Chapter 4
Product spaces. Conditional Distribution and Conditional Expectation

In this chapter we look at the following type of problems: If we know something extra about the experiment, how does that change our probability calculations. An important part of statistics (Bayesian statistics) is build on conditional distributions. However, what about the more complex and abstract notion of conditional expectation?

Why do we need conditional expectation?

Conditional expectation is a fundamental concept in the theory of stochastic processes. The simple idea is the following: suppose we have no information about a certain variable then our best guess about it would be some sort of regular expectation. However, in real life it often happens that we have some partial information about the random variable (or in time we come to know more about it). Then what we should do is every time there is new information the sample space Ω or the σ-algebra $\mathscr{F}$ is changing so they need to be recalculated. That will in turn change the probability $\mathbf{P}$ which will change the expectation of the variable. The conditional expectation provides a way to recalculate the expectation of the random variable given any new "consistent" information without going through the trouble of recalculating $(\Omega, \mathscr{F}, \mathbf{P})$ every time.

It is also easy to reason that since we calculate with respect to more precise information it will be depending on this more precise information, thus it is going to be a random variable itself, "adapted" to this information.

4.1 Product Spaces

Let $(\Omega_1, \mathscr{F}_1, \mu_1)$ and $(\Omega_2, \mathscr{F}_2, \mu_2)$ be two σ-finite measure spaces. Define:

$$\Omega = \Omega_1 \times \Omega_2 \text{ the cartesian product}$$
$$\mathscr{F} = \sigma(\{B_1 \times B_2 : B_1 \in \mathscr{F}_1, B_2 \in \mathscr{F}_2\})$$

Let $f : \Omega \to \mathbb{R}$ be $\mathscr{F}$ measurable such that

$$\forall\omega_1 \in \Omega_1 \; f(\omega_1,\cdot) \text{ is } \mathscr{F}_2 \text{ measurable on } \Omega_2$$
$$\forall\omega_2 \in \Omega_2 \; f(\cdot,\omega_2) \text{ is } \mathscr{F}_1 \text{ measurable on } \Omega_1$$

Then we define:

$$I_1^f(\omega_1) = \int_{\Omega_2} f(\omega_1,\omega_2)\mu_2(d\omega_2)$$
$$I_2^f(\omega_2) = \int_{\Omega_1} f(\omega_1,\omega_2)\mu_1(d\omega_1)$$

which are a kind of partial integrals, well defined by the measurability of the restrictions.

Theorem 4.1 (Fubini's theorem). *Define a measure:*

$$\mu(F) = \int_{\Omega_1}\int_{\Omega_2} 1_F(\omega_1,\omega_2)\mu_1(d\omega_1)\mu_2(d\omega_2).$$

Then μ is the unique measure defined on $(\Omega,\mathscr{F})$ called the product measure *with the property:*

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2) \qquad \forall A_i \in \mathscr{F}_i,$$

and as a consequence:

$$\int_\Omega f d\mu = \int_{\Omega_1} I_1^f(\omega_1)\mu(d\omega_1) = \int_{\Omega_2} I_2^f(\omega_2)\mu(d\omega_2)$$

Proof. Skipped. Apply the standard argument. Note that the first step is already given.

Example 4.1 (Application of Fubini's Theorem). Let X be a positive r.v. on $(\Omega,\mathscr{F},P)$. Consider $P\times\lambda$ on $(\Omega,\mathscr{F})\times([0,\infty),\mathscr{B}((0,\infty]))$, where λ is the Lebesgue measure. Let $A := \{(\omega,x) : 0 \le x < X(\omega)\}$. Note that A is the region under the graph of the random variable X. Let the indicator of this set be denoted with $h = 1_A$. Then:

$$I_1^h(\omega) = \int_{[0,\infty)} \mathbf{1}_A(\omega,x)d\lambda(x) = \int_0^\infty \mathbf{1}_{\{0\le x<X(\omega)\}}(x)d\lambda(x) = \int_0^{X(\omega)} d\lambda(x) = X(\omega)$$
$$I_2^h(x) = \int_\Omega \mathbf{1}_A(\omega,x)d\mathbf{P}(\omega) = \int_\Omega \mathbf{1}_{\{0\le x<X(\omega)\}}(\omega)d\mathbf{P}(\omega) = \mathbf{P}\{\omega : X(\omega) > x\},$$

since X is a positive r.v.

We now apply Fubini's Theorem and we get :

$$\mu(A) = \int_\Omega \int_{[0,\infty)} 1_A(x,\omega)d\mu(x)d\mathbf{P}(\omega)$$
$$= \int_\Omega X(\omega)d\mathbf{P}(\omega) = \int_0^\infty \mathbf{P}(X > x)dx$$

Thus reading the last line above:

$$\mathbf{E}(X) = \int_0^\infty \mathbf{P}(X > x)dx$$

This result is actually so useful that we will state it separately.

