Math 130B

Liam Hardiman

September 2, 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]. Some chunks of exposition and some of the examples come from Anna Ma's notes from when I TA'd under her in the Spring of 2022. 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.

$$\begin{aligned} p(M, L) &= 4/100 & \qquad p(M, R) &= 44/100 \\ p(F, L) &= 9/100 & \qquad p(F, R) &= 43/100. \end{aligned}$$

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

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

For Y we have

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

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

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

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.

[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.

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.

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,

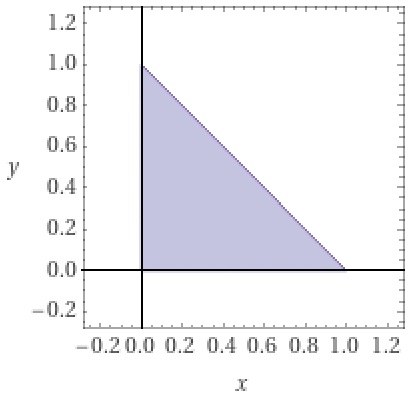

Figure 1: The joint density function $f(x,y)$ is nonzero in the shaded region.

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

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

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

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

□

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 $\{1, 2, \dots, n\}$ 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 $1 \le j \le n$ *and* $1 \le z - j \le n$ to both hold. Isolating j gives

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

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

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

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

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

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.

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

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 $X + Y \sim \mathit{Pois}(\lambda_1 + \lambda_2)$, 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 $\Longleftrightarrow$ 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 $x \ge 0$.

1.6 Joint Probability Distribution of Functions of Random Variables

Suppose $X \sim \text{Unif}[0,1]$. What is the distribution of $Y = X^2$? The range of Y is still $[0,1]$, but we shouldn't expect Y to be a uniform random variable – "most" numbers in $[0,1]$ have small squares. We can directly compute the distribution function.

$$\begin{aligned} F_Y(t) &= \Pr[Y \le t] \\ &= \Pr[X^2 \le t] \\ &= \Pr[X \le \sqrt{t}] \\ &= \sqrt{t}. \end{aligned}$$

Differentiating both sides gives the density of Y, $f_Y(t) = \frac{1}{2\sqrt{t}}$. In total, we inverted the function $y = x^2$ and then took a derivative.

Proposition 1.34. *Let X be a continuous random variable and let $Y = g(X)$ for some continuously differentiable function g. Then the density of Y is given by*

$$f_Y(y) = \frac{f(x)}{|dg/dx|},$$

where $x = g^{-1}(y)$. (Technically, this holds for x such that $|dg/dx|$ is nonzero)

Proof. We repeat the previous example, just in more generality. Start by computing the distribution function.

$$\begin{aligned} F_Y(y) &= \Pr[Y \le y] \\ &= \Pr[g(X) \le y] \\ &= \Pr[X \le g^{-1}(y)] \\ &= F_X(g^{-1}(y)). \end{aligned}$$

Now we differentiate both sides, recalling that $[g^{-1}(y)]' = 1/g'(x)$ and $x = g^{-1}(y)$.

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(x)} = \frac{f_X(x)}{|dg/dx|}.$$

□

Example 1.35. Suppose $X \sim \mathcal{N}(0,1)$. Let's find the density function of $Y = X^2$. In terms of the above proposition, $g(x) = x^2$ and we have

$$f_Y(y) = \frac{f_X(x)}{g'(x)} = \frac{1}{2x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Now we should get the right-hand side in terms of y. Since $y = x^2$ here, $x = \sqrt{y}$ and we have

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

For those familiar with statistics, Y is a χ^2 random variable with one degree of freedom.

The multivariable version of the above proposition is spiritually the same and has pretty much the same proof.

Theorem 1.36. *Suppose X_1 and X_2 are continuous random variables with joint density function f_{X_1,X_2}. Then if g_1 and g_2 are continuously differentiable functions $\mathbb{R}^2 \to \mathbb{R}$, and*

$$Y_1 = g_1(X_1, X_2) \qquad \textit{and} \qquad Y_2 = g_2(X_1, X_2),$$

then the joint density of Y_1 and Y_2 is given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J(x_1, x_2)|^{-1},$$

where J is the Jacobian determinant,

$$J(x_1, x_2) = \det \begin{bmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 \end{bmatrix}.$$

Example 1.37. Suppose X_1 and X_2 are independent exponential random variables with parameter $\lambda = 1$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Let's find the joint density function of Y_1 and Y_2.

In the context of Theorem 1.36, $g_1(x_1, x_2) = x_1 + x_2$ and $g_2(x_1, x_2) = x_1 - x_2$. Our Jacobian determinant is

$$J(x_1, x_2) = \det \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = -2.$$

Consequently,

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2)|J(x_1,x_2)|^{-1} = f_{X_1}(x_1)f_{X_2}(x_2)\cdot\frac{1}{2} = \frac{e^{-x_1-x_2}}{2} = \frac{e^{-y_1}}{2}.$$

We used the independence of X_1 and X_2 to split the joint density of X_1 and X_2. Now what values of y_1 and y_2 does this hold for? To do this, we find x_1 and x_2 in terms of y_1 and y_2. We can easily check that

$$x_1 = \frac{y_1+y_2}{2} \qquad \text{and} \qquad x_2 = \frac{y_1-y_2}{2}.$$

Since $0 \le x_1 \le 1$ and $0 \le x_2 \le 1$, we have

$$0 \le y_1 + y_2 \le 2 \qquad \text{and} \qquad 0 \le y_1 - y_2 \le 2.$$

How would we find the marginal density functions of Y_1 and Y_2?

Example 1.38. Suppose we pick a random point (X, Y) in the plane by choosing its coordinates one at a time, independently from a standard normal distribution. Let's find the distribution of the point's polar coordinates.

Remember that the polar coordinates of (X, Y) are

$$R = \sqrt{X^2+Y^2}, \qquad \Theta = \arctan(Y/X),$$

when X and Y are both positive (we need to change Θ when we're in the other quadrants). In the context of Theorem 1.36, we'll let $g_1(x, y) = \sqrt{x^2+y^2}$ and $g_2(x, y) = \arctan(y/x)$. Our Jacobian determinant is then

$$J(x,y) = \det\begin{bmatrix} x/\sqrt{x^2+y^2} & y/\sqrt{x^2+y^2} \\ -y/(x^2+y^2) & x/(x^2+y^2)\end{bmatrix} = 1/\sqrt{x^2+y^2}.$$

Now the conditional joint density of X and Y, *given that X and Y are both positive*, is

$$f(x, y \mid X > 0, Y > 0) = \frac{f(x,y)}{\Pr[X>0, Y>0]} = \frac{f(x,y)}{4} = \frac{2}{\pi}e^{-(x^2+y^2)/2},$$

provided that $x, y > 0$. By our theorem, the conditional joint density of R and Θ, given that X and Y are both positive, is

$$f_{R,\Theta}(r,\theta \mid X>0, Y>0) = f(x,y \mid X>0, Y>0)|J(x,y)|^{-1} = \frac{2}{\pi}\sqrt{x^2+y^2}e^{-(x^2+y^2)/2}.$$

In terms of r and θ, this becomes

$$f_{R,\Theta}(r,\theta \mid X>0, Y>0) = \frac{2}{\pi}re^{-r^2/2},$$

provided that x and y are both positive, i.e., that $r > 0$ and $0 \le \theta \le \pi/2$.

The exact same argument gives us the conditional distributions for the other possible values of X and Y. The only difference is that in each quadrant, we have to slightly modify our expression for θ. For example, in quadrant II where $x < 0$ and $y > 0$, $\theta = \pi - \arctan(y/x)$. Importantly, this doesn't change the value of $|J(x, y)|$. In total, we have

$$\begin{aligned} f_{R,\Theta}(r,\theta \mid X<0, Y>0) &= \frac{2}{\pi}re^{-r^2/2} \\ f_{R,\Theta}(r,\theta \mid X<0, Y<0) &= \frac{2}{\pi}re^{-r^2/2} \\ f_{R,\Theta}(r,\theta \mid X>0, Y<0) &= \frac{2}{\pi}re^{-r^2/2}. \end{aligned}$$

Putting it all together, we see that

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi} r e^{-r^2/2}, \quad 0 < \theta < 2\pi,\ 0 < r < \infty.$$

In particular, since the joint density of R and Θ factors into a product of functions, one depending on r and the other on θ (really, there is no function of θ since the density doesn't depend on θ at all), we see that R and Θ are independent!

Example 1.39. Suppose we really want to sample from a standard normal distribution, but we only know how to generate uniform $[0,1]$ random variables. We'll use the previous example as our starting point. If we could somehow generate the variables R and Θ (as in the previous example), then since

$$X = R\cos\Theta \qquad \text{and} \qquad Y = R\sin\Theta,$$

we could generate X and Y, which are independent standard normal random variables. It's definitely easy to generate Θ, since this variable is $\mathrm{Unif}(0, 2\pi)$. Unfortunately, it's not super clear how to generate R.

If, in the previous example, we instead looked for the joint density of $R^2 = X^2 + Y^2$ and $\Theta = \arctan(Y/X)$ (assuming for now that $X, Y \geq 0$), then we should take $d = g_1(x,y) = x^2 + y^2$ and $\theta = g_2(x,y) = \arctan(y/x)$ in Theorem 1.36. In this case, our Jacobian becomes.

$$J(x,y) = \det \begin{bmatrix} 2x & 2y \\ -y/(x^2+y^2) & x/(x^2+y^2) \end{bmatrix} = 2.$$

Following the same procedure as before, we get the joint density

$$f_{R^2,\Theta}(d,\theta) = f(x,y)|J(x,y)|^{-1} = \frac{1}{2}e^{-d/2} \cdot \frac{1}{2\pi}.$$

In particular, R^2 and Θ are independent and R^2 is an exponential random variable with parameter $1/2$.

Now we claim that we can easily generate R^2. Suppose U_1 is a $\mathrm{Unif}[0,1]$ random variable and consider the quantity $-2\ln U_1$. Since U_1 is a $\mathrm{Unif}[0,1]$ variable, its distribution function is $F_{U_1}(u_1) = u_1$ and we have

$$\begin{aligned} \Pr[-2\ln U_1 \leq u_1] &= \Pr[U_1 \geq e^{-u_1/2}] \\ &= 1 - e^{-u_1/2}. \end{aligned}$$

If we differentiate both sides with respect to u_1, we see that the density of $-2\ln U_1$ is $\frac{1}{2}e^{-U_1/2}$, so it's an exponential random variable with parameter $1/2$ – just like R^2. Putting it all together, if U_2 is a $\mathrm{Unif}[0,1]$ random variable independent from U_1, then

$$X = \sqrt{-2\ln U_1}\cos(2\pi U_2), \qquad Y = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

are independent standard normal random variables.

2 Properties of Expectation

2.1 Introduction

Recall the definitions of expected value.

