Chapter 4

Joint distributions

So far we have studied one random variable at a time, but things get interesting and useful when we study the relationships between many random variables. To begin, we will focus on capturing the dependence between a pair of random variables (X, Y). For example:

- X is the number of predators, and Y is the number of prey;
- X is the cholesterol level of a husband, and Y is the cholesterol level of his wife;
- X is the price of Google stock, and Y is the price of Apple stock;
- X is the amount of a fire insurance claim for a home in Pacific Palisades, Los Angeles, and Y is the amount of a fire insurance claim for the home next door.

In each case, the quantities are random, and there are forces that effect the two variables both separately and together, and they may affect each other. How do we capture this? The same way we capture everything: with a probability measure!

Definition 4.1. If X and Y are **jointly distributed** random variables, then their **joint range** is the set of values $\text{Range}(X, Y) \subseteq \mathbb{R}^2$ they can take on, and their **joint distribution** is a probability measure on $\mathbb{R}^2$:

$$A \mapsto P\left((X, Y) \in A\right), \quad A \subseteq \mathbb{R}^2.$$

As in the univariate case, there are many types of bivariate distributions, but we will focus on two special cases that are easy to work with: jointly discrete and jointly absolutely continuous.

4.1 Jointly discrete random pairs

Definition 4.2. A pair (X, Y) of **jointly discrete** random variables has a finite or countably infinite joint range, and the joint distribution is captured by the **joint probability mass function** (**joint PMF**), which tells us the individual probability of each pair:

$$P(X = x \cap Y = y), \quad (x, y) \in \text{Range}(X, Y).$$

Given the joint PMF, we can compute the joint distribution by adding:

$$P((X, Y) \in A) = \sum_{(x,y)\in A\cap \text{Range}(X,Y)} \sum P(X = x \cap Y = y), \quad A \subseteq \mathbb{R}^2. \tag{4.1}$$

This is a consequence of *countable additivity*.

Definition 4.3. If X and Y are jointly distributed, then the distribution of X or Y by itself in isolation is its **marginal distribution**. In the jointly discrete case, these distributions are summarized by the **marginal probability mass functions** (**marginal PMFs**) given by:

$$P(X = x) = \sum_{y\in \text{Range}(Y)} P(X = x \cap Y = y) \tag{4.2}$$

$$P(Y = y) = \sum_{x\in \text{Range}(X)} P(X = x \cap Y = y) \tag{4.3}$$

This is a consequence of *the law of total probability*, where we take the events $\{Y = y\}$ as the components of our partition.

Definition 4.4. If X and Y are jointly distributed, and we learn, for instance, that $Y = y$, what does that teach us about X? To answer this question, we need the **conditional distribution** of X given $Y = y$. In the jointly discrete case, this is summarized by the **conditional probability mass function** (**conditional PMF**):

$$P(X = x \mid Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)} \tag{4.4}$$

$$P(Y = y \mid X = x) = \frac{P(X = x \cap Y = y)}{P(X = x)} \tag{4.5}$$

This is just the definition of conditional probability applied to the events $A = \{X = x\}$ and $B = \{Y = y\}$.

Theorem 4.1. (**Marginal-conditional decomposition**) From Definition 4.4, we get "two identities for the price of one:"

$$P(X = x \cap Y = y) = P(X = x \mid Y = y)P(Y = y) = P(Y = y \mid X = x)P(X = x). \tag{4.6}$$

Corollary 4.2. (**Hierarchical representation**) In order to specify a joint distribution for (X, Y), you can do so *hierarchically*:

$$X \sim P_X$$
$$Y \mid X \sim P_{Y|X}.$$

Theorem 4.1 guarantees that we can stitch these two pieces together to get a valid joint distribution.

Theorem 4.3. (**Bayes' theorem, again**) Combine Definition 4.4 and Theorem 4.1, and you get Bayes' theorem for the conditional PMFs:

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x)P(X = x)}{P(Y = y)} \tag{4.7}$$

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}. \tag{4.8}$$

$$\tag{4.9}$$

Example 4.1. Consider rolling a pair of fair, four-sided dice. Let $X \in \{1, 2, 3, 4\}$ be the number that appears on Die 1, and let $Y \in \{1, 2, 3, 4\}$ be the number that appears on Die 2. We have a finite sample space with equally-likely outcomes, so we can derive the joint PMF by counting, and it is displayed in Table 4.1. If we sum down the rows, we get the marginal PMF of Y at the bottom, and if we sum across the columns, we get the marginal PMF of X on the right. If we divide the joint probabilities by the marginals, we get the conditional probabilities in the bottom two tables. We see in this case that the conditional PMFs are equal to the marginal PMFs *everywhere*, and this motivates a proper mathematical definition of *independence* for jointly discrete random variables.

