

HYPOTHESIS TESTING FOR THE MULTINOMIAL DISTRIBUTION CASE

Dr. Özcan BAYTEKİN

Marmara University,

Economic and Administrative Sciences

ÖZET:

Parametreleri n, P_1, P_2, \dots, P_k olan bir (Y_1, Y_2, \dots, Y_k) çok kategorili dağılım ile, yine parametreleri $m, p_1^*, p_2^*, \dots, p_k^*$ olan (X_1, \dots, X_k) çok kategorili dağılım göz önüne alındığında, bu iki çok kategorili dağılımın özdeş olup olmadıklarının test edilmesi makalenin amacını oluşturur. (Makalenin yazarının bildiği kadarıyla böyle bir test mevcut değildir) Bu makalenin amacı, bu iki çok kategorili dağılımın özdeş olup olmadıklarını test eden bir sıfır hipotezi yapılandırmaktır. Şöyleki,

$$H_0: p_1 = p_1^*, p_2 = p_2^* \dots p_k = p_k^*$$

ABSTRACT:

Suppose that (Y_1, Y_2, \dots, Y_k) has a multinomial distribution with parameters n, P_1, P_2, \dots, P_k , and (X_1, \dots, X_k) has a multinomial distribution[9] with parameters $m, p_1^*, p_2^*, \dots, p_k^*$. Construct a test of the null hypothesis[5] that the two multinomial distributions[1] are identical, that is, test

$$H_0: p_1 = p_1^*, p_2 = p_2^* \dots p_k = p_k^*$$

1. CONTINGENCY TABLES AND AN EXAMPLE

A problem frequently encountered in the analysis of count data concerns the

independence [12] of two methods of classification of observed events. For example, we might wish to classify defects found on furniture produced in a manufacturing [5]

Table 1.1: A Contingency Table

Type Of Defect					
Shift	A	B	C	D	Total
1	15(22,51)	21(20,99)	45(38,94)	13(11,56)	94
2	26(22,99)	31(21,44)	34(39,77)	5(11,81)	96
3	33(28,50)	17(26,57)	49(49,29)	20(14,63)	119
Total	74	69	128	38	309

Let p_A equal the unconditional probability that a defect will be type A. Similarly, define p_B , p_C , and p_D as the probabilities of observing the three other types of defects. Then these probabilities, which we call the column probabilities [19] of Table 1.1, will satisfy the requirement

$$p_A + p_B + p_C + p_D = 1$$

In like manner, let p_i ($i = 1, 2$ or 3) equal to the row probability [18] that a defect will have occurred on shift i , where

$$p_1 + p_2 + p_3 = 1$$

If the two classifications are independent [7] of each other, a cell probability [17] will equal the product of its respective row and column probabilities in accordance with the multiplicative law of probability. For example, the probability that a plant according to (1) the type of defect and (2) the production shift. We wish to investigate a contingency, a dependence between the two classifications. Do the proportions of various types of defects vary from shift to shift?

A total of $n = 309$ furniture defects were recorded and the defects were classified according to one of four types: A, B, C or D. At the same time, each piece of furniture was identified according to the production shift in which it was manufactured. These counts are presented in Table 1.1, which is known as a contingency table. Numbers in parenthesis

are the estimated expected cell frequencies. Particular defect will occur on shift I and be of type A is $(p_1)(p_A)$. We observe that the numerical values of the cell probabilities are unspecified in the problem under consideration. The null hypothesis specifies only that each cell probability will equal the product of its respective row and column probabilities and therefore imply independence of the two classifications.[3]

The analysis of the data obtained from a contingency table [13] consists in estimating the row and column probabilities in order to estimate the expected cell frequencies.

As we have noted, the estimated expected cell frequencies may be substituted for the $E(n_{ij})$ in χ^2 , and χ^2 will continue to possess a distribution in repeated sampling that is approximated by the chi-square probability distribution.

