Chris Piech
CS109

Handout #29
May 4th, 2016

Correlation

Calling the Election

Last night the Republican Indiana primary was called in favor of the Donald after only 9% of votes were reported. How does that work? A student asked, I looked into it, and it brought a tear to my already crying eyes. This example is not related to correlation, but it is right after the midterm so we are taking time to crystallize course concepts.

In good probability style let us start by declaring random variables. Let X be the total number of votes that candidate one gets and let Y be the total number of votes that candidate two gets. In order to call an election we want to calculate the probability $P(X > Y)$. If that probability is large enough we call the election. First let's rearrange terms and see that it is a problem we can solve:

$$P(X > Y) = P(X - Y > 0) = P(Y - X < 0)$$

Ok great. Now all we need is a distribution for X and Y and find the convolution of X and $-Y$. Since X and Y are similar we can just focus on finding the distribution for one of them. Let's look into X.

In a US election, a state is broken down into counties. It would be nice to think that all counties vote using the same distribution, but alas that is too incorrect an assumption. So it goes. The simple way forward is to declare a new random variable for each county X_i. Now we can write an expression for X:

$$X = \sum_i X_i$$

Decomposition is the best isn't it? Now we can think about just one county.

For each county I would like to work out what is the probability of a single person in that county voting for candidate X. Check this out boss. Let V_i be an indicator variable which is 1 if a voter in the county votes for candidate one. We only have 9% of precincts reporting. Of those precincts a few will be from county i. What we have are many samples of the distribution underlying V_i.

Assume each reported vote Z_j from county i is an IID sample of V_i. Let n be the number of voters reported. We can calculate the sample mean, $\bar{Z}_i$ and the variance of the sample mean $\text{Var}(\bar{Z}_i)$.

$$\bar{Z}_i = \sum_{j=1}^{n} \frac{Z_j}{n}$$

If the variance of the sample mean is small enough, we have enough voters in the precinct to include it. Since the sample mean is an unbiased estimate of the true mean *and* the true mean of an indicator variable is the probability of the underlying event:

$$P(V_i) = E[V_i] = \bar{Z}_i$$

The story keeps getting better. Now for a given county i we have a probability of a candidate getting a single vote from that county. Let m_i be the expected number of voters in the county. If we assume that they are voting independently the sum of votes, X_i is a binomial with a moderate p and a gigantic n. Yes, you guessed it, we can use the Normal approximation of the Binomial! We fit the mean and the variance of the binomial and get a distribution for X_i. Thus:

$$X_i \sim N(m_i\bar{Z}_i, m_i\bar{Z}_i(1-\bar{Z}_i))$$

You know what is so great about fitting a normal? The convolution of independent normals is very straightforward. Note that in real life they might not be more independent. In this case assuming independence is going to make us more conservative in our estimate. Thus we have all the pieces, we just need to bring it on home. So far we have been ignoring Y. Y is basically the same as X, but I am going to use different notation for its sample variance of county i, $\bar{W}_i$, to make it clear that that it is different:

$$X_i \sim N(m_i\bar{Z}_i, m_i\bar{Z}_i(1-\bar{Z}_i))$$
$$Y_i \sim N(m_i\bar{W}_i, m_i\bar{W}_i(1-\bar{W}_i))$$
$$Y - X \sim N\left(\sum_i m_i\bar{W}_i - \sum_i m_i\bar{Z}_i,\ \sum_i m_i\bar{W}_i(1-\bar{W}_i) + \sum_i m_i\bar{Z}_i(1-\bar{Z}_i)\right)$$
$$P(X > Y) = \phi\left(\frac{0 - \sum_i m_i\bar{W}_i - \sum_i m_i\bar{Z}_i}{\sqrt{\sum_i m_i\bar{W}_i(1-\bar{W}_i) + \sum_i m_i\bar{Z}_i(1-\bar{Z}_i)}}\right)$$

That last expression is something that we can calculate. Note that in all the sums we should only include counties with low enough variance of the sample mean for both candidates. Another aside on this calculation is that never assumes that all people vote, or that all voters vote for either candidate one or two. That lack of assumption makes it a more realistic model.

Wham-blam-alakazam! (Is that a real term? It felt right). I really like this example. It brings together so many concepts.

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 measure 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."

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.