

1 Conditional Distributions

Let X, Y be two random variables. For clarity, we will assume that X and Y are both discrete with values in a denumerable space S , but definitions are possible so that the results here hold more generally. For further concreteness we will often identify S with $\mathbb{N} = \{0, 1, \dots\}$.

Let us fix notation. The joint distribution of X, Y is determined by the probabilities $\mathbb{P}\{X = x, Y = y\}$ for x and y in S . The probability of all interesting events involving X and Y can be found by adding probabilities of this form. For example,

$$\begin{aligned}\mathbb{P}\{X < Y\} &= \sum_{(x,y):x < y} \mathbb{P}\{X = x, Y = y\} \\ \mathbb{P}\{X + Y = k\} &= \sum_{j=0}^k \mathbb{P}\{X = j, Y = j - k\} .\end{aligned}$$

Thus it is natural to introduce the joint probability mass function, $p_{X,Y}(x, y)$, which is defined as

$$p_{X,Y}(x, y) \stackrel{\text{def}}{=} \mathbb{P}\{X = x, Y = y\} .$$

The distribution of Y alone is determined by the probabilities $\mathbb{P}\{Y = y\}$, for y in S . Thus we introduce $p_Y(y)$, the probability mass function for Y :

$$p_Y(y) \stackrel{\text{def}}{=} \mathbb{P}\{Y = y\} .$$

We can obtain p_Y from $p_{X,Y}$ by summing over possible values of X :

$$p_Y(y) = \sum_{x \in S} p_{X,Y}(x, y) .$$

Suppose that the event $\{Y = y\}$ has positive probability. The *conditional distribution* of X given $\{Y = y\}$ is the probability distribution on S defined by assigning the probability $\mathbb{P}\{X = x \mid Y = y\}$ to each point x in S . We will denote the conditional probability mass function by $p_{X|Y}(\cdot|y)$.

If the joint distribution of X, Y is known, then it is possible to find the conditional distribution of X given $Y = y$: Use the definition of conditional probability to obtain

$$p_{X|Y}(x|y) \stackrel{\text{def}}{=} \mathbb{P}\{X = x \mid Y = y\} = \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}} = \frac{p_{X,Y}(x, y)}{p_Y(y)} . \quad (1)$$

Sometime it is not the joint distribution that is known, but rather, for each y , one knows the conditional distribution of X given $Y = y$. If one also knows the distribution of Y , then one can recover the joint distribution using (1). We also mention one more use of (1):

$$\begin{aligned} p_X(x) &= \sum_y p_{X,Y}(x,y) \\ &= \sum_y p_{X|Y}(x|y)p_Y(y). \end{aligned} \tag{2}$$

Thus, given the conditional distribution of X given $Y = y$ for each possible value y , and the (marginal) distribution of Y , one can compute the (marginal) distribution of X , using (2).

Example 1.1. Let Y be a Poisson(λ) random variable. Let $\{B_k\}_{k=1}^\infty$ be a sequence of independent and identically distribution Bernoulli(p) random variables, independent also of Y . That is,

$$B_k = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Let $X = \sum_{k=1}^Y B_k$. Thus X is a sum of random variables, where the number of terms in the sum is itself a random variable.

The joint distribution of X is not immediately obvious. But we can compute the conditional distribution of X given that $Y = y$. Some reflection reveals that this conditional distribution is Binomial(y, p). Then to determine the (marginal) distribution of X , use (2):

$$\begin{aligned} p_X(x) &= \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \left(\frac{e^{-\lambda} \lambda^y}{y!} \right) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{k=0}^{\infty} \frac{((1-p)\lambda)^k}{k!} \\ &= \frac{(\lambda p)^x e^{-\lambda p}}{x!}. \end{aligned}$$

We conclude that X has a Poisson($p\lambda$) distribution.

2 Conditional Expectations

Since we now know what the meaning of the conditional distribution of X given that $Y = y$, we can define the conditional expectation of X given that $Y = y$. This is the expectation computed according to the conditional distribution:

$$\mathbb{E}\{X \mid Y = y\} \stackrel{\text{def}}{=} \sum_x xp_{X|Y}(x|y). \quad (3)$$

Example 2.1. The set-up here is the same as in Example 1.1. Then $\mathbb{E}\{X \mid Y = y\} = yp$ since the conditional distribution of X given $Y = y$ is Binomial(y, p).

