Handout 6

Test for independence in a two-way table

Suppose we wish to test ideology and opinion on death penalty are independent. The hypothesis can be stated as $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j, vs H_1 : at least one $\pi_{ij} \neq \pi_{i+}\pi_{+j}$.

Let μ_{ij} be $E(n_{ij})$ when H_0 is true. We have considered two sampling schemes: joint multinomial, and independent multinomials. Fortunately, it turns out that the estimate of μ_{ij} (under H_0) equals $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ for the both the sampling schemes (reasons given in the Appendix). The most well-known test statistics are

$$\begin{aligned} X^2 &= \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \quad \text{[Pearson's chi-square]}, \\ G^2 &= \sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) \quad \text{[the LR]}. \end{aligned}$$

Fact 1. Under H_0, both X^2 and G^2 are approximately distributed as $\chi^2_{(I-1)(J-1)}$. [Rule of thumb: n_{ij}'s should be 5 or larger.]

Example 1 (Same as in Example 1 in Handout 5)

A random sample of 250 adults is taken in a state, and each person's political ideology and opinion on death penalty are noted. The observed counts as well as the estimated expected counts (in the brackets) are given below.

	Supports Death Penalty		Total
Political Ideology	Yes	No	
Liberal	37 (55.596)	76 (57.404)	113 (113)
Conservative	86 (67.404)	51 (69.596)	137 (137)
Total	123 (123)	127 (123)	250

We want to test if ideology and opinion on death penalty are independent, test $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j, vs H_1 : at least one $\pi_{ij} \neq \pi_{i+}\pi_{+j}$.

Note that for the sampling scheme here, the counts are jointly multinomial.

The estimated expected counts (under H_0) are obtained as follows

$$\begin{aligned} \hat{\mu}_{11} &= n_{1+}n_{+1}/n = (123)(113)/250 = 55.596, \\ \hat{\mu}_{12} &= n_{1+}n_{+2}/n = (123)(127)/250 = 57.404, \\ \hat{\mu}_{21} &= n_{2+}n_{+1}/n = (137)(123)/250 = 67.404, \\ \hat{\mu}_{22} &= n_{2+}n_{+2}/n = (137)(127)/250 = 69.596. \end{aligned}$$

The values of X^2 and G^2 are

$$\begin{aligned} X^2 &= \frac{(37 - 55.596)}{55.596} + \frac{(76 - 57.404)^2}{57.404} + \frac{(86 - 67.404)^2}{67.404} + \frac{(51 - 69.596)}{69.596} \\ &= 6.2207 + 6.0242 + 5.1304 + 4.9688 = 22.3441. \\ G^2 &= 2\left[(37)\log(37/55.596) + (76)\log(76/57.404) + (86)\log(86/67.404) + (51)\log(51/69.596)\right] \\ &= 2[-15.0662 + 21.3271 + 20.9533 - 15.8550] = 22.7184. \end{aligned}$$

The df for the chi-square test is $(I-1)(J-1) = (2-1)(2-1) = 1$. If the level of significance is $\alpha = 0.01$, then $\chi^2(0.01) = 6.635$. Both X^2 and G^2 are larger than $\chi^2(0.01)$, and hence we reject H_0. [the p-values for both the tests are smaller than 0.00001.]

Tests for Independence for 2×2 Tables

For 2×2 tables, in addition to the chi-square tests, there are two more tests which can be used to test for independence. Let π_1 be the proportion of 'Yes' among liberals and π_2 be the proportion of 'Yes' among conservatives. The following fact (discussed in the Appendix) provide the justification for these two additional tests.

Fact 2. For a 2×2 table, the following are equivalent
(a) $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j, (b) $\pi_1 = \pi_2$, (c) $\theta = 1$.

Thus in order to test for independence, we may test $H_0 : \pi_1 = \pi_2$ vs $H_1 : \pi_1 \neq \pi_2$, or we can carry out a test $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$.

