Anna Janicka

Probability Calculus 2021/2022
Lecture 10

1. Conditional Expected Value

When dealing with conditional probability, we have seen how to recalculate our assessment of the probability of an event given additional knowledge that we had about the results of the experiment. We did this by "rescaling" the initial probability within Ω to a smaller sample space, within the conditional event. Now that we have random variables defined over the sample spaces, we may wish to determine what is the effect that additional knowledge may have on our assessment of the distribution of a random variable; in most cases, we will be interested in the *average* effect, i.e. in the expected value of the random variable of interest under the condition (described by a different random variable). For example, if we wish to assess the expected value of the sum of points obtained in two dice rolls, we expect an average value of 7; if, however, we knew that in the first roll we obtained a six, our assessment should be modified – now we intuitively expect that, on average, we will obtain a total of 9.5. Similarly, when drawing a point randomly from a unit square, we expect that the product of the two coefficients of the point will be equal to $\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$; if, however, we knew that $X=\frac{1}{5}$, we would intuitively expect that $\mathbb{E}XY=\frac{1}{5}\cdot\frac{1}{2}=\frac{1}{10}$.

These considerations lead to the definition of a conditional expected value, which we will define only in two cases: for a discrete distribution and for a continuous distribution. The definition in the discrete case is intuitive and strictly corresponds to the definition of conditional probability. If we wish to define probability conditional on a discrete variable X being equal to x (given that it makes sense, i.e. $\mathbb{P}(X=x)>0$), we can treat the new distribution as a distribution resulting from assuming probability conditional on the event $\{X=x\}$. Then we will be in a position to define the conditional expected value as an expected value calculated using this conditional probability:

Definition 1. *Let (X,Y) be a discrete random vector such that $\mathbb{E}Y$ exists. For any $x\in\mathbb{R}$ such that $\mathbb{P}(X=x)>0$, we define the* **conditional expected value of variable Y given** *$X=x$ as the expected value of a random variable with distribution*

$$\mu(A)=\mathbb{P}(Y\in A|X=x).$$

That is, if $S_x=\{y\in\mathbb{R}:\mathbb{P}(X=x,Y=y)>0\}$, we have

$$\mathbb{E}(Y|X=x)=\sum_{y\in S_x} y\mathbb{P}(Y=y|X=x).$$

Examples:

(1) We toss a coin twice. Let X be the number of heads in two tosses, and let Y be equal to 1 if we obtained a head in the first toss and 0 otherwise. We have that

$X\backslash Y$	0	1	m. X
0	$\frac{1}{4}$	0	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
2	0	$\frac{1}{4}$	$\frac{1}{4}$
m. Y	$\frac{1}{2}$	$\frac{1}{2}$	

We have:

$$\mathbb{P}(X=0|Y=0)=\frac{1}{2},\mathbb{P}(X=1|Y=0)=\frac{1}{2},\mathbb{P}(X=2|Y=0)=0,$$

so

$$\mathbb{E}(X|Y=0)=0\cdot\frac{1}{2}+1\cdot\frac{1}{2}+2\cdot 0=\frac{1}{2}.$$

Similarly,

$$\mathbb{E}(X|Y=1)=0\cdot 0+1\cdot\frac{1}{2}+2\cdot\frac{1}{2}=\frac{3}{2}.$$

(2) If Y is a function of X, i.e. $Y = f(X)$, then we have that the set S_x consists only of one point: $y = f(x)$, which means that $\mathbb{E}(Y|X = x) = f(x) \cdot \mathbb{P}(Y = f(x)|X = x) = f(x) \cdot 1 = f(x)$.

We may also wish to calculate the conditional expected value of a function of variable Y instead of Y; in which case, we may use the following theorem:

Theorem 1. *Let (X, Y) be a discrete random vector, and $\varphi : \mathbb{R} \to \mathbb{R}$ a Borel function such that $\mathbb{E}|\varphi(Y)| < \infty$. We then have that for any x such that $\mathbb{P}(X = x) > 0$:*

$$\mathbb{E}(\varphi(Y)|X = x) = \sum_{y \in S_x} \varphi(y)\mathbb{P}(Y = y|X = x),$$

where $S_x = \{y \in \mathbb{R} : \mathbb{P}(X = x, Y = y) > 0\}$.