Notice that for each value y , we obtain a number $\mathbb{E}\{X \mid Y = y\}$. For the moment, we will write $\varphi_X(y)$ for the number $\mathbb{E}\{X \mid Y = y\}$. We now have a *function* from S (the possible values of the random variable Y) to the real numbers:

$$\varphi_X : S \rightarrow \mathbb{R}.$$

This function returns the value $\mathbb{E}\{X \mid Y = y\}$ when evaluated at the point $y \in S$.

Recall that applying a function to a random variable returns a new random variable. Hence we can apply the function $\varphi_X(\cdot)$ to the random variable Y to get a new random variable $\varphi_X(\cdot)$. This random variable is denoted by $\mathbb{E}\{X \mid Y\}$.

Example 2.2. The set-up here is the same as in Example 1.1. We computed in 2.1 that $\mathbb{E}\{X \mid Y = y\} = yp$. Hence, the random variable $\mathbb{E}\{X \mid Y\}$ equals Yp .

What is the point? One answer to this question is the following theorem, which is very useful:

Theorem 2.3. *Let X, Y be two random variables, and suppose that $\mathbb{E}\{X\}$ is finite. Then*

$$\mathbb{E}\{X\} = \mathbb{E}\{\mathbb{E}\{X \mid Y\}\}.$$

Often one is able to compute $\mathbb{E}\{X \mid Y = y\}$ for each possible value y . Then to compute $\mathbb{E}\{X\}$, one computes the expectation of $\mathbb{E}\{X \mid Y\}$.

Example 2.4. Let N_k be the first time that k success appear in a row in independent Bernoulli(p) trials, $\{X_k\}_{k=1}^\infty$.

$$\begin{aligned} \mathbb{E}\{N_k \mid (N_{k-1}, X_{N_{k-1}+1})\} &= \begin{cases} N_{k-1} + 1 & \text{if } X_{N_{k-1}+1} = 1 \\ N_{k-1} + 1 + \mathbb{E}\{N_k\} & \text{if } X_{N_{k-1}+1} = 0 \end{cases} \\ &= N_{k-1} + 1 + (1 - X_{N_{k-1}})\mathbb{E}\{N_k\}. \end{aligned}$$

Taking expectations on both sides give

$$\begin{aligned}\mathbb{E}\{N_k\} &= \mathbb{E}\{N_{k-1}\} + 1 + (1-p)\mathbb{E}\{N_k\} \\ &= \frac{1}{p} + \frac{\mathbb{E}\{N_{k-1}\}}{p}.\end{aligned}$$

Since N_1 is a Geometric(p) r.v., an inductive argument shows that

$$\mathbb{E}\{N_k\} = \frac{1}{p} + \cdots + \frac{1}{p^k}.$$

Example 2.5. Consider a branching process: each individual in a population reproduces independently, and the mean number of children of a single individual is μ . Let Z_n be the number of descendants of a single ancestor after n generations. If $Z_{n-1} = m$, then Z_n is the sum of m i.i.d. random variables, each with expectation μ . Thus,

$$\mathbb{E}\{Z_n \mid Z_{n-1} = m\} = m\mu,$$

and so

$$\begin{aligned}\mathbb{E}\{Z_n\} &= \mathbb{E}\{\mathbb{E}\{Z_n \mid Z_{n-1}\}\} \\ &= \mathbb{E}\{\mu Z_{n-1}\} \\ &= \mu \mathbb{E}\{Z_{n-1}\}.\end{aligned}$$

By induction, we have then $\mathbb{E}\{Z_n\} = \mu^n$.

3 Markov Chains

We now consider a *stochastic process*, a sequence of random variables $\{X_n\}_{n=0}^{\infty}$. We think of this dynamically: X_n is the position or state of some randomly evolving system after n units of time.

3.1 Definitions

We will assume that each X_n takes values in a countable state space, which as before we will often identify with $\mathbb{N} = \{0, 1, 2, \dots\}$.

Definition 3.1. A stochastic process $\{X_n\}_{n=0}^\infty$ with values in a countable state space obeys the *Markov property* if

$$\mathbb{P}\{X_{n+1} = j \mid X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i\} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}. \quad (4)$$

If in addition, the right-hand side of (4) does not depend on n , we will say that $\{X_n\}$ is a *Markov chain*.

