Random Variables Pt. 2

A Little More Complexity

Elliot Pickens

May 25, 2021

1 Intro

The content of this post will center around slightly more advanced distributions than the univariate ones I covered in part one. We'll begin by taking the natural step of combining two univariate distributions into a single bivariate distribution and then work towards more general multivariate distributions. Along the way we'll touch on marginal and conditional distributions, which are key to understanding multivariate distributions and the many methodologies that use them.

2 Bivaraite Distributions

Some processes cannot be modeled with a distribution based upon a single random variable. When problems become complicated and data gets more varied we need to incorporate multiple random variables at the same time. It's easy enough to work with a number of different random variables at the same time, but what if we want to unify them under a single distribution? To do that we have to explore the world of joint distributions. We'll start that search with the simplest joint distribution: the joint distribution of two random variables, which is also known as a bivariate distribution.

2.1 What is a Bivariate Distribution?

Recall that the distribution of a random variable is the collection of probabilities tied to its potential events. Once we have two random variables their *joint* or *bivariate* distribution is the set of probability values for every possible pair of events that our variables could produce. More precisely, the distribution consists of all the values of the form $P((X, Y) \in C)$ where C is the set of all pairs of reals (X, Y) that are events.

2.1.1 Discrete Bivariate Distributions

We have discrete and continuous *joint* distributions, just as we do for solo random variables. For example, if we have only a finite (or possibly countable) number of order pairs (X, Y) for two random variables X & Y then their joint distribution is a discrete one. Naturally, such a joint distribution is most likely to occur when both X & Y are discrete, since both X & Y must have finite or countable number of possible values and therefore the two of them cannot have more than a countable number of pairs.

We'll get into continuous joint distributions soon, but before we can get there we need to introduce joint probability functions. These functions which are often shortened to "joint *pf*" are defined on the xy-plane as follows:

$$f(x, y) = P(X = x \,\&\, Y = y) \tag{1}$$

For discrete joint pfs the probability of a set of ordered pairs (where $f(x, y) = 0$ if (x, y) is not a possible pair) is

$$P((x, y) \in C) = \sum_{(x,y) \in C} f(x, y) \tag{2}$$

And the probability of all possible pairs results in the equality

$$\sum_{All\ (x,y)} f(x,y) = 1 \tag{3}$$

You might recognize that the equations we laid out for discrete joint pfs are near identical to those of their univariate cousins. The only difference in these equations is that they depend on ordered pairs rather than single points. In both cases we are summing over a set of events to find probabilities, but I would like to draw attention to our use of a single summation in these definitions. Does it make sense to frame the task of solving for a probability via a joint pf as one summation? The answer is sometimes, but not always. If we are summing over multiple values from each random variable then it is much easier in practice to write things out as a nested summation where each variable get its own term. I won't get in too much detail on this here, but it should become more explicitly clear why a double sum is useful when we restate things in the continuous context.

2.1.2 Continuous Bivariate Distributions

We say that two random variables X & Y have a continuous bivariate distribution if they can be used to define a function f on the xy-plane with the property that the probability of any subset of the plane C is

$$P((x,y) \in C) = \int_C \int f(x,y)dxdy \tag{4}$$

When this integral exists we call f the joint probability density function, or joint pdf of X & Y. As was the case of single dimensional pdfs, joint pdfs have a closure (or support) defined as $\{(x,y)|f(x,y) > 0\}$.

A valid joint pdf must abide by two conditions:

$$f(x,y) \geq 0\ for\ -\infty < x < \infty\ and\ -\infty < y < \infty \tag{5}$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1 \tag{6}$$

These conditions should be reminiscent of those for single variable pdfs. The conditions have just been extended to include the second variable present in bivariate pdfs, while otherwise remaining unchanged - as has been the case for all of the bivariate definitions we've encountered.

The consequences of these definitions also carry over. The use of integrals in these definitions still implies that any finite or countable set of points on the plane will have a probability value of 0. When dealing with bivariate distributions, however, we must remember that when integrating to find probability we are solving for a "volume" rather than an area. This implies that slices of the distribution with 0 "volume" have 0 probability. Therefore integration over any set of points where we hold the value of one of our two variables constant will have a probability of 0. More formally, any sets of the form $\{(x,y)|y = f(x), a < x < b\}$ or $\{(x,y)|x = f(y), a < y < b\}$ have a probability of 0.