Corollary 4.1. *If X is a* ***positive*** *random variable with distribution function $F(x)$ and we denote $\overline{F}(x) = 1 - F(x)$, we have:*

$$\mathbf{E}(X) = \int_0^\infty \overline{F}(x)dx$$

4.2 Conditional distribution and expectation. Calculation in simple cases

We shall give a general formulation of conditional expectation that will be most useful in the second part of this textbook. But, until then we will present the cases that actually allow the explicit calculation of conditional distribution and expectation.

Let X and Y be two discrete variables on $(\Omega, \mathscr{F}, P)$.

Definition 4.1 (Discrete Conditional Distribution). The conditional distribution of Y given $X = x$: $F_{Y|X}(\cdot|x)$ is:

$$F_{Y|X}(y|x) = \mathbf{P}(Y \leq y|X = x)$$

The conditional probability mass function of $Y|X$ is:

$$f_{Y|X}(y|x) = \mathbf{P}(Y = y|X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Note: In the case when $\mathbf{P}(X = x) = 0$ we cannot define the conditional probability.

Definition 4.2 (Discrete Conditional Expectation). Let $\psi(x) = \mathbf{E}(Y|X = x)$ then $\psi(X) = \mathbf{E}[Y|X]$ is called the conditional expectation.

Remark 4.1. The conditional expectation is a random variable.

Definition 4.3 (Continuous Conditional Distribution). Let X, Y be two continuous random variables. The conditional distribution is defined as:

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} dv$$

The function $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ is the conditional probability density function.

Definition 4.4 (Continuous Conditional Expectation). The conditional expectation for two continuous random variables is $\psi(X) = \mathbf{E}[Y|X]$ where the function ψ is calculated:

$$\psi(x) = \mathbf{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

Example 4.2. A point is picked uniformly from the surface of the unit sphere. Let $L =$longitude angle θ and let $l =$ latitude angle ϕ. Let us find the distribution functions of $\theta|\phi$ and $\phi|\theta$.

Let C be a set on the sphere (or generally in $\mathbb{R}^3$). The surface area of the sphere is $4\pi r^2 = 4\pi$. The set of points from which we sample is $S(0,1) = \{(x,y,z) : x^2 + y^2 + z^2 = 1\}$. Then, since we pick the points uniformly the position of a point chosen has distribution:

$$\mathbf{P}((x,y,z) \in C) = \int_C \frac{1}{4\pi} \mathbf{1}_{\{x^2+y^2+z^2=1\}}(x,y,z)dxdydz$$

Since we are interested in longitude and latitude we change to polar coordinates to obtain the distribution of these variables. We take the transformation: $X = r\cos\theta\cos\phi$, $Y = r\sin\theta\cos\phi$ and $Z = r\sin\phi$. To obtain the distribution we calculate the new integral. The Jacobian of the transformation is:

$$\begin{aligned} J &= \begin{vmatrix} -r\sin\theta\cos\phi & -r\cos\theta\sin\phi & \cos\theta\cos\phi \\ r\cos\theta\cos\phi & -r\sin\theta\sin\phi & \sin\theta\cos\phi \\ 0 & r\cos\phi & \sin\phi \end{vmatrix} \\ &= r^2\cos^3\phi + r^2\sin^2\phi\cos\phi = r^2\cos\phi \end{aligned}$$

Note that the indicator is 1 if and only if $r = 1$. We conclude that

$$\mathbf{P}((x,y,z) \in C) = \int_{\mathrm{Im}\,C} \frac{1}{4\pi}|\cos\phi|d\theta d\phi,$$

where *ImC* is the set of *polar* coordinates that make the set C. Therefore, the joint distribution function is

$$f(\theta,\phi) = \frac{1}{4\pi}|\cos\phi|, \qquad \phi \in [-\pi/2, \pi/2], \theta \in [0, 2\pi].$$

Now, we get the marginal of ϕ:

$$f_\phi(\phi) = \int_0^{2\pi} \frac{1}{4\pi}|\cos\phi|d\theta = \frac{|\cos\phi|}{2},$$

and the marginal of θ:

$$f_\theta(\theta) = \int_{-\pi/2}^{\pi/2} \frac{1}{4\pi}|\cos\phi|d\phi = \int_{-\pi/2}^{\pi/2} \frac{1}{4\pi}\cos\phi d\phi = \frac{1}{2\pi}$$

Thus, we calculate immediately the conditional distributions:

$$f_{\theta|\phi}(\theta|\phi) = \frac{1}{2\pi}, \qquad \theta \in [0, 2\pi]$$
$$f_{\phi|\theta}(\phi|\theta) = \frac{\cos\phi}{2}, \qquad \phi \in [-\pi/2, \pi/2]$$

We note that θ and ϕ are independent (the product of marginals is equal to the joint distribution) but the conditionals are different due to the parameterizations (this particular example is known as *the Borel paradox*). Also note that the conditional expectations are equal to the regular expectations, this is of course because the variables are independent. We will obtain this property in general in the following section.