We mention that some authors call any stochastic process with values in a countable space and satisfying (4) a Markov chain, and if right-hand side of (4) does not depend on n , say that the chain is *time homogenous*. We will assume that all of our chains are time homogenous, and include it in the definition.

A matrix \mathbf{P} is called *stochastic* if $P_{i,j} \geq 0$ for all i and j , and if $\sum_{j \in S} P_{i,j} = 1$. Thus each row of \mathbf{P} specifies a probability distribution.

Thus, a Markov chain specifies a stochastic matrix \mathbf{P} (whose rows and columns are indexed by S), by

$$P_{i,j} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}. \quad (5)$$

This matrix is called the *transition matrix* of $\{X_n\}_{n=0}^\infty$.

Conversely, given a stochastic matrix \mathbf{P} , there exists a Markov chain $\{X_n\}_{n=0}^\infty$ which has a transition matrix equal to \mathbf{P} . We will not prove this last assertion.

Hence, due to this correspondence, we will sometimes identify a Markov chain with its transition matrix \mathbf{P} , and sometimes call a transition matrix \mathbf{P} a Markov chain.

Given the distribution of X_0 , the position of the Markov chain at time 0, and the transition matrix \mathbf{P} , one can recover the entire joint distribution of (X_0, X_1, \dots) :

$$\begin{aligned} \mathbb{P}\{X_0 = i_0, \dots, X_n = i_n\} &= \mathbb{P}\{X_n = i_n \mid X_0 = i_0, \dots, X_{n-1} = i_{n-1}\} \\ &\quad \times \mathbb{P}\{X_0 = i_0, \dots, X_{n-1} = i_{n-1}\} \\ &= P_{i_{n-1}, i_n} \mathbb{P}\{X_{n-1} = i_{n-1} \mid X_0 = i_0, \dots, X_{n-2} = i_{n-2}\} \\ &\quad \times \mathbb{P}\{X_0 = i_0, \dots, X_{n-2} = i_{n-2}\} \\ &= P_{i_{n-1}, i_n} P_{i_{n-2}, i_{n-1}} \mathbb{P}\{X_0 = i_0, \dots, X_{n-2} = i_{n-2}\} \\ &\quad \vdots \\ &= P_{i_{n-1}, i_n} P_{i_{n-2}, i_{n-1}} \cdots P_{i_0, i_1} \mathbb{P}\{X_0 = i_0\}. \end{aligned}$$

Thus, any interesting probability concerning the chain $\{X_n\}_{n=0}^\infty$ can be determined from \mathbf{P} and the distribution of the initial state X_0 . (The distribution of X_0 is called the *initial distribution*.)

3.2 n -step Transition Probabilities

Recall that if \mathbf{A} is an $r \times s$ matrix and \mathbf{B} is an $s \times t$ matrix (where any or all of r, s, t are all allowed to equal ∞), then we can form the matrix \mathbf{AB} , an $r \times t$ matrix whose (i, j) th entry is given by

$$(\mathbf{AB})_{i,j} = \sum_{\ell} A_{i,\ell} B_{\ell,j}.$$

We wish to express $\mathbb{P}\{X_{m+n} = j \mid X_m = i\}$ in terms of the matrix \mathbf{P} . Thus, we wish to determine the matrix $\mathbf{M}^{m,n}$ whose (i, j) th entry is

$$M_{i,j}^{m,n} = \mathbb{P}\{X_{m+n} = j \mid X_m = i\}.$$

We have

$$\begin{aligned} \mathbb{P}\{X_{m+n} = j \mid X_m = i\} &= \frac{\mathbb{P}\{X_m = i, X_{m+n} = j\}}{\mathbb{P}\{X_m = i\}} \\ &= \sum_{\ell} \frac{\mathbb{P}\{X_m = i, X_{m+n-1} = \ell, X_{m+n} = j\}}{\mathbb{P}\{X_m = i\}} \\ &= \sum_{\ell} \mathbb{P}\{X_{m+n} = j \mid X_m = i, X_{m+n-1} = \ell\} \\ &\quad \times \frac{\mathbb{P}\{X_m = i, X_{m+n-1} = \ell\}}{\mathbb{P}\{X_m = i\}} \\ &= \sum_{\ell} P_{\ell,j} \mathbb{P}\{X_{m+n-1} = \ell \mid X_m = i\} \\ &= \sum_{\ell} M_{i,\ell}^{m,n-1} P_{\ell,j} \\ &= (M^{m,n-1} \mathbf{P})_{i,j}. \end{aligned}$$