Definition 2.1. If X is a discrete real-valued random variable with probability mass function p, then its expected value (also called its expectation) is defined to be

$$E[X] = \sum_x xp(x).$$

If X is a continuous random variable with density f, then its expected value is defined to be

$$E[X] = \int_{\mathbb{R}} xf(x)\ dx.$$

One basic fact is that if the (discrete) variable X takes values between a and b with probability 1, then

$$\begin{aligned} E[X] &= \sum_{x:p(x)>0} xp(x) \\ &\geq \sum_{x:p(x)>0} ap(x) \\ &= a \sum_{x:p(x)>0} p(x) \\ &= a. \end{aligned}$$

The same idea shows that $E[X] < b$ and the same argument applies to continuous random variables.

2.2 Expectation of Sums of Random Variables

Let's recall a couple of basic facts about expectation and then generalize them to multiple random variables.

Lemma 2.2. *If X is a continuous random variable, then*

$$E[X] = \int_0^\infty \Pr[X > t]\ dt - \int_{-\infty}^0 \Pr[X < t]\ dt.$$

Proof. If we assume that X has density f, then we just change the order of integration (this is called Fubini's theorem).

$$\begin{aligned} \int_0^\infty \Pr[X > t]\ dt - \int_{-\infty}^0 \Pr[X < t]\ dt &= \int_0^\infty \int_t^\infty f(x)\ dxdt - \int_{-\infty}^0 \int_{-\infty}^t f(x)\ dxdt \\ &= \int_0^\infty f(x) \int_0^x dtdx - \int_{-\infty}^0 f(x) \int_x^0 dtdx \\ &= \int_0^\infty xf(x)\ dx - \int_{-\infty}^0 (-x)f(x)\ dx \\ &= E[X]. \end{aligned}$$

□

Using this lemma, we can prove the following obvious-looking theorem.

Theorem 2.3. *If X is a continuous random variable with density f and $g : \mathbb{R} \to \mathbb{R}$ is a continuous function, then*

$$E[g(X)] = \int_{\mathbb{R}} g(x)f(x)\ dx.$$

Proof. The idea is to use the previous lemma on the random variable $g(X)$.

$$\begin{aligned}
E[g(X)] &= \int_0^\infty \Pr[g(X) > t]\ dt - \int_{-\infty}^0 \Pr[g(X) < t]\ dt \\
&= \int_0^\infty \int_{x:g(x)>t} f(x)\ dxdt - \int_{-\infty}^0 \int_{x:g(x)<t} f(x)\ dxdt \\
&= \int_{x:g(x)>0} f(x) \int_0^{g(x)} dt - \int_{x:g(x)<0} f(x) \int_{g(x)}^0 dt \\
&= \int_{x:g(x)>0} g(x)f(x)\ dx + \int_{x:g(x)<0} g(x)f(x)\ dx \\
&= \int_{\mathbb{R}} g(x)f(x)\ dx.
\end{aligned}$$

□

This theorem maybe looks obvious since taking $g(x) = x$ just gives us the definition of expectation. This is why this theorem is sometimes called *LOTUS (the law of the unconscious statistician)*. We also remark that this theorem is definitely true in the case of discrete random variables. If X and Y are discrete with joint pmf p, then the expected value of $g(X, Y)$ is the sum over all (x, y) of $g(x, y)$ times the probability that $X = x$ and $Y = y$. But this is just

$$\sum_{x,y} g(x, y)p(x, y).$$

The same-ish proof of the above theorem gives us an analogue of this theorem for multiple variables.

Theorem 2.4. *If X and Y are continuous random variables with joint density f and $g : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function, then*

$$E[g(X, Y)] = \int_{\mathbb{R}^2} g(x, y)f(x, y)\ dydx.$$

Proof. We apply Lemma 2.2 to the variable $g(X, Y)$

$$\begin{aligned}
E[g(X, Y)] &= \int_0^\infty \Pr[g(X, Y) > t]\ dt - \int_{-\infty}^0 \Pr[g(X, Y) < t]\ dt \\
&= \int_0^\infty \int_{(x,y):g(x,y)>t} f(x, y)\ dydxdt - \int_{-\infty}^0 \int_{(x,y):g(x,y)<t} f(x, y)\ dydxdt \\
&= \int_{(x,y):g(x,y)>0} f(x, y) \int_0^{g(x,y)} dt - \int_{(x,y):g(x,y)<0} f(x, y) \int_{g(x,y)}^0 dt \\
&= \int_{(x,y):g(x,y)>0} g(x, y)f(x, y)dydx + \int_{(x,y):g(x,y)<0} g(x, y)f(x, y)\ dydx \\
&= \int_{\mathbb{R}^2} g(x, y)f(x, y)\ dydx.
\end{aligned}$$

□

This theorem allows us to prove another seemingly obvious result.

Proposition 2.5 (Linearity of Expectation)**.** *If X and Y are random variables, then*

$$E[X+Y] = E[X] + E[Y].$$

Proof. If X and Y are discrete, then

$$\begin{aligned} E[X+Y] &= \sum_{x,y}(x+y)p(x,y) \\ &= \sum_{x,y} xp(x,y) + \sum_{x,y} yp(x,y) \\ &= \sum_{x} xp_X(x) + \sum_{y} yp_Y(y) \\ &= E[X] + E[Y]. \end{aligned}$$

In the continuous case, we just apply the previous theorem and do the same thing.

$$\begin{aligned} E[X+Y] &= \int_{\mathbb{R}^2} (x+y)f(x,y)\ dydx \\ &= \int_{\mathbb{R}^2} xf(x,y)\ dydx + \int_{\mathbb{R}^2} yf(x,y)\ dydx \\ &= \int_{\mathbb{R}} xf_X(x)\ dx + \int_{\mathbb{R}} yf_Y(y)\ dy \\ &= E[X] + E[Y]. \end{aligned}$$

□

Note that this proposition doesn't require independence!

Example 2.6. Suppose n married couples are living in a town and m deaths occur at random. What is the expected number of "intact" couples?

Let X be the number of intact couples. Instead of computing $E[X]$ directly, we'll break X up into smaller pieces whose expectations are easier to compute. For each couple, let X_i be the random variable

$$X_i = \begin{cases} 1, & \text{if couple } i \text{ is intact} \\ 0, & \text{otherwise.} \end{cases}$$

Each X_i is a Bernoulli random variable with some success probability p (no couple should be more likely to survive than any other). What's more is that $X_1 + \cdots X_n = X$. Indeed, if x couples survive, then each surviving couple corresponds to some i such that $X_i = 1$. By the linearity of expectation we have

$$E[X] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = np.$$

Now we just need to find p. For any $i \leq n$ we ha

$$\begin{aligned} p &= \Pr[\text{couple } i \text{ is intact}] \\ &= \Pr[\text{the } m \text{ deaths happen among the other } 2n-2 \text{ people}] \\ &= \frac{\binom{2n-2}{m}}{\binom{2n}{m}}. \end{aligned}$$

So in total we have

$$E[X] = n\frac{\binom{2n-2}{m}}{\binom{2n}{m}}.$$

This idea of breaking up a random variable into a sum of Bernoulli random variables (sometimes called *indicator random variables*) comes up a lot. Did we really need the linearity of expectation to solve this problem? Let's do the same example again, but directly.

Example 2.7. If X is the number of intact couples, then it looks like we need to find the probability mass function for X,

$$p(x) = \Pr[X = x] = \Pr[\text{exactly } x \text{ couples remain intact}].$$

What values of x do we need to consider. Let's consider the extreme cases. If each death affects a different couple, then each of the m deaths breaks a couple and there are $n - m$ couples left intact. On the other hand, if the deaths come in pairs (with maybe one leftover if there is an odd number of deaths) where both people in a couple die at once, then there are $n - \lceil m/2 \rceil$ couples left intact. So we need to compute $\Pr[X = k]$ for $k = n - m, \ldots, n - \lceil m/2 \rceil$.

(a) For each k, choose which of the k couples are to be left intact. There are $\binom{n}{k}$ ways to do this.

(b) For each of the $n - k$ remaining couples, we need to assign death to one of its members. There are 2^{n-k} ways to do this.

(c) Part (b) assigned $n - k$ deaths, but we need m in total. If $m > n - k$, then this leaves $2n - 2k - (n - k)$ people left (the $2k$ comes from part (a)) we could potentially assign death to. There are $\binom{n-k}{m-(n-k)}$ ways to do this. The number of *couples* actually affected by this is $\binom{n-k}{m-(n-k)}2^{-(m-(n-k))}$ though, since marking both members of a couple for death only gets rid of one couple.

(d) In total, there were $\binom{2n}{m}$ ways we could assign the deaths.

Putting these pieces together, we obtain

$$\Pr[X = k] = \binom{n}{k}2^{n-k}\binom{n-k}{m-(n-k)}2^{-(m-(n-k))}\Big/\binom{2n}{m}.$$

So the expected value of X is

$$E[X] = \binom{2n}{m}^{-1}\sum_{k=n-m}^{n-\lceil m/2 \rceil}\binom{n}{k}2^{2(n-k)-m}\binom{n-k}{m-(n-k)}.$$

Proposition 2.8 (The union bound)**.** *If $A_1, \ldots, A_n$ are some events, then*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n}\Pr[A_i].$$

Proof. Of course we'd have strict equality if the events were disjoint, but we aren't assuming that here.

For each event A_i, let X_i be the indicator random variable indicating whether or not A_i happened (that is, $X_i = 1$ if A_i happened, and $X_i = 0$ otherwise). Then if we let $X = X_1 + \cdots + X_n$, we have

$$E[X] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = \sum_{i=1}^{n} \Pr[A_i],$$

which is the quantity we're trying to bound from below. Now let Y be another random variable defined by

$$Y = \begin{cases} 1, & \text{if at least one of the } A_i\text{'s happens} \\ 0, & \text{otherwise.} \end{cases}$$

Notice that $X \geq Y$ with probability 1. That is, if $X = k$ for some integer k, then we must have $X_i = k$ for k different i's. But then the corresponding events A_i all happened, in which case $Y = 1$. This implies that $E[X] \geq E[Y]$. Since $E[Y] = \Pr[\cup A_i]$, the proposition follows. □

Example 2.9. Suppose N people throw their hats into the center of a room. The hats are mixed up and each person randomly selects one hat. What is the expected number of people who pick their own hat?

If we let X be the number of people who get their hat back, we can break X into a sum of indicators $X = X_1 + \cdots + X_N$, where X_i indicates whether or not the i-th person gets their hat back. Since each person has a $1/N$ chance of grabbing their own hat, we have

$$E[X] = \sum_{i=1}^{N} E[X_i] = \sum_{i=1}^{N} \frac{1}{N} = 1.$$

Here's a property of the expectation that comes in handy when working with discrete random variables.

Proposition 2.10. *Suppose X is a discrete random variable and $E[X] \geq a$. Then $\Pr[X \geq a] > 0$. In other words, if the expectation of X is at least a, then X is at least a with positive probability.*

Proof. Suppose the claim were false. Then $E[X] \geq a$ but $X < a$ with probability 1. But then

$$E[X] = \sum_{x} xp(x) = \sum_{x:p(x)>0} xp(x) < \sum_{x:p(x)>0} ap(x) = a \sum_{x:p(x)>0} p(x) = a,$$

a contradiction. □

Remark 2.11. The same proof idea shows that if $E[X] \leq a$, then $\Pr[X \leq a] > 0$.

