36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 27: November 4

Lecturer: Siva Balakrishnan

27.1 Review and Outline

In the last class we discussed hypothesis testing:

1. FDR and FDR control

2. Inverting hypothesis tests

Today we will discuss independence testing. This is Chapter 15 of the Wasserman book.

There are many applications of independence testing: in drug trials a natural question is whether the outcome is independent of the treatment or not, in machine learning we often want to do "feature selection" i.e. find features that are "associated" with a variable we want to predict and so on. In all of these cases, we either want to test if two random variables are independent or not, or to measure the strength of their dependence or association.

27.2 Testing Independence and Measuring Association

Given two random variables X and Y, it is often of interest to understand if the two random variables are dependent, and to estimate the strength of their dependence.

Concretely, we would like to study the hypothesis testing question:

$$\begin{aligned} H_0 : &\quad X \perp\!\!\!\perp Y \\ H_1 : &\quad X \not\perp\!\!\!\perp Y, \end{aligned}$$

where we use the symbols $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ to denote independent and not independent respectively.

Broadly, one of the key ideas is to use the fact that if X and Y are independent then their joint distribution will be equal to the product of their marginals. We can use the "distance" between the joint and product of marginals as a measure of association.

27.3 Two binary variables

To begin with let us consider the simplest possible setting. Suppose that (X, Y) are two binary random variables and we observe n i.i.d pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.

We can represent this binary data in a two-by-two table, i.e. a table of the form:

	Y= 0	Y = 1	
X = 0	C_{00}	C_{01}	$C_{0\cdot}$
X = 1	C_{10}	C_{11}	$C_{1\cdot}$
	$C_{\cdot 0}$	$C_{\cdot 1}$	

Here C_{ij} denotes the number of times we observed the pair (i, j). This is a special case of something known as a contingency table.

With all of this in place we can define our first test for independence. This is the Pearson χ^2 test. The test statistic is:

$$T = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(C_{ij} - E_{ij})^2}{E_{ij}},$$

where we define the expected counts as

$$E_{ij} = \frac{C_{i\cdot} C_{\cdot j}}{n}.$$

In this case, under the null the statistic will have an asymptotic χ^2_1 distribution, so we can obtain a size α test by using the appropriate quantile of the χ^2 distribution, i.e. we reject the null hypothesis if:

$$T \geq \chi^2_{1,\alpha}.$$

27.3.1 Measuring Association of Two Binary Variables

We can think about the population version of the two-by-two table:

	Y= 0	Y = 1	
X = 0	p_{00}	p_{01}	$p_{0\cdot}$
X = 1	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	

Note that the counts C are simply a multinomial with the probabilities in the population table. We can define the odds ratio:

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}},$$

where the basic observation is that under independence $\psi = 1$. It is also common to use the log-odds ratio $\gamma := \log \psi$ which is 0 under independence.

In order to measure association between two binary RVs we could then estimate the log odds ratio. The MLE for it is:

$$\widehat{\gamma} = \log \left(\frac{C_{00}C_{11}}{C_{01}C_{10}} \right).$$

It is also easy to estimate the variance of $\widehat{\gamma}$ (see the Wasserman book).

27.4 Two discrete variables

The generalization from two binary variables to two discrete variables is straightforward. Suppose that the two variables take values $X \in \{1, \ldots, I\}$, and $Y \in \{1, \ldots, J\}$. In this case we again have an $I \times J$ contingency table.

	Y= 0	$\ldots$	Y = J	
X = 0	p_{00}	$\ldots$	p_{0J}	$p_{0\cdot}$
$\vdots$				
X = I	p_{I0}	$\ldots$	p_{IJ}	$p_{I\cdot}$
	$p_{\cdot 0}$	$\ldots$	$p_{\cdot J}$	

Again we would use the same χ^2 statistic which is:

$$T = \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{(C_{ij} - E_{ij})^2}{E_{ij}},$$

where

$$E_{ij} = \frac{C_{i\cdot}C_{\cdot j}}{n},$$

as before. The general rule is that T will be asymptotically χ^2 distributed with degrees of freedom $(I - 1)(J - 1)$, so we can use this distribution to determine the cutoff.

27.5 Two continuous variables

The case of two continuous variables is dealt with somewhat briefly in the Wasserman book. There are many important measures of association for two continuous RVs (some of the ones we talk about are more general). Many of them do not lead to valid tests of independence, but they are widely used in statistical practice.

