Math 130B

Liam Hardiman

August 9, 2022

Abstract

I'm writing these lecture notes for UC Irvine's Math 130B course, taught in the summer of 2022. This is a five-ish week course where I plan to get through chapters 6-8 of Ross' book [1]. The class structure consists of a two hour lecture followed by a one hour discussion section three days a week. I'm aiming to get through one or two sections of the book per lecture with a midterm soon after chapter 6, maybe partway into chapter 7.

Contents

1 Jointly Distributed Random Variables

Many of life's more interesting problems are multifaceted. For example, in a clinical trial for a cholesterol drug, we might be interested in a patient's cholesterol levels *and* how many hours they exercise each week. Or if we're interested in California's gas consumption, we'd be interested in how much gas each station sells *and* its price of gas.

In this section, we address how to look at more than one random variable at the same time.

1.1 Joint Distribution Functions

Remember that we can define the *probability mass function* of a discrete random variable X to be the function that takes in a value and returns the probability that X attains that value.

$$p(a) = \Pr[X = a].$$

Examples 1.1. (a) Suppose we roll a pair of dice and X is the sum of the values shown. Then X

can take any integer value between 2 and 12 and its probability mass function is

$$\begin{aligned} p(2) = p(12) &= \frac{1}{36} & p(3) = p(11) &= \frac{2}{36} \\ p(4) = p(10) &= \frac{3}{36} & p(5) = p(9) &= \frac{4}{36} \\ p(6) = p(8) &= \frac{5}{36} & p(7) &= \frac{6}{36}. \end{aligned}$$

(b) Suppose Alice is communicating with Bob by sending him bits (0's or 1's) one by one. Suppose each bit Alice sends has probability p of successfully getting to Bob and each transmission is independent of the others. If X is the first time a bit fails to transmit properly (maybe there's too much noise on the channel), then X is a geometric random variable with probability mass function

$$p(n) = p^{n-1}(1-p).$$

Situations naturally arise where we might want to look at two discrete random variables at the same time. For example, if we roll two dice and X is the smaller roll and Y is the larger one, can we define an analogue of the probability mass function?

Definition 1.2. Suppose X and Y are two discrete random variables taking values in the sets A and B, respectively. Then their *joint probability mass function* is the function $p : A \times B \to [0, 1]$ defined by

$$p(a, b) = \Pr[X = a, Y = b].$$

Example 1.3. Say we roll two dice and X is the largest value shown and Y is the sum of the two values. Let's compute a few values of the joint probability mass function of X and Y. We have

$$\begin{aligned} p(3, 5) &= \Pr[X = 3, Y = 5] \\ &= \Pr[\{(3, 2), (2, 3)\}] \\ &= \frac{2}{36}. \end{aligned}$$

This is because the only way for the largest value to be 3 and the sum to be 5 is for one of the dice to show 2 and the other to show 3. We also have

$$p(1, 8) = 0$$

since there's no way for two dice to sum to 8 and the largest value be a 1.

How does the joint mass function of X and Y relate to the *marginal* probability mass functions? Well if we just specify that $X = a$, then we haven't put any restrictions on Y. This gives us

$$\begin{aligned} p_X(a) &= \Pr[X = a] \\ &= \Pr[X = a, Y < \infty] \\ &= \sum_{y \in B} \Pr[X = a, Y = y] \\ &= \sum_{y \in B} p(a, y). \end{aligned} \tag{1}$$

Similarly, we have

$$p_Y(b) = \sum_{x \in A} p(x, b).$$

Example 1.4. Say 100 people are asked for their handedness (right-handed or left-handed) and sex (male or female). The survey produces the following table.

	L	R
M	4	44
F	9	43

If we randomly select one of these people and let X be their sex and Y be their handedness, then we can obtain the joint probability mass function by just reading off values from the table.

$$p(M,L) = 4/100 \qquad p(M,R) = 44/100$$
$$p(F,L) = 9/100 \qquad p(F,R) = 43/100.$$

Let's compute the marginal probability mass functions. For X we have

$$p_X(M) = p(M,L) + p(M,R) = \frac{4}{100} + \frac{44}{100} = \frac{48}{100}$$
$$p_X(F) = p(F,L) + p(F,R) = \frac{9}{100} + \frac{43}{100} = \frac{52}{100}.$$

For Y we have

$$p_Y(L) = p(M,L) + p(F,L) = \frac{4}{100} + \frac{9}{100} = \frac{13}{100}$$
$$p_Y(R) = p(M,R) + p(F,R) = \frac{44}{100} + \frac{43}{100} = \frac{87}{100}.$$

Example 1.5. In the previous example we determined the marginal mass function from the joint mass function. Can we go the other way? That is, if we know the marginal mass functions for X and Y, can we determine the joint mass function? Well here's another possible outcome of the same survey from the previous example.

	L	R
M	3	45
F	10	42

It's easy to check that we get the same marginal mass functions in this modified example. So if we started with the marginals, how would we know whether the survey outcome was given by this table or the previous one? Since we can't really tell, it looks like the marginals don't determine the joint.