Here's an example that uses the linearity of expectation to prove something that looks like it has nothing to do with probability.

Example 2.12. Let $v_1, v_2, \ldots, v_n$ be vectors in $\mathbb{R}^d$ of unit length (i.e. $\|v_i\| = 1$ for each i). Then we can always pick signs $\varepsilon_i \in \{-1, 1\}$ such that

$$\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\| \leq \sqrt{n}.$$

To see why, let's choose the ε_i *randomly* and independently of each other. Now let X be the random variable

$$X = \|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|^2.$$

If we expand this out, we get

$$X = \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j v_i \cdot v_j,$$

where $v_i \cdot v_j$ is the usual dot product of the vectors v_i and v_j. Now if we take the expectation of both sides, we get

$$E[X] = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i \cdot v_j E[\varepsilon_i \varepsilon_j].$$

If $i \neq j$, then $\varepsilon_i \varepsilon_j$ is 1 or -1 with equal probability, so these terms contribute nothing to the expectation. On the other hand, if $i = j$, then $\varepsilon_i \varepsilon_j = \varepsilon_i^2 = 1$, and we have

$$E[X] = \sum_{i=1}^{n} v_i \cdot v_i = \sum_{i=1}^{n} \|v_i\|^2 = n.$$

By the previous proposition, there must be a positive probability that we randomly chose signs such that $X \leq n$, in which case $\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\| \leq \sqrt{n}$. Of course, this also shows that we can pick signs that reverse this inequality.

This idea of showing the *deterministic* existence of some object by *probabilistic* means is sometimes called the probabilistic method. It comes up a lot in combinatorics and graph theory. One of the first uses of this idea was due to Szele in 1943.

Example 2.13. A *tournament on n vertices* is a complete directed graph on n vertices. In other words, each unordered pair of vertices $\{u, v\}$ is connected by either an edge going from u to v, denoted (u, v), or an edge from v to u, (v, u). You can think of it as an actual game tournament with n players who all play each other once, and we draw an arrow from the winner to the loser (we assume there are no ties).

Now start at some player v_1 in the tournament and hop to some player that they beat (if there is such a player). That is, walk along some edge out of v_1 and arrive at some new player v_2. Now repeat this process (if possible), walking along an edge out of v_2. If we can keep doing this and eventually arrive at each player, then we say the tournament has a *Hamiltonian path.* In other words, a Hamiltonian path is a permutation of the vertices $(v_{i_1}, v_{i_2}, \ldots, v_{i_n})$ such that player v_{i_1} beats player v_{i_2} and so on.

How many Hamilton paths can a tournament have? Let's choose a tournament at random. That is, for each pair of players, flip a fair coin to determine who beats whom. Now there are $n!$ possible Hamiltonian paths – one for each permutation of the players. If σ is a permutation, let X_σ be the random variable that indicates whether or not σ induces a path in our tournament. If we let X be the number of Hamiltonian paths in our random tournament, then $X = \sum_\sigma X_\sigma$ and

$$E[X] = \sum_{\sigma} E[X_\sigma].$$

The permutation σ induces a path if the $n - 1$ edges between the players all point in the correct direction. This happens with probability $2^{-(n-1)}$, so we have

$$E[X] = n!2^{-(n-1)}.$$

There must then be *some* tournament with at least $n!/2^{n-1}$ Hamiltonian paths in it.

2.3 Moments of the Number of Events that Occur

In the last section, we solved most of the examples by letting the random variable X be the number of occurrences of something we cared about. That is, if $\{A_i\}_{i\geq 1}$ were some events and X was the number of events that occurred, we reasoned about X by letting X_i be the variable that indicates whether or not A_i happened, and writing $X = \sum X_i$. Then the linearity of expectation told us that

$$E[X] = E\left[\sum X_i\right] = \sum E[X_i] = \sum \Pr[A_i].$$

This is advantageous since it's usually a whole lot easier to compute $\Pr[A_i]$ than directly computing $E[X]$.

We can actually get more out of this method. For example, if we wanted the number of *pairs* of events A_i, A_j that occur simultaneously, $\binom{X}{2} = \frac{X(X-1)}{2}$, then

$$\binom{X}{2} = \sum_{i<j} X_i X_j.$$

If we take the expectation of both sides, we get

$$E\left[\binom{X}{2}\right] = \sum_{i<j} E[X_i X_j] = \sum_{i<j} \Pr[A_i, A_j]. \tag{8}$$

If we multiply both sides by 2, we get $E[X^2] - E[X]^2 = 2\sum_{i<j}\Pr[A_i, A_j]$. More generally, $\binom{X}{k}$ is the number of distinct subsets of k events that all occur simultaneously and we have

$$\binom{X}{k} = \sum_{i_1<\ldots<i_k} X_{i_1}\cdots X_{i_k},$$

so the expected number of subsets of k events that all happen simultaneously is

$$E\left[\binom{X}{k}\right] = \sum_{i_1<\ldots<i_k} \Pr[A_{i_1}, \ldots, A_{i_k}].$$

Example 2.14. Let's return to the hats example from before. There are N people, each of whom tosses their hat into the center of the room. Each person then selects a hat at random.

If X is the number of people who select their own hat and A_i is the event that the i-th person gets their own hat, then the expected number of pairs of people who simultaneously get their own hats is

$$\sum_{1\leq i<j\leq n} \Pr[A_i, A_j] = \sum_{1\leq i<j\leq n} \Pr[A_i \mid A_j]\Pr[A_j] = \sum_{1\leq i<j\leq n} \frac{1}{n-1}\cdot\frac{1}{n} = \binom{n}{2}\frac{1}{n-1}\frac{1}{n} = \frac{1}{2}.$$

In particular, $E[X(X-1)] = 1$. Since we calculated $E[X] = 1$ in the previous example, we conclude that $\mathrm{Var}[X] = 1$ as well.

Example 2.15. This famous example is sometimes called the coupon-collector problem. The setup is that there are n types of coupons and each box of cereal contains a single coupon, where coupon i appears with probability p_i and $\sum_i p_i = 1$.

Suppose we buy T boxes of cereal. Let's find the expected number of distinct coupons we get. If we let A_i, $1 \le i \le n$, be the event that we see coupon i in any of our T boxes, then

$$E[X] = \sum_{i=1}^{n} \Pr[A_i] = N - \sum_{i=1}^{n} (1 - p_i)^T.$$

As for the variance, we have by (8)

$$E[X(X-1)] = 2 \sum_{1 \le i < j \le n} \Pr[A_i, A_j].$$

Now, arguably, the easiest way to proceed is to write

$$\Pr[A_i, A_j] = 1 - \Pr[A_i^C] - \Pr[A_j^C] + \Pr[A_i^C, A_j^C] = 1 - (1-p_i)^T - (1-p_j)^T + (1 - p_i - p_j)^T,$$

but this will make the summation a lot uglier. Our way out is to instead let Y be the number of coupons we *don't* see in any of our cereal boxes. Then $X = N - Y$ and $\mathrm{Var}[X] = \mathrm{Var}[Y]$, so we can calculate the variance of this instead. The expectation is easy since

$$E[Y] = N - E[X] = \sum_{i=1}^{n} (1 - p_i)^T.$$

For the variance, we again use (8) to get

$$E[Y(Y-1)] = 2 \sum_{1 \le i < j \le n}^{n} \Pr[A_i^C, A_j^C] = 2 \sum_{1 \le i < j \le n} (1 - p_i - p_j)^T.$$

We then have

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{Var}[Y] \\ &= E[Y^2] - E[Y]^2 \\ &= 2 \sum_{i<j} (1 - p_i - p_j)^T + \sum_{i=1}^{n} (1-p_i)^T - \left(\sum_{i=1}^{n} (1-p_i)^T \right)^2.\end{aligned}$$

Let's look at the specific case where each coupon is equally likely to show up in each box, i.e. $p_i = 1/n$ for each i. Then this all reduces to

$$E[X] = n \left[1 - \left(1 - \frac{1}{n} \right)^T \right],$$

and

$$\mathrm{Var}[X] = n(n-1) \left(1 - \frac{2}{n} \right)^T + n \left(1 - \frac{1}{n} \right)^T - n^2 \left(1 - \frac{1}{n} \right)^{2T}.$$

Remember from calculus that

$$\lim_{n \to \infty} \left(1 + \frac{x}{n} \right)^n = e^x.$$

Using this, if we look at the more specific case where $T = n$, i.e. we open as many boxes as there are coupons, and let the number of coupons grow to infinity, then

$$E[X] \to n \left(1 - \frac{1}{e} \right)$$

and

$$\mathrm{Var}[X] \to n \left(\frac{1}{e} - \frac{1}{e^2} \right).$$

2.4 Covariance, Variance of Sums, Correlations

Sometimes we care about random variables that aren't independent. Here's a definition that helps us quantify how close they are to being independent.

Definition 2.16. If X and Y are random variables, then their *covariance* is

$$\operatorname{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])].$$

Notice that $\operatorname{Cov}(X, X)$ just gives us the variance. Remember that we usually use the more convenient expression for variance,

$$\operatorname{Var}[X] = E[X^2] - E[X]^2.$$

We have a similar expression for covariance.

Proposition 2.17. *For random variables X and Y,*

$$Cov(X, Y) = E[XY] - E[X]E[Y].$$

Proof. It's pretty much just symbol pushing. Remember that $E[X]$ and $E[Y]$ are just constants.

$$\begin{aligned}\operatorname{Cov}(X, Y) &= E\big[XY - E[X]Y - XE[Y] + E[X]E[Y]\big] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y].\end{aligned}$$

□

We advertised covariance as a measure of independence, so let's justify this.

Definition 2.18. If X and Y are independent random variables, then

$$E[XY] = E[X]E[Y],$$

so $\operatorname{Cov}(X, Y) = 0$.

Proof. If X and Y are continuous, remember that independence tells us that the density function splits:

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

Then we have

$$\begin{aligned}E[XY] &= \int_{\mathbb{R}^2} xyf(x, y)\ dxdy \\ &= \int_{\mathbb{R}} xf_X(x)\ dx \int_{\mathbb{R}} yf_Y(y)\ dy \\ &= E[X]E[Y].\end{aligned}$$

Of course, the same proof works for discrete random variables – just change the integrals to sums and the density functions to mass functions. □

Example 2.19. Consider the following study of families having two children. In the following table, X represents the number of children the older sibling has and Y is the number of children the younger sibling has. Given this joint mass function, compute the covariance of X and Y.