2.1.3 Mixed Bivariate Distributions

What if we want to find the joint distribution of multiple random variables when some are continuous and others are discrete. At first, it might seem odd to be blending these different types of random variables, but it's far more common and far less complicated than you might think. To make it work we just need to divide things along the lines the two variable types. That is to say that we integrate where we need to integrate, and sum where we need to sum.

To put it all together, we'll build a mixed bivariate distribution by modifying the integral in equation 7 we used to define the bivariate continuous pdf. Assume we have a discrete random variable X and a continuous random variable Y. And imagine we also have a function $f(x,y)$ that is defined on the xy-plane such that for any pair A & B of sets of real numbers,

$$P(X \in A \, and \, Y \in B) = \int_B \sum_{x \in A} f(x,y)dy \tag{7}$$

Then f is the joint pdf of X & Y if such an integral exists. Of course, if we were to switch things up and make Y continuous and X discrete then we could simply rearrange the equation above to match the integral and sum to their correct variables. It is also possible to switch the order of the sum and integral should it be easier one way or another (and proper care is taken during the rearrangement).

Now that we have our new joint pdf we can quickly restate the properties 5 & 6 by saying that for a discrete X and continuous Y $f(x,y) \geq 0 \, \forall \, x, y$ and

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} f(x_i, y)dy = 1 \tag{8}$$

since the total probability must be 1 as always.

2.1.4 Bivariate CDFs

Having introduced several different types of pfs and pdfs it's time to touch on their cdfs. The bivariate cumulative distribution function of two random variables X & Y is the function F defined on all values of x & y ($-\infty < x < \infty$ & $-\infty < y < \infty$) as

$$F(x,y) = P(X \leq x \, and \, Y \leq y) \tag{9}$$

When both variables are continuous we can define this function as

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(r,s)drds \tag{10}$$

where r and s are dummy integration variables. We can then use this definition of the function to derive f by taking the derivative of F as

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \frac{\partial^2 F(x,y)}{\partial y \partial x} \tag{11}$$

where the this derivative exists.

Since probability cannot decrease as we include more points in our region of interest, bivariate cdfs are monotonically increasing in both X and Y (when the other is fixed). This allows use to find the probability of an interval by subtracting pdfs like

$$\begin{aligned} P(a < X \leq b \,\&\, c < Y \leq d) &= P(a < X \leq b \,\&\, Y \leq d) - P(a < X \leq b \,\&\, Y \leq c) && (12) \\ &= P(X \leq b \,\&\, Y \leq d) - P(X \leq a \,\&\, Y \leq d) && (13) \\ &\;\; - P(X \leq b \,\&\, Y \leq c) - P(X \leq a \,\&\, Y \leq c) && (14) \\ &= F(b,d) - F(a,d) - F(b,c) - F(a,c) && (15) \end{aligned}$$

It is also possible to derive the single variable cdfs from the bivariate ones they're included in by using limits. If we want the single variable cdf of X, F_1 then we can derive by taking the limit

$$F_1(x) = lim_{y \to \infty} F(x,y) \tag{16}$$

Alternatively we could get F_2 for Y by taking the limit

$$F_2(y) = lim_{x \to \infty} F(x,y) \tag{17}$$

I'm going to leave out the proof that such limits do in fact produce the single variable versions of the cdf, but it can easily be worked out by observing that the limit $lim_{m \to \infty} P(\{X \leq x \,\&\, Y \leq m\})$ is equal to the sum of the probability of $\{X \leq x \,\&\, n-1 < Y < n\} \, \forall \, n$.

3 Marginal Distributions

We just saw how the cdf of a (bivariate) joint distribution can be used to derive the cdf of just one of the random variables used to construct it. Oddly enough, working backwards from a joint distribution to isolate a single random variable is often necessary. When we derive the distribution of a random variable X from a joint distribution we call it the marginal distribution of X. All random variables present in a joint distribution have a marginal distribution, and each of those distributions comes complete with its own marginal cdf and pf/pdf.