In order to carry out a test for the null hypothesis of equality of proportions π_1 and π_2, consider

$$\begin{aligned} \hat{\pi}_1 &= n_{11}/n_{1+}, \hat{\pi}_2 = n_{21}/n_{2+}, \\ z &= \frac{\hat{\pi}_1 - \hat{\pi}_2}{SE}, \text{ where} \\ SE &= \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_{1+} + \hat{\pi}_2(1-\hat{\pi}_2)/n_{2+}}. \end{aligned}$$

When $\pi_1 = \pi_2$, the z-statistic is approximately $N(0,1)$. Thus for testing $H_0 : \pi_1 = \pi_2$ vs $H_1 : \pi_1 \neq \pi_2$, we reject H_0 if $|z| > z_{\alpha/2}$, where area to the right of $z_{\alpha/2}$ under the standard normal curve is $\alpha/2$ (and α is the given level of significance). The p-value of this test is smaller than 0.00001.

For the data in Example 1,

$$\begin{aligned} \hat{\pi}_1 &= n_{11}/n_{1+} = 37/113 = 0.3274, \\ \hat{\pi}_2 &= n_{21}/n_{2+} = 86/137 = 0.6277, \\ SE &= \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_{1+} + \hat{\pi}_2(1-\hat{\pi}_2)/n_{2+}} \\ &= \sqrt{(0.3274)(1-0.3274)/113 + (0.6277)(1-.0.6277)/137} \\ &= 0.0605, \\ z &= \frac{\hat{\pi}_1 - \hat{\pi}_2}{SE} = \frac{0.3274 - 0.6277}{0.0605} = -4.964. \end{aligned}$$

The p-value of this test is less than 0.0001. Thus we reject the null hypothesis $\pi_1 = \pi_2$ which is equivalent to rejecting the null hypothesis of independence.

Now let us carry out the test $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$, we can calculate the test statistic

$$z = \log(\hat{\theta})/SE, \text{ where } SE = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}},$$

and reject H_0 if $|z| > z_{\alpha/2}$ where area to the right of $z_{\alpha/2}$ under the normal curve is $\alpha/2$.

For Example 1 in this Handout,

$$\begin{aligned}
\hat{\pi}_1 &= n_{11}/n_{1+} = 37/113 = 0.3274, \\
\hat{\pi}_2 &= n_{21}/n_{2+} = 86/137 = 0.6277, \\
\hat{\theta} &= \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{0.3274/(1-0.3274)}{0.6277/(1-0.6277)} = 0.2887, \\
SE &= \sqrt{1/37 + 1/76 + 1/86 + 1/51} = 0.2672, \\
z &= \frac{\log(\hat{\theta})}{SE} = \frac{\log(0.2887)}{0.2672} = \frac{-1.2424}{0.2672} == -4.6497.
\end{aligned}$$

If $\alpha = 0.01$, then $z_{\alpha/2} = 2.576$. We reject at level $\alpha = 0.01$ since $|z| > 2.576$. Thus we reject the null hypothesis of independence. The p-value is smaller than 0.00001.

Residuals.

As a diagnostic tool, we can obtain residual of table counts. These residuals help us understand if the deviations of the observed counts from the (estimated) expected counts are substantial or not. Two commonly used residuals are

$$\begin{aligned}
\text{Pearson residual:} \qquad & e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}, \\
\text{Standardized residual} \quad : \quad & r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - n_{i+}/n)(1 - n_{+j}/n)}}.
\end{aligned}$$

Under H_0, $e_{ij} \overset{approx}{\sim} N(0, (I-1)(J-1)/(IJ))$ and $r_{ij} \overset{approx}{\sim} N(0,1)$. Since the variance of e_{ij} is smaller than 1, it is preferable to use the standardized residuals $\{r_{ij}\}$. For the data in Example 1, we have the counts as well as the standardized residuals.