Let's be a little more specific. Suppose the marginals are specified by these equations

$$p_X(M) = p_{ML} + p_{MR} = 48/100$$
$$p_X(F) = p_{FL} + p_{FR} = 52/100$$
$$p_Y(L) = p_{ML} + p_{FL} = 13/100$$
$$p_Y(R) = p_{MR} + p_{FR} = 87/100.$$

Finding the joint mass function amounts to solving this system for the variables $p_{ML}, p_{MR}, p_{FL}, p_{FR}$. This is a linear system with four equations and four unknowns, so this sounds promising. The corresponding matrix equation is

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_{ML} \\ p_{MR} \\ p_{FL} \\ p_{FR} \end{bmatrix} = \begin{bmatrix} 48/100 \\ 52/100 \\ 13/100 \\ 87/100 \end{bmatrix}.$$

If we go through the usual procedure of row-reduction, the coefficient matrix reduces to

$$\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

This matrix doesn't have full rank, so the system does *not* have a unique solution. In particular, there isn't just one joint mass function corresponding to these marginals.

Let's move on to continuous random variables. Remember that every (real-valued) random variable X gives us a function $F_X : \mathbb{R} \to [0, 1]$ called its *(cumulative) distribution function*:

$$F_X(t) = \Pr[X \leq t]. \tag{2}$$

Likewise, if we have two random variables X and Y, we can define their *joint (cumulative) distribution function*.

Definition 1.6. Let X and Y be two random variables. Then their *joint cumulative distribution function*, $F : \mathbb{R}^2 \to [0, 1]$ is defined by

$$F(a, b) = \Pr[X \leq a, Y \leq b].$$

If there's any possibility for ambiguity, we might write $F_{X,Y}$ to remind us that F is the cumulative distribution function for X and Y.

How is the joint distribution function related to the *marginal* distribution functions of X and Y? Like in the discrete case, if we just specify that $X \leq a$, then we haven't put any restrictions on Y. This gives us

$$\begin{aligned} F_X(a) &= \Pr[X \leq a] \\ &= \Pr[X \leq a, Y < \infty]. \end{aligned} \tag{3}$$

Now the events $\{X \leq a, Y \leq t\}$ form an increasing sequence of events as t increases. That is, if $t_1 < t_2$, then we have the inclusion

$$\{X \leq a, Y \leq t_1\} \subseteq \{X \leq a, Y \leq t_2\}.$$

This is helpful because probabilities play nicely with increasing (or decreasing) sequences of events. Namely, if $E_1 \subseteq E_2 \subseteq \cdots$ is an increasing sequence of events, then

$$\Pr\left[\bigcup_{n=1}^{\infty} E_n\right] = \lim_{n \to \infty} \Pr[E_n].$$

Using this, (3) becomes

$$\begin{aligned} F_X(a) &= \Pr[X \leq a, Y < \infty] \\ &= \Pr\left[\bigcup_{b \geq 0} \{X \leq a, Y \leq b\}\right] \\ &= \lim_{b \to \infty} \Pr\left[X \leq a, Y \leq b\right] \\ &= \lim_{b \to \infty} F(a, b) \end{aligned}$$

The same idea tells us that

$$F_Y(b) = \lim_{a \to \infty} F(a, b).$$

When working with continuous random variables, we often work with their *density functions*. Specifically, if X is a continuous random variable, there is some function f such that for (pretty much)[1] any set $B \subseteq \mathbb{R}$,

$$\Pr[X \in B] = \int_B f(x)\ dx.$$

Here's the analogue for multiple variables.

Definition 1.7. Let X and Y be continuous random variables. We say X and Y have a *continuous joint distribution* if there is some function $f : \mathbb{R}^2 \to [0, 1]$ such that

$$\Pr[(X, Y) \in C] = \int_C f(x, y)\ dydx.$$

In this case, we call f the *joint probability density function (pdf)* of X and Y.

In the discrete case we were able to start with a joint mass function and sum over one of the variables to obtain the marginal of the other variable. Here's the analogue for continuous random variables.

Proposition 1.8. *Suppose X and Y are jointly continuous random variables with joint probability density function f. Then X and Y are continuous random variables with density functions*

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\ dy$$
$$f_Y(y) = \int_{\mathbb{R}} f(x, y)\ dx,$$

respectively.

Proof. Suppose $B \subseteq \mathbb{R}$ is measurable (don't worry too much about this assumption). Then

$$\begin{aligned}\Pr[X \in B] &= \Pr[X \in B, Y \in \mathbb{R}] \\ &= \int_B \left(\int_{\mathbb{R}} f(x, y)\ dy \right) dx.\end{aligned}$$

So the function

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\ dy$$

plays the role of the density function for X. The same idea gives the density function for Y. □

Example 1.9. Let X and Y be random variables with joint pdf

$$f(x, y) = \begin{cases} kxy, & \text{if } x, y \geq 0,\ x + y \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

where k is some constant.

