} p_{Y|X=x}(y) \quad \text{for all } y \in \mathbb{R}.$$

Since also the conditional distribution of Y given $X = x$ is discrete, it is straightforward to compute its expected value.

Definition 3. Let (X, Y) be a discrete random vector. For every $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ we define the *conditional expectation* of Y given $X = x$ as

$$\mathbb{E}[Y|X = x] := \sum_{y\in\mathbb{R}} y \cdot p_{Y|X=x}(y),$$

given that the sum is well defined.

Note that just as $\mathbb{E}[Y]$ is the expected value of the random variable Y with respect to the measure $\mathbb{P}$, the conditional expectation $\mathbb{E}[Y|X = x]$ is the expectation of Y with respect to the measure $\mathbb{P}(\cdot\,|X = x)$.

Example 4. Roll two fair dice. Let X denote the sum of the dice and let Y denote the lowest of their values. Then (X, Y) is a discrete random vector. The conditional distribution of Y given $X = 7$ satisfies

$$p_{Y|X=7}(y) = \frac{\mathbb{P}(X = 7, Y = y)}{\mathbb{P}(X = 7)} = \frac{2/36}{6/36} = 1/3$$

for $y = 1, 2, 3$ and is zero otherwise. The conditional expectation of Y given $X = 7$ is thus

$$\mathbb{E}[Y|X = 7] = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} = 2.$$

□

The following elementary properties of conditional expectations are useful.

Proposition 4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and X, Y and Y' discrete random variables. Let $a, b \in \mathbb{R}$ be constants and $g : \mathbb{R}^2 \to \mathbb{R}$ some function. Then, for every $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$, we have*

(i) $\mathbb{E}[c|X = x] = c$;
(ii) $\mathbb{E}[aY + bY'|X = x] = a\,\mathbb{E}[Y|X = x] + b\,\mathbb{E}[Y'|X = x]$;
(iii) $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$ *if X and Y are independent;*
(iv) $\mathbb{E}[g(X, Y)|X = x] = \mathbb{E}[g(x, Y)|X = x]$.

Proof. Properties (i)-(ii) hold for 'regular' expectations, and thus also for the conditional expectation as it is simply the expectation with respect to a different probability measure. Hence, properties (i)-(ii) need no proofs, but properties (iii)-(iv) do.

In order to prove (iii), suppose that X and Y are independent. Then

$$p_{Y|X=x}(y) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y).$$

That is, the distributions of Y with respect to $\mathbb{P}$ and $\mathbb{P}(\,\cdot\,|X = x)$ are the same, so also the expectations coincide. This proves (iii).

Finally, in order to prove (iv), note that the probability that the vector (X, Y) attains the value (z, y) with respect to $\mathbb{P}(\,\cdot\,|X = x)$ equals

$$\mathbb{P}(X = z, Y = y|X = x) = \frac{\mathbb{P}(X = z, X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)} = p_{Y|X=x}(y)$$

if $z = x$ and is otherwise zero. In particular, we obtain

$$\mathbb{E}[g(X, Y)|X = x] = \sum_{y \in \mathbb{R}} g(x, y) \cdot p_{Y|X=x}(y) = \mathbb{E}[g(x, Y)|X = x],$$

as required. □

Conditional distributions: continuous vectors

We proceed to define conditional distributions for continuous random vectors (X, Y). The continuous case is problematic, as we cannot directly condition on the event $\{X = x\}$, as it is an event which occurs with probability zero. To circumvent this problem, we shall work directly with the density function. Since density functions can be thought of as a continuum analogue to mass functions, a natural candidate for a conditional density may be obtained from (2). At this point we provide the relevant definitions, and save their justification for later.

Definition 5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and (X, Y) a continuous random vector. Let $f_{X,Y}$ be a density function for (X, Y) and let $f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ be the marginal density of X. For $x \in \mathbb{R}$ such that $f_X(x) > 0$ we define the *conditional distribution* of Y given $X = x$ as the continuous distribution given by the density function

$$f_{Y|X=x}(y) := \frac{f_{X,Y}(x, y)}{f_X(x)} \quad \text{for } y \in \mathbb{R}. \tag{3}$$

We note that the expression in (3) integrates to 1 for every $x \in \mathbb{R}$ such that $f_X(x) > 0$, so the expression in (3) defines indeed a probability distribution. We note also a distinction between the discrete and the continuous settings: In the discrete setting, the conditional distribution was defined as the distribution of our variable with respect to a different probability measure. In the continuous setting we have simply defined a new probability distribution without connection to the original random variable.

