

Stat 312: Lecture 23

Categorical Data

Moo K. Chung
mchung@stat.wisc.edu

April 22, 2003

Concepts

1. Multinomial data. There are k outcomes (categories): $C_1, C_2, \dots, C_k$. The probability of sample X_i belonging to category C_j is p_j , i.e.

$$P(X_i \in C_1) = p_1, \dots, P(X_i \in C_k) = p_k.$$

2. The null hypothesis of interest is

$$H_0 : p_1 = c_1, \dots, p_k = c_k$$

for given constants $c_1, \dots, c_k$ ($c_1 + \dots + c_k = 1$).

$$H_1 : p_j \neq c_j \text{ for some } j.$$

3. Let $x_1, \dots, x_n$ be the observations. Suppose that there are N_j items in category C_j . Note that $N_1 + \dots + N_k = n$. We assume $N_1, \dots, N_k$ to be distributed as *multinomial* with parameter $p_1, \dots, p_k$. The *expected* number of elements in category C_j is $EN_j = np_j$.
4. *Chi-squared goodness-of-fit test*. The test procedure involves measuring the relative discrepancy between the expectation and observation using

$$\chi^2 = \left(\frac{N_1 - EN_1}{EN_1} \right)^2 + \dots + \left(\frac{N_k - EN_k}{EN_k} \right)^2.$$

Assuming $np_j \geq 5$ for every j ,

$$\chi^2 \sim \chi_{k-1}^2.$$

In-class 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.

Self-study problems

Read p. 616-619. Example 14.1.