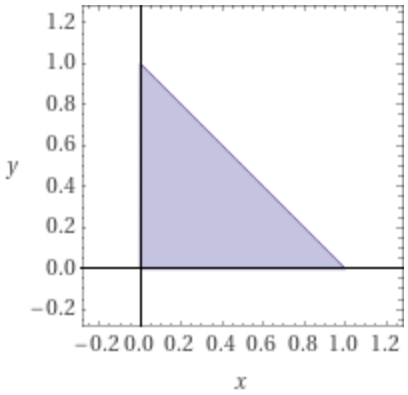

Figure 1: The joint density function $f(x, y)$ is nonzero in the shaded region.

Let's determine the actual value of k. We must have the following

$$1 = \Pr[(X, Y) \in \mathbb{R}^2] = \int_{\mathbb{R}^2} f(x, y)\ dydx,$$

so we're just going to have to evaluate this integral and then solve for k. It's usually a good idea to draw the region in question when computing double integrals like this. As x ranges from 0 to 1, y ranges from 0 to $1 - x$. To see this, draw a vertical slice upwards from any point on the x-axis until it intersects the line $x + y = 1$. Our integral then becomes

$$\begin{aligned} k \int_0^1 \int_0^{1-x} xy\ dydx &= k \int_0^1 x \left[\frac{1}{2}y^2\right]_{y=0}^{1-x} dx \\ &= \frac{k}{2} \int_0^1 x(1-x)^2\ dx \\ &= \frac{k}{2} \left[\frac{1}{4}x^4 - \frac{2}{3}x^3 + \frac{1}{2}x^2\right]_{x=0}^{x=1} \\ &= k/24. \end{aligned}$$

Since this expression must be equal to 1, we have $k = 24$.

Now let's compute the marginal pdf's of X and Y. To get the marginal for X, we "integrate out the y." For any fixed x, the value of y ranges between 0 and $1 - x$, so we have

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\ dy = \int_0^{1-x} kxy\ dy = \frac{k}{2}x(1-x)^2 = 12x(1-x)^2.$$

Similarly, for any fixed y, the value of x ranges between 0 and $1 - y$.

$$f_Y(y) = \int_{\mathbb{R}} f(x, y)\ dx = \int_0^{1-y} kxy\ dx = \frac{k}{2}y(1-y)^2 = 12y(1-y)^2.$$

Example 1.10. The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y}, & \text{if } 0 < x < \infty, 0 < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

[1]Technically, B needs to be what's called a *measurable* set. Pretty much any set you'd care about is measurable, but we need this restriction for the theory to hold up.

Let's compute $\Pr[X > 1, Y < 1]$. To compute the probability of *any* event, we simply integrate the joint density function over that event. In this case we have

$$\begin{aligned}\Pr[X > 1, Y < 1] &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}\, dxdy \\ &= \int_0^1 2e^{-2y}\left[-e^{-x}\right]_{x=1}^{x=\infty} dy \\ &= 2e^{-1}\int_0^1 e^{-2y}dy \\ &= e^{-1}(1-e^{-2}).\end{aligned}$$

In the case of a single random variable, the density and distribution functions are related by differentiation. That is, if X has density function f and distribution function F, then

$$f(x) = \frac{d}{dx}F(x).$$

The multivariable analogue is what you would probably expect. If X and Y are jointly continuous with density function $f(x, y)$ and distribution function $F(x, y)$, then we have by Fubini's theorem

$$f(x,y) = \frac{\partial^2}{\partial x \partial y}F(x,y) = \frac{\partial^2}{\partial y \partial x}F(x,y). \tag{4}$$

We aren't going to angst over the proof here, but this technically only holds for the values of (x, y) where the partial derivatives are defined and continuous.

1.2 Independent Random Variables

Remember that we said that two *events* A and B are independent if

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B].$$

We can carry this definition over to random variables.

Definition 1.11. Let X and Y be random variables. Then X and Y are *independent* if for any measurable sets A and B we have

$$\Pr[X \in A, Y \in B] = \Pr[X \in A] \cdot \Pr[Y \in B].$$

Let's show that this definition plays nicely with the machinery we defined in the previous section.

Proposition 1.12. *The discrete random variables X and Y, taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, are independent if and only if*

$$p(x, y) = p_X(x)p_Y(y) \tag{5}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Proof. First let's suppose that X and Y are independent. Then we can consider the singleton sets $\{x\}$ and $\{y\}$ for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

$$p(x, y) = \Pr[X = x, Y = y] = \Pr[X = x]\Pr[Y = y] = p_X(x)p_Y(y).$$

Now suppose that equation (5) holds. Then for any sets $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$ we have

$$\begin{aligned}\Pr[X \in A, Y \in B] &= \sum_{x \in A, y \in B} p(x,y) \\ &= \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) \\ &= \Pr[X \in A] \Pr[Y \in B],\end{aligned}$$

so X and Y are independent. □

Example 1.13. Suppose we perform $n+m$ independent trials, each having common success probability p. Let X be the number of successes in the first n trials and let Y be the number of successes in the next m trials. Are X and Y independent? Intuitively, knowing what happens in the first n trials shouldn't tell us anything about what happens in the next m trials, so we hope that X and Y are independent. Indeed, we have

$$\Pr[X = x, Y = y] = \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{y} p^y (1-p)^{m-y} = \Pr[X = x] \Pr[Y = y].$$