X \Y	0	1	2
0	.05	.12	.03
1	.07	.1	.08
2	.02	.26	.27

First let's get the marginal mass functions by summing along rows or columns.

$$\begin{aligned} p_X(0) &= .05 + .12 + .03 = .2 & p_Y(0) &= .05 + .07 + .02 = .14 \\ p_X(1) &= .07 + .1 + .08 = .25 & p_Y(1) &= .12 + .1 + .26 = .48 \\ p_X(2) &= .02 + .26 + .27 = .55 & p_Y(2) &= .03 + .08 + .27 = .38 \end{aligned}$$

Now we have

$$\begin{aligned} E[X] &= 0 \cdot 0.2 + 1 \cdot 0.25 + 2 \cdot 0.55 = 1.35 \\ E[Y] &= 0 \cdot 0.14 + 1 \cdot 0.48 + 2 \cdot 0.38 = 1.24 \\ E[XY] &= 0 \cdot 0.05 + 1 \cdot 0.1 + 2 \cdot (.08 + .26) + 4 \cdot 0.27 = 1.86, \end{aligned}$$

So the covariance is

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 1.86 - 1.35 \cdot 1.24 = 0.186.$$

While covariance gives us a way to measure independence, in practice, statisticians often use the correlation coefficient instead.

Definition 2.20. If X and Y are random variables, then the *correlation* of X and Y is

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}},$$

so long as Var$[X]$ and Var$[Y]$ are positive (which happens whenever X and Y are both non-constant).

One advantage of the correlation over the covariance is that is unitless (covariance has units of (units of X)×(units of Y)) and it is always between -1 and 1 (which can be proven using the Cauchy-Schwarz inequality). Roughly speaking, $\rho(X, Y)$ measures how strong the linear relationship between X and Y is. If $|\rho(X, Y)|$ is close to 1, then there is a strong linear relationship, and if this quantity is closer to 0, then this relationship is absent. If $\rho(X, Y)$ is positive, then X and Y tend to increase or decrease together. If $\rho(X, Y)$ is negative, then an increase in X usually corresponds to a decrease in Y.

Example 2.21. In the previous example,

$$\begin{aligned} E[X^2] &= 0^2 \cdot 0.2 + 1^2 \cdot 0.25 + 2^2 \cdot 0.55 = 2.45 \\ E[Y^2] &= 0^2 \cdot 0.14 + 1^2 \cdot 0.48 + 2^2 \cdot 0.38 = 2, \end{aligned}$$

so

$$\begin{aligned} \mathrm{Var}[X] &= E[X^2] - E[X]^2 = 2.45 - 1.35^2 = 0.6275 \\ \mathrm{Var}[Y] &= E[Y^2] - E[Y]^2 = 2 - 1.24^2 = 0.4624, \end{aligned}$$

and

$$\rho(X, Y) = \frac{.186}{\sqrt{0.6275 \cdot 0.4624}} \approx .345$$

We conclude that there is a decent linear relationship between X and Y, where an increase in Y tends to follow an increase in X.

Example 2.22. Suppose X and Y are uniformly distributed on the region

$$R = \{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1, x^2 \leq y \leq 1\}.$$

Let's find $\rho(X,Y)$.

For *any* region $R \subseteq \mathbb{R}^2$, if X and Y are uniformly distributed on R, then X and Y have joint probability mass function

$$f(x,y) = 1/\text{Area}(R), \quad \text{if } (x,y) \in R.$$

In our case,

$$\text{Area}(R) = \int_R dydx = \int_0^1 \int_{x^2}^1 dydx = \int_0^1 (1-x^2)dx = \frac{2}{3}.$$

Now

$$\begin{aligned} E[X] &= \int_R xf(x,y)\ dydx = \frac{3}{2}\int_0^1 \int_{x^2}^1 x\ dydx = \frac{3}{2}\int_0^1 x - x^3\ dx = \frac{3}{8} \\ E[Y] &= \int_R yf(x,y)\ dydx = \frac{3}{2}\int_0^1 \int_{x^2}^1 y\ dydx = \frac{3}{4}\int_0^1 1 - x^4\ dx = \frac{3}{5} \\ E[XY] &= \int_R xyf(x,y)\ dydx = \frac{3}{2}\int_0^1 \int_{x^2}^1 xy\ dydx = \frac{3}{4}\int_0^1 x - x^5\ dx = \frac{1}{4}. \end{aligned}$$

Hence the covariance is

$$\text{Cov}(X,Y) = \frac{1}{4} - \frac{3}{8}\cdot\frac{3}{5} = \frac{1}{40}.$$

We also have

$$\begin{aligned} E[X^2] &= \int_R x^2f(x,y)\ dydx = \frac{3}{2}\int_0^1 \int_{x^2}^1 x^2\ dydx = \frac{3}{2}\int_0^1 x^2 - x^4\ dx = \frac{1}{5} \\ E[Y^2] &= \int_R y^2f(x,y)\ dydx = \frac{3}{2}\int_0^1 \int_{x^2}^1 y^2\ dydx = \frac{1}{2}\int_0^1 1 - x^6\ dx = \frac{3}{7}, \end{aligned}$$

so the variances are

$$\begin{aligned} \text{Var}[X] &= E[X^2] - E[X]^2 = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = \frac{19}{320} \\ \text{Var}[Y] &= E[Y^2] - E[Y]^2 = \frac{3}{7} - \left(\frac{3}{5}\right)^2 = \frac{12}{175} \end{aligned}$$

and the correlation is

$$\rho(X,Y) = \frac{1/40}{\sqrt{(19/320)\cdot(12/175)}} \approx .392.$$

In particular, X and Y have a decent positive linear relationship.

Let's record some basic properties of covariance.

Proposition 2.23. *(i) If X and Y are random variables, then $Cov(X,Y) = Cov(Y,X)$.*

(ii) If X and Y are random variables and a and b are real numbers, then $Cov(aX,bY) = abCov(X,Y)$.

(iii) If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are random variables, then

$$Cov\left(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} Cov(X_i, Y_j).$$

Proof. Part (i) is immediate from the definition of covariance. Part (ii) is similarly clear:

$$\mathrm{Cov}(aX, bY) = E[(aX)(bY)] - E[aX]E[bY] = abE[XY] - abE[X]E[Y] = ab\mathrm{Cov}(X, Y).$$

Part (iii) follows from just expanding the product and using the linearity of expectation:

$$\begin{aligned}\mathrm{Cov}\left(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j\right) &= E\left[\left(\sum_{i=1}^{m} X_i\right)\left(\sum_{j=1}^{n} Y_j\right)\right] - E\left[\left(\sum_{i=1}^{m} X_i\right)\right] E\left[\left(\sum_{j=1}^{n} Y_j\right)\right] \\ &= E\left[\sum_{i=1}^{m}\sum_{j=1}^{n} X_i Y_j\right] - E\left[\left(\sum_{i=1}^{m} X_i\right)\right] E\left[\left(\sum_{j=1}^{n} Y_j\right)\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(E[X_iY_j] - E[X_i]E[Y_j]\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Cov}(X_i, Y_j).\end{aligned}$$

□

This proposition basically says that covariance is *bilinear*, that is, linear in both of its arguments – kind of like the dot product from calculus/linear algebra. Since variance is just the covariance of a variable with itself, we can use this proposition to get something like a linearity of variance – almost.

Corollary 2.24. *If $X_1, \ldots, X_n$ are random variables, then*

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i] + 2\sum_{1 \le i < j \le n} Cov(X_i, X_j).$$

In particular, if X_i and X_j are uncorrelated for all $i \neq j$, then the variance of the sum is the sum of the variances.

We might have hoped that variance would be linear just like expectation, but this can't be the case if we try to think about it intuitively. Variance is a measure of how "spread out" a random variable is. It's possible that two relatively spread out random variables cancel each other out when we add them together in a way that the variance of their sum is small. For example, X might have large variance. Consequently, $-X$ also has large variance, but $X + (-X) = 0$ has no variance.

Example 2.25. Say we have m balls that we throw at n bins. Each ball lands in a random bin, independent of each other ball. Let's find the mean and variance of the number of empty bins.

Let X be the number of empty bins. We can compute $E[X]$ using the linearity of expectation. Foe each bin $1 \le i \le n$, we let X_i be the random variable indicating whether or not bin i is empty after we've thrown all our balls. Each ball has a $1 - \frac{1}{n}$ chance of missing bin i, so we have

$$E[X_i] = \left(1 - \frac{1}{n}\right)^m,$$

and therefore the expected number of empty bins is

$$E[X] = n\left(1 - \frac{1}{n}\right)^m .$$

Now each X_i is a Bernoulli random variable, so its variance is

$$\text{Var}[X_i] = \left(1 - \frac{1}{n}\right)^m \left[1 - \left(1 - \frac{1}{n}\right)^m\right].$$

But variance isn't exactly linear, so we need to look at the covariance terms. Intuitively, X_i and X_j shouldn't be uncorrelated – if we know that bin i is empty, then then it's less likely that bin j is also empty since all those ball that miss bin i have to go somewhere. More quantitatively, the probability that bins i and j are both empty, $E[X_iX_j]$ is

$$E[X_iX_j] = \left(1 - \frac{2}{n}\right)^m$$

since each ball has a $1-\frac{2}{n}$ chance of missing bins i and j. Since $\text{Cov}(X_i, X_j) = E[X_iX_j]-E[X_i]E[X_j]$, the variance of X is

$$\begin{aligned}
\text{Var}[X] &= \sum_{i=1}^{n} \text{Var}[X_i] + 2\sum_{i<j} \text{Cov}(X_i, X_j) \\
&= n\left(1 - \frac{1}{n}\right)^m \left[1 - \left(1 - \frac{1}{n}\right)^m\right] + n(n-1)\left[\left(1 - \frac{2}{n}\right)^m - \left(1 - \frac{1}{n}\right)^{2m}\right].
\end{aligned}$$

Let's look at the specific case where there are just as many balls as there are bins, i.e. $m = n$. In this case, we have

$$\begin{aligned}
E[X] &\to \frac{n}{e} \\
\text{Var}[X] &\to \frac{n}{e}\left(1 - \frac{1}{e}\right).
\end{aligned}$$

Here's an example that hopefully clarifies what the correlation coefficient really measures.

Example 2.26. Let X be a random variable with mean μ and variance σ^2. Set $Y = a + bX$ for some real numbers a and b. Let's calculate the correlation between X and Y.

First we calculate the covariance.

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(X, a + bX) \\
&= \text{Cov}(X, a) + \text{Cov}(X, bX) \\
&= 0 + b\text{Var}[X] \\
&= b\sigma^2.
\end{aligned}$$

We also need the variance of Y. Note that a constant is independent of any random variable, so

$$\text{Var}[Y] = \text{Var}[a + bX] = \text{Var}[a] + \text{Var}[bX] = 0 + b^2\text{Var}[X] = b^2\sigma^2.$$

The correlation is then

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}} = \frac{b\sigma^2}{\sqrt{\sigma^2}\sqrt{b^2\sigma^2}} = \frac{b}{|b|} = \begin{cases} 1, & \text{if } b > 0 \\ -1, & \text{if } b < 0. \end{cases}$$

In other words, a random variable is always perfectly correlated with a linear transformation of itself.

Let's justify a claim about correlation that we made earlier.