3.1 PF/PDFs of Marginal Distributions

Although they weren't introduced as marginal cdfs, the single variable cdfs we derived in 16 and 17 are marginal cdfs of their respective random variables. The pf/pdfs tied to these cdfs are the marginal pf/pdfs of their random variables.

Stepping away from cdfs, if we want to analyze marginal pdfs and pfs in isolation it's best to start from a joint pf or pdf. To do this let's assume that we have a discrete joint distribution of the random variables X and Y with a joint pf f. Then the marginal pf of X is

$$f_1(x) = \sum_{All\ y} f(x,y) \tag{18}$$

and the marginal pf of Y is

$$f_2(y) = \sum_{All\ x} f(x,y) \tag{19}$$

The implication of these sums is that the probability that $X = x$ or $Y = y$ is found by inspecting the union of all events where $X = x$ or $Y = y$. In other words we have to look at all the ordered pairs where the single variable event of interest occurs.

Finding the marginal pdfs of continuous distributions follows the same process, but with the standard swapping of a sum for an integral. Thus, for a joint continuous distribution of X and Y the marginal pdf of a random variable X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad for\ -\infty < x < \infty \tag{20}$$

and the marginal pdf of Y is the same as 20 with y substituted for x and dy converted to dx. We can easily show that f_1 is a marginal pdf by noticing that $P(X \leq x) = P((X,Y) \in C)$ where $C = \{(r,s)|r \leq x\}$ which is

$$\begin{aligned} P((X,Y) \in C) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(r,s)dsdr \qquad (21) \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(r,s)ds \right] dr \qquad (22) \end{aligned}$$

and that $\int_{-\infty}^{\infty} f(r,s)ds = f_1(r)$. Then $P(X < x) = \int_{-\infty}^{x} f_1(r)dr$ and f_1 is our marginal pdf.

In the case of a mixed bivariate distribution, where one random variable is continuous and the other is discrete the marginal distributions maintain the nature of their parent variable. This means that if we have a mixed joint distribution with a continuous X and discrete Y the pdf of X is

$$P(X = x) = f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \forall\, x \tag{23}$$

and the pf of Y is

$$f_2(y) = \sum_{All\ x} f(x,y) \quad \forall\, y \tag{24}$$

3.2 Independence

Many probability questions and statements have included the necessary note that "independence is assumed" or something of the like. Intuitively we understand that saying two random variables are independent means that they have no relation to one another and that the state of one doesn't impact the state of the other. If, however, we want to be more precise we can say that two random variables are independent if for any two sets of real A and B where $\{X \in A\}$ and $\{Y \in B\}$

$$P(X \in A \,\&\, Y \in B) = P(X \in A)P(Y \in B) \tag{25}$$

By this equation we get that X and Y are independent so long as all possible events of the sort like $\{X \in A\}$ and $\{Y \in B\}$ are independent. And we can also say that the independence of X and Y implies that

$$P(X \leq x \,\&\, Y \leq y) = P(X \leq x)P(Y \leq y) \tag{26}$$

for all reals x and y. We can also see that 26 implies 25, so long as 26 is true for all real numbers.

Independence also impacts cdfs. The joint cdf of two independent random variables is the product of their solo cdfs. We can strengthen this statement by asserting that two variables are independent if and only if $F(x, y) = F_1(x)F_2(y)$ for all reals x and y.

The joint pf/pdf of independent random variables mirrors what we just stated was true of joint cdfs. Two random variables are only independent if and only if their joint pf/pdf f can be written as

$$f(x, y) = h_1(x)h_2(y) \tag{27}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$, where h_1 is a nonnegative function of x and h_2 is a nonnegative function of y. Although we used generic nonnegative functions in 27 we can narrow the statement with the corollary that

$$f(x, y) = f_1(x)f_2(y) \tag{28}$$

for any two independent random variables X and Y.