Let's define a new random variable Z to be the total number of successes in all $m+n$ trials. Are X and Z independent? Well if we know there are some successes in the first n trials, then we definitely know that there are at least that many successes in total, so we suspect that these aren't independent. We have

$$\begin{aligned}p(x,z) &= \Pr[x \text{ successes in the first } n \text{ trials, } z \text{ successes total}] \\ &= \Pr[x \text{ successes in the first } n \text{ trials, } z-x \text{ successes in the next } m \text{ trials}] \\ &= \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{z-x} p^{z-x} (1-p)^{m-z+x}.\end{aligned}$$

However,

$$\begin{aligned}p_X(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ p_Z(z) &= \binom{n+m}{z} p^z (1-p)^{n+m-z}.\end{aligned}$$

It's easily seen that the product of these two quantities does not match up with the previous quantity, so X and Z are *not* independent.

So we can check to see if two discrete random variables are independent by looking at their mass functions. What about continuous random variables? We should think of a discrete random variable's mass function as being analogous to a continuous random variable's density function, and this informs the next proposition.

Proposition 1.14. *If X and Y are continuous random variables, then they are independent if and only if*

$$f(x,y) = f_X(x) f_Y(y).$$

Proof. Suppose X and Y are independent. This is really a statement about *distribution* functions, not density functions, so we have

$$\Pr[X \leq x, Y \leq y] = \Pr[X \leq x] \Pr[Y \leq y], \tag{6}$$

which is equivalent to

$$F(x, y) = F_X(x)F_Y(y).$$

Now let's take the mixed x and y partial derivatives of both sides to obtain

$$\frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial^2}{\partial x \partial y}\big(F_X(x)F_Y(y)\big) = \frac{\partial}{\partial x} F_X(x) \frac{\partial}{\partial y} F_Y(y).$$

We arrived at the last equality using the fact that $F_X(x)$ is constant with respect to y and $F_Y(y)$ is constant with respect to x. Since differentiation distributions gives us densities, we have

$$f(x, y) = f_X(x)f_Y(y)$$

as desired.

Conversely, suppose that equation (6) holds. We pretty much copy the proof of the previous proposition with integrals in place of sums. For any sets A and B we have

$$\begin{aligned} \Pr[X \in A, Y \in B] &= \int_A \int_B f(x, y)\ dydx \\ &= \int_A f_X(x)\ dx \int_B f_Y(y)\ dy \\ &= \Pr[X \in A] \Pr[Y \in B]. \end{aligned}$$

□

So we have have independence if our joint density (or mass) function factors into the product of the marginals. We can actually say more – factoring into *any* product of functions, each depending on just one variable, is enough.

Proposition 1.15. *Let X and Y be continuous random variables. Then X and Y are independent if and only if the joint density function can be factored as*

$$f(x, y) = h(x)g(y) \tag{7}$$

for some functions g and h.

Proof. If X and Y are independent, then the previous proposition tells us that we can just take $g = f_X$ and $h = f_Y$.

Conversely, suppose that $f(x, y) = h(x)g(y)$. Then

$$\begin{aligned} 1 &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)\ dxdy \\ &= \int_{\mathbb{R}} h(x)\ dx \cdot \int_{\mathbb{R}} g(y)\ dy. \end{aligned}$$

But these last two integrals have to each be equal to some constants, c_1 and c_2, respectively. We also have

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f(x,y)\ dy \\ &= \int_{\mathbb{R}} h(x)g(y)\ dy \\ &= c_1 h(x). \end{aligned}$$

Similarly, $f_Y(y) = c_2 g(y)$. Putting it all together, we have

$$\begin{aligned} f_X(x)f_Y(y) &= c_1 h(x) \cdot c_2 g(y) \\ &= (c_1 c_2) h(x) g(y) \\ &= f(x,y). \end{aligned}$$

□

Example 1.16. Suppose X and Y are jointly continuous random variables with joint density

$$f(x,y) = \begin{cases} kxy, & \text{if } (x,y) \in [0,1]^2 \\ 0, & \text{otherwise.} \end{cases}$$

Are X and Y independent? They'd definitely be independent it we could factor the joint density, but how does the piecewise nature of the density play into this? Let's define the *indicator function* for the set our density actually lives on. That is, let $I(x,y)$ be defined by

$$I(x,y) = \begin{cases} 1, & \text{if } (x,y) \in [0,1]^2 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $I(x,y)$ indicates whether or not (x,y) lives in the set $[0,1]^2$, the unit square in $\mathbb{R}^2$. We go through the trouble of defining this function because we can use it to write our density as

$$f(x,y) = kxy \cdot I(x,y).$$

The kxy part is clearly a product of a function of only x with a function of only y. If we could factor the indicator function I into functions of just x and just y, then we will have shown independence. Our ability to do this is going to come down to the nature of the set $[0,1]^2$. Notice that $(x,y) \in [0,1]^2$ if and only if both coordinates live in $[0,1]$. That is, if we define $\tilde{I}(x)$ by

$$\tilde{I}(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{otherwise,} \end{cases}$$

then we have $I(x,y) = \tilde{I}(x)\tilde{I}(y)$, so our joint density factors as

$$f(x,y) = (kx \cdot \tilde{I}(x))(y \cdot \tilde{I}(y)),$$

so X and Y are independent.

