Independence in Variables

Discrete

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

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

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:

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

This can be stated equivalently using either the CDF or the PDF:

$$F_{X,Y}(a,b) = F_X(a)F_Y(b) \text{ for all } a, b$$
$$f(X=x, Y=y) = f(X=x)f(Y=y) \text{ for all } x, y$$

More generally, if you can factor the joint density function then your random variable are independent (or the joint probability function for discrete random variables):

$$f(X=x, Y=y) = h(x)g(y)$$
$$\mathrm{P}(X=x, Y=y) = h(x)g(y)$$

Example: Showing Independence

Let N be the # of requests to a web server/day and that $N \sim \mathrm{Poi}(\lambda)$. Each request comes from a human with probability = p or from a "bot" with probability = $(1–p)$. Define X to be the # of requests from humans/day and Y to be the # of requests from bots/day. Show that the number of requests from humans, X, is independent of the number of requests from bots, Y.

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

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

To get started we need to first write an expression for the joint probability of X and Y. To do so, we use the chain rule:

$$\mathrm{P}(X=x, Y=y) = \mathrm{P}(X=x, Y=y|N=x+y)\,\mathrm{P}(N=x+y)$$

We can calculate each term in this expression. The first term is the PMF of the binomial $X|N$ having x "successes". The second term is the probability that the Poisson N takes on the value $x+y$:

$$\mathrm{P}(X=x, Y=y|N=x+y) = \binom{x+y}{x} p^x (1-p)^y$$
$$\mathrm{P}(N=x+y) = e^{-\lambda}\frac{\lambda^{x+y}}{(x+y)!}$$

Now we can put those together we have an expression for the joint:

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

At this point we have derived the joint distribution over X and Y. In order to show that these two are independent, we need to be able to factor the joint:

$$\begin{aligned}
&\mathrm{P}(X = x, Y = y) \\
&= \binom{x+y}{x} p^x (1-p)^y e^{-\lambda} \frac{\lambda^{x+y}}{(x+y)!} \\
&= \frac{(x+y)!}{x! \cdot y!} p^x (1-p)^y e^{-\lambda} \frac{\lambda^{x+y}}{(x+y)!} \\
&= \frac{1}{x! \cdot y!} p^x (1-p)^y e^{-\lambda} \lambda^{x+y} && \text{Cancel (x+y)!} \\
&= \frac{p^x \cdot \lambda^x}{x!} \cdot \frac{(1-p)^y \cdot \lambda^y}{y!} \cdot e^{-\lambda} && \text{Rearrange}
\end{aligned}$$

Because the joint can be factored into a term that only has x and a term that only has y, the random variables are independent.

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 $X_1, X_2, \dots$. 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.

Expectation of Products

Lemma: Product of Expectation for Independent Random Variables:
If two random variables X and Y are independent, the expectation of their product is the product of the individual expectations.

$$\begin{aligned}
&E[X \cdot Y] = E[X] \cdot E[Y] && \text{if } X \text{ and } Y \text{ are independent} \\
&E[g(X)h(Y)] = E[g(X)]E[h(Y)] && \text{where } g \text{ and } h \text{ are functions}
\end{aligned}$$

Note that this assumes that X and Y are independent. Contrast this to the sum version of this rule (expectation of sum of random variables, is the sum of individual expectations) which does **not** require the random variables to be independent.