Proposition 2.27. *If X and Y are random variables, then $-1 \leq \rho_{X,Y} \leq 1$.*

Proof. For concreteness, let $\sigma_X = \sqrt{\operatorname{Var}[X]}$ and $\sigma_Y = \sqrt{\operatorname{Var}[Y]}$. Now let's look at the variance of $(X/\sigma_X) + (Y/\sigma_Y)$.

$$\begin{aligned} \operatorname{Var}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] &= \operatorname{Var}\left[\frac{X}{\sigma_X}\right] + \operatorname{Var}\left[\frac{Y}{\sigma_Y}\right] + 2\operatorname{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2}\operatorname{Var}[X] + \frac{1}{\sigma_Y^2}\operatorname{Var}[Y] + \frac{2}{\sigma_X\sigma_Y}\operatorname{Cov}(X,Y) \\ &= 2 + 2\rho_{X,Y}. \end{aligned}$$

Now variance is always nonnegative (it's the expectation of the square of something), so we have

$$0 \leq 2 + 2\rho_{X,Y},$$

and $\rho_{X,Y}$ is always at least -1. If we use the same argument on the variable $(X/\sigma_X) - (Y/\sigma_Y)$ then we can show that $\rho_{X,Y} \leq 1$. □

2.5 Conditional Expectation

If X and Y are discrete random variables, then

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

If they're continuous, then

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

Example 2.28. Say X and Y are jointly continuous with joint density

$$f(x,y) = \frac{e^{-x/y}e^{-y}}{y}, \quad x \geq 0,\ y \geq 0.$$

Let's compute $E[X \mid Y = y]$.

We get the conditional density $f_{X|Y}(x \mid y)$ by computing the marginal density $f_Y(y)$ and then

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

To get the marginal density, we integrate the x away.

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\ dx = e^{-y}, \quad y \geq 0.$$

So we have

$$f(x \mid y) = \frac{e^{-x/y}}{y}, \quad x, y \geq 0.$$

Now we can compute the expectation.

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

Theorem 2.29 (The Law of Total Expectation)**.** *Let X and Y be random variables. Then*

$$E_Y[E_X[X \mid Y = y]] = E[X].$$

Proof. If X and Y are discrete, then

$$\begin{aligned}
E_Y[E_X[X \mid Y = y]] &= \sum_i E[X \mid Y = y_i] \Pr[Y = y_i] \\
&= \sum_i \sum_j x_j \Pr[X = x_j \mid Y = y_i] \Pr[Y = y_i] \\
&= \sum_{i,j} x_j \Pr[X = x_j, Y = y_i] \\
&= \sum_j x_j \Pr[X = x_j] \\
&= E[X].
\end{aligned}$$

Like usual, replacing sums with integrals and mass functions with densities gives the proof for the continuous case. □

We can use this to compute expectations in the following way. Say we want the expectation of X, but this variable is kind of complicated. If X conditioned on some outcome of *another* random variable Y is simpler, then we can use the law of total expectation to get the expectation of X.

Example 2.30. A coal miner is trapped in an underground room with three doors. Door 1 leads to the surface after two hours. Door 2 leads back to the room after five hours, and Door 3 also leads back to the room after three hours. Let's find the expected number of hours it takes for the miner to reach the surface.

If we were to just use the definition of expectation and no other technology, this problem appears hard. It looks like we'd need to sum over all possible times and find the probability that it takes the miner exactly that long to leave. Conditional expectation gives us another way thought.

If we let X be the number of hours until we reach the surface, we can condition X on the *first* door the miner takes, Y. Then by the law of total expectation we have

$$\begin{aligned}
E[X] &= E_Y[E_X[X \mid Y]] \\
&= E[X \mid Y = 1] \Pr[Y = 1] + E[X \mid Y = 2] \Pr[Y = 2] + E[X \mid Y = 3] \Pr[Y = 3] \\
&= \frac{1}{3}\big(E[X \mid Y = 1] + E[X \mid Y = 2] + E[X \mid Y = 3]\big).
\end{aligned}$$

Now $E[X \mid Y = 1] = 2$ because if the miner chooses door 2 initially, then they just walk the two hours to the surface and they're done. On the other hand, $E[X \mid Y = 2] = 5 + E[X]$. To see this, if the miner chooses the second door, then they spend five hours walking back to the room, at which point the problem basically starts over, and we expect it will take them $E[X]$ hours to reach the surface. Applying the same logic to door 3 gives

$$E[X] = \frac{1}{3}\big(2 + (5 + E[X]) + (3 + E[X])\big) \implies E[X] = 10/3.$$

Example 2.31. Say we flip a coin with head probability p until we get k consecutive heads. Let's find the expected number of flips we need in order for this to happen.

Let N_k be the number of flips to get k heads. Let's condition N_k on N_{k-1}. Then

$$E[N_k \mid N_{k-1}] = p(1 + N_{k-1}) + (1-p)(1 + N_{k-1} + E[N_k]).$$

The first term comes from the fact that if we have $k-1$ consecutive heads, then the next flip is heads with probability p and we'll have k consecutive heads. On the other hand, the next flip will be tails with probability $1-p$, in which case we expect $E[N_k]$ flips for k consecutive heads.

By the law of total expectation, we have

$$E[N_k] = E[E[N_k \mid N_{k-1}]] = p(1 + E[N_k]) + (1-p)(1 + E[N_{k-1}] + E[N_k]).$$

Rearranging terms, we get

$$E[N_k] = \frac{1}{p} + \frac{1}{p}E[N_{k-1}].$$

Applying the same logic to $E[N_{k-1}]$ gives

$$E[N_k] = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^2}E[N_{k-2}].$$

Inductively, we have

$$E[N_k] = \sum_{i=1}^{k} \frac{1}{p^i}.$$

2.6 Moment Generating Functions

Recall that calculating the mean and variance of a random variable requires us to compute $E[X]$ and $E[X^2]$, respectively. These are the first two *moments* of X, and in general we have the following definition.

Definition 2.32. If X is a random variable, then the k-th moment of X is $E[X^k]$.

Example 2.33. Let's calculate the third and fourth moments of $X \sim \mathcal{N}(0,1)$. The third moment is given by

$$E[X^3] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^3 e^{-x^2/2} dx.$$

Now if this integral exists, it's equal to the sum of the integrals over the half-lines.

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^3 e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x^3 e^{-x^2/2} dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} x^3 e^{-x^2/2} dx.$$

Since $x^3 e^{-x^2/2}$ is an odd function, if one of these integrals is finite, then so is the other and they'll be equal in magnitude but opposite in sign. These integrals are indeed finite since $x^3 \le e^{x^2/4}$ for all x sufficiently large, say $x > M$. Then we have

$$\begin{aligned}
\int_0^\infty x^3 e^{-x^2/2} dx &= \int_0^M x^3 e^{-x^2/2} dx + \int_M^\infty x^3 e^{-x^2/2} dx \\
&\le \int_0^M x^3 e^{-x^2/2} dx + \int_M e^{-x^2/4} dx \\
&< \infty.
\end{aligned}$$

Thus, $E[X^3] = 0$ in this case. We could have done integration by parts, but this way has fewer calculations, arguably.

As for the fourth moment,

$$E[X^4] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^4 e^{-x^2/2} dx,$$

we integrate by parts. Setting $u = x^3$ and $dv = xe^{-x^2/2}dx$, we get

$$\begin{aligned}
E[X^4] &= -\frac{1}{\sqrt{2\pi}} x^3 e^{-x^2/2} \Big|_{x=-\infty}^{x=\infty} + \frac{3}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-x^2/2} dx \\
&= 3E[X^2].
\end{aligned}$$

Since we know the mean and variance of X are 0 and 1, respectively, $E[X^2] = \mathrm{Var}[X] + E[X]^2 = 1$, so $E[X^4] = 3$.

We'll see soon that a random variable's moments carry a lot of information. It might not be obvious at first glance, but the following device, in a way, keeps track of all of the moments of a random variable.

Definition 2.34. If X is a random variable, then the *moment generating function of* X is the function

$$M(t) = E[e^{tX}].$$

If we want to emphasize that this is the moment generating function of X and not some other random variable, we write $M_X(t)$.

Where does this thing get its name from? In what sense does it "generate" the moments of X? The Taylor series expansion for e^x gives us some intuition. Recall that for any real x,

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{k=0}^\infty \frac{x^k}{k!}.$$

Now set $x = tX$ and take the expectation of both sides. Assuming that we can split the expectation over an *infinite* sum, we have

$$E[e^{tX}] = 1 + tE[X] + \frac{1}{2!}t^2 E[X^2] + \frac{1}{3!}t^3 E[X^3] + \cdots = \sum_{k=0}^\infty \frac{t^k}{k!} E[X^k].$$

Let's take the derivative of both sides with respect to t. Assuming that we can distribute the derivative over an *infinite* sum, this gives

$$\frac{d}{dt} E[e^{tX}] = E[X] + tE[X^2] + \frac{1}{2!} E[X^3] + \cdots = \sum_{k=1}^\infty \frac{t^{k-1}}{(k-1)!} E[X^k].$$

If we set $t = 0$, we just get $\frac{d}{dt}E[e^{tX}]|_{t=0} = E[X]$. But notice that if we repeat this process, we get $\frac{d}{dt}E[e^{tX}]|_{t=0} = E[X^2]$, and so on. We record this observation as a proposition and provide another proof.

Proposition 2.35. *Let X be a random variable and let $M(t)$ be its moment generating function. Then for any nonnegative integer n,*

$$M^{(n)}(t) = E[X^n].$$

Here $M^{(n)}$ is the n-th derivative of M.

Proof. We assume that we can switch the order of differentiation and summation/integration. Whether or not we can actually do this depends on the random variable X. This is okay for pretty much every random variable we'll see in this class. A more thorough investigation of when this is justified requires some *measure theory*, a fundamental area of math that's closely connected to real analysis, upon which pretty much all of modern probability theory rests. Under this assumption, we have

$$\begin{aligned}\frac{d^n}{dt^n}M(t) &= \frac{d^n}{dt^n}E[e^tX] \\ &= E\left[\frac{d^n}{dt^n}e^{tX}\right] \\ &= E[X^n e^{tX}].\end{aligned}$$

The claim follows from setting $t = 0$. It's easy to check that taking the first two derivatives gives us the first two moments of X. □

Example 2.36. Let's compute the moment generating function of $X \sim \text{Pois}(\lambda)$. We have

$$\begin{aligned}M(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} e^{tn}\frac{e^{-\lambda}\lambda^n}{n!} \\ &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^t)^n}{n!} \\ &= e^{\lambda(e^t-1)}.\end{aligned}$$

Example 2.37. Let's compute the moment generating function of $X \sim \mathcal{N}(0,1)$. We complete the square inside the exponential.

$$\begin{aligned}M_X(t) &= E[e^{tX}] \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{tx}e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(x^2-2tx)/2}dx \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(x-t)^2/2}dx \\ &= e^{t^2/2}.\end{aligned}$$

We can use this to easily compute the moment generating function of an arbitrary normal random variable. If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then Y and $\mu + \sigma X$ have the same distribution. Then we have

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t(\mu+\sigma X)}] \\ &= e^{t\mu} E[e^{t\sigma X}] \\ &= e^{t\mu} M_X(t\sigma) \\ &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \end{aligned}$$

So differentiating the moment generating function gives the moments. Is that it? Well it turns out that moment generating functions play nicely with sums of independent random variables.