Example 1.17. Suppose X and Y are jointly continuous random variables with joint density

$$f(x,y) = \begin{cases} kxy, & \text{if } x,y \geq 0, x+y \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Are X and Y independent? This density looks a lot like the one from the previous exercise. The difference here is that $f(x,y)$ behaves like kxy on a different set this time. Now f lives on a triangle in the first quadrant rather than an axis-aligned square. Proceeding in the same way as before, if we define

$$I(x,y) = \begin{cases} 1, & \text{if } x,y \geq 0, x+y \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

Then we still have $f(x,y) = kxy \cdot I(x,y)$ like before. Can we still factor it? Well if we specify the x coordinate to live between 0 and 1, then the y coordinate needs to satisfy $0 \leq y \leq 1-x$. If we define the functions I_1 and I_2 by

$$I_1(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{otherwise,} \end{cases} \qquad I_2(x,y) = \begin{cases} 1, & \text{if } 0 \leq y \leq 1-x \\ 0, & \text{otherwise,} \end{cases}$$

then we do get the factorization

$$f(x,y) = kxy \cdot I_1(x) \cdot I_2(x,y),$$

but this isn't helpful since $I_2(x,y)$ is a function of both x and y.

This is *not* a proof that X and Y aren't independent. For all we know, there's some weird factorization of $f(x,y)$ we just haven't found yet. Let's approach it a bit differently. This is the same density from Example 1.9 and we found the marginal density functions to be

$$f_X(x) = \begin{cases} (k/2)x(1-x)^2, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise,} \end{cases} \qquad f_Y(y) = \begin{cases} (k/2)y(1-y)^2, & \text{if } 0 \leq y \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Notice that we do *not* have that $f(x,y) = f_X(x)f_Y(y)$ for all x, y where these three functions are defined, so X and Y are *not* independent.

1.3 Sums of Independent Random Variables

Say we have two real-valued random variables X and Y. We'll assume they're discrete for now. Their sum $Z = X + Y$ is clearly a random variable as well. How do the mass functions of X and Y, p_X and p_Y, and the joint mass function p relate to the mass function of $X + Y$? Well $p_Z(z) = \Pr[Z = z]$ and we can break the event $\{Z = z\}$ into the events

$$\{Z = z\} = \bigcup_x \{X = x, Y = z - x\}.$$

To see this, note that in order for $X + Y = z$ to be true, X can be anything so long as $Y = z - X$. Moreover, these events are clearly disjoint since X and Y can only take one value at a time. Since the probability of a *disjoint* union is just the sum of the probabilities of the constituent events, we have

$$p_Z(z) = \sum_x \Pr[X = x, Y = z - x] = \sum_x p(x, z - x).$$

Now if X and Y are independent, the joint density factors as $p(x,y) = p_X(x)p_Y(y)$ and we've proven the following proposition.

Proposition 1.18. *If X and Y are discrete random variables, then the probability mass function of $Z = X + Y$ is*

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x).$$

Example 1.19. Suppose X and Y are independent random variables, both taking values in $\{1, 2, \ldots, n\}$ uniformly at random (if $n = 6$, then you can think X and Y as the outcomes of dice rolls). If we set $Z = X + Y$, then the previous proposition tells us that

$$p_Z(z) = \sum_{j=1}^{n} p_X(j) p_Y(z - j).$$

Now it might be tempting to just set $p_X(j)$ and $p_Y(z - j)$ to $1/n$. If this were the case, then we would have $p_Z(z) = 1/n$ for each z. But this definitely doesn't line up with our intuition – when we roll two dice, some outcomes are more likely than others (there's only one way to roll a 2, but six ways to roll a 7). The problem is that $p_Y(z - j)$ isn't always $1/n$. Indeed, if $z - j < 1$, then Y never takes the value $z - j$.

We can fix this by looking at the conditions that make $p_X(j)$ and $p_Y(z - j)$ *both* positive. In order for this to happen, we need $1 \leq j \leq n$ *and* $1 \leq z - j \leq n$ to both hold. Isolating j gives

$$z - n \leq j \leq z - 1 \qquad \text{and} \qquad 1 \leq j \leq n.$$

If z is between 2 and n this becomes

$$p_Z(z) = \sum_{j=1}^{z-1} p_X(j) p_Y(z - j) = \sum_{j=1}^{z-1} \frac{1}{n^2} = \frac{z - 1}{n^2}.$$

On the other hand, if z is between $n + 1$ and $2n$ we have

$$p_Z(z) = \sum_{j=z-n}^{n} p_X(j) p_Y(z - j) = \sum_{j=z-n}^{n} \frac{1}{n^2} = \frac{2n - z + 1}{n^2}.$$

So in total we have

$$p_Z(z) = \begin{cases} (z - 1)/n^2, & \text{if } 2 \leq z \leq n \\ (2n - z + 1)/n^2, & \text{if } n + 1 \leq z \leq 2n. \end{cases}$$

Importantly, we have that the sum of two uniform random variables is *not* another uniform random variable.