Before wrapping up this section I would like to end with a few comments on the interpretation of independence. All the definitions I've laid out state that any entanglement that would normally be present in a joint distribution can be completely negated through factorization. That factorization is a mathematical way of saying that knowing that event that one of the random variables takes on tells us nothing about what event the other might take on. For example, for any y we might learn that $Y = y$, but this has zero impact on any of the events $\{X = x\}$. Alternatively we could say that the probability $X = x$ given $Y = y$ is just $P(X = x)$, or in the grammar of conditional probability $P(X = x|Y = y) = P(X = x)$.

4 Conditional Distributions

When we ask for the probability of some event A given some event B has already taken place we're asking to solve for the conditional probability of A given B. This sort of calculation is incredibly useful, but what if we want to generalize it to work on the level of distributions? Then we instead ask how some random variable Y affects another random variable X, and answer by finding the conditional distribution of X given Y.

Recall that the definition of conditional probability is $P(A|B) = \frac{P(A \cap B)}{P(B)}$. We can generalize this definition to work for two random variables X and Y with a joint distribution f by defining the conditional distribution of X given Y (where $f_2(y) > 0$) as

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} \tag{29}$$

Similarly, $g_2(y|x) = \frac{f(x,y)}{f_1(x)}$ (where $f_1(x) > 0$) would be the conditional distribution (or conditional pf/pdf) of Y given X.

4.1 Discrete Conditional Distributions

When X and Y have a discrete joint distribution we can confirm that their conditional distributions are true pfs. Looking at $g_1(x|y)$, assume $f_2(y) > 0$, so $g_1(x|y) \geq 0 \,\forall\, x$. Then

$$\sum_x g_1(x|y) = \frac{1}{f_2(y)} \sum_x f(x,y) = \frac{1}{f_2(y)} f_2(y) = 1 \tag{30}$$

Thus, we know that g_1 is a proper pf. $g_2(y|x)$ abides by the same logic and can be verified by exchanging variables. We should also note that for g_1, g_2, and all other conditional distributions we can define them as needed for values that would result in a zero in the denominator.

4.2 Continuous Conditional Distributions

The formula for the pdf of a continuous conditional distribution is no different than that of its discrete counter part. In fact, it is exactly the same as 29, but the formula does not tell us everything. For continuous conditional distributions we cannot just define our pdf at the points where $f_2(y) > 0$. Instead we must define it over all possible values of our random variable of interest. Thus, $g_1(x|y)$ is defined for $-\infty < x < \infty$, and can even be defined for points where $f_2(y) = 0$ (so long as g_1 remains a valid pdf). We can verify that 29 is a valid pdf by noticing that if $f_2(y) = 0$ then we can just choose a pdf for g_1, and for all other cases where $f_2(t) > 0$

$$\int_{-\infty}^{\infty} g_1(x|y)dx = \int_{-\infty}^{\infty} \frac{f(x,y)}{f_2(y)} dx = \frac{1}{f_2(y)} \int_{-\infty}^{\infty} f(x,y)dx = \frac{f_2(y)}{f_2(y)} = 1 \tag{31}$$

In might seem mundane to have to read through me repeatedly state and then restate each of these definitions for both discrete and continuous distributions, and in practice much of this is somewhat inconsequential sans a swapping of a sum for a integral here or there. But if we want to really understand what is going on behind the curtain we need to pay as close attention as possible. An example of why this matters is present here when we ask the question of what it means to condition X on Y when $Y = y$. Each individual event $\{Y = y\}$ has zero probability if Y is a continuous distribution, so what are we doing when we condition on it? The answer is that we aren't really conditioning on $\{Y = y\}$, but rather $\{y - \epsilon < Y < y + \epsilon\}$. This allows us to work with an interval in Y that is sufficiently large for a (possibly) non zero-probability, while not being large enough to make a visible difference to us as users of the formula (given we can choose ϵ to be as small as we want). Therefore, in the continuous case we are really just conditioning X on Y where Y is incredibly close to y.

4.2.1 Mixed Conditional Distributions

The conditional distribution of a mixed distribution follows 29 just like pure joint distributions. We just have to remember to swap variables and summation/integration terms as needed.