Proposition 2.38. *If X and Y are independent random variables, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Proof. We have

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX} e^{tY}].$$

Now since X and Y are independent, so are e^{tX} and e^{tY} (why?). the above expression is then equal to $E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$. □

Arguably the most important fact about moment generating functions is that they completely determine the random variable (under moderate assumptions).

Theorem 2.39. *Suppose X and Y are two random variables. If the moment generating functions $M_X(t)$ and $M_Y(t)$ agree on some open interval $(-\epsilon, \epsilon)$, then X and Y have the same distribution.*

A rigorous proof of this statement isn't actually that difficult, but it requires some complex analysis ("complex" as in complex numbers), so we won't prove it here. This theorem is important because it lets us make statements about random variables by looking at their moment generating functions.

Example 2.40. Let X_1 and X_2 be normal random variables with means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively. Here's another proof that $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. By Example 2.37 we have

$$\begin{aligned} M_{X_1+X_2}(t) &= M_{X_1}(t) M_{X_2}(t) \\ &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right) \exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right). \end{aligned}$$

But this is the moment generating function of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Since the moment generating function determines the distribution, we have that $X_1 + X_2$ must have this distribution.

The above example gives us a recipe for how to prove the following proposition.

Proposition 2.41. *Suppose $X_1, \ldots, X_n$ are independent normal random variables with means μ_i and variances σ_i^2 for $1 \le i \le n$. Then*

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

We can also prove things that look obvious, but might be more annoying to prove without moment generating functions.

Proposition 2.42. *Let X, Y and Z be independent random variables. If $X+Y$ has the same distribution as $X+Z$, then Y and Z must have the same distribution.*

Proof. If $X+Y$ and $X+Z$ have the same distribution, then they have the same moment generating functions, i.e.

$$M_{X+Y}(t) = M_{X+Z}(t).$$

But we can split the moment generating function of a sum of independent random variables, so

$$M_X(t)M_Y(t) = M_X(t)M_Z(t).$$

Now since e^x is positive for any x, the moment generating function $M_X(t)$ is positive for all t and we can cancel it from both sides of the above equation to get $M_Y(t) = M_Z(t)$. Since the moment generating function determines the distribution, we conclude that Y and Z have the same distribution. □

This class is all about looking at multiple random variables at the same time. This motivates us to define a moment generating function for joint distributions.

Definition 2.43. The joint moment generating function for the random variables $X_1, \ldots, X_n$ is

$$M(t_1, \ldots, t_n) = E[\exp(t_1X_1 + \cdots + t_nX_n)].$$

Remark 2.44. This is *not* the same thing as the moment generating function of the sum $X_1 + \cdots + X_n$. The joint moment generating function is a function of n variables, but the MGF of the sum is just a single variable function:

$$M_{X_1+\cdots+X_n}(t) = E\big[\exp\big(t(X_1 + \cdots + X_n)\big)\big].$$

Example 2.45. Let Z_1 and Z_2 be independent standard normal random variables. Let's find the joint moment generating function for

$$X_1 = Z_1 + 2Z_2, \quad X_2 = 2Z_1 + Z_2.$$

By definition we have

$$M(t_1, t_2) = E[\exp(t_1X_1 + t_2X_2)].$$

Now it might be tempting to split the exponential into a product and then split the expectation over that product, but X_1 and X_2 aren't independent (check the covariance). But Z_1 and Z_2 are independent, so we do a little bit of algebra:

$$\begin{aligned} M(t_1, t_2) &= E\big[\exp\big((t_1 + 2t_2)Z_1 + (2t_1 + t_2)Z_2\big)\big] \\ &= E\big[\exp\big((t_1 + 2t_2)Z_1\big)\big]E\big[\exp\big((2t_1 + t_2)Z_2\big)\big] \\ &= M_{Z_1}(t_1 + 2t_2)M_{Z_2}(2t_1 + t_2). \end{aligned}$$

Of course, Z_1 and Z_2 have the same distribution, so they have the same MGF,

$$M_{Z_1}(t) = M_{Z_2}(t) = \exp(t^2/2),$$

so the joint MGF is then

$$M(t_1, t_2) = \exp\left(\frac{(t_1 + 2t_2)^2 + (2t_1 + t_2)^2}{2}\right).$$

Let's record some basic facts about the joint MGF.

Proposition 2.46. *Let $X_1, \ldots, X_n$ be random variables. Then*

(i) The joint MGF determines the joint distribution. That is, if $Y_1, \ldots, Y_n$ have the same joint MGF as $X_1, \ldots, X_n$, then they have the same joint distribution.

(ii) $X_1, \ldots, X_n$ are independent if and only if

$$M(t_1, \ldots, t_n) = \prod_{i=1}^{n} M_{X_i}(t_i).$$

(iii) We can recover the MGF of X_i from the joint MGF:

$$M_{X_i}(t_i) = M(0, \ldots, 0, t_i, 0, \ldots, 0),$$

where t_i appears in the i-th argument.

The proof for part (i) is beyond the scope of this course, but it isn't too hard if you have the required machinery. Part (ii) is easy to prove: if the X_i's are independent, then so are the $e^{t_i X_i}$'s, so we can split the expectation. Conversely, if the MGF factors, then it has the same MGF as a set of independent random variables. Since the MGF determines the distribution by part (i), the variables must be independent. Part (iii) is immediate. Moment generating functions will come in handy in the next section.

2.7 The Multivariate Normal Distribution

Definition 2.47. The random variables $X_1, \ldots, X_n$ have *multivariate normal distribution* (sometimes we say they are *jointly normal*) if we can write

$$\begin{aligned} X_1 &= \mu_1 + a_{1,1}Z_1 + a_{1,2}Z_2 + \cdots + a_{1,m}Z_m \\ X_2 &= \mu_2 + a_{2,1}Z_1 + a_{2,2}Z_2 + \cdots + a_{2,m}Z_m \\ \vdots &= \vdots \\ X_n &= \mu_n + a_{n,1}Z_1 + a_{n,2}Z_2 + \cdots + a_{n,m}Z_m \end{aligned}$$

for independent standard normal random variables $Z_1, \ldots Z_m$ and constants $\mu_i, a_{i,j}$ with $1 \leq i \leq n$, $1 \leq j \leq m$.

In other words, $X_1, \ldots, X_n$ are jointly normal if the vector $\vec{X} = [X_1, \ldots, X_n]^T$ is an affine linear transformation of the vector $\vec{Z} = [Z_1, \ldots, Z_m]^T$, i.e.

$$\vec{X} = \vec{\mu} + A\vec{Z},$$

where $\vec{\mu} = [\mu_1, \ldots, \mu_n]^T$ and $[A]_{i,j} = a_{i,j}$. The random variables X_1 and X_2 in Example 2.45 are jointly normal.

What information do we need in order to specify a multivariate normal distribution? For example, we need to know the number of trials n and success probability p to specify a binomial distribution. In order to specify a one-dimensional normal random variable, we need to know its mean and its variance.

Are mean and variance enough to determine a multivariate normal distribution? If we go back to Example 2.45, then we can see that $X_1 \sim \mathcal{N}(0, 5)$ and $X_2 \sim \mathcal{N}(0, 5)$. But X_1 and X_2 aren't independent since

$$\mathrm{Cov}(X_1, X_2) = 2\mathrm{Var}[Z_1] + 5\mathrm{Cov}(Z_1, Z_2) + 2\mathrm{Var}[Z_2] = 4.$$

On the other hand, the variables $Y_1 = \sqrt{5}Z_1$ and $Y_2 = \sqrt{5}Z_2$ are also jointly normal and Y_1, Y_2 both have distribution $\mathcal{N}(0, 5)$. But Y_1 and Y_2 are independent unlike X_1 and X_2. So it looks like knowing all the means and variances isn't enough to determine a multivariate normal distribution. The solution is going to come from looking at the moment generating function.

If $X_1, \ldots, X_n$ are jointly normal, then their joint moment generating function is given by

$$M(t_1, \ldots, t_n) = E[\exp(t_1X_1 + \cdots + t_nX_n)].$$

Now if we fix $t_1, \ldots, t_n$, then $Y := t_1X_1 + \cdots + t_nX_n$ is a normal random variable with mean μ and variance σ^2 given by

$$\mu = \sum_{i=1}^{n} t_i\mu_i, \quad \sigma^2 = \sum_{i,j} t_it_j\mathrm{Cov}(X_i, X_j).$$

We could write $\mathrm{Cov}(X_i, X_j)$ in terms of the $a_{i,j}$'s, but let's not do that now. We know the moment generating function for Y is given by

$$M_Y(t) = E[\exp(tY)] = \exp\left(\mu t + \frac{\sigma^2t^2}{2}\right).$$

But the joint MGF of $X_1, \ldots, X_n$ is

$$M(t_1, \ldots, t_n) = E[\exp(t_1X_1 + \cdots + t_nX_n)] = E[\exp Y],$$

which is just $M_Y(1)$. We then have

$$M(t_1, \ldots, t_n) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp\left(\sum_{i=1}^{n} t_i\mu_i + \sum_{i,j} t_it_j\mathrm{Cov}(X_i, X_j)\right).$$

Since the moment generating function determines the distribution, and the moment generating function is determined by the μ_i's and the $\mathrm{Cov}(X_i, X_j)$'s, we conclude that the multivariate normal distribution is determined by the means of the X_i's and all of the pairwise covariances $\mathrm{Cov}(X_i, X_j)$. We record this in the following theorem.

Theorem 2.48. *The multivariate normal distribution is completely determined by $E[X_i]$ and $Cov(X_i, X_j)$ for $i, j = 1, \dots, n$.*

Example 2.49. Suppose X and Y are jointly normal with $E[X] = \mu_X$, $E[Y] = \mu_Y$, $\text{Var}[X] = \sigma_X^2$, $\text{Var}[Y] = \sigma_Y^2$. Suppose further that X and Y have correlation ρ. Let's find $\Pr[X < Y]$.

The "old" way we have of doing this problem uses the fact that

$$\Pr[X < Y] = \int_{\{x<y\}} f(x, y)\ dydx,$$

where f is the joint pdf of X and Y. The issue here is that we don't know what the joint density is (not yet at least, we'll find it later).

Note that $X - Y$ is a normal random variable. This sounds obvious since it's the sum (or difference) of two normal random variables, but the complicating factor is that X and Y aren't independent. The key here is that X and Y are *jointly normal*, so X and Y are both linear combinations of independent normal random variables. Consequently, their difference is also a linear combination of normal random variables, and is thus normal as well.

Since $X - Y$ is normal, it's determined by its mean and variance, so we'll need those

$$\begin{aligned} E[X - Y] &= E[X] - E[Y] = \mu_X - \mu_Y, \\ \text{Var}[X - Y] &= \text{Var}[X] + \text{Var}[Y] - 2\text{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho_{X,Y}\sigma_X\sigma_Y. \end{aligned}$$

Recall that $\Phi(t) = \Pr[Z < t]$, where Z is a standard normal random variable. Since $\frac{(X-Y)-E[X-Y]}{\text{Var}[X-Y]^{1/2}}$ is a standard normal random variable, we have

$$\begin{aligned} Pr[X - Y < 0] &= \Pr\left[\frac{(X - Y) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\rho_{X,Y}\sigma_X\sigma_Y}} < \frac{-(\mu_X - \mu_Y)}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\rho_{X,Y}\sigma_X\sigma_Y}}\right] \\ &= \Phi\left(\frac{\mu_Y - \mu_X}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\rho_{X,Y}\sigma_X\sigma_Y}}\right). \end{aligned}$$

Remember that independent random variables are always uncorrelated, but the converse isn't necessarily true (take $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$ – these are uncorrelated but not independent). It turns out to be true for jointly normal random variables though.