Example 1.20. If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, let's compute the distribution of $X + Y$.

We have that

$$\begin{aligned} \Pr[X + Y = n] &= \sum_{k=0}^{n} \Pr[X = k, Y = n - k] \\ &= \sum_{k=0}^{n} \Pr[X = k] \Pr[Y = n - k] \\ &= \sum_{n=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n - k)!} \\ &= e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n - k)!}. \end{aligned}$$

Now the sum at the end should remind us of the binomial theorem since it has a product of two terms whose powers sum to n. We almost have the correct binomial coefficient too. We just need to multiply and divide by $n!$.

$$\begin{aligned}\Pr[X+Y=n] &= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n.\end{aligned}$$

This is the mass function of a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

Let's look at the case of continuous random variables. The analogy (mass functions $\iff$ densities) and (sums $\iff$ integrals) leads us to the following proposition.

Proposition 1.21. *Suppose X and Y are jointly continuous real-valued random variables with joint density function f. Then the variable $Z = X + Y$ has density function*

$$f_Z(z) = \int_{\mathbb{R}} f(x, z-x)\ dx = \int_{\mathbb{R}} f(z-y, y)\ dy.$$

In particular, if X and Y are independent, then this becomes

$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\ dx = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\ dy.$$

Proof. Let's look at the distribution function for Z. That is, for any real a we have

$$F_Z(a) = \Pr[Z \le a] = \Pr[X+Y \le a] = \int_{x+y \le a} f(x,y)\ dxdy.$$

Upon looking at a diagram of this region, we turn this into an iterated integral.

$$F_Z(a) = \int_{\mathbb{R}} \int_{-\infty}^{a-x} f(x,y)\ dxdy.$$

Now if we do the substitution $y = z - x$, this becomes

$$F_Z(a) = \int_{\mathbb{R}} \int_{-\infty}^{a} f(x, z-x)\ dzdx = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(x, z-x)\ dxdz.$$

We used Fubini's theorem to switch the order of integration at the end. Now we can use the fundamental theorem of calculus to take the derivative of both sides with respect to a to get the probability density function of Z,

$$f_Z(z) = \int_{\mathbb{R}} f(x, z-x)\ dx.$$

Now if X and Y are independent, we can factor the joint density function to get $f(x, z-x) = f_X(x) f_Y(z-x)$. □

Example 1.22. Let's do the continuous version of the previous example. That is, suppose X and Y are independent random variables taking values in the interval $[0,1]$ uniformly. By the above proposition, the density of $Z = X + Y$ is given by

$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\, dx.$$

To actually compute this integral, we need to know the values of x that make $f_X(x)$ and $f_Y(z-x)$ positive. Since the density of the uniform distribution is given by

$$f_X(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{otherwise,} \end{cases}$$

we have that $f_X(x)$ is positive if and only if $0 \leq x \leq 1$ and $f_Y(z-x)$ is positive if and only if $0 \leq z - x \leq 1$. When we combine these, we see that the values of x that make both inequalities true depend on what z is. In particular, when $0 \leq z \leq 1$, we need $0 \leq x \leq z$, and when $1 \leq z \leq 2$, we need $z \leq x \leq 1$. So if $0 \leq z \leq 1$, we have

$$f_Z(z) = \int_0^z 1\, dz = z$$

and when $1 \leq z \leq 2$ we have

$$f_Z(z) = \int_z^1 1\, dz = 1 - z.$$

Example 1.23. Let X and Y be independent standard normal random variables (that is, they both have mean 0 and variance 1). Recall that the density of X is then

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Let's compute the density of the sum $Z = X + Y$. We have

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_X(x) f_Y(z-x)\, dx \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(z-x)^2/2}\, dx \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-x^2 + zx - z^2/2}\, dx. \end{aligned}$$

At this point, we complete the square in the exponent.

$$-x^2 + zx - z^2/2 = -x^2 + zx - z^2/4 + z^2/4 - z^2/2 = -(x - z/2)^2 - z^2/4.$$

So the density becomes

$$f_Z(z) = \frac{1}{2\pi} e^{-z^2/4} \int_{\mathbb{R}} e^{-(x-z/2)^2} dx = \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} e^{-z^2/4}.$$

Here we've used the fact that $\int_{\mathbb{R}} e^{-x^2}\, dx = \sqrt{\pi}$ along with a simple substitution. In class, I mistakenly said this integral was $\sqrt{2\pi}$. Note that this is the density of a normal random variable having mean 0 and variance 2.

So the sum of two independent standard normal random variables is again a normal random variable. More generally, the sum of arbitrarily many independent normal random variables is again a normal random variable. The proof of this more general fact is pretty much the same, but the algebra is a little messier.