4.3 Multiplication Rule for Conditional Distributions (and More)

The rule that $P(A \cap B) = P(A)P(B|A)$ for individual events, and can be easily expanded to work for pf/pdfs of distributions. To show this, let X and Y have pf/pdfs f_1 and f_2 respectively, and combine to form the joint distribution f with conditional distributions g_1 and g_2. Then for each x and y such that $f_2(y) > 0$

$$f(x,y) = g_1(x|y)f_2(y) \tag{32}$$

and

$$f(x,y) = g_2(y|x)f_1(x) \tag{33}$$

4.3.1 Law of Total Probability for Distributions

Given the relationship between conditional distributions and joint pf/pdfs shown in 32 and 33 we can easily generalize the law of total probability to the distribution level by noticing that we can get the marginal distributions f_1 from f as

$$f_1(x) = \sum_y g_1(x|y) f_2(y) \tag{34}$$

for a discrete random variable Y and

$$f_1(x) = \int_{-\infty}^{\infty} g_1(x|y) f_2(y) dy \tag{35}$$

And we can also get $f_2(y)$ if we wanted to by simply swapping things around to set things up correctly.

4.3.2 Bayes' Theorem for Random Variables

Recall that Bayes' Theorem for two events is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. For distributions it remains the same, but with individual probabilities replaced by pf/pdfs to create

$$g_1(x|y) = \frac{g_2(y|x) f_1(x)}{f_2(y)} \tag{36}$$

by equating 32 and 33 and then rearranging them as needed.

4.4 Conditional Distributions of Independent Random Variables

What if X is conditioned on Y when the two random variables are independent? Well we stumble upon yet another definition of independence. That is that two random variables are independent if and only if for all values x and all values of y where $f_2(y) > 0$ we get

$$g_1(x|y) = f_1(x) \tag{37}$$

We can show this by remember that two random variables are independent if and only if $f(x, y) = f_1(x)f_2(y)$ for $-\infty < x < \infty$ and $-\infty < y < \infty$. Then for $f_2(y) > 0$ we get $f_1(x) = \frac{f(x,y)}{f_2(y)}$ and since $\frac{f(x,y)}{f_2(y)} = g_1(x|y)$ then $f_1(x) = g_1(x|y)$ hold for all x & y where $f_2(y) > 0$.

4.5 Conditional Distributions are Distributions!

Although the setup of conditional distributions can be a little confusing it is important to remember that they are distributions and act like them! Everything we know about the behavior of regular distributions also applies here whether we're solving for probability on an interval or doing anything else!

5 Multivariate Distributions

Now that we've had a look at joint, marginal, and conditional distributions we can generalize them to an arbitrary number of random variables. In applied settings (especially in the era of big data) we often want to build models based off of many different variables, or make inferences from complex data sets where intricate relationships are present. To make these sorts of models and analyses work we need to make sure we can manipulate and combine as many random variables as we could possibly need.

In this section I'm going to spend some time revising what I've shown leading up to this point in the post to make sure they work in a multivariate setting. Please beware that this section is going to cover quite a few different definitions in rapid succession without very much exposition unless otherwise necessary. Hopefully it'll become clear why that is the case while reading through it, but the basic fact of importance here is that these definitions are almost identical to their bivariate siblings.

5.1 Vector Notation for Random Variables

When workings with a large sequence of random variables it is not convenient or expeditious to continually state something along the lines of "the joint distributions of $X_1, X_2, ..., X_n$..." each time we'd like to talk about a problem with n random variables involved. Instead we can use vector notation. A "random vector" $\mathbf{X}$ is the vectorized form of the n random variables $X_1, ..., X_n$, which can be written as $\mathbf{X} = (X_1, ..., X_n)$. We can use these vectors to shorten common statements like $F(X_1, ..., X_n)$ down to a concise $F(\mathbf{x})$. Keep in mind that each random vector $\mathbf{X}$ is n-dimensional and therefore whatever function it is inserted into (like $F(x)$) must be defined on that same n-dimensional space.

Throughout the remainder of this section I'll be primarily using vector notation. I'll try my best to put these vectors in bold.