Thus

$$\mathbf{M}^{m,n} = \mathbf{M}^{m,n-1} \mathbf{P}.$$

By induction we have that

$$\mathbf{M}^{m,n} = \mathbf{M}^{m,0} \mathbf{P}^n.$$

But $\mathbf{M}^{m,0}$ is the identity matrix \mathbf{I} , and so

$$\mathbf{M}^{m,n} = \mathbf{P}^n.$$

Now let μ be the row vector

$$\mu \stackrel{\text{def}}{=} [\mathbb{P}\{X_0 = 0\} \ \mathbb{P}\{X_0 = 1\} \ \mathbb{P}\{X_0 = 2\} \ \dots].$$

Thus μ_k , the k th component of μ , is $\mathbb{P}\{X_0 = k\}$. The vector μ specifies the initial distribution of the Markov chain.

We will write $\mathbb{P}_\mu\{\cdot\}$ for probabilities concerning the Markov chain with transition matrix \mathbf{P} and initial distribution μ . We write $\mathbb{P}_i\{\cdot\}$ when the initial distribution is given by $\mathbb{P}\{X_0 = i\} = 1$, that is, when the chain is started in position i . When no assumption is made on the initial distribution, we will write simply $\mathbb{P}\{\cdot\}$.

We would like an expression for $\mathbb{P}_\mu\{X_n = i\}$ in terms of \mathbf{P} and μ :

$$\begin{aligned}\mathbb{P}_\mu\{X_n = i\} &= \sum_{\ell} \mathbb{P}_\mu\{X_n = i \mid X_0 = \ell\} \mathbb{P}_\mu\{X_0 = \ell\} \\ &= \sum_{\ell} \mu_{\ell} P_{\ell,i}^n \\ &= (\mu \mathbf{P}^n)_i.\end{aligned}$$

That is, the row vector $\mu \mathbf{P}^n$ gives the distribution of X_n when the initial distribution is μ .

3.3 Classification of States

We say state j is *accessible* from state i if there exists some n (possibly depending on i and j) so that $P_{i,j}^n > 0$. That is, if the chain has positive probability of reaching state j , starting in state i , after n steps.

States i and j *communicate* if j is accessible from i and i is accessible from j .

A state i is called *absorbing* if $P_{i,i} = 1$. Thus if the chain ever lands in state i , then it stays there forever. Clearly an absorbing state communicates with no other state.

A Markov chain is called *irreducible* if any two states i and j communicate. Thus in an irreducible chain, any state is reachable from any other state with positive probability.

We will mostly restrict attention to irreducible chains.

Let N_i be the number of visits of a markov chain to state i . That is

$$N_i = \sum_{n=1}^{\infty} \mathbf{1}\{X_n = i\}.$$

Also, define

$$f_i = \mathbb{P}_i\{X_n = i \text{ for some } n\} = \mathbb{P}_i\{N_i \geq 1\}.$$

Proposition 3.2. *Given $X_0 = i$, the random variable N_i has a Geometric distribution with parameter $1 - f_i$.*

Let $T_{i,m}$ be the time of the m th visit of X_n to state i . Notice that the event that $T_{i,m} = n$ depends only on X_1, \dots, X_n and forces $X_n = i$.

Exercise 3.3. Show that (4) holds more generally:

$$\begin{aligned} \mathbb{P}\{X_{n+1} = b_1, \dots, X_{n+m} = b_m \mid X_0 = a_0, \dots, X_n = a_n\} \\ = \mathbb{P}\{X_1 = b_1, \dots, X_m = b_m \mid X_0 = a_n\}. \end{aligned} \quad (6)$$