By definition, the conditional distribution is a continuous distribution, and it is straightforward to define its expectation.

Definition 6. Let (X, Y) be a continuous random vector. For all $x \in \mathbb{R}$ such that $f_X(x) > 0$ we define the *conditional expectation* of Y given $X = x$ as

$$\mathbb{E}[Y|X = x] := \int_{-\infty}^{\infty} y \cdot f_{Y|X=x}(y)\, dy,$$

given that the integral is well defined.

As in the discrete setting, we have the following elementary properties.

Proposition 7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (X, Y, Y') a continuous random vector. Let $a, b \in \mathbb{R}$ be constants and $g : \mathbb{R}^2 \to \mathbb{R}$ some function. Then, for every $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ we have*

(i) $\mathbb{E}[c|X = x] = c$;
(ii) $\mathbb{E}[aY + bY'|X = x] = a\,\mathbb{E}[Y|X = x] + b\,\mathbb{E}[Y'|X = x]$;
(iii) $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$ *if X and Y are independent;*
(iv) $\mathbb{E}[g(X, Y)|X = x] = \mathbb{E}[g(x, Y)|X = x]$.

Proof. Proving these properties is left as an exercise. Since we no longer may interpret $\mathbb{P}(\,\cdot\,|X = x)$ as a probability measure we need to derive all properties from the definition of the continuous distribution. □

Conditional distributions: justification

Conditioning on a discrete random variable is straightforward, since conditioning on an event of positive probability is well defined. When X is discrete, the conditional probability $\mathbb{P}(\,\cdot\,|X = x)$ defines a new probability measure, and the conditional distribution and expectation of some random variable Y (defined on the same probability space) is simply the law and expectation with respect to this conditional measure.

Conditioning on a continuous random variable is more subtle, since it is not clear how to condition on an event of probability zero. In the case of continuous vectors (X, Y), we chose a different approach. Instead of dealing with the difficulty of defining a probability measure $\mathbb{P}(\,\cdot\,|X = x)$, we chose to define the conditional distribution and expectation of Y by means of a density function. We shall next motivate why this is a reasonable approach.

Let X a continuous random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with density function f_X. Given $\Delta > 0$, let $x_k := k \cdot \Delta$ and set $I_k := (x_{k-1}, x_k]$. Recall that the (Riemann) integral is defined as the limit of a sum as the mesh size goes to zero. That is,

$$\mathbb{P}\big(X \in (a, b]\big) = \int_a^b f_X(x)\, dx = \lim_{\Delta \to 0} \sum_k f_X(x_k) \cdot \Delta, \tag{4}$$

where the sum ranges over all k such that $x_k \in (a, b]$. That is, to compute probabilities involving X we essentially discretise the integral and sum the contributions of X over small non-zero intervals. While X has zero probability to equal any given point in the interval $(a, b]$, the probability that X belongs to a small interval is typically non-zero.

Let A be some event. Motivated by the discretisation procedure above, it would be reasonable to try to circumvent the problem of division by zero when conditioning on the event $\{X = x\}$ by conditioning on the non-zero event $\{X \in (x - \Delta, x]\}$. That is, it would be reasonable to identify the conditional probability with respect to $X = x$ with the limit

$$\mathbb{P}(A|X = x) = \lim_{\Delta \to 0} \mathbb{P}(A|X \in (x - \Delta, x]), \tag{5}$$

given that the limit exists. However, making sense of such limits for *all* $x \in \mathbb{R}$ in a consistent manner is problematic, as the following example shows.