The maximum likelihood estimator[17] for any row or column probability is found as follows. Let n_{ij} denote the observed frequency in row i and column j of the contingency table, and let p_{ij} denote the probability of an observation falling into this cell. If observations are independently selected, then the cell frequencies have a multinomial distribution and the maximum

likelihood estimator of p_{ij} is simply observed relative frequency for that cell. That is

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, c$$

Likewise, viewing row i as a single cell, the probability for row i is given by p_i , and hence

$$\hat{p}_i = \frac{r_i}{n}$$

(where r_i denotes the number of observations in row i) is the maximum likelihood estimator of p_i .

By analogous arguments, the maximum likelihood estimator of the j th-column probability is c_j/n .

(where c_j denotes the number of observations in column j).

Now let us compute the maximum likelihood estimator of the row and column probabilities

$$\begin{aligned} \hat{p}_A &= \frac{c_1}{n} = \frac{74}{309} & \hat{p}_B &= \frac{c_2}{n} = \frac{69}{309} \\ \hat{p}_C &= \frac{c_3}{n} = \frac{128}{309} & \hat{p}_D &= \frac{c_4}{n} = \frac{38}{309} \end{aligned}$$

The row probabilities p_1, p_2, p_3 can be estimated using the row totals, r_1, r_2 and r_3 .

$$\begin{aligned} \hat{p}_1 &= \frac{r_1}{n} = \frac{94}{309} & \hat{p}_2 &= \frac{r_2}{n} = \frac{96}{309} \\ \hat{p}_3 &= \frac{r_3}{n} = \frac{119}{309} \end{aligned}$$

Under the null hypothesis, the estimated expected value of n_{11} is

$$\hat{E}(n_{11}) = n(\hat{p}_1 \cdot \hat{p}_A) = n \frac{r_1}{n} \frac{c_1}{n} = \frac{r_1 c_1}{n}$$

In other words, we observe that the estimated expected value of the observed cell frequency, n_{ij} , for a contingency table is equal to the product of its respective row and column totals divided by the total frequency; that is,

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$$

The estimated expected cell frequencies for our example are shown in parenthesis in Table 1.1.

We may use the expected and observed cell frequencies shown in Table 1.1 to calculate the value of the test statistic: [15]

$$\begin{aligned} \chi^2 &= \sum_{j=1}^4 \sum_{i=1}^3 \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})} \\ &= \frac{(15 - 22,51)^2}{22,51} + \frac{(26 - 22,99)^2}{22,99} + \dots \\ &\dots + \frac{(20 - 14,63)^2}{14,63} = 19,17 \end{aligned}$$

The only remaining obstacle involves the determination of the appropriate number of degrees of freedom [13] associated with the test statistic. We will give this as a rule which we will attempt to justify. The degrees of freedom associated with a contingency table possessing r rows and c columns will always equal $(r-1)(c-1)$. For example, we will compare χ^2 with the critical value of χ^2 with $(r-1)(c-1) = (3-1)(4-1) = 6$ degrees of freedom.

You will recall that the number of degrees of freedom associated with the χ^2 statistic will equal the number of cells (in this case, $k = rc$) less one degree of freedom for each independent linear restriction [10] placed upon the observed cell frequencies. [2] The total number of cells for the data of the Table 1.1 is $k = 12$. From this we subtract one degree of freedom because the sum of the observed cell frequencies must equal n ; that is

$$n_{11} + n_{12} + \dots + n_{34} = 309$$

In addition, we used the cell frequencies to estimate three of the four column probabilities. Note that the estimate of the fourth column probability will be

determined once we have estimated $p_A, p_B, p_C,$ because

$$p_A + p_B + p_C + p_D = 1$$

Thus we lose $c-1 = 3$ degrees of freedom for estimating the column probabilities.

Finally, we used the cell frequencies to estimate $(r-1) = 2$ row probabilities, and therefore we lose $r-1 = 2$ additional degrees of freedom. The total number of degrees of freedom remaining will be

$$d.f = 12 - 1 - 3 - 2 = 6$$

And, in general, we see that the total number of degrees of freedom associated with an $r \times c$ contingency table will be

$$d.f = rc - 1 - (c-1) - (r-1) \\ = (r-1)(c-1)$$

Therefore, if we use $\alpha = 0,05$, we will reject the null hypothesis[14] that the two classifications are independent if $\chi^2 > 12,592$. Since the value of the test statistic[6], $\chi^2 = 19,17$, exceeds the critical value of χ^2 , we will reject the null hypothesis. The data presents sufficient evidence to indicate that the proportion of the various types of defects varies from shift to shift. A study of the production operations for the three shifts would probably reveal the cause.