5.2 Multivariate Joint Distributions

For n random variables $X_1, ..., X_n$ the joint cdf of all n variables is the function F whose output is defined in $\mathbb{R}^n$ as

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n) \tag{38}$$

or in vector notation

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) \tag{39}$$

This multivariate cdf operates in an identical manner to the univariate and bivariate cdfs we've already seen. It also satisfies all of the properties we laid out earlier in this post for bivariate cdfs.

5.2.1 Discrete Distributions

For discrete joint distributions where our random vector $\mathbf{X}$ can only take on a finite or countable number of values we define the joint pf as

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) \tag{40}$$

for all $\mathbf{x} \in \mathbb{R}^n$. We can also find the probability of a subset C of $\mathbb{R}^n$ by computing

$$P(\mathbf{X} \in C) = \sum_{x \in C} f(\mathbf{x}) \tag{41}$$

5.2.2 Continuous Distributions

The pdf of purely continuous joint distributions is defined as a non-negative function for which the integral

$$P((X_1, ..., X_n) \in C) = \int_C ... \int f(x_1, ..., x_n) dx_1 ... dx_n \tag{42}$$

exists on all intervals $C \subset \mathbb{R}^n$. In it's vectorized form we can write it as

$$P(\mathbf{X} \in C) = \int_C ... \int f(\mathbf{x}) d\mathbf{x} \tag{43}$$

We can also define f as the derivative of its cdf (at least all points where its derivative exists). To do this we just need to take the partial derivative of F for all n random variables to get

$$f(\mathbf{x}) = \frac{\partial^n F(\mathbf{X})}{\partial x_1 ... \partial x_n} \tag{44}$$

5.2.3 Mixed Distributions

Some problems require a mixture of continuous and discrete random variables. For a joint distributions of this sort we might have j continuous random variables and k discrete random variables (with $j+k=n$). Under such conditions we just have to make sure to mix our sums and integrals correctly. With bivariate mixed distributions we always had one summation term associated with the discrete variable and an integral tied to the continuous one. The same rule applies in the general context. Summations go with the discrete variables and integrals go with the continuous ones. Returning to our example of solving for $P(\mathbf{X} \in C)$, if $\mathbf{X}$ is mixed then our solution will include j summations and k integrals rather than just n integrals.

5.3 Marginal Distributions

A joint distribution of n random variables also has marginal distributions for each of the random variables. Let's assume we have a joint distribution of the random variables $X_1, X_2, ..., X_n$ and we want to find the marginal distribution of X_1. Then we can use the same process we used in 20, and integrate over all variables except X_1 to find its marginal distribution f_1. The sort of nested integral we would need to find f_1 would look like

$$f_1(x_1) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{n-1} f(x_1...x_n)dx_2...dx_n \tag{45}$$

Marginal distributions of just one random variable are fairly straight forward, but it is possible to generalize the concept of a marginal distribution further when working with more than two random variables. We can find the marginal distribution of any k random variables (where $k < n$). All we need to do is integrate over the $n-k$ variables we are not interested in including in our marginal distribution. Imagine once again we have n random variables $X_1, ..., X_n$, but this time we want to find a marginal distribution of the first k random variables. That is to say we want to find the marginal distribution of all X_i where $i \leq k$. To do this we could use the calculation

$$f_{1...k}(x_1, x_2, ..., x_k) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{n-k} f(x_1...x_n)dx_{k+1}...dx_n \tag{46}$$

We can also find a marginal cdf of a joint distribution in the same way we did in 16. By expanding upon that relation to work in the context of having an arbitrary number of variables we can find the marginal cdf F_1 of X_1 from a joint cdf F via the relation

$$F_1(x_1) = P(X_1 \leq x_1) = P(X_1 \leq x_1, X_2 < \infty, ..., X_n < \infty) \tag{47}$$
$$= \lim_{x_2,...,x_n \to \infty} F(x_1, x_2, ..., x_n) \tag{48}$$

5.4 Independence

Independence can be generalized to greater quantities of random variables if we modify our original definitions. The first way we can do this is to say that n random variables are independent of one another if for every n set of real numbers: $A_1, A_2, ..., A_3$ the following relationship holds:

$$P(X_1 \in A_1, X_2 \in A_2, ..., X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)...P(X_n \in A_n) \tag{49}$$

Second, we can say that if our n random variables have a joint cdf F then they can only be independent if and only if

$$F(x_1, ..., x_2) = F_1(x_1)...F_n(x_n) \tag{50}$$

is true for all points $(x_1, ..., x_n) \in \mathbb{R}^n$. In plain terms, our collection of random variables can only be independent if their joint cdf can be written as the product of all of their marginal cdfs.

