BIOS 6244 Analysis of Categorical Data
October 10, 2005 Lecture

Chi-Squared Tests of Independence (Sec. 2.4)

The chi-squared (χ^2) test can be used to test null hypotheses about the cell probabilities $\{\pi_{ij}\}$ from *any* size contingency table, even multi-dimensional ones. In this section, however, we focus on 2-way contingency tables with I rows and J columns.

For a sample of size n with cell counts $\{n_{ij}\}$, the values $\{\mu_{ij} = n\pi_{ij}\}$ are called the *expected cell frequencies*. They represent the expected values $\{E(n_{ij})\}$ when H_0 is true.

The basic idea behind the χ^2 test is to compare the observed (sample) cell frequencies with the expected cell frequencies. If H_0: $\{E(n_{ij}) = \mu_{ij}\}$ is true, the expected frequencies under H_0 should be close to the observed frequencies, i.e., $n_{ij} \approx \mu_{ij}$ in each cell. Large differences of $n_{ij} - \mu_{ij}$ would indicate that it is likely that H_0 is not true and that we should reject H_0.

Pearson's χ^2 (Sect. 2.4.1)

In 1900, Karl Pearson proposed the following test statistic for comparing the observed and expected frequencies:

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}. \qquad (4)$$

Note that the minimum value of X^2 is 0, obtained when and only when $n_{ij} = \mu_{ij}$ for all i,j. Thus, we are only interested in an upper-tailed test of H_0 based on X^2.

Pearson showed that X^2 has an approximate χ^2 distribution when the values of n_{ij} are "large enough." The general rule of thumb is that the χ^2 approximation is adequate as long as $\mu_{ij} = n\pi_{ij} \geq 5$ for all i,j. (Note that the values of $n\pi_{ij}$ are calculated using the hypothesized values of π_{ij} under the null hypothesis.) There are exact methods that are preferable to the χ^2 test that can be used to test hypotheses about the μ_{ij}. We will cover these in Section 2.6.

χ^2 Distribution

The χ^2 distribution is characterized by a parameter called the *degrees of freedom* (df). [More on why it is called the degrees of freedom later.] It can be shown that if $W \sim \chi^2(df)$, then $E(W) = df$ and $Var(W) = 2df$. The χ^2 distribution is defined only for $W \geq 0$ and is positively skewed, becoming more bell-shaped as df increases. Figure 2.1, p. 29, of our text illustrates the shape of the χ^2 distribution for df = 1, 5, 10, 20. Note that at df = 20, the χ^2 distribution is already starting to look somewhat like the normal.

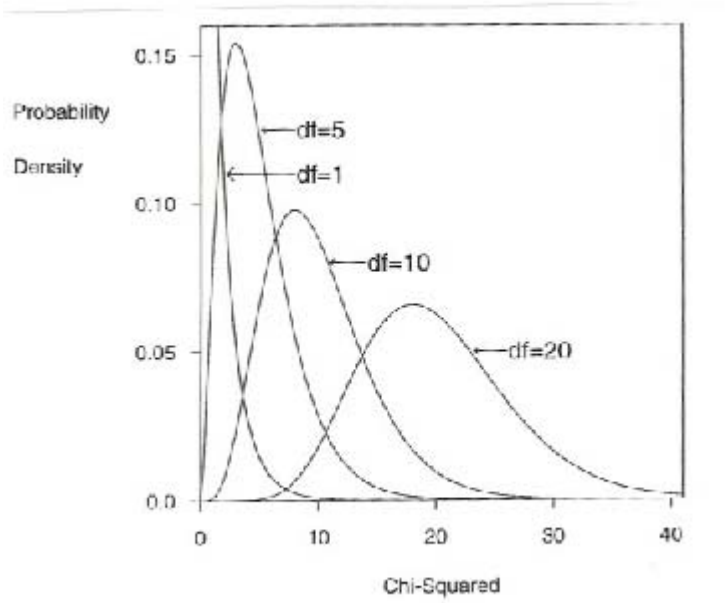

Figure 2.1 Examples of chi-squared distributions.