2. CONTINGENCY TABLE AND THE MAXIMUM LIKELIHOOD ESTIMATOR

In this section we will refer to the $r \times c$ contingency table of section I, to show that the maximum likelihood estimator of the probability for row $i, p_i,$ is

$$\hat{p}_i = \frac{r_i}{n}, i = 1, 2, \dots, r.$$

In order to find the maximum likelihood estimator of p_i , the probability of falling in row i , consider row i as a single cell with r_i observations falling in this cell. Then the variables r_1, r_2, \dots, r_r follow a multinomial

distribution[8] with parameters $n, p_1, p_2, \dots, p_r.$ [1] Hence the likelihood function is

$$L = \frac{n!}{r_1! r_2! \dots r_r!} p_1^{r_1} p_2^{r_2} \dots p_r^{r_r} = K \prod_{j=1}^r p_j^{r_j} \text{ so that}$$

$$\text{with } LnL = LnK + \sum_{j=1}^r r_j Ln p_j \text{ with } \sum_{j=1}^r p_j = 1$$

Notice that, because of the above restriction, we may write

$$p_r = 1 - \sum_{j=1}^{r-1} p_j$$

and that p_r is really a function of p_i for $i = 1, 2, \dots, r-1.$ Hence,

$$LnL = LnK + \sum_{j=1}^{r-1} r_j Ln p_j + \left(n - \sum_{j=1}^{r-1} r_j \right) Ln \left(1 - \sum_{j=1}^{r-1} p_j \right)$$

Now,

$$\frac{d(LnL)}{dp_i} = \frac{r_i}{p_i} - \frac{\left(n - \sum_{j=1}^{r-1} r_j \right)}{\left(1 - \sum_{j=1}^{r-1} p_j \right)} \text{ for } i = 1, 2, \dots, r-1$$

setting these $r-1$ equations equal to zero we have, for $i = 1, 2, \dots, r-1$

$$r_i \left(1 - \sum_{j=1}^{r-1} p_j \right) = \hat{p}_i \left(n - \sum_{j=1}^{r-1} r_j \right) \quad (1)$$

In order to solve the $r-1$ equations simultaneously, add them together to obtain

$$\sum_{j=1}^{r-1} r_j \left(1 - \sum_{j=1}^{r-1} \hat{p}_j \right) = \sum_{j=1}^{r-1} \hat{p}_j \left(n - \sum_{j=1}^{r-1} r_j \right)$$

$$n \sum_{j=1}^{r-1} \hat{p}_j = \sum_{j=1}^{r-1} r_j$$

$$\sum_{j=1}^{r-1} \hat{p}_j = \frac{1}{n} \left(\sum_{j=1}^{r-1} r_j \right)$$

Substituting in (1) we have

$$r_i \left(1 - \frac{\sum_{j=1}^{r-1} r_j}{n} \right) = \hat{p}_i \left(n - \sum_{j=1}^{r-1} r_j \right)$$

$$r_i \left(n - \sum_{j=1}^{r-1} r_j \right) = n \hat{p}_i \left(n - \sum_{j=1}^{r-1} r_j \right)$$

$$\hat{p}_i = \frac{r_i}{n}$$

Also, by using the method of Lagrange undetermined multipliers[20], the maximization can be made simpler by letting

$$\Psi = \text{Ln}K + \sum_{j=1}^{r-1} r_j \text{Ln}p_j + \left(\sum_{j=1}^r p_j - 1 \right)$$

and solving

$$\frac{d\Psi}{dp_i} = 0 \quad i = 1, 2, 3 \dots r$$

for the estimates \hat{p}_i .