Third, if our n random variables have a joint pf/pdf then they are independent if and only if their joint pf/pdf f can be written as the product of their individual marginal pf/pdfs at all points $(x_1, ..., x_n) \in \mathbb{R}^n$. We can write out this relation as

$$f(x_1, ..., x_2) = f_1(x_1)...f_n(x_n) \tag{51}$$

5.4.1 Random Samples

Let's assume we have a distribution with a pf/pdf f. Then we say that n random variables $X_1, ..., X_n$ are a random sample if they are all independent of one another and all have the marginal pf/pdf f. We can also say that these random variables are independent and identically distributed or i.i.d. with a sample size of n.

The joint distribution of our i.i.d. sample is then

$$g(x_1, ..., x_2) = f(x_1)...f(x_n) \tag{52}$$

5.5 Conditional Distributions

In 46 we found the marginal distribution of $k < n$ random variables. Let's call this marginal distribution $f_{1...k}$. Now how might we use that marginal distribution to find the conditional distribution of the remaining $n-k$ random variables? The answer lies in 16. Using that relation (conditional $= \frac{\text{total pf/pdf}}{\text{marginal}}$) we can then solve for the conditional distribution of $X_{k+1}, ..., X_n$ given $X_1 = x_1, ..., X_k = x_k$ using

$$g_{k+1...n}(x_{k+1}, ..., x_n | x_1, ..., x_k) = \frac{f(x_1, ..., x_n)}{f_{1...k}(x_1, ..., x_k)} \tag{53}$$

The definition of a multivariate conditional pf/pdf follows directly from these example. In this example we have the random vector $\mathbf{X} = (X_1, ..., X_n)$, and the two sub-vectors $\mathbf{Y} = (X_1, ..., X_k)$ and $\mathbf{Z} = (X_{k+1}, ..., X_n)$. We also have the joint pf/pdf of $\mathbf{X}$ (which is also $(\mathbf{Y}, \mathbf{Z})$) f and the marginal pf/pdf of $\mathbf{Y}$ $f_{1...k}$. Then the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ for every point $\mathbf{y} \in \mathbb{R}^k$ where $f_{1...k}(\mathbf{y}) > 0$ is

$$g_1(\mathbf{z}|\mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{z})}{f_{1...k}(\mathbf{y})} \tag{54}$$

And can be rewritten as

$$g_1(\mathbf{z}|\mathbf{y}) f_{1...k}(\mathbf{y}) = f(\mathbf{y}, \mathbf{z}) \tag{55}$$

It is also safe to assume that $f(\mathbf{y}, \mathbf{z}) = 0$ when $f_{1...k}(\mathbf{y}) = 0$, so 54 holds for all $\mathbf{y}$ and $\mathbf{z}$, but it also means that g_1 is not unique given our interference at points where $f_{1...k}(\mathbf{y}) = 0$.

5.5.1 Bayes' Theorem and the Law of Total Probability

Using the definition of a multivariate conditional distribution shown in 54 we can show the marginal pdf of $\mathbf{z}$ is

$$f_{k+1...n}(\mathbf{z}) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{k} g_1(\mathbf{z}|\mathbf{y}) f_{1...k}(\mathbf{y}) d\mathbf{y} \tag{56}$$

using the total law of probability. And the conditional pdf of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ will be

$$g_2(\mathbf{y}|\mathbf{z}) = \frac{g_1(\mathbf{z}|\mathbf{y}) f_{1...k}(\mathbf{y})}{f_{k+1...n}(\mathbf{z})} \tag{57}$$

by Bayes' Theorem.