The case of continuous random vectors is somewhat more complicated. We can not define the conditional probability as before, due to the fact that none of the points in the support of the density function of a continuous vector satisfy the condition that their probability is non-negative; on the contrary, the probability of taking on any specific value for a continuous distribution is always 0. We will be able to define, however, conditional density:

Definition 2. *Let (X, Y) be a continuous random vector with density $g : \mathbb{R}^2 \to [0, \infty)$. Let $g_X(x) = \int_{-\infty}^{\infty} g(x, y)dy$ be the marginal density of X. For all $x \in \mathbb{R}$, we define the* **conditional density** *of variable Y given $X = x$ as the function*

$$g_{Y|X}(y|x) = \begin{cases} \frac{g(x,y)}{g_X(x)} & \text{if } g_X(x) > 0 \\ f(y) & \text{otherwise,} \end{cases}$$

where $f : \mathbb{R} \to [0, \infty)$ is any density function of our choice.

The density function f is needed only for completeness; it is never used (and so the shape of f is totally unimportant).

Note that this definition of conditional density corresponds to the definition of conditional probability, where we normalize the probability of a product of events by dividing by the probability of the condition; the conditional density is obtained in a similar way: we take the joint density function and normalize it by the marginal density of the variable defining the condition. The conditional density fulfills all the requirements for a density function, so it may be thought of as the density of a conditional distribution.

Note also that the conditional density is not defined unequivocally; we have several reasons for that. One reason is the arbitrary assumption of the density function f; the other reasons are due to the fact that each of the densities (joint and marginal) may also be modified in particular points without consequence for the distribution.

A third note that is worth making is that the conditional density "behaves" as expected in the case of independent random variables: if the variables are independent, then the joint density function may be presented as the product of marginal density functions, in which case the division by one of them gives the (unconditional) marginal density of the other function as the conditional density. That is, the value of one variable has no impact upon our assessment of the density of the other value.

Examples:

(1) Let (X, Y) be a variable with uniform distribution over a square with vertices at points $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$. The joint density of (X, Y) is

$$g(x, y) = \frac{1}{2}\mathbf{1}_{\{|x|+|y| \leqslant 1\}}(x, y).$$

The marginal density of X is equal to $\int_{-\infty}^{\infty} g(x, y)dy = (1 - |x|)\mathbf{1}_{(-1,1)}(x)$. The conditional density of Y, given $X = x$, may therefore be written as

$$g_{Y|X}(y|x) = \begin{cases} \frac{\mathbf{1}_{\{|y| \leqslant 1-|x|\}}(x,y)}{2(1-|x|)} = \frac{\mathbf{1}_{(-1+|x|,1-|x|)}(y)}{2(1-|x|)} & \text{for } x \in (-1, 1) \\ \text{any density} & \text{otherwise.} \end{cases}$$

Given such a density function, we may calculate, for example, the conditional probability $\mathbb{P}(Y \geqslant \frac{1}{2}|X = x)$:

$$\mathbb{P}(Y \geqslant \frac{1}{2}|X = x) = \int_{\frac{1}{2}}^{\infty} g_{Y|X}(y|x)dy = \begin{cases} \frac{1/2-|x|}{2(1-|x|)} & \text{if } |x| \leqslant \frac{1}{2} \\ 0 & \text{if } x \in (-1,1)\backslash(-1/2,1/2). \end{cases}$$

For other values of x we do not define the conditional probability.