3. HYPOTHESIS TESTING FOR THE MULTINOMIAL DISTRIBUTION CASE

Suppose that $(Y_1 \dots Y_k)$ has a multinomial distribution with parameters $n, p_1, p_2 \dots p_k$, and $(X_1, X_2 \dots X_k)$ has a multinomial distribution with parameters $m, p_1^*, p_2^* \dots p_k^*$. [4] The purpose of this paper is to construct a test of the null hypothesis that the two multinomial distributions are identical; that is, test $H_0: p_1 = p_1^* \dots p_k = p_k^*$

Now, suppose that we have two multinomial experiments[16], each with k cells. The cell counts and cell probabilities are n_i and p_i, m_i and p_i^* , respectively and

$$\sum_{i=1}^k p_i = \sum_{i=1}^k p_i^* = 1$$

The hypothesis of interest is

$$H_0: p_1 = p_1^* \dots p_k = p_k^*$$

If H_0 is true, then $p_i = p_i^*$, so that the expected cell counts could be obtained if we could estimate p_i , the probability of falling in cell i for each of the two experiments. In general, the likelihood function can be written as:

$$L = f(n_1, n_2 \dots n_k, m_1, m_2 \dots m_k)$$

$$= \frac{n!}{\pi n_j!} \pi p_j^{n_j} \frac{m!}{\pi m_j!} \pi p_j^{*m_j}$$

$$= K \pi p_j^{n_j} \cdot p_j^{*m_j}$$

Under H_0 ,

$$L = K \pi p_j^{n_j+m_j} \quad \text{and}$$

$$\text{Ln}L = \text{Ln}K + \sum_{j=1}^k (n_j + m_j) \text{Ln}p_j$$

Maximizing $\text{Ln}L$ subject to the restriction

$$\sum_{j=1}^k p_j = 1 \quad \text{we obtain as in Section II}$$

$$\hat{p}_i = \frac{n_i + m_i}{n + m} \quad \text{for } i = 1, 2, 2 \dots k-1$$

$$\hat{p}_k = 1 - \sum_{i=1}^{k-1} \hat{p}_i \quad \text{then}$$

$$\hat{E}(n_i) = n \cdot \hat{p}_i = n \left(\frac{n_i + m_i}{n + m} \right) \quad i = 1, 2, 3 \dots k$$

$$\hat{E}(m_i) = m \cdot \hat{p}_i = m \left(\frac{n_i + m_i}{n + m} \right) \quad i = 1, 2, 3 \dots k$$

and the test statistic is

$$\chi^2 = \sum_{i=1}^k \frac{\left[n_i - n \left(\frac{n_i + m_i}{n + m} \right) \right]^2}{n \left(\frac{n_i + m_i}{n + m} \right)} + \sum_{i=1}^k \frac{\left[m_i - m \left(\frac{n_i + m_i}{n + m} \right) \right]^2}{m \left(\frac{n_i + m_i}{n + m} \right)}$$

χ^2 will have an approximate chi-square distribution[4] with degree of freedom $2k - 2 - (k-1) = k - 1$

Note that there are $2k$ cells. Two degrees of freedom are lost since

$$\sum_{i=1}^k n_i = n \quad \sum_{i=1}^k m_i = m$$

and $k-1$ cells probabilities have been estimated using the observed cell counts, n_i and m_i . Hence, a rejection region[11] for the test will be based upon $k-1$ degrees of freedom.

CONCLUSIONS

The material in this paper has been concerned with a test of a hypothesis regarding the cell probabilities associated with a multinomial experiment. When the number of observations, n , is large, the test statistic, χ^2 , can be shown to possess, approximately, a chi-square probability

distribution in repeated sampling, the number of degrees of freedom being dependent upon the particular application. In general we assume that n is large and that the minimum expected cell frequency is equal to or is greater than 5.

Also, in this paper, it is successfully constructed a test of the null hypothesis that the two multinomial distributions are identical.

distribution in repeated sampling, the number of degrees of freedom being dependent upon the particular application. In general we assume that n is large and that the minimum expected cell frequency is equal to or is greater than 5.

Also, in this paper, it is successfully constructed a test of the null hypothesis that the two multinomial distributions are identical.