Proposition 2.50. *If X and Y are jointly normal, then they are independent if and only if they are uncorrelated.*

Proof. Say $E[X] = \mu_X$, $E[Y] = \mu_Y$, $\text{Var}[X] = \sigma_X^2$ and $\text{Var}[Y] = \sigma_Y^2$. Independent always implies uncorrelated, so we just need to prove the converse. The joint moment generating function of X and Y is given by

$$M(t_1, t_2) = \exp\left((t_1\mu_1 + t_2\mu_2) + \frac{(\sigma_X^2 t_1^2 + \sigma_Y^2 t_2^2 + 2t_1t_2\text{Cov}(X, Y))}{2}\right).$$

If X and Y are uncorrelated, then $\text{Cov}(X, Y) = 0$ and we can factor the MGF:

$$M(t_1, t_2) = \exp\left(t_1\mu_1 + \frac{\sigma_X^2 t_1^2}{2}\right)\exp\left(t_2\mu_2 + \frac{\sigma_Y^2 t_2^2}{2}\right),$$

which is the MGF of two independent normal random variables. Since the MGF determines the distribution, we conclude that X and Y are independent. □

2.8 Random Vectors

The technology we've developed so far sometimes looks a whole lot nice when it's written in terms of vectors. If $X_1, \ldots, X_n$ are random variables, we can encode them as a single vector in $\mathbb{R}^n$ as $\vec{X} = [X_1, \ldots, X_n]^T$ (we transpose to make it a column vector by convention). We can make sense of the expectation of a vector – just take the expectation of each component.

$$\vec{\mu} = E[\vec{X}] := [E[X_1], E[X_2], \ldots, E[X_n]]^T.$$

What about variance? We *could* define the variance of $\vec{X}$ to just be the vector of variances, $[\text{Var}[X_1], \ldots, \text{Var}[X_n]]^T$, but as we saw with normal random variables, knowing the variance of each individual variable doesn't tell the whole story of joint distribution. The covariances of each pair of variables are important, but how should we encode them?

Definition 2.51. If $X_1, \ldots, X_n$ are random variables, then their *covariance matrix* is the $n \times n$ matrix Σ whose entries are given by

$$[\Sigma]_{i,j} = \text{Cov}(X_i, X_j).$$

For example, if $n = 2$, then

$$\Sigma = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}[X_2] \end{bmatrix}.$$

Example 2.52. Say we flip a fair coin 30 times and let X_1 be the number of heads in the first 20 flips and let X_2 be the number of heads in the last 20 flips. Let's find the covariance matrix of X_1 and X_2.

X_1 and X_2 are $\text{Bin}(20, 1/2)$ random variables, so their variances are the same and equal to

$$\text{Var}[X_1] = \text{Var}[X_2] = np(1-p) = 5$$

X_1 and X_2 aren't independent since they have some flips in common. If we let

$$\begin{aligned} Y_1 &= \text{number of heads in the first 10 flips} \\ Y_2 &= \text{number of heads in flips 11-20} \\ Y_3 &= \text{number of heads in flips 21-30.} \end{aligned}$$

Then the Y_i's are all independent and $X_1 = Y_1 + Y_2$ and $X_2 = Y_2 + Y_3$. The covariance is given by

$$\text{Cov}(X_1, X_2) = \text{Cov}(Y_1+Y_2, Y_2+Y_3) = \text{Cov}(Y_1, Y_2)+\text{Cov}(Y_1, Y_3)+\text{Var}[Y_2]+\text{Cov}(Y_2, Y_3) = \text{Var}[Y_2] = 5/2.$$

So the covariance matrix is

$$\Sigma = \begin{bmatrix} 5 & 5/2 \\ 5/2 & 5 \end{bmatrix}.$$

Let's record some basic facts about the covariance matrix.

Proposition 2.53. *Suppose Σ is the covariance matrix for the random variables $X_1, \ldots, X_n$. Let $\vec{X} = [X_1, \ldots, X_n]^T$ and let $\vec{\mu} = E[\vec{X}]$.*

(i) $\Sigma = E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T]$.

(ii) The matrix Σ is positive semidefinite. That is, for any $y \in \mathbb{R}^n$, we have $y^T \Sigma y \geq 0$. Some elementary linear algebra shows that this is equivalent to all of Σ's eigenvalues being nonnegative.

(iii) Σ is diagonal if and only if the X_i's are uncorrelated.

Proof. Part (i) looks a lot like the formula for covariance, $\operatorname{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$. In fact, that's what each entry in $E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T]$ is.

$$\left[E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T]\right]_{i,j} = E[(\vec{X} - \vec{\mu})_i(\vec{X} - \vec{\mu})_j] = E[(X_i - E[X_i])(X_j - E[X_j])] = \operatorname{Cov}(X_i, X_j).$$

Part (i) helps us prove part (ii). If y is any vector in $\mathbb{R}^n$, then

$$y^T \Sigma y = y^T E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T] y = E[y^T(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T y] = E\left[\left(y \cdot (\vec{X} - \vec{\mu})\right)^2\right] \geq 0.$$

Part (iii) follows immediately from the definition. The non-diagonal entries of Σ are $\operatorname{Cov}(X_i, X_j)$ with $i \neq j$ and if the X_i's are uncorrelated, then these are all zero. □

Let's finish this off by using our new linear algebra notation to find the density function for the multivariate normal distribution.

Theorem 2.54. *Let $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ (that is, $\vec{X}$ has multivariate normal distribution with $E[\vec{X}] = \vec{\mu}$ and the covariance matrix of $\vec{X}$ is Σ). Then the pdf of $\vec{X}$ is*

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right].$$

Proof. By definition of the multivariate normal distribution, there is some matrix A (which we can assume is $n \times n$) such that, for some independent standard normal random variables $Z_1, \dots, Z_n$, we have

$$\vec{X} = \vec{\mu} + A\vec{Z}.$$

The pdf of $\vec{X}$, by the change of variables formula is then given by

$$f_{\vec{X}}(\vec{x}) = f_{\vec{Z}}(\vec{z}) \cdot |J(\vec{z})|^{-1}.$$

Since the Z_i's are independent, we can factor $f_{\vec{Z}}$ into

$$f_{\vec{Z}}(\vec{z}) = f_{Z_1}(z_1) \cdots f_{Z_n}(z_n) = \frac{1}{\sqrt{(2\pi)^n}} \prod_{i=1}^n e^{-z_i^2/2} = \frac{1}{\sqrt{(2\pi)^n}} e^{-\vec{z}^T \vec{z}/2}.$$

We also have the following relationship between the covariance matrix Σ and A.

$$\begin{aligned}
\Sigma &= \operatorname{Cov}(\vec{X}) \\
&= E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T] \\
&= E[(A\vec{Z})(A\vec{Z})^T] \\
&= E[A\vec{Z}\vec{Z}^T A^T] \\
&= AE[\vec{Z}\vec{Z}^T]A^T \\
&= AA^T.
\end{aligned}$$

The last line follows from the fact that the Z_i's are independent, so $E[\vec{Z}\vec{Z}^T] = I$. In particular, we have

$$\det \Sigma = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2$$

The determinant of A also happens to be the determinant of the Jacobian matrix. To see this, notice that since $X_i = a_{i,1}Z_1 + \cdots + a_{i,n}Z_n$, we have

$$\frac{\partial x_i}{\partial z_i} = a_{i,j}.$$

The determinant of the Jacobian is then the determinant of A. Finally, since $\vec{Z} = A^{-1}(\vec{X} - \vec{\mu})$ and

$$\vec{z}^T\vec{z} = [A^{-1}(\vec{x} - \vec{\mu})]^T[A^{-1}(\vec{x} - \vec{\mu})] = (\vec{x} - \vec{\mu})^T(A^{-1})^TA^{-1}(\vec{x} - \vec{\mu}) = (\vec{x} - \vec{\mu})^T\Sigma^{-1}(\vec{x} - \vec{\mu}),$$

we have

$$\begin{aligned} f_{\vec{X}}(\vec{x}) &= f_{\vec{Z}}(\vec{z}) \cdot |J(\vec{z})|^{-1} \\ &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T\Sigma^{-1}(\vec{x} - \vec{\mu})\right]. \end{aligned}$$

□

3 Limit Theorems

People generally have a hard time connecting probability to reality. The tools we'll develop in this section, arguably, solidify some of the intuitions that we have.

3.1 Chebyshev's Inequality and the Weak Law of Large Numbers

The following theorem lets us estimate the probability that a nonnegative random variable is "big."

Theorem 3.1 (Markov's Inequality). *Let X be a nonnegative random variable. Then for any $t \geq 0$,*

$$\Pr[X \geq t] \leq \frac{E[X]}{t}.$$

Proof. Consider the random variable I that indicates whether or not X is at least t. Clearly $E[I] = \Pr[X \geq t]$. We also have that $I \leq X/t$ with probability 1. To see this, if $X \geq t$, then $I = 1 \leq X/t$. If $X < t$, then $I = 0 \leq X/t$ since X is nonnegative. We then have

$$\Pr[X \geq t] = E[I] \leq E[X/t] = \frac{E[X]}{t}.$$

□

In other words, the probability that X is larger than t falls off at least as fast as $1/t$. This is interesting because we don't need to know a whole lot about our distribution to apply Markov's inequality. we just need to know its mean. It turns out though that this bound usually isn't that great for solving problems involving real-world statistics.

Example 3.2. Suppose the average college basketball player is six feet tall. If we wanted to estimate the probability that the height of a randomly chosen player, X, was greater than twelve feet, Markov's inequality would tell us that

$$\Pr[X \geq 12] \leq \frac{E[X]}{12} = \frac{6}{12} = \frac{1}{2}.$$

Of course, our intuition tells us there's almost no chance that a person, basketball player or otherwise, is taller than twelve feet!

Markov is pretty good for more abstract results though.

Example 3.3. A *graph* $G = (V, E)$ consists of a set V of *vertices* and a set E of *edges*. An edge $e = \{u, v\} \in E$ consists of a pair of vertices. In graph theory, we like to investigate the properties of graphs. For example, the *diameter* of a graph is the longest distance between two vertices in the graph. Here distance means exactly what you think it means – it's the length of the shortest path between the vertices (again, path means exactly what you think).

Here are some examples.