(2) Using the chain rule, we may transform conditional densities in the same way we transformed conditional probabilities. For example, let us draw a number Λ uniformly from $(0,1)$, and then, knowing the value of $\Lambda = \lambda$, let us draw X from an exponential distribution with parameter λ. We can easily find the joint density of the vector (Λ, X): since the density of Λ is equal to $g_\Lambda(\lambda) = 1_{(0,1)}(\lambda)$, and the conditional density of X given that $\Lambda = \lambda$ is equal to $g_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x} 1_{(0,\infty)}(x)$, we have that

$$g_{(\Lambda,X)}(\lambda, x) = g_{X|\Lambda}(x|\lambda) \cdot g_\Lambda(\lambda) = \lambda e^{-\lambda x} 1_{(0,1)}(\lambda) 1_{(0,\infty)}(x).$$

Knowing the joint density of (Λ, X), we can now find the unconditional marginal density of X, g_X:

$$g_X(x) = \int_{-\infty}^{\infty} g_{(\Lambda,X)}(\lambda, x) d\lambda = 1_{(0,\infty)}(x) \int_0^1 \lambda e^{-\lambda x} d\lambda = 1_{(0,\infty)}(x) \left[-\frac{\lambda}{x} e^{-\lambda x} - \frac{1}{x^2} e^{-\lambda x} \right|_0^1$$
$$= 1_{(0,\infty)}(x) \left(\frac{1}{x^2} - \frac{1}{x} e^{-x} - \frac{1}{x^2} e^{-x} \right).$$

Having defined the continuous equivalent of the conditional probability, we can now define the conditional expected value in the continuous case – as the expected value of a variable with the conditional, rather than unconditional, density:

Definition 3. *Let (X, Y) be a continuous random vector with density $g : \mathbb{R}^2 \to [0, \infty)$, such that $\mathbb{E}|Y| < \infty$. For all $x \in \mathbb{R}$ we define the* **conditional expected value of variable Y given $X = x$** *as the expected value of a random variable with density $f_x(y) = g_{Y|X}(y|x)$, i.e.*

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

In the first example above, we had $g_{Y|X}(y|x) = \frac{\mathbf{1}_{(-1+|x|,1-|x|)}(y)}{2(1-|x|)}$ for $x \in (-1, 1)$, so that

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y \frac{\mathbf{1}_{(-1+|x|,1-|x|)}(x,y)}{2(1-|x|)} = \int_{-1+|x|}^{1-|x|} \frac{y}{2(1-|x|)} dy = 0.$$

Also in the continuous conditional expectation case, the “normal” properties of expected values are maintained:

Theorem 2. *Let (X, Y) be a continuous random vector with density $g : \mathbb{R}^2 \to [0, \infty)$, and $\varphi : \mathbb{R} \to \mathbb{R}$ be a Borel function such that $\mathbb{E}|\varphi(Y)| < \infty$. Then, we have that for any $x \in \mathbb{R}$,*

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

It is often convenient to treat the conditional expected value, which as it has been defined is a function of the value of the unconditional variable, as a random variable itself. We shall use the following definition, for both the discrete and continuous cases:

Definition 4. *Let (X, Y) be a random vector, such that $\mathbb{E}|Y| < \infty$. The* **conditional expected value of Y given X**, *denoted as $\mathbb{E}(Y|X)$, is a random variable such that*

$$\mathbb{E}(Y|X) = m(X),$$

where $m(x) = \mathbb{E}(Y|X = x)$.

Examples:

(1) In the first discrete example, we had $\mathbb{E}(X|Y=0)=\frac{1}{2}$ and $\mathbb{E}(X|Y=1)=\frac{3}{2}$. We can therefore say that $\mathbb{E}(X|Y=y)=y+\frac{1}{2}$, in which case we have that $\mathbb{E}(X|Y)=Y+\frac{1}{2}$.

(2) In the continuous example with a uniform distribution over the square with area 2, we had $\mathbb{E}(Y|X=x)=0$, which means that $\mathbb{E}(Y|X)=0$ (i.e. a random variable which is always equal to 0).

The conditional expected value has all the basic properties of "ordinary" expected values, for example:

Theorem 3. *Let $X, Y, Z : \Omega \to \mathbb{R}$ be random variables such that $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$. We have:*

(i) *If $X \geqslant 0$, then $\mathbb{E}(X|Z) \geqslant 0$.*

(ii) $|\mathbb{E}(X|Z)| \leqslant \mathbb{E}(|X||Z)$.

