Properties of Multiple Random Variables

1 Expectation with Multiple RVs

Expectation over a joint isn't nicely defined because it is not clear how to compose the multiple variables. However, expectations over functions of random variables (for example sums or multiplications) are nicely defined: $E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$ for any function $g(X,Y)$. When you expand that result for the function $g(X,Y) = X+Y$ you get a beautiful result:

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

This can be generalized to multiple variables:

$$
E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]
$$

2 Independence with Multiple RVs

Discrete

Two discrete random variables X and Y are called independent if:

$$
P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x, y
$$

Intuitively: knowing the value of X tells us nothing about the distribution of Y. If two variables are not independent, they are called dependent. This is a similar conceptually to independent events, but we are dealing with multiple *variables*. Make sure to keep your events and variables distinct.

Continuous

Two continuous random variables X and Y are called independent if:

$$
P(X \le a, Y \le b) = P(X \le a)P(Y \le b) \text{ for all } a, b
$$

This can be stated equivalently as:

$$
F_{X,Y}(a,b) = F_X(a)F_Y(b) \text{ for all } a, b
$$
$$
f_{X,Y}(a,b) = f_X(a)f_Y(b) \text{ for all } a, b
$$

More generally, if you can factor the joint density function then your continuous random variable are independent:

$$
f_{X,Y}(x,y) = h(x)g(y) \text{ where } -\infty < x, y < \infty
$$

Example 2

Let N be the # of requests to a web server/day and that $N \sim Poi(\lambda)$. Each request comes from a human (probability = p) or from a "bot" (probability = $(1-p)$), independently. Define X to be the # of requests from humans/day and Y to be the # of requests from bots/day.

Since requests come in independently, the probability of X conditioned on knowing the number of requests is a Binomial. Specifically:

$$(X|N) \sim Bin(N,p)$$
$$(Y|N) \sim Bin(N,1-p)$$

Calculate the probability of getting exactly i human requests and j bot requests. Start by expanding using the chain rule:

$$P(X = i, Y = j) = P(X = i, Y = j|X+Y = i+j)P(X+Y = i+j)$$

We can calculate each term in this expression:

$$P(X = i, Y = j|X+Y = i+j) = \binom{i+j}{i} p^i(1-p)^j$$

$$P(X+Y = i+j) = e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}$$

Now we can put those together and simplify:

$$P(X = i, Y = j) = \binom{i+j}{i} p^i(1-p)^j e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}$$

As an exercise you can simplify this expression into two independent Poisson distributions.

Symmetry of Independence

Independence is symmetric. That means that if random variables X and Y are independent, X is independent of Y and Y is independent of X. This claim may seem meaningless but it can be very useful. Imagine a sequence of events $X_1, X_2, \ldots$. Let A_i be the event that X_i is a "record value" (eg it is larger than all previous values). Is A_{n+1} independent of A_n? It is easier to answer that A_n is independent of A_{n+1}. By symmetry of independence both claims must be true.

3 Convolution (aka adding)

Convolution is the result of adding two different random variables together. For some particular random variables computing convolution has intuitive closed form equations. Importantly convolution is the sum of the random variables themselves, not the addition of the probability density functions (PDF)s that correspond to the random variables.

Independent Binomials with equal p

For any two Binomial random variables with the same "success" probability: $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$ the sum of those two random variables is another binomial: $X+Y \sim Bin(n_1 + n_2, p)$. This does not hold when the two distribution have different parameters p.

Independent Poissons

For any two Poisson random variables: $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$ the sum of those two random variables is another Poisson: $X+Y \sim Poi(\lambda_1 + \lambda_2)$. This holds when λ_1 is not the same as λ_2.

Independent Normals

For any two normal random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ the sum of those two random variables is another normal: $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

General Independent Case

For two general independent random variables (aka cases of independent random variables that don't fit the above special situations) you can calculate the CDF or the PDF of the sum of two random variables using the following formulas:

$$F_{X+Y}(a) = P(X + Y \leq a) = \int_{y=-\infty}^{\infty} F_X(a - y) f_Y(y) dy$$

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a - y) f_Y(y) dy$$

There are direct analogies in the discrete case where you replace the integrals with sums and change notation for CDF and PDF.

Example 1

Calculate the PDF of $X + Y$ for independent uniform random variables $X \sim Uni(0, 1)$ and $Y \sim Uni(0, 1)$? First plug in the equation for general convolution of independent random variables:

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a - y) f_Y(y) dy$$

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a - y) dy \qquad\qquad \text{Because } f_Y(y) = 1$$

It turns out that is not the easiest thing to integrate. By trying a few different values of a in the range $[0, 2]$ we can observe that the PDF we are trying to calculate is discontinuous at the point $a = 1$ and thus will be easier to think about as two cases: $a < 1$ and $a > 1$. If we calculate f_{X+Y} for both cases and correctly constrain the bounds of the integral we get simple closed forms for each case:

$$f_{X+Y}(a) = \begin{cases} a & \text{if } 0 < a \leq 1 \\ 2 - a & \text{if } 1 < a \leq 2 \\ 0 & \text{else} \end{cases}$$