Chi-square Analysis

Charles Nicholson
cnicholson@ou.edu

January 31, 2014

Concepts

1. The chi-square (χ^2) test is useful for analyzing *discrete* variables.
2. χ^2-tests evaluate deviations of observed frequencies from expected frequencies.
   Two primary uses are to:
   (a) evaluate goodness-of-fit to a distribution
   (b) test for association between variables
3. The observed test statistic is computed as

$$\chi^2_{\text{obs}} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

   Where O_i and E_i are the observed and expected frequencies for category $i \in C = \{1, 2, \ldots, k\}$.
4. The observed statistic is a measure of the deviation between the observed and expected values.
5. If $E_i \geq 5 \; \forall i \in C$, then $\chi^2 \sim \chi^2_{df}$. Where df is the number of degrees of freedom.
   If $\exists i \in C$ s.t. $E_i < 5$, the chi-square test is inappropriate.
6. The deviation between observed and expected is "large" (i.e. statistically significant) if $\chi^2 > \chi^2_{df,\alpha}$ where $\chi^2_{df,\alpha}$ is the *critical value* for a given level of significance (i.e. 1-α) and degrees of freedom.
   - Common values for α include 0.05 and 0.1.
   - Degrees of freedom are related to the number of categories.
7. The probability p that χ^2 obtains a value as large as computed can be calculated from the inverse of the χ^2_{df} distribution.
8. If $\chi^2 > \chi^2_{df,\alpha}$, or if p is small, then the deviations are significant:
   (a) χ^2 goodness-of-fit tests whether a sample data follows a particular distribution. If $\chi^2 > \chi^2_{k-1,\alpha}$ then the evidence does *not* support the claim that the observed frequencies in the k bins fit the expected frequencies.
   (b) χ^2 test for association can be used to test whether two discrete variables are independent. If $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$ then the null hypothesis is rejected: the frequencies in the r rows and c columns of data in the two variable contingency table do not appear to be independent.

Example problems

Example 1. There are four blood types (ABO system). It is believed that 34, 15, 23 and 28% of people have blood type A, B, AB and O, respectively.

The blood samples of 100 students were collected: A(12), B(56), AB(2), O(30). Test if the collected blood sample contradicts the previous belief.

Solution. The null hypothesis would be

$$H_0 : p_A = 0.34,\ p_B = 0.15,\ p_{AB} = 0.23,\ p_O = 0.28.$$

Since the expected percentages are given, the expected frequencies are calculated as $E_i = p_i N$ for $i = \{A, B, AB, O\}$ and where N is the total number of samples collected.

The observed statistic is computed as:

$$\chi^2_{\text{obs}} = \frac{(12-34)^2}{34} + \frac{(56-15)^2}{15} + \frac{(2-23)^2}{23} + \frac{(30-28)^2}{28} = 1445$$

In R, the probability associated with the χ^2 value can be obtained as follows:

```
chi <-(12-34)^2/34+(56-15)^2/15+(2-23)^2/23+(30-28)^2/28
chi
[1] 145.6187

1-pchisq(chi,3)  # there are 4 groups, therefore df = 4 - 1 = 3
[1] 0
```

Since the p-value is almost 0, we reject H_0. We cannot conclude the ABO observed distribution is the same as the expected distribution.

Example 2. A contingency table is a table showing the relationship between two discrete variables. For example, consider two variables: *Diseased* and *Exposed*, with two factor levels each: *Yes* and *No*.

	Exposed		Row
Diseased	Yes	No	Totals
Yes	35	42	77
No	65	17	82
Column Totals	100	59	159

Determine if the two variables are associated or independent.

Solution. If the variables are independent, then the *Exposed* proportions should not be impacted by *Diseased* factor levels, and visa-versa. That is, under the independence assumption, the expected cell proportions should be a function of the row and column totals, i.e.:

$$\text{Expected Frequency for cell } c_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{\text{Total}}$$

In our example, this calculation would look as follows:

	Exposed	
Diseased	Yes	No
Yes	$\frac{(77)(100)}{159} = 48.42$	$\frac{(77)(59)}{159} = 28.57$
No	$\frac{(82)(100)}{159} = 51.57$	$\frac{(82)(59)}{159} = 30.43$

and the corresponding χ^2 statistic is then given by,

$$\chi^2_{\text{obs}} = \frac{(35 - 48.42)^2}{48.42} + \frac{(42 - 28.57)^2}{28.57} + \frac{(65 - 51.57)^2}{51.57} + \frac{(17 - 30.43)^2}{30.43} = 19.457$$

In R, the probability associated with the χ^2 value can be obtained as follows:

```
chi <- (35 - 48.42)^2/48.42  +
       (42 - 28.57)^2/28.57 +
       (65 - 51.57)^2/51.57 +
       (17 - 30.43)^2/30.43
chi
[1] 19.45723

> 1-pchisq(chi,1)  # num rows = 2, num cols = 2; df = (r-1)(c-1)=(2-1)(2-1) = 1
[1] 1.028775e-05
```

Since the p-value is almost 0, we reject H_0. The evidence suggest that the variables *Diseased* and *Exposed* are related, i.e. independence is rejected.