Correlation

Covariance

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:

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

> **Lemma: Correlation of Independent Random Variables**:
> If two random variables X and Y are independent, than their covariance must be 0.
>
> $$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[Y]E[X] \qquad &&\text{Def of Cov} \\ &= E[X]E[Y] - E[Y]E[X] \qquad &&\text{Lemma Product of Expectation} \\ &= 0\end{aligned}$$
>
> Note that the reverse claim is not true. Covariance of 0 does not prove independence.

Using this equation (and the product lemma) it is 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:

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

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:

$$\text{Cov}(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

$$\text{Cov}(X, X) = \text{Var}(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j)$$

That last property gives us a third way to calculate variance. We can use it to, again, show how to get the variance of a Binomial.

Correlation

We left off last class talking about covariance. Covariance was interesting because it was a quantitative measurement of the relationship between two variables. Today we are going to extend that concept to correlation. 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:

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

Correlation measures linearity between X and Y.

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

If $\rho(X, Y) = 0$ we say that X and Y are "uncorrelated."

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.