Lisa Yan
CS109

Lecture Notes #13
May 4, 2020

Statistics of Multiple RVs

Based on a chapter by Chris Piech and Lisa Yan

As you can imagine, reporting probability mass functions or distributions is often not ideal: We either have to find a common distribution that fits our experiment, or we have to report a probability table or a bar graph. In the single random variable case, we often report expectation or variance as *statistics* that characterize our randomness. A similar paradigm applies for the multiple random variable case! In this section, we discuss statistics of two random variables; in particular, (1) how to easily calculate the expectation of the sum of multiple random variables, and (2) how to report how two random variables vary with one another.

1 Expectation with Multiple RVs

Expectation over a joint distribution is not nicely defined because it is not clear how to compose the multiple variables. However, expectations over functions of random variables (for example sums or products) 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:

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

Let's go back to our old friends—the Binomial and Negative Binomial RVs—and show how we could have derived expressions for their expectation.

Expectation of Binomial

First let's start with some practice with the sum of expectations of indicator variables. Let $Y \sim \text{Bin}(n, p)$, in other words if Y is a Binomial random variable. We can express Y as the sum of n Bernoulli random indicator variables $X_i \sim \text{Ber}(p)$. Since X_i is a Bernoulli, $E[X_i] = p$

$$Y = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

Let's formally calculate the expectation of Y:

$$E[Y] = E[\sum_{i}^{n} X_i] = \sum_{i}^{n} E[X_i] = E[X_0] + E[X_1] + \ldots E[X_n] = np$$

Expectation of Negative Binomial

Recall that a Negative Binomial is a random variable that semantically represents the number of trials until r successes. Let $Y \sim \text{NegBin}(r, p)$.

Let X_i = # trials to get success after the $(i-1)$-th success. We can then think of each X_i as a Geometric RV: $X_i \sim \text{Geo}(p)$. Thus, $E[X_i] = \frac{1}{p}$. We can express Y as:

$$Y = X_1 + X_2 + \cdots + X_r = \sum_{i=1}^{r} X_i$$

Let's formally calculate the expectation of Y:

$$E[Y] = E[\sum_{i=1}^{r} X_i] = \sum_{i=1}^{r} E[X_i] = E[X_1] + E[X_2] + \ldots E[X_r] = \frac{r}{p}$$

Coupon Collector's Problem

There are several versions of the coupon collector's problem in probability theory, but the most common formulation is as follows: You would like to collect coupons from cereal boxes, but you must purchase a box of cereal to open and discover what coupon type you have. More formally, suppose you buy n boxes of cereal, and there are k different types of coupons. For each box you buy, you "collect" a coupon of type i. What is the expected number of boxes that you must purchase until you have at least one coupon of each type?

How does this relate to computer science? Suppose you are a big cloud provider, and you have to service n web requests with a limited number of k servers. Each web request is a request to server i. What is the expected number of utilized servers after n requests?

Example: Hash Tables

Problem: Yes, hash table problems can be a variation of the coupon collector's problem! Consider a hash table with k buckets. You hash each string to bucket i. What is the expected number of strings to hash until each bucket has at least 1 string?

Solution: Define Y as the number of strings to hash until each bucket has at least 1 string. We want to compute $E[Y]$. Let us also define Y_i to be the number of trials (strings) until the next success, *after* we've seen our i-th success.

For example, Y_0 is the number of strings hashed until our first hash into an empty bucket (we start with k empty buckets), Y_1 is the number of additional strings to hash until we hash into an empty bucket (we have 1 non-empty bucket and $k-1$ empty buckets, etc.). In the general case, we have i non-empty buckets and $k-i$ empty buckets after the i-th success, and we are successful if we hash a string to one of the $k-i$ empty buckets bucket. The probability of success p is then $p_i = \frac{k-i}{k}$. With this definition of Y_i, $Y_i \sim Geo(p)$, and $E[Y_i] = \frac{1}{p_i} = \frac{k}{k-i}$.

Note that $Y = Y_0 + Y_1 + Y_2 + \cdots + Y_{n-1}$. We can show the following:

$$
\begin{aligned}
E[Y] &= E[\sum_{i=0}^{n} Y_i] = \sum_{i=0}^{n} E[Y_i] && \text{Expectation of sum is sum of expectations} \\
&= \sum_{i=0}^{n} \frac{k}{k-i} = \frac{k}{k} + \frac{k}{k-1} + \frac{k}{k-2} + \cdots + \frac{k}{1} \\
&= k\left[\frac{1}{k} + \frac{1}{k-1} + \cdots + 1\right] = O(k \log k)
\end{aligned}
$$