Theorem 1.24. *If $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean μ_i and variance σ_i^2, respectively, then the sum $\sum_{i=1}^n X_i$ is normally distributed with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$.*

Example 1.25. The number of candies in a standard bag of plain M&M's is normally distributed with a mean of 55 candies and a standard deviation of 2 candies. The number of candies in a sharing size bag of plain M&M's is also normally distributed with a mean of 340 candies and a standard deviation of 3 candies. What is the probability that six standard bags of M&M's together contain more candies than one sharing bag?

Let $X_1, \ldots, X_6$ be the number of candies in the six standard bags and let Y be the number of candies in the sharing sized bag. Then we want the probability that $X_1 + \cdots + X_6 - Y > 0$. By the above theorem, this sum is a normal random variable with mean $6 \cdot 55 - 340 = -10$ candies and variance $6 \cdot 2 + 3 = 15$.

We can compute the desired probability by using a z-table as follows. If we let $W = X_1 + \cdots + X_6 - Y$, then

$$\Pr[W > 0] = \Pr\left[\frac{W - (-10)}{\sqrt{15}} > \frac{0 - (-10)}{\sqrt{15}}\right].$$

Now $Z = (W + 10)/\sqrt{15}$ is a standard normal random variable, and a z-table lets us look up the probability that such a random variable is less that t for many values of t. Using table 5.1 in the textbook, we see that $\Pr[Z < 10/\sqrt{15}] \approx \Pr[Z < 2.58] \approx .9951$. Thus,

$$\Pr[Z > 2.58] \approx 1 - .9951 = .0049.$$

1.4 Conditional Distributions – Discrete Random Variables

Let's briefly recall the basics of conditional probability. If E and F are events, then $\Pr[E \mid F]$ (read "the probability of E given F") is defined to be

$$\Pr[E \mid F] = \frac{\Pr[E \cap F]}{\Pr[F]}.$$

Intuitively, $\Pr[E \mid F]$ is the probability that E happens with the additional information that F happened. The above definition quantitatively captures this idea when you think of it as shrinking your probability space to just F. If the events E and F are independent, then we have

$$\Pr[E \mid F] = \frac{\Pr[E \cap F]}{\Pr[F]} = \frac{\Pr[E] \cdot \Pr[F]}{\Pr[F]} = \Pr[E].$$

Intuitively this makes sense – if E and F are independent, then learning that E happened shouldn't tell you anything about whether or not F happened. Quantitatively, this means that $E \cap F$ makes up the same fraction of F as E does in the original probability space.

Now let's think about conditional probability in the context of random variables.

Definition 1.26. Let X and Y be discrete random variables. Then the *conditional probability mass function of X, given that $Y = y$* is

$$p_{X|Y}(x \mid y) = \Pr[X = x \mid Y = y] = \frac{p(x, y)}{p_Y(y)}.$$

This definition corresponds to plugging the events $\{X = x\}$ and $\{Y = y\}$ into the definition of conditional probability.

Example 1.27. Say 100 people are asked for their handedness (right-handed or left-handed) and sex (male or female). The survey produces the following table.

	L	R
M	4	44
F	9	43

Select a person from this population at random and let X be their handedness and Y be their sex. Let's find the conditional pmf of X given that we selected a female.

$$p_{X|Y}(L \mid F) = \frac{\frac{9}{100}}{\frac{9}{100} + \frac{43}{100}} = \frac{9}{52}$$
$$p_{X|Y}(R \mid F) = \frac{\frac{43}{100}}{\frac{9}{100} + \frac{43}{100}} = \frac{43}{52}.$$

Example 1.28. Suppose X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2. Let's find the conditional distribution of X given that $X + Y = n$.

It's important to note that just because X and Y are independent, it does not follow that X and $X + Y$ are independent. Indeed, if we know that $X = n$, then $X + Y$ must be at least n. Now we have that

$$\begin{aligned}\Pr[X = k \mid X + Y = n] &= \frac{\Pr[X = k, X + Y = n]}{\Pr[X + Y = n]} \\ &= \frac{\Pr[X = k, Y = n - k]}{\Pr[X + Y = n]} \\ &= \frac{\Pr[X = k] \cdot \Pr[Y = n - k]}{\Pr[X + Y = n]}.\end{aligned}$$

Now it looks like we need to know the distribution of the sum $X + Y$. We figured this out in Example 1.20 where we saw what $X + Y \sim Pois(\lambda_1 + \lambda_2)$, so we have

$$\begin{aligned}\Pr[X = k \mid X + Y = n] &= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^n} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}.\end{aligned}$$

This is a binomial distribution with n trials and success probability $\lambda_1/(\lambda_1 + \lambda_2)$.

If X and Y are discrete random variables, we can condition on a particular outcome $Y = y$ to obtain a new random variable. That is, the random variable $(X \mid Y = y)$ is itself a random variable. In particular, we can compute its expected value.

$$E[X \mid Y = y] = \sum_x x \cdot p_{X|Y}(x \mid y).$$

1.5 Conditional Distributions – Continuous Random Variables

Using the mass function $\iff$ density function analogy, we arrive at the following definition.