Example 5. Let X be a continuous random variable and let $A = \{X \geq 0\}$. Note that when defined as in (5) we obtain $\mathbb{P}(A|X = 0) = 0$. If we were to define $\mathbb{P}(A|X = 0)$ as the limit of $\mathbb{P}(A|X \in [0, \Delta))$, then we would obtain $\mathbb{P}(A|X = 0) = 1$. That is, the conditional probability could depend on the limiting procedure by which it is defined. □

Suppose, however, that we believe that (5) can be made sense of for *most* $x \in \mathbb{R}$. Moreover, note that for small values of $\Delta > 0$ we have

$$\mathbb{P}\big(A, X \in (a, b]\big) \approx \sum_k \mathbb{P}(A, X \in I_k) = \sum_k \mathbb{P}(A|X \in I_k)\mathbb{P}(X \in I_k),$$

where the sum goes over all k such that $x_k \in (a, b]$. Then, approximating $\mathbb{P}(X \in I_k)$ by $f_X(x_k) \cdot \Delta$ and sending $\Delta \to 0$ would, as in (4), give

$$\mathbb{P}\big(A, X \in (a, b]\big) = \int_a^b \mathbb{P}(A|X = x) \cdot f_X(x)\, dx \quad \text{for all } a < b. \tag{6}$$

In particular, the conditional probability (if well defined) should satisfy the following continuous version of the law of total probability

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A|X = x) \cdot f_X(x)\, dx.$$

As we saw in the example, making sense of (5) in a general setting involves some problems. The standard approach is to instead define the conditional probability of A given X as *any* function $x \mapsto \mathbb{P}(A|X = x)$ that satisfies (6). Just as the density function of a continuous random variable is not unique, (6) does not define a function uniquely.[4] However, it is possible to show that any two functions that satisfy (6) have to coincide in *almost all* points x.

Let us finally verify that our choice in (3) is indeed compatible with the above discussion. For a continuous random vector (X, Y), the condition in (6) is equivalent to conditional distribution of Y given $X = x$ satisfying

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} F_{Y|X=u}(y) \cdot f_X(u)\, du \quad \text{for all } x, y \in \mathbb{R}. \tag{7}$$

Inserting our choice in (3) into the right-hand side of (7) leaves us with

$$\begin{aligned}\int_{-\infty}^{x} F_{Y|X=u}(y) \cdot f_X(u)\, du &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{Y|X=u}(v) f_X(u)\, du dv \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)\, du dv,\end{aligned}$$

which is equal to $F_{X,Y}(x, y)$. That is, our choice in (3) is compatible with (7), and our definition of conditional distribution for continuous vectors is justified.

Conditional expectation as a random variable

Conditioning is a useful technique which in more complex settings may simplify the calculation of probabilities and expectations. Suppose first that (X, Y) is a discrete random vector and set $S = \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$. We may define a function $h : S \to \mathbb{R}$ where $h(x) := \mathbb{E}[Y|X = x]$. Since X is supported on S, also $h(X)$ defines a random variable.

Suppose instead that (X, Y) is a continuous random vector and set $S = \{x \in \mathbb{R} : f_X(x) > 0\}$. We may then define the function $h : S \to \mathbb{R}$ where $h(x) := \mathbb{E}[Y|X = x]$ as before. Since X is again supported on S, again $h(X)$ defines a random variable.

We often write $E[Y|X]$ for the random variable $h(X)$.

[4] Modifying its value in a few points does not change the value of the integral.

Theorem 8. *Let (X, Y) be a random vector satisfying $\mathbb{E}[|Y|] < \infty$. Then*

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y].$$

Proof. We prove the theorem in the case that (X, Y) is a discrete random vector; for a proof of the continuous case see [1, p. 34].

By definition of expectation and the function h we have

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[h(X)] = \sum_{x\in\mathbb{R}} h(x)\cdot p_X(x) = \sum_{x\in\mathbb{R}} \mathbb{E}[Y|X=x]\cdot p_X(x).$$

By definition of conditional distribution and expectation, the above is equals

$$\sum_{x\in\mathbb{R}}\left(\sum_{y\in\mathbb{R}} y\cdot p_{Y|X=x}(y)\right)\cdot p_X(x) = \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} y\cdot p_{X,Y}(x,y) = \sum_{y\in\mathbb{R}} y\cdot p_Y(y),$$

which we recognise as $\mathbb{E}[Y]$. Note above that we in the second-to-last equality have interchanged the order of summation, which is allowed under the assumption that $\mathbb{E}[|Y|] < \infty$. □

The following example illustrates how conditioning can be used to break up the computation of a probability or expectation, based on the outcome of some auxiliary random variable.