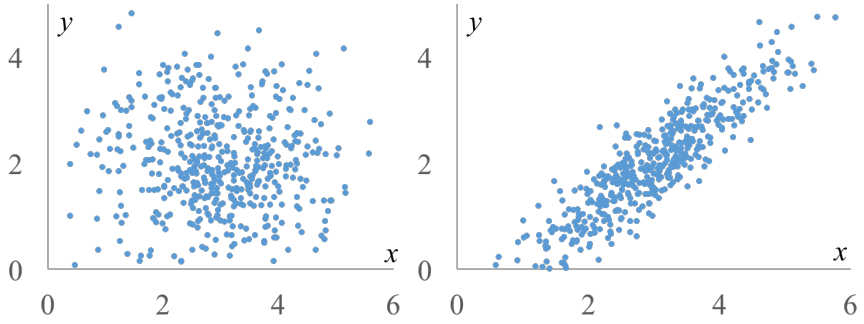
Expectations of Products Lemma

We know that the expectation of the sum of two random variables is equal to the sum of the expectations of the two variables. However, the expectation of the product of two random variables only has a nice decomposition in the case where the random variables are *independent* of one another.

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \qquad \textit{if } X \text{ and } Y \text{ are independent}$$

Here's a proof for independent discrete random variables X and Y. If you would like to prove this for independent continuous random variables, just interchange the summations with integrals.

$$
\begin{aligned}
E[g(X)h(Y)] &= \sum_y \sum_x g(x)h(y)p_{X,Y}(x,y) \\
&= \sum_y \sum_x g(x)h(y)p_X(x)p_Y(y) \\
&= \sum_y \left(h(y)p_Y(y) \sum_x g(x)p_X(x) \right) = \left(\sum_x g(x)p_X(x) \right) \left(\sum_y h(y)p_Y(y) \right) \\
&= E[g(X)]E[h(Y)]
\end{aligned}
$$



2 Covariance and Correlation

Consider the two multivariate distributions shown bellow. In both images I have plotted one thousand samples drawn from the underlying joint distribution. Clearly the two distributions are different. However, the mean and variance are the same in both the x and the y dimension. What is different?

Covariance is a quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean. It is a mathematical relationship that is defined as:

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

That is a little hard to wrap your mind around (but worth pushing on a bit). The outer expectation will be a weighted sum of the inner function evaluated at a particular (x, y) weighted by the probability of (x, y). If x and y are both above their respective means, or if x and y are both below their respective means, that term will be positive. If one is above its mean and the other is below, the term is negative. If the weighted sum of terms is positive, the two random variables will have a positive correlation. We can rewrite the above equation to get an equivalent equation:

$$\text{Cov}(X, Y) = E[XY] - E[Y]E[X]$$

Using this equation (and the product lemma) is it easy to see that if two random variables are independent their covariance is 0. The reverse is *not* true in general.

Properties of Covariance

Say that X and Y are arbitrary random variables:

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$
$$\text{Cov}(X, X) = E[X^2] - E[X]E[X] = \text{Var}(X)$$
$$\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$$

Let $X = X_1 + X_2 + \cdots + X_n$ and let $Y = Y_1 + Y_2 + \cdots + Y_m$. The covariance of X and Y is:

$$\begin{aligned}
\text{Cov}(X, Y) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j) \\
\text{Cov}(X, X) = \text{Var}(X) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{Cov}(X_i, X_j)
\end{aligned}$$

That last property gives us a third way to calculate variance. You could use this definition to calculate the variance of the binomial.

Correlation

Covariance is interesting because it is a quantitative measurement of the relationship between two variables. Correlation between two random variables, $\rho(X, Y)$ is the covariance of the two variables normalized by the variance of each variable. This normalization cancels the units out and normalizes the measure so that it is always in the range [0, 1]:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)Var(Y)}}$$

Correlation measures linearity between X and Y.

$\rho(X, Y) = 1$	$Y = aX + b$ where $a = \sigma_y / \sigma_x$
$\rho(X, Y) = -1$	$Y = aX + b$ where $a = -\sigma_y / \sigma_x$
$\rho(X, Y) = 0$	absence of linear relationship

If $\rho(X, Y) = 0$ we say that X and Y are “uncorrelated." If two varaibles are independent, then their correlation will be 0. However, it doesn’t go the other way. A correlation of 0 does not imply independence.

When people use the term correlation, they are actually referring to a specific type of correlation called “Pearson" correlation. It measures the degree to which there is a linear relationship between the two variables. An alternative measure is “Spearman" correlation which has a formula almost identical to your regular correlation score, with the exception that the underlying random variables are first transformed into their rank. “Spearman" correlation is outside the scope of CS109.