Math 3215: Lecture 23

Will Perkins

April 17, 2012

1 Review of Regression

For a linear regression we assume the model:

$$y_i = \alpha + \beta(x_i - \overline{x}) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

We then estimate the parameters α, β, σ^2 and form confidence intervals for the paramters. This gives us a confidence interval for our predictions.

How about a quadratic regression? Same idea.

$$y_i = \alpha + \beta(x_i - \overline{x}) + \gamma(x_i - \overline{x})^2 + \epsilon_i$$

2 Uses of the Chi Square Test

Recall the Chi Square statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} = n \sum_{i=1}^{k} \frac{(\overline{p}_i - p_i)^2}{p_i}$$

As $n \to \infty$ this statistic has $k-1$ degrees of freedom if no paramters are being estimated (we lose one degree of freedom since we know $\sum Y_i = n$). For every parameter we estimate, we lose a degree of freedom.

So if we have integer-valued data and want to test whether the data comes from a Poisson distribution, we first estimate the paramter λ, using the MLE: $\overline{x}$. We then calculate the p_i's according to a Poisson distribution with mean $\overline{x}$: $p_i = e^{-\overline{x}}\overline{x}^i/i!$.

But now there's a problem: for large i, p_i is very small, and we probably have very few observations with value i. So we have to group the data together, say by putting all $i \geq 10$ into one 'bin'. The rule of thumb for the chi squared distribution is to make sure that the expected number of observations for each bin is at least 5.

Once we bin the data (into k bins say), we can compute the χ^2 statistic and then a p-value, using the fact that under the null hypothesis the statistic is asympotically χ^2_{k-2}. One degree of freedom is lost from the data summing to n, the other is lost by estimating λ.

2.1 Tests of homogneity

Another use of the chi square statistic. Say two graders are grading final exams. You assign exams to each randomly and want to know whether their have the same grade distribution. Say there are k possible grades. How can we statistically test this?

The data:

We get $2k$ peices of data: $X_1, \ldots X_k$ are the number of exams graded $1, \ldots k$ by the first grader and $Y_1 \ldots Y_k$ are the number of exams graded $1, \ldots k$ by the second grader. We also know the number of exams given to each grader: $n_1 = \sum X_i$, $n_2 = \sum Y_i$.

What are we trying to test? We assume there are underlying true probabilities for each grader, $p_1, \ldots p_k$ for grader 1 and $q_1, \ldots q_k$ for grader 2. We want to test the null hypothesis that $p_i = q_i$ for all i.

To form the chi square statistic, we need to estimate the value of $p_i = q_i$. The MLE is $\frac{X_i+Y_i}{n_1+n_2}$. How many of the probabilities do we need to estimate? Only $k-1$, since the kth is determined by all the rest.

Now we form the chi squared statistic with the estimated probabilities, and compute the p-values using the χ^2_{k-1} distribution. There are $k-1$ degrees of freedom since we begin with $2k$ probabilites (or data points), but lose one degree of freedom for each grader since the totals add up to n_1 and n_2. That leaves us with $2k-2$, but then we have estimated $k-1$ probabilities, so in all we have $2k-2-(k-1) = k-1$ degrees of freedom.

- How can we test whether 6 different graders all have the same grade distribution?

2.2 Test of Independence

Another application of the chi square test is to test whether two or more attributes are independent. Say we conduct a poll before an election. Each respondent lists their job (1 of k possible jobs) and their preferred candidate (1 of l possible candidates). We want to know if a person's job and voting preference are independent variables.

- What is the null hypothesis?
- What is the alternate hypothesis?
- How many different observed categories are there?
- If the variables were independent, what relationship would exist between the underlying probabilities?
- Which probabilities should we estimate to form the chi square test statistic?
- Write down the formula for the test statistic
- To what distribution should this converge as $n \to \infty$ under the null hypothesis?