Example 6. A stick of length one is cut in two at a point chosen uniformly at random. The remainder of the stick is once more cut in two at a uniformly chosen point. We shall let Y denote that length of the resulting stick, and aim to compute $\mathbb{E}[Y]$.

Let X denote the first cut point. By assumption, X is uniformly distributed on the interval $[0, 1]$, and the conditional distribution of Y given $X = x$ is the uniform distribution on the interval $[0, x]$. By Theorem 8 and the definition of expectation

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right] = \int_{-\infty}^{-\infty} \mathbb{E}[Y|X=x]\cdot f_X(x)\,dx.$$

For $x \in [0, 1]$ the conditional expectation equals

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y\cdot f_{Y|X=x}(y)\,dy = \int_0^x y\cdot\frac{1}{x}\,dy = \left[\frac{y^2}{2x}\right]_0^x = \frac{x}{2}.$$

Putting the two together gives us that

$$\mathbb{E}[Y] = \int_0^1 \frac{x}{2}\cdot f_X(x)\,dx = \int_0^1 \frac{x}{2}\,dx = \left[\frac{x^2}{4}\right]_0^1 = \frac{1}{4}.$$

Note that we, using Theorem 8, have computed the expectation of Y without first deriving its density function. □

The above discussion justifies an interpretation of the conditional expectation $\mathbb{E}[Y|X]$ as a random variable itself. Using properties of conditional expectation derived above, we deduce the following additional properties of this random variable.

Theorem 9. *Let (X, Y) be a random vector and $g : \mathbb{R} \to \mathbb{R}$ a function. Then*

(i) $\mathbb{E}[g(X)Y|X] = g(X)\,\mathbb{E}[Y|X]$*;*
(ii) $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ *if X and Y are independent.*

Proof. The statements follow from the properties of conditional expectation established in Propositions 4 and 7. By definition $\mathbb{E}[g(X)Y|X]$ is a random variable which on the event $\{X = x\}$ takes the value $\mathbb{E}[g(X)Y|X = x]$. By properties (iv) and (ii) of the propositions, we find that

$$\mathbb{E}[g(X)Y|X = x] = \mathbb{E}[g(x)Y|X = x] = g(x)\,\mathbb{E}[Y|X = x].$$

Since also $g(X)\,\mathbb{E}[Y|X]$ is a random variable which on the event $\{X = x\}$ takes the same value, the two random variables are equal (almost surely). This proves the first statement.

For the second statement, note that if X and Y are independent, then by property (iii) of the propositions the random variable $\mathbb{E}[Y|X]$ takes the value $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$ on the event $\{X = x\}$. Hence $\mathbb{E}[Y|X] = \mathbb{E}[Y]$. □

Conditional probabilities and conditional distributions

Conditional distributions describe the distribution of a random variable given the outcome of another. At occasions we are interested in the distribution of a random variable Y given an event of the form $\{X \in I\}$. If the event occurs with positive probability, then, also in the case that (X, Y) is continuous, the conditional probability $\mathbb{P}(Y \leq y|X \in I)$ should be understood in the sense of (1). This raises the question how conditional probabilities of this form relate to the conditional distribution of Y given $X = x$.

Suppose first that (X, Y) is a discrete random vector. From (1) we obtain

$$\mathbb{P}(Y = y|X \in I) = \frac{\sum_{x \in I} \mathbb{P}(X = x, Y = y)}{\mathbb{P}(X \in I)} = \frac{1}{\mathbb{P}(X \in I)} \sum_{x \in I} p_{X,Y}(x, y).$$

The above expression sums up to 1, so that Y is again a discrete random variable with respect to $\mathbb{P}(\,\cdot\,|X \in I)$.