Definition 4.5. A jointly discrete pair (X, Y) are **independent** if the joint PMF factors into the product of marginals:

$$P(X = x \cap Y = y) = P(X = x)P(Y = y) \quad \text{for all } (x, y) \in \mathbb{R}^2. \tag{4.10}$$

Alternatively but equivalently, X and Y are independent if

$$P(X = x \mid Y = y) = P(X = x) \quad \text{for all } (x, y) \in \mathbb{R}^2 \tag{4.11}$$

$$P(Y = y \mid X = x) = P(Y = y) \quad \text{for all } (x, y) \in \mathbb{R}^2. \tag{4.12}$$

In other words, learning about one variable teaches you nothing about the other.

Example 4.2. Take X and Y from Example 4.1 and define a new pair of jointly discrete random variables $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$. Table 4.2 enumerates the joint range and lists out the joint PMF. Summing across rows or columns gives the marginal PMFs, and dividing the joint probabilities by the marginal probabilities gives the conditional PMFs in the bottom two tables. Unlike Example 4.1, U and V are an example of **dependent** random variables. Consider for instance $U = 1$ and $V = 3$:

$$P(U = 1) \times P(V = 3) = \frac{7}{16} \times \frac{5}{16} \approx 0.136 \neq 0.125 = P(U = 1 \cap V = 3).$$

		Y				
	$P(X = x \cap Y = y)$	1	2	3	4	$P(X = x)$
X	1	1 / 16	1 / 16	1 / 16	1 / 16	1 / 4
	2	1 / 16	1 / 16	1 / 16	1 / 16	1 / 4
	3	1 / 16	1 / 16	1 / 16	1 / 16	1 / 4
	4	1 / 16	1 / 16	1 / 16	1 / 16	1 / 4
	$P(Y = y)$	1 / 4	1 / 4	1 / 4	1 / 4	1

		Y			
	$P(X = x \mid Y = y)$	1	2	3	4
X	1	1 / 4	1 / 4	1 / 4	1 / 4
	2	1 / 4	1 / 4	1 / 4	1 / 4
	3	1 / 4	1 / 4	1 / 4	1 / 4
	4	1 / 4	1 / 4	1 / 4	1 / 4
		1	1	1	1

		Y				
	$P(Y = y \mid X = x)$	1	2	3	4	
X	1	1 / 4	1 / 4	1 / 4	1 / 4	1
	2	1 / 4	1 / 4	1 / 4	1 / 4	1
	3	1 / 4	1 / 4	1 / 4	1 / 4	1
	4	1 / 4	1 / 4	1 / 4	1 / 4	1

Table 4.1: Joint, marginal, and conditional PMFs in Example 4.1. If you sum the joint probabilities down the rows or columns, you get the marginal probabilities. If you divide the joint probabilities by the marginal probabilities, you get the conditional probabilities. This example is not actually super interesting because X and Y are independent, so the conditionals are always equal to the marginals.

	$P(U=u \cap V=v)$	V 1	2	3	4	$P(U=u)$
U	1	1 / 16	2 / 16	2 / 16	2 / 16	7 / 16
	2	0	1 / 16	2 / 16	2 / 16	5 / 16
	3	0	0	1 / 16	2 / 16	3 / 16
	4	0	0	0	1 / 16	1 / 16
	$P(V=v)$	1 / 16	3 / 16	5 / 16	7 / 16	1

	$P(U=u \mid V=v)$	V 1	2	3	4
U	1	1	2 / 3	2 / 5	2 / 7
	2	0	1 / 3	2 / 5	2 / 7
	3	0	0	1 / 5	2 / 7
	4	0	0	0	1 / 7
		1	1	1	1

	$P(V=v \mid U=u)$	V 1	2	3	4	
U	1	1 / 7	2 / 7	2 / 7	2 / 7	1
	2	0	1 / 5	2 / 5	2 / 5	1
	3	0	0	1 / 3	2 / 3	1
	4	0	0	0	1	1

Table 4.2: Joint, marginal, and conditional PMFs in Example 4.2. If you sum the joint probabilities down the rows or columns, you get the marginal probabilities. If you divide the joint probabilities by the marginal probabilities, you get the conditional probabilities. U and V are not independent in this example.