Example 4.3. Many clustering algorithms are based on random projections. For simplicity we consider the direction of the first coordinate unit vector $\overrightarrow{e}_1$ as the best possible projection. However, the probability of finding this direction exactly is zero so we consider a tolerance angle α_e and we say that a projection is "good enough" if it makes an angle less than α_e with $\overrightarrow{e}_1$.

We want to calculate the probability that a random direction makes an angle less than α with $\overrightarrow{e}_1$.

The example is in $\mathbb{R}^3$ but we can easily generalize it to any dimension. We assume that $0 < \alpha_e < \pi/2$, otherwise the problem becomes trivial.

Directions in $\mathbb{R}^3$ are equivalent to points on the unit sphere. Therefore, the probability to be calculated is twice the probability that a point chosen at random on the sphere belongs to the cone of angle α_e centered at the origin. Why twice? Because we do not care if the angle formed by the random direction is with $\overrightarrow{e}_1$ or $-\overrightarrow{e}_1$. Thus, we calculate the probability by taking the ratio of the area of the intersection of the said cone and the sphere and the total surface area of the sphere.

The area of the unit sphere in $\mathbb{R}^d$ is readily calculated as $\frac{2\pi^{d/2}}{\Gamma(d/2)}$ (e.g., Kendall (2004), $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function). In the particular case when $d = 3$ ($\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$) we obtain the well known area 4π.

To compute the support area of the cone we switch to polar coordinates:

$$x_1 = r\cos\theta_1$$
$$x_2 = r\sin\theta_1\cos\theta_2$$
$$x_3 = r\sin\theta_1\sin\theta_2$$

where$r \in [0, \infty), \theta_1 \in [0, \pi], \theta_2 \in [0, 2\pi]$.

The points of interest can be found when $r = 1$ and $\theta_1 \in [0, \alpha_e]$, and we need to double the final area found to take into account symmetric angles with respect to $\overrightarrow{e}_1$.

One can check immediately, that the Jacobian of this change of variables is $r^2\sin\theta_1$ and that the probability needed is easily calculated as:

$$2\sin^2\frac{\alpha_e}{2}$$

If we now consider K projections then the probability that at least one is a “good enough projection” is:

$$1-\left(1-2\sin^2\frac{\alpha_e}{2}\right)^K$$

Note that the example is extendable to the more interesting R^d case but in that case we do not obtain an exact formula instead only bounds. See **Ion: give citation once it exists.**

4.3 Conditional expectation. General definition

To summarize the previous section, if X and Y are two random variables we have defined the conditional distribution and conditional expectation of one **with respect to the other**. In fact, we have defined more: the conditional expectation of one **with respect to the information contained in the other**.

More precisely, in the previous subsection we defined the expectation of X conditioned by the σ-algebra generated by Y: $\sigma(Y)$. Thus, we may write without a problem:

$$\mathbf{E}[X|Y]=\mathbf{E}[X|\sigma(Y)].$$

This notion may be generalized to define conditional expectation with respect to any kind of of information (σ-algebra). As definition we shall use the following theorem. We will skip the proof.

Theorem 4.2. *Let $(\Omega,\mathscr{F},\mathbf{P})$ be a probability space, and let $\mathscr{K}\subseteq\mathscr{F}$ a sub-σ-algebra. Let X be a random variable on $(\Omega,\mathscr{F},\mathbf{P})$ such that either X is positive or $X\in L^1(\Omega)$. Then, there exists a random variable Y, measurable with respect to $\mathscr{K}$ with the property:*

$$\int_A Y\,dP=\int_A X\,dP \quad ,\forall A\in\mathscr{K}$$

This Y is defined to be the conditional expectation of X with respect to $\mathscr{K}$ and is denoted $\mathbf{E}[X|\mathscr{K}]$.

Remark 4.2. We note that the conditional expectation, unlike the regular expectation is a random variable measurable with respect to the sigma algebra under which is conditioned. In simple language it has adapted itself to the information contained in the σ-algebra $\mathscr{K}$. In the simple cases presented in the previous section the conditional expectation is measurable with respect to $\sigma(Y)$. But since this is a very simple sigma algebra then it has to be in fact a function of Y.

Note: We will take this theorem as a definition.