Suppose next that (X, Y) is continuous. Then, using (1),

$$\mathbb{P}(Y \leq y|X \in I) = \frac{\mathbb{P}(X \in I, Y \leq y)}{\mathbb{P}(X \in I)} = \frac{1}{\mathbb{P}(X \in I)} \int_I \int_{-\infty}^{y} f_{X,Y}(u, v)\, du dv.$$

That is, the conditional distribution of Y given $X \in I$ is the continuous distribution with density function given by

$$f_{Y|X \in I}(y) := \frac{1}{\mathbb{P}(X \in I)} \int_I f_{X,Y}(x, y)\, dx \quad \text{for } y \in \mathbb{R}. \tag{8}$$

By definition of conditional distribution, we may rewrite the expression in (8) to obtain

$$f_{Y|X\in I}(y) = \frac{1}{\mathbb{P}(X \in I)} \int_I f_{Y|X=x}(y) \cdot f_X(x)\, dx.$$

We see that the density of the conditional distribution of Y given $X \in I$ can be interpreted as the average of the conditional density for Y given $X = x$ over the set I.

Having determined the distribution of Y with respect to the measure $\mathbb{P}(\,\cdot\,|X \in I)$, it is straightforward to compute its expectation. In both the discrete and continuous case the expectation satisfies the following formula.

Proposition 10. *Let (X, Y) be a random vector and let I be a set such that $\mathbb{P}(X \in I) > 0$. Then,*

$$\mathbb{E}[Y|X \in I] = \frac{\mathbb{E}\left[Y \cdot \mathbf{1}_{\{X\in I\}}\right]}{\mathbb{P}(X \in I)},$$

where $\mathbf{1}_{\{X\in I\}}$ is the random variable that equals 1 if $X \in I$ and 0 otherwise.

Proof. We consider the case when (X, Y) is continuous, and leave the case of discrete vectors as an exercise. By definition of expectation, we have that

$$\mathbb{E}[Y|X \in I] = \int_{-\infty}^{\infty} y \cdot f_{Y|X\in I}(y)\, dy = \frac{1}{\mathbb{P}(X \in I)} \int_I \int_{-\infty}^{\infty} y \cdot f_{X,Y}(x, y)\, dxdy.$$

On the other hand, we have

$$\mathbb{E}\left[Y \cdot \mathbf{1}_{\{X\in I\}}\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \cdot \mathbf{1}_{\{x\in I\}} \cdot f_{X,Y}(x, y)\, dxdy,$$

which coincides with the above integral. □

Epilogue

We have in this texted looked closer at the elementary treatment of conditional distributions and expectation. We have considered the cases when (X, Y) is a discrete random vector and a continuous random vector in detail. However, the theory extends to cover other cases too. For instance, for any random vector (X, Y), as long as $X = x$ occurs with positive probability, the conditional distribution of Y given $X = x$ is again well defined. The conditional distribution is simply the distribution of Y with respect to the measure $\mathbb{P}(\,\cdot\,|X = x)$. Moreover, the conditional expectation is the expectation of Y with respect to this measure. In the case that the conditional distribution is continuous, with density $g_{Y|X=x}$ say, we have, by the usual definition of expectation, that

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y \cdot g_{Y|X=x}(y)\, dy.$$

In the case that (X, Y) is a random vector and X is continuous, then the conditional distribution of Y given $X = x$ should again satisfy (6) and (7).

Of course, solving these equations is non-trivial. However, given the distribution of X and the conditional distribution of Y given $X = x$, these equations give us an expression for the joint distribution of the vector (X, Y). Replacing probability by expectation in (6) shows that the expectation of Y should likewise satisfy the formula

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] \cdot f_X(x)\, dx. \tag{9}$$

Note that (9) is analogous to Theorem 8. A thorough treatment of the topic is able to prove these statements rigorously, but that lies beyond the scope of this text.

A more formal treatment of conditional distributions and expectations requires the use of measure theory, as presented e.g. in [2]. In the case that X is continuous, conditional probabilities and expectation with respect to X should satisfy (6) and (9) respectively. As mentioned, there is not a unique solution to these equations. However, any two solutions will coincide for *almost every* outcome of the variable X. This favours the view of conditional probabilities $\mathbb{P}(A|X)$ and expectations $\mathbb{E}[Y|X]$ as random variables. In the discrete setting this is equivalent to the elementary treatment considered here. In the continuous setting this makes sense since we have no reason to insist on a precise definition of these objects for *every* outcome of the variable X, since any specific outcome occurs with probability zero.

References

[1] Allan Gut. *An intermediate course in probability.* 2nd edition, Springer, 2009.

[2] David Williams. *Probability with martingales.* Cambridge university press, 1991.