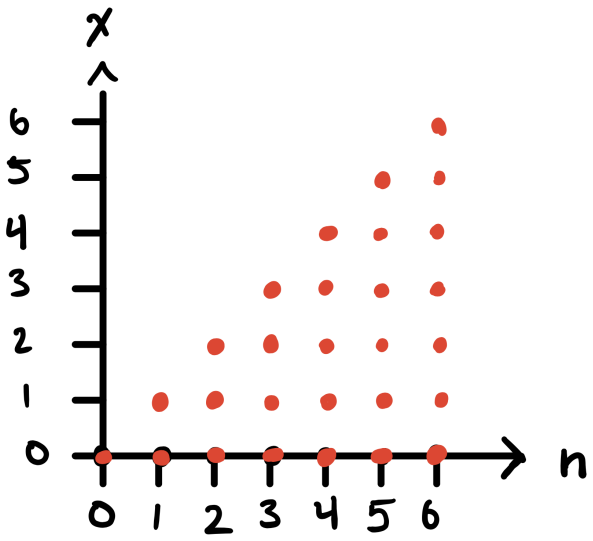

Figure 4.1: The set of pairs that (N, X) can take on in Example 4.3.

To compute joint, marginal, or conditional probabilities, you sum the relevant PMF:

$$\begin{aligned}
P(U \leq 2 \cap V \geq 3) &= 8/16 = 1/2 \\
P(U \geq 3) &= P(U = 3) + P(U = 4) = 1/4 \\
P(V \geq 3) &= P(V = 3) + P(V = 4) = 3/4 \\
P(U \geq 3 \mid V = 3) &= P(U = 3 \mid V = 3) + P(U = 4 \mid V = 3) = 1/5 \\
P(U \geq 3 \mid V = 4) &= P(U = 3 \mid V = 4) + P(U = 4 \mid V = 4) = 3/7 \\
P(U \leq 2 \mid V \geq 3) &= \frac{P(U \leq 2 \cap V \geq 3)}{P(V \geq 3)} = \frac{1/2}{3/4} = 2/3.
\end{aligned}$$

Example 4.3. Consider the joint distribution of random variables N and X, written in hierarchical form:

$$\begin{aligned}
N &\sim \text{Poisson}(\lambda) \\
X \mid N = n &\sim \text{Binomial}(n, p).
\end{aligned}$$

This is a model of the number of successes (X) in a random number of trials (N). For instance, N could represent the number of particles emitted in a minute, and X is the number of those particles you successfully detect with an imperfect measurement device. Or N could represent the number of car accidents that occur in an hour at a fraught intersection, and X is the number of those accidents that are fatal. The joint range $\text{Range}(N, X)$ is displayed in Figure 4.1.

We derive the marginal pmf of X. Since $\text{Range}(N) = \mathbb{N}$, then marginally, $\text{Range}(X) = \mathbb{N}$. So

fix $k \in \mathbb{N}$. Noting initially that $P(X = k \cap N = n) = 0$ for $n < k$, we have

$$\begin{aligned}
P(X = k) &= \sum_{n=0}^{\infty} P(X = k \cap N = n) \\
&= \sum_{n=k}^{\infty} P(X = k \cap N = n) \\
&= \sum_{n=k}^{\infty} P(X = k \mid N = n) P(N = n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= \sum_{n=k}^{\infty} \frac{\cancel{n!}}{(n-k)!k!} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{\cancel{n!}} \\
&= \frac{p^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{(1-p)^{n-k}}{(n-k)!} \lambda^n && \text{pull out constants} \\
&= \frac{p^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{(1-p)^{n-k}}{(n-k)!} \lambda^n \frac{\lambda^k}{\lambda^k} && \text{multiply by 1} \\
&= \frac{p^k}{k!} e^{-\lambda} \lambda^k \sum_{n=k}^{\infty} \frac{(1-p)^{n-k}}{(n-k)!} \lambda^{n-k} \\
&= \frac{(p\lambda)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{[\lambda(1-p)]^{n-k}}{(n-k)!} \\
&= \frac{(p\lambda)^k}{k!} e^{-\lambda} \sum_{i=0}^{\infty} \frac{[\lambda(1-p)]^i}{i!} && \text{reindex} \\
&= \frac{(p\lambda)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} && \text{recognize Taylor series} \\
&= \frac{(p\lambda)^k}{k!} e^{-p\lambda}, \quad k \in \mathbb{N}.
\end{aligned}$$

So marginally, X has the Poisson distribution with rate $p\lambda$. Furthermore, as p gets closer to 1, the distribution of X gets closer to just being the original distribution of N. So the higher the probability of success, the less of a wedge there is between the number of trials that occurred and the number of trials that succeeded. If $p = 1$ and every trial is guaranteed go succeed, these are literally the same.