Proof of Proposition 3.2. By (6)

$$\begin{aligned} \mathbb{P}_i\{(X_{n+1}, X_{n+2}, \dots) \text{ hits } i \mid T_{i,m} = n\} &= \mathbb{P}\{(X_1, \dots) \text{ hits } i \mid X_0 = i\} \\ &= \mathbb{P}_i\{(X_1, \dots) \text{ hits } i\} \\ &= f_i. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}_i\{N_i > m\} &= \sum_{n=1}^{\infty} \mathbb{P}_i\{N_i > m, T_{i,m} = n\} \\ &= \sum_{n=1}^{\infty} \mathbb{P}_i\{(X_{n+1}, X_{n+2}, \dots) \text{ hits } i \mid T_{i,m} = n\} \mathbb{P}_i\{T_{i,m} = n\} \\ &= \mathbb{P}_i\{(X_1, X_2, \dots) \text{ hits } i\} \sum_{n=0}^{\infty} \mathbb{P}_i\{T_{i,m} = n\} \\ &= f_i \mathbb{P}_i\{T_{i,m} < \infty\} \\ &= f_i \mathbb{P}_i\{N_i > m - 1\} \end{aligned}$$

This implies that N_i is a Geometric random variable with parameter $1 - f_i$. ■

We conclude that

$$\mathbb{E}_i\{N_i\} = \frac{1}{1 - f_i}.$$

In particular $f_i = 1$ if and only if $\mathbb{E}_i\{N_i\} = \infty$.

If $f_i = 1$, we say that state i is *recurrent*. If $f_i < 1$, we say that state i is *transient*.

We have proved the following proposition:

Proposition 3.4. *State i is recurrent if and only if*

$$\sum_n P_{i,i}^n = \infty.$$

We comment that the previous analysis was a possible because of the *strong Markov property*. A *stopping time* is a random time T so that for each n , the event $\{T = n\}$ is determined by the r.v.s (X_0, \dots, X_n) . Thus to check at time n whether it is time to stop (that is, if $\{T = n\}$), it is enough to have observed the chain up until time n . An example is $T_{m,i}$, the time of the m th return to state i . The strong Markov property says that if T is a stopping time, the new process $(X_{T+1}, X_{T+2}, \dots)$ is a Markov chain whose distribution is the same as (X_1, X_2, \dots) , depending on (X_0, X_1, \dots) only through the random variable X_T .

Suppose that state i and state j communicate, and suppose that state i is recurrent. There exists r and s so that $P_{j,i}^r > 0$ and $P_{i,j}^s > 0$. Notice that for $n > r + s$,

$$\mathbb{P}_j \{X_n = j\} \geq \mathbb{P}_j \{X_r = i, X_{n-s} = i, X_n = j\} \quad (7)$$

$$P_{j,j}^n \geq P_{j,i}^r P_{i,i}^{n-s-r} P_{i,j}^s \quad (8)$$

and so

$$\sum_n P_{j,j}^n \geq P_{j,i}^r P_{i,j}^s \sum_{n>r+s} P_{i,i}^{n-s-r} = P_{j,i}^r P_{i,j}^s \sum_{m=1}^{\infty} P_{i,i}^m = \infty. \quad (9)$$

We conclude that state j is recurrent as well. Thus, in an irreducible Markov chain, either all states are transient, or all states are recurrent.

Example 3.5 (Random Walk on \mathbb{Z}). Consider the random walk on \mathbb{Z} which increases by 1 with probability p and decreases by 1 with probability $1 - p$. That is,

$$P_{i,j} = \begin{cases} p & \text{if } j = i + 1 \\ 1 - p & \text{if } j = i - 1. \end{cases}$$

Let us consider $P_{0,0}^{2n}$:

$$\begin{aligned} P_{0,0}^{2n} &= \binom{2n}{n} (p(1-p))^n \\ &= \frac{(2n!)}{(n!)^2} (p(1-p))^n \\ &\sim \frac{\sqrt{2\pi}(2n)^{2n+1/2} e^{-2n}}{(2\pi)n^{2n+1} e^{-2n}} (p(1-p))^n \\ &= \frac{1}{\sqrt{2\pi n}} (4p(1-p))^n. \end{aligned}$$

Notice that $0 \leq 4p(1-p) \leq 1$, with equality if and only if $p = \frac{1}{2}$. Thus for $p \neq 1/2$, the right-hand side above is summable, as it is dominated by a geometric series. On the other hand, for $p = 1/2$, the right-hand side is not summable. Thus the random walk is recurrent if and only if $p = \frac{1}{2}$.