1 The Chi-squared Test for Differences in Probabilities, $r \times c$

**Data** There are $r$ populations in all, and one random sample is drawn from each population. Let $n_i$ represent the number of observations in the $i$th sample (from the $i$th population) for $1 \leq i \leq r$. Each observation in each sample may be classified into one of $c$ different categories. Let $O_{ij}$ be the number of observations from the $i$th sample that fall into category $j$, so

$$n_i = O_{i1} + O_{i2} + \cdots + O_{ic} \quad \text{for all } i$$

(1)

The data are arranged in the following $r \times c$ contingency table.

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>\cdots</th>
<th>Class $c$</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 1</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
<td>\cdots</td>
<td>$O_{1c}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Population 2</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
<td>\cdots</td>
<td>$O_{2c}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
<td>\cdots</td>
</tr>
<tr>
<td>Population $r$</td>
<td>$O_{r1}$</td>
<td>$O_{r2}$</td>
<td>\cdots</td>
<td>$O_{rc}$</td>
<td>$n_r$</td>
</tr>
<tr>
<td>Totals</td>
<td>$C_1$</td>
<td>$C_2$</td>
<td>\cdots</td>
<td>$C_c$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

The total number of observations from all samples is denoted by $N$.

$$N = n_1 + n_2 + \cdots + n_r$$

(2)

The number of observations in the $j$th column is denoted by $C_j$. That is, $C_j$ is the total number of observations in the $j$th category, or class, from all samples combined.

$$C_j = O_{1j} + O_{2j} + \cdots + O_{rj}, \quad \text{for } j = 1, 2, \ldots, c$$

(3)

**Assumptions**

1. Each sample is a random sample.

2. The outcomes of the various samples are all mutually independent (particularly among samples, because independence within samples is part of the first assumption).

3. Each observation may be categorized into exactly one of the $c$ categories or classes.

**Test Statistic** The test statistic $T$ is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{where } E_{ij} = \frac{n_i C_j}{N}$$

(4)

While the term $O_{ij}$ represents the observed number in cell $(i,j)$, the term $E_{ij}$ represents the expected number of observations in cell $(i,j)$, if $H_0$ is really true. That is, if $H_0$ is true the number of observations in cell $(i,j)$ should be close to the $i$th sample size $n_i$ multiplied by the proportion $C_j/N$ of all observations in category $j$. Note that in the $2 \times 2$ case this
The $r \times c$ Contingency Table

Nonparametric Statistics  
Fall 2003

$T$ equals $T^2$ from the previous section, because only the two-sided alternative hypothesis is considered.

An equivalent expression for $T$, more suited for hand calculator use, is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$$  

(5)

**Null Distribution** The null distribution of $T$ is given approximately by the chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom, whose quantiles are given in Table A2. The exact distribution of $T$ is very difficult to find and therefore is almost never used.

This chi-squared approximation is satisfactory if the $E_{ij}$s in the test statistic are not too small. The chi-squared approximation appears to be satisfactory in most cases if all $E_{ij}$s are greater than 0.5 and at least half are greater than 1.0. However, if and $E_{ij}$s are less than 0.5, or if most $E_{ij}$s are less than 1.0, the chi-squared approximation may not be accurate, and thought should be given to combining several similar rows or columns to eliminate the very small row or column totals, or omitting a row or column with very few observations in it.

**Hypotheses** Let the probability of a randomly selected value from the $i$th population being classified in the $j$th class be denoted by $p_{ij}$.

- $H_0$: All of the probabilities in the same column are equal to each other
- $H_1$: At least two of the probabilities in the same column are not equal to each other

Note that it is not necessary to stipulate the various probabilities. The null hypothesis merely states the probability of being in class $j$ is the same for all populations, no matter what that probability might be.

Because of the difficulties involved in tabulating the exact distribution of $T$, the approximation based on the large sample distribution (where $E_{ij}$s are large) is used to find the critical region. The critical region of approximate size $\alpha$ corresponds to values of $T$ larger than $\text{qchisq}(1 - \alpha, (r - 1)(c - 1))$.

