MATH 3080 Lab Lecture 13

Curtis Miller

02/10/20

Lecture 13

Goodness-of-Fit Tests

A statistical test used to decide whether a data set came from a paricular distribution or not is known as a goodness-of-fit test since it decides whether the suggested distribution is a "good fit" for the data. Many goodness-of-fit tests exist, and here we will study the chi-square (or χ^2) tests. Here we look at chi^2 tests of two flavors: one deciding whether a categorical variable follows a particular distribution or not, and one deciding whether two categorical variables are independent or not.

I walked to 7-Eleven and bought a share-size bag of regular M&Ms, then counted how many M&Ms there were of each color. Below is the data set.

Red	Brown	Green	Yellow	Orange	Blue	Total
21	5	10	16	15	36	103

The official distribution of M&M candy colors in 1997 is listed below:

Red	Brown	Green	Yellow	Orange	Blue
.2	.3	.1	.2	.1	.1

Our question: Based off of our sample, should we believe that this is the distribution of colors in the bag? To be more precise, there are $K = 6$ possible categories (colors), we observe n_k candies for color k and there are $N = \sum_{k=1}^{K}$ n_k$ candies total. We have probabilities p_k of observing each of these colors for a randomly sampled candy and these probabilities are specified under the null hypothesis: denote the null-hypothesis probabilities as p_{k0}. We wish to decide between the null hypothesis

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \dots, p_K = p_{K0}$$

and the alternative hypothesis which simply states that the null hypothesis is false (two or more of the null hypothesis probabilities are incorrect). To do this we compare the observed count of each category, n_k, against the expected count if the null hypothesis were true, given by Np_{k0}. We make these comparisons using the χ^2 statistic:

$$\sum_{k=1}^{K} \frac{(n_k - Np_{k0})^2}{Np_{k0}}.$$

If the null hypothesis is true, N is large, and $Np_{k0} > 10$, this statistic is well approximated by a χ^2 distribution with $K - 1$ degrees of freedom. If the null hypothesis is false, we would expect to see at least one observed count far away from its expected count, causing that term in the statistic to be large and thus the overall statistic to be large. Thus large values of the χ^2 statistic are evidence against the null hypothesis and thus suggest rejecting it. This procedure together is the χ^2 test for goodness of fit, and the test can be performed in R using the function `chisq.test()`. We give this function first the observed counts, then the hypothesized distribution under then null hypothesis. It will then return the results of the χ^2 test.

```r
counts <- c(21, 5, 10, 16, 15, 36)
dist <- c(.2, .3, .1, .2, .1, .1)

chisq.test(counts, dist)
```

```
## Warning in chisq.test(counts, dist): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  counts and dist
## X-squared = 12, df = 10, p-value = 0.2851
```

In this case the null hypothesis was not rejected, though the function warned that the assumptions (generally related to sample size) may not be satisfied and thus the asymptotic approach to computing the statistic (that is, using the χ^2 distribution) may not work. If this is in fact a concern then we can tell the function to use simulation methods to get a distribution that may be better for computing p-values. We can set the parameter B to tell the function how many simulations to do, the larger the better (but the default should be good).

```r
chisq.test(counts, dist, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  counts and dist
## X-squared = 12, df = NA, p-value = 1
```

In any case it seems we can't reject the null hypothesis based on this data set (though another statistician reached a different conclusion).

If instead we wanted to test whether each color was equally likely or not, we can leave the second parameter unspecified, as the default null hypothesis is equal probability for all categories.

```r
chisq.test(counts)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  counts
## X-squared = 33.485, df = 5, p-value = 3.014e-06
```

In this case the null hypothesis was rejected; the true distribution of the data certainly does not appear to assign equal probability to all colors.

Let's now suppose we have two categorical variables and we want to determine whether those random variables are independent or not. In R we could refer to the `Titanic` data set, which tracks how many individuals died on the ship *Titanic* along with perhaps relevant information such as their age, sex, and class on the ship. We will have a two-way table tracking the number of people in each class and how many did or did not survive. We will get this table like so:

```
(class_survive_titanic <- apply(Titanic, c(1, 4), sum))
```

```
##       Survived
## Class   No Yes
##   1st  122 203
##   2nd  167 118
##   3rd  528 178
##   Crew 673 212
```

We wish to decide between the null hypothesis stating that the two variables are independent and the alternative hypothesis stating that the null hypothesis is false (they are not independent). In this case the observed counts will be the counts in each cell. The expected counts are estimated counts if the null hypothesis of independence were in fact true. Suppose that for variable A we have J possible categories and the probability of the observation belonging to category j is p_j, and for variable B there are K possible categories and the probability of observing category k is q_k. If the independence hypothesis is true then the probability that both category j and category k are observed is $p_j q_k$. We do not know these probabilities, though. Let n_{jk} be the number of observations falling in both categories j and k. Let $N_{j\cdot} = \sum_{k=1}^{K} n_{jk}$ and $N_{\cdot k} = \sum_{j=1}^{J} n_{jk}$. Let $N = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}$. Then we would estimate p_j with $\hat{p}_j = \frac{N_{j\cdot}}{N}$ and $\hat{q}_k = \frac{N_{\cdot k}}{N}$. Our estimated count for the number of observations that belong both to category j for variable A and category k for variable B if the null hypothesis is true is $N\hat{p}_j\hat{q}_k = \frac{N_{j\cdot}N_{\cdot k}}{N} = o_{jk}$. Our test statistic will then be

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{jk} - o_{jk})^2}{o_{jk}}.$$

If the null hypothesis is true the approximate distribution of this statistic will be a χ^2 distribution with $(J-1)(K-1)$ degrees of freedom. As before, if the alternative is true, then we expect to see a cell count far away from what it should be if the null hypothesis were true, and the statistic will be large.

The function `chisq.test()` can also perform the test for independence. We can decide if class matters to surviving the *Titanic* disaster like so:

```
chisq.test(class_survive_titanic)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  class_survive_titanic
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

In this case the null hypothesis is soundly rejected; class does seem to matter.