	Supports Death Penalty	
Political Ideology	Yes	No
Liberal	37 (-4.727)	76 (4.727)
Conservative	86 (4.727)	51 (-4.727)

If the null were true, then we would expect the calculated standardized residuals to be between -3 and 3. Here they are not indicating the possibility that the null hypothesis of independence may not be true.

Appendix

Discussion on Fact 2.

We will show that (a) $\Longleftrightarrow$ (b) $\Longrightarrow$ (c) $\Longrightarrow$ (a).
(a)$\Longrightarrow$(b).

Note that under independence $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. Hence

$$\begin{aligned}
\pi_1 &= \pi_{11}/\pi_{1+} = \pi_{1+}\pi_{+1}/\pi_{1+} = \pi_{+1}, \; 1 - \pi_1 = 1 - \pi_{+1} = \pi_{+2}, \\
\pi_2 &= \pi_{21}/\pi_{2+} = \pi_{2+}\pi_{+1}/\pi_{2+} = \pi_{+1}, \; 1 - \pi_2 = 1 - \pi_{+1} = \pi_{+2},
\end{aligned}$$

and hence $\pi_1 = \pi_2$.

(b) $\Longrightarrow$ (a).

If $\pi_2 = \pi_1$, we have

$$\begin{aligned} \pi_{21}/\pi_{2+} &= \pi_{11}/\pi_{1+}, \text{ ie, } \pi_{21}/\pi_{11} = \pi_{2+}/\pi_{1+}, \text{ ie,} \\ 1 + \pi_{21}/\pi_{11} &= 1 + \pi_{2+}/\pi_{1+}, \text{ ie, } (\pi_{11} + \pi_{21})/\pi_{11} = (\pi_{1+} + \pi_{2+})/\pi_{1+}, \text{ ie,} \\ \pi_{+1}/\pi_{11} &= 1/\pi_{1+}, \text{ ie, } \pi_{11} = \pi_{1+}\pi_{+1}. \end{aligned}$$

Since $\pi_{11} = \pi_{1+}\pi_{+1}$, we have

$$\pi_{12} = \pi_{1+} - \pi_{11} = \pi_{1+} - \pi_{1+}\pi_{+1} = \pi_{1+}(1 - \pi_{+1}) = \pi_{1+}\pi_{+2}.$$

Similar arguments will show that

$$\pi_{21} = \pi_{2+}\pi_{+1}, \ \pi_{22} = \pi_{2+}\pi_{+2}.$$

(b) $\Longrightarrow$ (c).

When $\pi_1 = \pi_2$, we clearly have

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1.$$

(c) $\Longrightarrow$ (b).

$$\begin{aligned} 1 &= \theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \Longrightarrow \pi_2/(1-\pi_2) = \pi_1/(1-\pi_1), \\ 1 + \pi_2/(1-\pi_2) &= 1 + \pi_1/(1-\pi_1), \text{ ie,} \\ 1/(1-\pi_2) &= 1/(1-\pi_1), \text{ ie, } 1 - \pi_2 = 1 - \pi_1, \text{ ie } \pi_1 = \pi_2. \end{aligned}$$

Discussion on Joint multinomial and independent multinomials (also called product multinomial).

Fact 3. Let $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ (the hypothesis of independence), and denote $\mu_{ij} = E(n_{ij})$ under H_0.
(a) Under H_0, for joint multinomial, $\mu_{ij} = n\pi_{i+}\pi_{+j}$, and $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n(n_{i+}/n)(n_{+j}/n) = n_{i+}n_{+j}/n$.
(b) Under H_0, for independent multinomials (independent rows), $\mu_{ij} = n_{i+}\pi_{+j}$, and $\hat{\mu}_{ij} = n_{i+}\hat{\pi}_{+j} = n_{i+}(n_{+j}/n) = n_{i+}n_{+j}/n$.