The $p$-value is the probability of a chi-squared random variable with $(r - 1)(c - 1)$ degrees of freedom exceeding the observed value of $T$.

2 The Chi-squared Test for Independence

**Data** A random sample of size $N$ is obtained. The observations in the random sample may be classified according to two criteria. Using the first criterion each observation is associated with one of the $r$ rows, and using the second criterion each observation is associated with one of the $c$ columns. Let $O_{ij}$ be the number of observations associated with row $i$ and
column $j$ simultaneously. The cell counts $O_{ij}$ may be arranged in an $r \times c$ contingency table. The total number of observations in row $i$ is designated by $R_i$, and in column $j$ by $C_j$. The sum of the numbers in all of the cells is $N$.

**Assumptions**

1. The sample of $N$ observations is a random sample. (Each observation has the same probability as every other observation of being classified in row $i$ and column $j$, independently of the other observations.)

2. Each observation may be classified into exactly one of $r$ different categories according to one criterion and into exactly one of $c$ different categories according to a second criterion.

**Test Statistic** Let $E_{ij}$ equal $R_iC_j/N$. Then the test statistic is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

or, for more convenient use with hand calculators,

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$$

where the summation is taken over all cells in the contingency table.

**Null Distribution** The null distribution of the test statistic is given approximately by the chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

**Hypotheses** $H_0$: The event “an observation is in row $i$” is independent of the event “that same observation is in column $j$”, for all $i$ and $j$.

By the definition of independence of events, $H_0$ may be stated as follows.

$$H_0 : P(\text{row } i, \text{column } j) = P(\text{row } i)P(\text{column } j), \quad \text{for all } i, j$$

The negation of $H_0$ stated as

$$H_1 : P(\text{row } i, \text{column } j) \neq P(\text{row } i)P(\text{column } j), \quad \text{for some } i, j$$

Reject $H_0$ if $T$ exceeds $qchisq(1 - \alpha, df = (r - 1)(c - 1))$. The $p$-value is $1 - pchisq(T, df = (r - 1)(c - 1))$.

### 3 The Chi-squared Test with Fixed Marginal Totals

**Data** The data are summarized in an $r \times c$ contingency table, as in the two previous applications, except that the row and column totals are determined beforehand, and are therefore
fixed, not random. The row and column totals are denoted by $n_i$ and $c_j$, respectively, to emphasize the fact that they are given and not random. The total number of observations is $N$.

**Assumptions**

1. Each observation is classified into exactly one cell.
2. The observations are observations on a random sample. Each observation has the same probability of being classified into cell $(i, j)$ as any other observation.
3. The row and column totals are given, not random.

**Test Statistic** Let $E_{ij} = n_i c_j / N$ be the expected number of observations in cell $(i, j)$. Then the test statistic, as before, is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where the summation is over all $rc$ cells.

**Null Distribution**

$$\chi^2_{df=(r-1)(c-1)}$$

**Hypotheses** The hypotheses may be either of the two sets of hypotheses introduced in the two previous applications in this section, under the condition that the row and column totals are fixed. Reject region and $p$-value are obtained the same way.

4 **Exercises: 4.2 #1, 3, 5, 7**

1. Test whether the following observations indicate a dependence between the two variables observed: (3.6,13), (4.7,19), (1.4,9), (5.5,15), (4.8,27), (4.3,14), (3.0,6), (4.2,11), (6.0, 24), (6.8,26), (4.1,18), (3.2,9), (4.0,8), (1.9,6), (0.4,7), (4.9,14), (5.6,18), and (5.6,20). Which test of this section is being used?

2. $T = 6.813678$, the $p$-value is 0.07807965. Fixed Marginal Totals.

3. B and C are combined together. $T = 2.496141$, the $p$-value is 0.1141256.

4. $T = 6.604837$, the $p$-value is 0.678185.

\[1^4 See R code\]