27.5.1 Pearson's correlation

We have seen this one before. A measure of association between two random variables is their Pearson correlation:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

It is straightforward to estimate the Pearson correlation. If X, Y have a joint normal distribution then the Pearson correlation is zero if the RVs are independent so in this special case it measures the dependence between two RVs.

In some sense, the Pearson correlation coefficient measures linear association between two random variables. The way to interpret this statement is simply that the Pearson correlation is 1 for two random variables that are related linearly (i.e. if $Y = aX + b$, and $a > 0$).

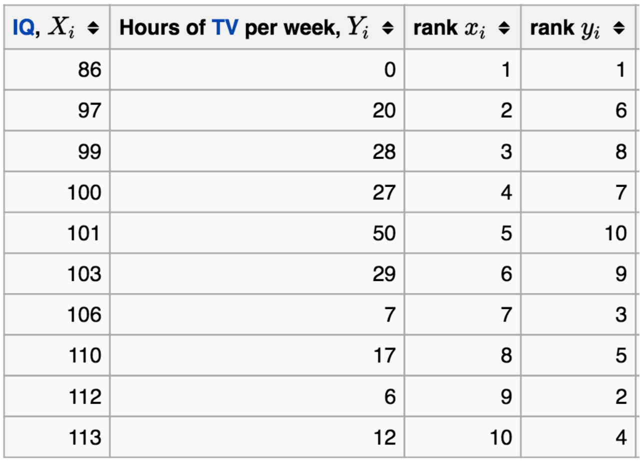
27.5.2 Spearman's rho

Spearman's rho is a version of Pearson's that measures monotonic association between two random variables, i.e. it is 1 if there is any monotonically increasing function that relates the two random variables.

The idea is quite simple: we transform the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to their *ranks*, i.e. assume you sorted the Xs in increasing order (assume there are no ties, if there are we usually average the ranks), and replaced each X_i with its rank r_{X_i} and did the same for the Ys.

Now you have data: $(r_{X_1}, r_{Y_1}), \ldots, (r_{X_n}, r_{Y_n})$. Spearman's rho is then defined as the Pearson correlation of this rank data.

Here is an example from Wikipedia:



IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i
86	0	1	1
97	20	2	6
99	28	3	8
100	27	4	7
101	50	5	10
103	29	6	9
106	7	7	3
110	17	8	5
112	6	9	2
113	12	10	4

27.5.3 Mutual information

It turns out that the above two measures although widely applied do not in general give valid tests for independence, i.e, they can be zero even when the RVs are dependent. They do however give one-sided tests, i.e. if the true (not estimated) Spearman rho or the Pearson correlation coefficient is non-zero then the RVs are dependent.

A different non-parametric measure of association that does lead to valid tests for independence is the *mutual information.* Recall, our original intuition that dependence means that the joint distribution is far from the product of the marginals. A natural idea is to use the KL divergence to measure this distance. This is called the mutual information. We assume the random variables have continuous densities, then we would compute:

$$\begin{aligned} I(X;Y) &= D_{\text{KL}}(p(X,Y)||p(X)p(Y)) \\ &= \int_X \int_Y p(X,Y) \log \left(\frac{p(X,Y)}{p(X)p(Y)} \right) dX dY. \end{aligned}$$

So in order to measure the strength of association we could estimate the MI via kernel density estimation (say):

$$\widehat{I}(X;Y) = \int_X \int_Y \widehat{p}_h(X,Y) \log \left(\frac{\widehat{p}_h(X,Y)}{\widehat{p}_h(X)\widehat{p}_h(Y)} \right) dX dY.$$

This is an example of a plugin estimator.

To perform a test for independence, we really need to understand the distribution of this estimator. This turns out to be quite challenging so people often estimate the cut-offs using an idea very similar to that of the permutation test.

27.6 One discrete and one continuous distribution

It turns out that there is a simplification that can be used in the case when we are testing for independence between a discrete $X \in \{1, \ldots, I\}$ and a continuous Y.

The idea is to notice that if we denote the conditional CDFs of Y as:

$$F_i(y) = P(Y \leq y | X = i),$$

then the two RVs are independent if and only if $F_1 = F_2 = \ldots = F_i$.

Testing if CDFs are equal is quite easy, for each pair (i, j) we use the Kolmogorov-Smirnov statistic:

$$T_{ij} = \sup_y |F_i(y) - F_j(y)|.$$

We can combine these statistics T_{ij} in different ways (take the largest, the sum, etc). Again determining the asymptotic distribution can be difficult (the Wasserman book gives this for binary X) but we can use simulation-based methods.