Chris Piech
CS109

Handout #28
May 2nd, 2016

Covariance and Sampling

Product of Expectations Lemma

Here is a lovely little lemma to get us started:

$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ if and only if X and Y are independent

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:

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

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:

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

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.

Estimating Mean and Variance from Samples

Let's say you are the king of Bhutan and you want to know the average happiness of the people in your country. You can't ask every single person, but you could ask a random subsample. In this next section we will consider principled claims that you can make based on a subsample. Assume we randomly sample 200

Bhutanese and ask them about their happiness. Our data looks like this: $72, 85, \ldots, 71$. You can also think of it as a collection of 200 I.I.D. (independent, identically distributed) random variables $X_1, X_2, \ldots, X_n$.

From this data we can calculate a sample mean ($\bar{X}$) and a sample variance (S^2).

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$
$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

The first question to ask is, are those biased estimates? Yes. We will prove that that is the case for $\bar{X}$. The proof for S^2 is in lecture slides.

$$\begin{aligned} E[\bar{X}] &= E[\sum_{i=1}^{n} \frac{X_i}{n}] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu \end{aligned}$$

Standard Error

Ok, you convinced me that they are not biased. But now I want to know how much my sample mean might vary relative to the true mean.

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}(\sum_{i=1}^{n} \frac{X_i}{n}) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n} \\ &\approx \frac{S^2}{n} && \text{Since S is an unbiased estimate} \\ \text{Std}(\bar{X}) &\approx \sqrt{\frac{S^2}{n}} && \text{Since Std is the square root of Var} \end{aligned}$$

That $\text{Std}(\bar{X})$ term has a special name. It is called the standard error and its how you report uncertainty of estimates of means in scientific papers (and how you get error bars). Great! Now we can compute all these wonderful statistics for the Bhutanese people. But wait! You never told me how to calculate the $\text{Std}(S^2)$. True, that is outside the scope of CS109. You can find it on wikipedia if you want.

Let's say we calculate the our sample of happiness has n = 200 people. The sample mean is $\bar{X} = 83$ (what is the unit here? happiness score?) and the sample variance is $S^2 = 450$. We can now calculate the standard error of our estimate of the mean to be 1.5. When we report our results we will say that the average happiness score in Bhutan is 83 ± 1.5 with variance 450.