(iii) *For any $a, b \in \mathbb{R}$ we have $\mathbb{E}(aX+bY|Z) = a\mathbb{E}(X|Z)+b\mathbb{E}(Y|Z)$.*

The conditional expected value also has many useful properties specific to the definition, for example:

Theorem 4. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables such that $\mathbb{E}|Y| < \infty$. We have that*

(i) $\mathbb{E}|\mathbb{E}(Y|X)| < \infty$ *and* $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$.

(ii) *If X and Y are independent, then $\mathbb{E}(Y|X) = \mathbb{E}Y$.*

(iii) *If $h(X)$ is a limited random variable, then $\mathbb{E}(h(X)\cdot Y|X) = h(X)\mathbb{E}(Y|X)$.*

Given the definition of the conditional expectation, we can also define probability conditional on random variables:

Definition 5. *Let X be a random variable. For any event $A \in \mathcal{F}$, we define*

$$\mathbb{P}(A|X) = \mathbb{E}(1_A|X).$$

2. Linear Regression

We will briefly touch upon a topic which is of great practical interest and can be derived from probability calculus considerations, but lies predominantly within the scope of econometrics (and statistics) and therefore will not be covered thoroughly by this course. This is the topic of optimal approximation of one random variable with another, in the most simple case – linear approximation.

Let us assume that we have two random variables defined over the same sample space Ω, with a given joint distribution. Let us also assume that one of them is much easier to observe, or that it may be observed earlier and therefore serve as a predictor of the other variable; or that we simply wish to find a "rule of thumb" for a relationship between two variables. In all of these cases, we will be interested in approximating one variable with another variable. The simplest possible form of this approximation – the easiest computationally – is the linear form, where we look for an approximation of variable Y with variable X of the type $aX+b$, where $a, b \in \mathbb{R}$.

When choosing the best possible approximation, we also have to choose the criterion of comparison of different formulae; in the case of approximating one random variable with a different random variable, it seems plausible to assume that our aim will be to minimize the average deviation of the approximation from the real value; the deviation will be computed in the natural metric in $\mathbb{R}^2$, namely – quadratic. In other words, the problem of finding the best linear approximation may be reduced to finding $a, b \in \mathbb{R}$ such that $f(a,b) = \mathbb{E}(Y-aX-b)^2$ is minimized.

Let us rephrase the expression slightly:

$$f(a,b) = \mathbb{E}(Y^2+a^2X^2+b^2-2aXY-2bY+2abX) = \mathbb{E}(Y^2)+a^2\mathbb{E}X^2-2b\mathbb{E}Y+2ab\mathbb{E}X+2a\mathbb{E}XY+b^2;$$

therefore, for a given value of a, $f(a,b)$ as a function of b is a quadratic function, with minimum at $b = \mathbb{E}Y - a\mathbb{E}X$. It will therefore suffice to find the minimum value of function

$$h(a) = f(a, \mathbb{E}Y - a\mathbb{E}X) = \mathbb{E}(Y-\mathbb{E}Y-a(X-\mathbb{E}X))^2 = \mathrm{Var}Y + a^2\mathrm{Var}X - 2a\mathrm{Cov}(X,Y).$$

This minimum (given that $\mathrm{Var}X \neq 0$) is equal to

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}X},$$

in which case

$$b = \mathbb{E}Y - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}X}\mathbb{E}X.$$

The value of the parameter a may be transformed slightly: $a = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X}$.

At the minimum, the value of the mean quadratic error of the approximation, referred to as the residual variance, is $\sigma_Y^2(1 - \rho_{X,Y}^2)$. Note that the residual variance is zero (the approximation is perfect) when the variables have perfect linear correlation (in which case $|\rho_{X,Y}| = 1$), and is equal to σ_Y^2 (there is no explanation of Y with X) if the variables are not (linearly) correlated.

Note that in order to find the best linear approximation, we did not need full information about the joint distribution of the two variables; the knowledge of the covariance was sufficient. This is very convenient, since in practice, more often than not, we will not have full knowledge about the two variables to be analyzed; in most cases, we will only have an empirical sample. In this case, however, it is sufficient to calculate the sample means, variances and covariance of the two variables (which can be done effectively), to be able to construct the linear approximation.