Remark. Recall that π_{i+} and π_{+j} are defined as: $\pi_{i+} = \pi_{i1} + \cdots + \pi_{iJ}$ (sum of the i^{th} row of the table of π_{ij}'s) and $\pi_{+j} = \pi_{1j} + \cdots + \pi_{Ij}$ (sum of the j^{th} column)

For the joint multinomial, π_{i+} and π_{+j} can always be estimated irrespective of whether H_0 is true or false, and the estimates are $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$.

For the product (or independent) multinomial case when the rows are independent, we can never estimate π_{i+}. We cannot estimate π_{+j} if H_0 is false (with the exceptional case $n_{1+} = \cdots = n_{I+}$). However, if H_0 is true, then the MLE of π_{+j} is n_{+j}/n.

Rule for obtaining the df of the chi-square distribution of $\mathbf{X}^2$ and G^2 under H_0.

The df of any chi-square test is

df=(# of parameters estimated in the full model)

- (# of parameters estimated for the model under H_0).

Let us consider the null hypothesis of independence: $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j.

Joint multinomial

For the joint multinomial

# of parameters estimated in the full model is $IJ - 1$,

# of parameters estimated for the model under H_0 is $(I - 1) + (J - 1)$.

Thus the df of the chi-square distribution under H_0 is (for the joint multinomial scheme)

$(IJ - 1) - [(I - 1) + (J - 1)] = (I - 1)(J - 1) = (I - 1(J - 1)$.

Independent multinomials.

For the independent multinomial scheme, each row is multinomial and the rows are independent. For instance, for row 1, $(n_{11}, \ldots, n_{1J}) \sim multinomial(n_{1+}; \pi_{11}/\pi_{1+}, \ldots, \pi_{1J}/\pi_{1+})$. So the parameters in row 1 are $\pi_{11}/\pi_{1+}, \ldots, \pi_{1J}/\pi_{1+}$ and they sum to 1. So the number of parameters to be estimated in row 1 is $J - 1$. Similarly, there are $J - 1$ parameters to be estimated in row 2 is $J - 1$, and so on.

Thus the # of parameters for the full model equals $I(J - 1)$.

Note that for row 1, $E(n_{1j}) = n_{1+}(\pi_{1j}/\pi_{1+}), j = 1, \ldots, J$. Similarly, for row 2, $E(n_{2j}) = n_{2+}(\pi_{2j}/\pi_{2+}), j = 1, \ldots, J$. In general, for row i, $E(n_{ij}) = n_{i+}(\pi_j/\pi_{i+}), j = 1, \ldots, J$.

Under the null hypothesis of independence, for any column $j = 1, \ldots, J$, all the I entries in that column $\pi_{1j}/\pi_{1+}, \ldots, \pi_{Ij}/\pi_{I\cdot}$ are they same, and the common value is π_{+j}. If H_0 were true, then it turns out that the MLE of π_{+j} is $\hat{\pi}_{+j} = n_{+j}/n$. When H_0 is true, $E(n_{ij}) = n_{i+}\pi_{+j}$.

Thus under H_0, for row 1, $E(n_{1j})$ is estimated by $n_{1+}\hat{\pi}_{+j} = n_{1+}(n_{+j}/n) = n_{1+}n_{+j}/n, j = 1, \ldots, J$. Similarly for row 2, under H_0, $E(n_{1j})$ is estimated by $n_{2+}\hat{\pi}_{+j} = n_{2+}(n_{+j}/n) = n_{2+}n_{+j}/n, j = 1, \ldots, J$.

In general, for row i, $E(n_{ij})$ is estimated by $n_{i+}\hat{\pi}_{+j} = n_{i+}(n_{+j}/n) = n_{i+}n_{+j}/n, j = 1, \ldots, J$.

Note that the number of parameters estimated for the model under H_0 equals $J - 1$.

Thus the df of the chi-square test under H_0 is

$$I(J - 1) - (J - 1) = IJ - I - J + 1 = (I - 1)(J - 1).$$