Definition 1.29. Let X and Y be continuous random variables with joint probability density function f. Then the *conditional probability density function of X given that $Y = y$ is*

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}.$$

Of course, this only makes sense for the values of y where $f_Y(y) > 0$.

To see why this definition is "correct," remember that $f_X(x)\,dx$ is roughly the probability that X lies between x and $x + dx$. In particular,

$$\begin{aligned} f_{X|Y}(x \mid y)\,dx &= \frac{f(x,y)\,dxdy}{f_Y(y)\,dy} \\ &\approx \frac{\Pr[x \le X \le x+dx, y \le Y \le y+dy]}{\Pr[y \le Y \le y+dy]} \\ &= \Pr[x \le X \le x+dx \mid y \le Y \le y+dy]. \end{aligned}$$

Just as in the case of discrete random variables, we can condition X on $Y = y$ and take the expectation.

$$E[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x \mid y)\,dx.$$

Remark 1.30. If X and Y are independent continuous random variables, then

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

and

$$E[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x \mid y)\,dx = \int_{\mathbb{R}} x f_X(x)\,dx = E[X].$$

That is, if X and Y are independent, conditioning X on the event $Y = y$ "doesn't do anything."

Example 1.31. The joint density of X and Y is given by

$$f(x,y) = \frac{12}{5}x(2-x-y) \qquad (x,y) \in [0,1]^2.$$

Let's compute the conditional density of X given that $Y = y$ then compute $E[X \mid Y = 1/3]$.

Since $f_{X|Y}(x \mid y) = f(x,y)/f_Y(y)$, it looks like we need to find the marginal density function $f_Y(y)$. We can find this by "integrating out the x".

$$\begin{aligned} f_Y(y) &= \int_0^1 f(x,y)\,dx \\ &= \int_0^1 \frac{12}{5}x(2-x-y)\,dx \\ &= \cdots \\ &= \frac{12}{5}\left(\frac{2}{3} - \frac{y}{2}\right). \end{aligned}$$

So we have

$$f_{X|Y}(x \mid y) = \frac{x(2-x-y)}{\frac{2}{3} - \frac{y}{2}} = \frac{6x(2-x-y)}{4-3y}$$

for any $x, y \in [0,1]^2$.

Now for the conditional expectation.

$$\begin{aligned} E[X \mid Y = 1/3] &= \int_0^1 x f_{X|Y}(x \mid 1/3)\ dx \\ &= \int_0^1 \frac{6x^2(\frac{5}{3} - x)}{3}\ dx \\ &= \cdots \\ &= \frac{11}{10}. \end{aligned}$$

Example 1.32. A particle with mass 1 splits into a smaller particle and some energy, where the mass of the smaller particle is a uniform $[0,1]$ random variable. The smaller particle then splits in the same way. What is the distribution of the mass of the final particle?

Let Y be the mass of the particle after the first split and let X be the mass of the final particle. Then $Y \sim unif(0,1)$ and $(X \mid Y = y) \sim unif(0,y)$. So we know these density functions

$$f_Y(y) = 1 \quad \text{if } y \in [0,1], \qquad f_{X|Y}(x \mid y) = \frac{1}{y} \quad \text{if } x \in [0,y].$$

Since we know the conditional and marginal densities, we can reconstruct the joint density.

$$f(x,y) = f_Y(y) \cdot f_{X|Y}(x \mid y) = \frac{1}{y} \quad \text{if } 0 \le x \le y,\ 0 \le x \le y.$$

Now we can get the density of x by "integrating out the y."

$$f_X(x) = \int_x^1 \frac{1}{y}\ dy = -\ln x \quad \text{for } x \in (0,1].$$

Example 1.33. The lifetime of a light bulb has conditional distribution $Exp(\Lambda)$, where $\Lambda \sim unif(a,b)$ (maybe the fuse in the bulb is randomly selected). Find the marginal distribution of the lifetime of the light bulb.

We're given the following densities

$$f_{X|\Lambda}(x \mid \lambda) = \lambda e^{-\lambda x} \quad \text{if } x \ge 0 \qquad f_\Lambda(\lambda) = \frac{1}{b-a} \quad \text{if } a \le \lambda \le b.$$

We want the density $f_X(x)$. Like in the previous example, we use the marginal and conditional densities to reconstruct the joint density.

$$f_{X,\Lambda}(x,\lambda) = f_\Lambda(\lambda) f_{X|\Lambda}(x \mid \lambda) = \frac{1}{b-a}\lambda e^{-\lambda x} \quad \text{if } a \le \lambda \le b,\ x \ge 0.$$

Now we can find the density of X by integrating out the λ.

$$\begin{aligned} f_X(x) &= \int_a^b \frac{\lambda e^{-\lambda x}}{b-a}\ d\lambda \\ &= \cdots \\ &= \frac{e^{-ax}(1+ax) - e^{-bx}(1+bx)}{x^2(b-a)}, \end{aligned}$$

for any $x \ge 0$.

References

[1] Ross, Sheldon. *A First Course in Probability*. Ninth Edition. Pearson. 2014.