Proposition 4.1 (Properties of the Conditional Expectation). *Let $(\Omega, \mathscr{F}, \mathbf{P})$ a probability space, and let $\mathscr{K}, \mathscr{K}_1, \mathscr{K}_2$ sub-σ-algebras. Let X and Y be random variables of the probability space. Then we have:*

(1) If $\mathscr{K} = \{\varnothing, \Omega\}$ then $\mathbf{E}[X|\mathscr{K}] = \mathbf{E}X = const$.
(2) $\mathbf{E}[\alpha X + \beta Y|\mathscr{K}] = \alpha\mathbf{E}[X|\mathscr{K}] + \beta\mathbf{E}[Y|\mathscr{K}]$ for α, β real constants.
(3) If $X \leq Y$ a.s. then $\mathbf{E}[X|\mathscr{K}] \leq \mathbf{E}[Y|\mathscr{K}]$ a.s.
(4) $\mathbf{E}\left[\mathbf{E}[X|\mathscr{K}]\right] = \mathbf{E}X$
(5) If $\mathscr{K}_1 \subseteq \mathscr{K}_2$ then

$$\mathbf{E}\left[\mathbf{E}[X|\mathscr{K}_1]|\,\mathscr{K}_2\right] = \mathbf{E}\left[\mathbf{E}[X|\mathscr{K}_2]|\,\mathscr{K}_1\right] = \mathbf{E}[X|\mathscr{K}_1]$$

(6) If X is independent of $\mathscr{K}$ then

$$\mathbf{E}[X|\mathscr{K}] = \mathbf{E}[X]$$

(7) If Y is measurable with respect to $\mathscr{K}$ then

$$\mathbf{E}[XY|\mathscr{K}] = Y\mathbf{E}[X|\mathscr{K}]$$

After proving these properties (see Problem 4.2) they will become essential in working with conditional expectation. In fact the definition is never used anymore.

Example 4.4. Let us obtain a weak form of the Wald's equation (an equation that serves a fundamental role in the theory of stochastic processes) right now by a simple argument. Let $X_1, X_2, \dots, X_n, \dots$ be i.i.d. with finite mean μ and let N be a random variable taking values in strictly positive integers and independent of X_i for all i. For example, X_i's may be the results of random experiments and N may be some stopping strategy established in advance. Let $S_N = X_1 + X_2 + \dots + X_N$. Find $\mathbf{E}(S_N)$.

Let

$$\begin{aligned}\varphi(n) = \mathbf{E}[S_N|N=n] &= \mathbf{E}[X_1 + X_2 + \dots + X_N|N=n] \\ &= \sum_{i=1}^{n}\mathbf{E}[X_i|N=n] = \sum_{i=1}^{n}\mathbf{E}[X_i] = n\mu\end{aligned}$$

by independence. Therefore, $\mathbf{E}[S_N|N] = \varphi(N) = N\mu$. Finally, using the properties of conditional expectation:

$$\mathbf{E}(S_N) = \mathbf{E}[\mathbf{E}[S_N|N]] = \mathbf{E}[N\mu] = \mu\mathbf{E}[N].$$

Note that we have not used any distribution form only the properties of the conditional expectation.

Problems

4.1. Prove the Fubini's Theorem 4.1 on page 80.

4.2. Using the Theorem-Definition 4.2 on page 84 prove the seven properties of the conditional expectation in Proposition 4.1.

4.3. Let X be a random variable on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Let a set $A \in \mathscr{F}$ and the sigma algebra generated by the set denoted $\sigma(A)$. What is $\mathbf{E}[X|\sigma(A)]$? Let $\mathbf{1}_A$ denote the indicator of A. What is $\mathbf{E}[X|\mathbf{1}_A]$?

4.4. Let X, Y, Z be three random variables with joint distribution

$$P(X = k, Y = m, Z = n) = p^3 q^{n-3}$$

for integers k, m, n satisfying $1 \leq k < m < n$, where $0 < p < 1$, $p + q = 1$. Find $E\{Z|X, Y\}$.

4.5. A circular dartboard has a radius of 1 foot. Thom throws 3 darts at the board until all 3 darts are sticking in the board. The locations of the 3 darts are independent and uniformly distributed on the surface of the board. Let T_1, T_2, and T_3 be the distances from the center to the closest dart, the next closest dart, and the farthest dart, respectively, so that $T_1 \leq T_2 \leq T_3$. Find $\mathbf{E}[T_2]$.

4.6. Let $X_1, X_2, \dots, X_{1000}$ be i.i.d. each taking on both 0 and 1 with probability $\frac{1}{2}$. Put $S_n = X_1 + \cdots + X_n$. Find $\mathbf{E}\left[(S_{1000} - S_{300})\mathbf{1}_{\{S_{700}=400\}}\right]$ and $\mathbf{E}\left[(S_{1000} - S_{300})^2\mathbf{1}_{\{S_{700}=400\}}\right]$.