Instead of looking at the diameter of some specific graph, let's pick a graph with n vertices *at random* in the following way. For each pair of vertices $\{u, v\}$, include it as an edge with probability p (independent of all other edges). Call this graph (which is a random variable!) $G(n, p)$. Let's investigate its diameter. We'll prove the following:

$$\lim_{n\to\infty} \Pr[G(n,p) \text{ has diameter } 2] \to 1.$$

This seems unintuitive: a graph with n vertices could have diameter close to n, but this is saying that's exceptionally rare. If the distance between two vertices is at least 3, then there's no path of length 2 between them. If we fix two vertices u and v, let $X_{u,v}$ indicate whether or not there's *no* 2-path between them. There are $n-2$ possible 2-paths between them – one for each other vertex in the graph. Each path appears with probability p^2 since it's made of two edges, so $E[X_{u,v}] = (1-p^2)^{n-2}$. If X is the number of pairs of vertices at distance at least 3 from each other, then

$$E[X] \leq \sum_{\{u,v\}\subseteq V} E[X_{u,v}] = \frac{n(n-1)}{2}(1-p^2)^{n-2}.$$

This goes to zero as $n \to \infty$ since $(1-p^2)^{n-2}$ is exponentially small in n. By Markov's inequality, we then have

$$\Pr[X \geq 1] \leq E[X] \to 0, \text{ as } n \to \infty.$$

So there no vertices at distance at least 3 from each other, which means the diameter is at most 2. Now the diameter is 1 if and only if all $\binom{n}{2}$ possible edges appear. This happens with probability $p^{\binom{n}{2}}$, which goes to zero as $n \to \infty$. Thus, the diameter is exactly 2 with probability that approaches 1 as $n \to \infty$.

We can use Markov's inequality to build the following stronger result.

Theorem 3.4 (Chebyshev's Inequality). *Let X be a random variable with mean μ and variance σ^2. Then for any t,*

$$\Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}.$$

Proof. Even though X might not be nonnegative, $(X-\mu)^2$ definitely is, so we can apply Markov to it to get

$$\Pr[|X-\mu| \geq t] = \Pr[(X-\mu)^2 \geq t^2] \leq \frac{E[(X-\mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

□

In a way, Chebyshev's inequality stronger than Markov's inequality. We need a little more information to use it – we need both the mean and the variance, but it gives us a better estimate, with probability falling off like $1/t^2$ instead of $1/t$.

Example 3.5. Letl's return to our basketball player example. Suppose the average height is still six feet, and there's a standard deviation of 0.25 feet. Then Chebyshev tells us that the probability that a randomly chosen player's height X, is taller than twelve feet with the following probability

$$\Pr[X \geq 12] = \Pr[|X-6| \geq 6] \leq \frac{0.25^2}{6^2} \approx 0.0017.$$

Much better than what Markov told us.

We can use Chebyshev's inequality to prove a version of one of the most well-known results in probability.

Theorem 3.6 (The Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables, each with mean μ. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \epsilon\right] = 0.$$

Proof. We'll prove this theorem under the additional assumption that the X_i's have finite variance σ^2. The idea is to just apply Chebyshev's inequality. The variance of X_i/n is σ^2/n^2 and since the X_i's are independent, the variance of $(X_1+\cdots+X_n)/n$ is σ^2/n and we have

$$\Pr\left[\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \epsilon\right] \leq \frac{\sigma^2/n}{\epsilon^2} \to 0 \text{ as } n\to\infty.$$

□

In words, the weak law of large numbers says that, for any margin of error ϵ, the sample mean is eventually likely to be within ϵ of the true average, μ.

Example 3.7. Flip a coin n times and let X_i indicate whether or not the i-th flip was heads. Then the weak law says that the probability of seeing at least $0.501n$ heads is

$$\Pr[X_1+\cdots+X_n \geq .501n] \leq \Pr\left[\left|\frac{X_1+\cdots+X_n}{n} - \frac{1}{2}\right| \geq .001\right] \to 0 \text{ as } n\to\infty.$$

In other words, as we flip more and more coins, it becomes less and less likely that the number of heads is far away from what we expect.

Example 3.8. Here's a geometric example that shows that our intuitions about geometry break down in higher dimensions. We'll show that for any fixed $\epsilon > 0$, at least a $(1 - \epsilon)$ proportion of the volume of the cube $[-1, 1]^n$ lives in the region $(1 - \epsilon)\sqrt{n/3} \leq |x| \leq (1 + \epsilon)\sqrt{n/3}$ for n sufficiently large. That is, we can fit a lot of volume in a small distance in high dimensions.

Let $X_1, \dots, X_n$ be independent random variables uniformly sampled from $[-1, 1]$. This way, the vector $\vec{X} = (X_1, \dots, X_n)$ is uniformly sampled on the cube $[-1, 1]^n$. Let's apply the weak law to the variables X_i^2. We have

$$E[X_i^2] = \int_{-1}^{1} x^2 \cdot \frac{1}{2}\, dx = \frac{1}{3}.$$

By the weak law, for n sufficiently large, $\frac{X_1^2+\cdots+X_n^2}{n}$ is close to 1/3. Taking square roots, $|\vec{X}|/\sqrt{n}$ is within, say, $\epsilon/\sqrt{3}$ with probability approaching 1, i.e.

$$\Pr\left[\left|\frac{\sqrt{X_1 + \cdots + X_n}}{\sqrt{n}} - \frac{1}{\sqrt{3}}\right| \leq \frac{\epsilon}{\sqrt{3}}\right] = \Pr[(1-\epsilon)\sqrt{n/3} \leq |\vec{X}| \leq (1+\epsilon)\sqrt{n/3}] \to 1$$

as $n \to \infty$. But this is the fraction of the volume of the cube inside the annulus $(1 - \epsilon)\sqrt{n/3} < |x| < (1 + \epsilon)\sqrt{n/3}$.

4 The Central Limit Theorem

The central limit theorem is one of the most important theorems in statistics. It's also one of the most important uses of moment generating functions as we'll see. It basically answers the question "what's so 'normal' about the normal distribution?"

Theorem 4.1 (The Central Limit Theorem). *Let $X_1, X_2, \dots$ be a sequence of iid random variables with mean μ and variance σ^2. Then for any t,*

$$\Pr\left[\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq t\right] \to \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx,$$

as $n \to \infty$.

Remark 4.2. Maybe it's clear that we subtract off $n\mu$ because this makes the numerator in the above expression have mean zero, but where's the $\sigma\sqrt{n}$ coming from? Since the X_i's are iid, the variance of $X_1 + \cdots + X_n$ is $n\sigma^2$. Dividing by $\sigma\sqrt{n}$ then makes $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ have variance 1.

So in other words, centering and dividing by the variance makes a sum of iid random variables approach a standard normal random variable.

The main technical tool we'll use to prove this is a result about moment generating functions.

Lemma 4.3. *Let $Z_1, Z_2, \dots$ be a sequence of random variables, where Z_n has distribution function F_n and moment generating function M_n. Moreover, let Z be a random variable with distribution function F and moment generating function M. If for all t,*

$$\lim_{n\to\infty} M_n(t) = M(t),$$

then

$$\lim_{n\to\infty} F_n(t) \to F(t)$$

for all t at which $F(t)$ is continuous.

So basically, to show that a family of distributions converges to some distribution, it suffices to show that the moment generating functions converge to the corresponding mgf.

Proof of the Central Limit Theorem. Let $S_n = X_1 + \cdots + X_n$. To make the notation a bit cleaner, we'll assume that $\mu = 0$ and $\sigma = 1$. The above lemma tells us that it's enough to prove that the mgf of $\frac{S_n}{\sqrt{n}}$ converges to that of a standard normal random variable, which is $e^{t^2/2}$. To this end, let $M(t)$ be the mgf of X_i. Then the mgf of $X_i/\sqrt{n}$ is $M(t/\sqrt{n})$. Since the X_i's are independent, the mgf of their sum is the product of the mgf's so the mgf of $S_n/\sqrt{n}$ is

$$M\left(\frac{t}{\sqrt{n}}\right)^n.$$

So our goal is to show that $M(t/\sqrt{n})^n \to e^{t^2/2}$ for all t. If we let $L(t) = \ln M(t)$, then this is equivalent to showing that $nM(t/\sqrt{n}) \to t^2/2$. Now as $n \to \infty$, $t/\sqrt{n} \to 0$ for any t.

We then have

$$\lim_{n\to\infty} L(t/\sqrt{n}) = L(0) = \ln M(0) = \ln 1 = 0.$$

This gives

$$\lim_{n\to\infty} nL(t/\sqrt{n}) = \lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} = \frac{0}{0}.$$

Looks like we'll need L'Hopital's rule. We have that

$$L'(t) = \frac{M'(t)}{M(t)} \to \frac{M'(0)}{M(0)} = \frac{E[X_i]}{1} = 0$$

as $t \to 0$. L'Hopital then says

$$\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} = \lim_{n\to\infty} \frac{-(1/2)tn^{-3/2}L'(t/\sqrt{n})}{-1/n^2} = \frac{0}{0}.$$

Let's try L'Hopital's rule one more time. The quotient rule says

$$L''(t) = \frac{M(t)M''(t) - M'(t)^2}{M(t)^2} \to M''(0) - M'(0)^2 = E[X_i]^2 - E[X_i]^2 = \operatorname{Var}[X_i] = 1.$$

This gives

$$\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} = \lim_{n\to\infty} \frac{tL(t/\sqrt{n})}{2\cdot 1/n^{1/2}} = \lim_{n\to\infty} \frac{-(1/2)t^2n^{-3/2}L''(t/\sqrt{n})}{2\cdot(-1/2)n^{-3/2}} = t^2/2.$$

Hence, $M(t/\sqrt{n})^n \to e^{t^2/2}$, so by our lemma, $S_n/\sqrt{n}$ approaches a standard normal random variable in distribution. □

Example 4.4. Let $X \sim \text{Pois}(100)$. Let's estimate $\Pr[X \geq 120]$ using the central limit theorem. The central limit theorem looks like it needs a sum of random variables though, and we don't have a sum. We can turn X into a sum though. Recall that if $Y_1 \sim \text{Pois}(\lambda_1)$ and $Y_2 \sim \text{Pois}(\lambda_2)$ are independent, then $Y_1 + Y_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$ (try to prove this if you don't remember it). Then we can write $X = X_1 + \cdots + X_{100}$, where $X_i \sim \text{Pois}(1)$. Now we can apply the central limit theorem. Since $E[X_i] = \operatorname{Var}[X_i] = 1$, we have

$$\Pr[X \geq 120] = \Pr[X_1 + \cdots + X_{100} \geq 120] = \Pr\left[\frac{X_1 + \cdots + X_{100} - 100}{10} \geq \frac{120 - 100}{10}\right] \approx 1 - \Phi(2) \approx .02275.$$

Example 4.5. Roll ten dice. Let's estimate the probability that their sum is between 30 and 40.

Let X_i be the value shown on the i-th die. Then $E[X_i] = 7/2$ and $\mathrm{Var}[X_i] = 35/12$. We then have

$$\begin{aligned}\Pr[30 \leq X_1 + \cdots + X_{10} \leq 40] &= \Pr\left[\frac{30-35}{\sqrt{350/12}} \leq \frac{X_1 + \cdots + X_{10} - 35}{\sqrt{350/12}} \leq \frac{40-35}{\sqrt{350/12}}\right] \\ &\approx \Phi(0.926) - \Phi(-0.926) \\ &\approx .64555.\end{aligned}$$

References

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