Once again 56 will need to be altered if we're working with a mixed or discrete joint distribution to include summations as needed.

5.5.2 Conditional Independence

We touched a little on how we can define two variables as being independent using conditional distributions in 37, but there is another type of independence that we have not yet touched upon. That is conditional independence. Conditional independence centers around whether some number of variables are independent given some random vector $\mathbf{Z}$. Thus, when checking for conditional independence we aren't testing whether the random variables are independent in the pure sense, but whether they're independent after being conditioned on some other set of random variables.

To define this, assume $\mathbf{Z}$ is a random vector with a joint pf/pdf $f_0(\mathbf{z})$. Then some set of n random variables $X_1, ..., X_n$ are conditionally independent given $\mathbf{Z}$ if we can say that for each $\mathbf{z}$ where $f_0(\mathbf{z}) > 0$

$$g(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{n} g_i(x_i|\mathbf{z}) \tag{58}$$

Therefore, our n random variables are conditionally independent if the conditional distribution ($g(\mathbf{x}|\mathbf{z})$) of $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$ is equal to the product of the conditional distributions of each individual random variables X_i.

Note: independence is a special case of conditional independence. To quickly sketch the proof, imagine that we have a set of independent random variables $X_1, ..., X_n$ that we condition on some constant random variable W where $P(W = c) = 1$. Then we want to show 58 is true. First, notice that W is independent to every X_i, so $g_i(x_i|w) = \dfrac{f(x_i, w)}{f_0(w)} = \dfrac{f_i(x_i)f_0(w)}{f_0(w)} = f_i(x_i)$ and $g(\mathbf{x}|w) = f(\mathbf{x})$ (by the same logic). Thus, $g(\mathbf{x}|w) = f(\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} g_i(x_i|w)$. As a result on it's own this is not particularly interesting, but it does give us the ability to say that anything we prove from conditionally independent random variables is also true of standard independent random vairables.

5.5.3 Conditional Editions of misc. Theorems

Before moving on from multivariate conditional distributions we should remind ourselves that just as we mentioned in section 4.5, conditional distributions are distributions. That means each and every one of the theorems we've touched on has a conditional version. Let's take 56 as an example. We can find it's conditional version by conditioning it on some other random vector $\mathbf{W} = \mathbf{w}$ to produce

$$f_{k+1...n}(\mathbf{z}|\mathbf{w}) = \underbrace{\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}}_{k} g_1(\mathbf{z}|\mathbf{y}, \mathbf{w}) f_{1...k}(\mathbf{y}|\mathbf{w}) d\mathbf{y} \tag{59}$$

where the structure of the equation remains the same, but with all terms being conditioned on $\mathbf{W} = \mathbf{w}$. Along the same lines we rewrite 57 using the say method to get the conditional version of Bayes' Theorem, which is

$$g_2(\mathbf{y}|\mathbf{z}, \mathbf{w}) = \frac{g_1(\mathbf{z}|\mathbf{y}, \mathbf{w}) f_{1...k}(\mathbf{y}|\mathbf{w})}{f_{k+1...n}(\mathbf{z}|\mathbf{w})} \tag{60}$$

Based on the examples 60 and 59 it might be apparent that generally we can find the conditional version of theorems of interest by conditioning all probabilistic terms of an equation on $\mathbf{W} = \mathbf{w}$. This should work on all probabilistic concepts from the most basic individual probability to expected values and beyond.

6 Conclusion

In this post we covered quite a bit. We touched on the most important facts about joint distributions without all too much detail, but hopefully enough to spark a little interest. If I didn't cover any particular topic in a satisfactory manner please feel free to reach out and let me know (I might just be inclined to go and rewrite it up). I plan on returning to a number of these topics in the future with a much finer toothed comb, but for now I hope this was a decent introduction.

In my next post (which I hope will be much shorter) I'll discuss functions of random variable. I think that'll more or less close out my series on the basics of random variables, but I might also include a bonus post if I'm feeling energetic.

7 Acknowledgments

These notes were based on *Probability and Statistics (Fourth Edition)* by DeGroot & Schervish.