To perform the upper-tailed test of H_0: $E(n_{ij}) = \mu_{ij}$ using the χ^2 test, calculate an approximate p-value using the χ^2 distribution with df degrees of freedom to calculate $Pr(X^2 \geq x^2)$ where x^2 denotes the observed value of X^2 calculated from (4) above. The value of df is hypothesis-dependent, and we will discuss how to determine the degrees of freedom for a given hypothesis in Section 2.4.3.

Likelihood Ratio Test (Sec. 2.4.2)

The approach here differs from the χ^2 test in that the test statistic is based on the ratio of the maxima of two likelihood functions:

$$\Lambda = \frac{\text{max. likelihood when } H_0 \text{ is true}}{\text{max. likelihood when all parameters are unrestricted}}. \qquad (5)$$

Note that $\Lambda \leq 1$ since the numerator can be no larger than the denominator. If the max. likelihood under H_0 is much smaller than the global max., i.e. $\Lambda << 1$, this indicates that H_0 is unreasonable and should be rejected.

It can be shown that $-2\log(\Lambda)$ has an approximate χ^2 distribution. [Note that $0 \leq \Lambda \leq 1$ implies $0 \leq -2\log(\Lambda) < \infty$.] For 2-way contingency tables, we use "G^2" as the symbol for the test statistic $-2\log(\Lambda)$. Using the method of maximum likelihood (ML) to maximize the likelihood functions, and then substituting the ML estimates of μ_{ij} into the likelihood functions in Equation (5), it turns out that

$$G^2 = -2\log(\Lambda) = 2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right).$$

G^2 is called the *likelihood-ratio chi-squared statistic*. Like the χ^2 test statistic, $G^2 \geq 0$, and $G^2 = 0$ when $n_{ij} = \mu_{ij}$ for all i,j. Large values of G^2 provide evidence against H_0.

The χ^2 and LR tests are similar in many ways and often give nearly identical results, especially when n is large.

Tests of Independence (Sec. 2.4.3)

Recall that, in our general set-up for a 2-way contingency table, we used X as the symbol for the row variable and Y as the symbol for the column variable. The hypothesis that X and Y are independent is equivalent to the hypothesis that Pr(X = i and Y = j) = Pr(X = i)Pr(Y = j) for all i,j. In terms of joint and marginal probabilities, this can be written as

$$H_0: \ \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all i,j,}$$

where π_{i+} and π_{+j} denote the marginal row and column probabilities, respectively. In other words, *the joint probabilities are determined by the marginal probabilities*.

Under the null hypothesis, the expected cell frequencies are given by

$$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}.$$

However, π_{i+} and π_{+j} are usually unknown when testing independence, so we replace them by their sample estimates, yielding the *estimated expected frequencies*:

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

Note that the $\hat{\mu}_{ij}$ have the same row and column totals as the n_{ij}.

To test

$$H_0: \{X \text{ \& } Y \text{ are independent}\},$$

the χ^2 and LR test statistics are calculated as follows:

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \text{ and}$$

$$G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

Computational Formula for χ^2 Test

For a 2x2 table,the following formula should be used to calculate the χ^2 test statistic (see Exercise 2.32, pp. 51-52):

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

Finding Degrees of Freedom for χ^2 and LR Tests of Independence

Both the χ^2 and LR tests of independence have an approximate χ^2(df) distribution. How do we find the degrees of freedom (df)?

We can apply the following general rule for finding df, which works with any null hypothesis involving parameters:

df = # of non-redundant parameters associated with the alternative hypothesis - # of non-redundant parameters associated with the null hypothesis

For the test of independence, which we saw was equivalent to H_0: $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i,j, we are hypothesizing that the cell probabilities are determined by the π_{i+} (I-1 non-redundant parameters) and the π_{+j} (J-1 non-redundant parameters). Why did we substract 1 from the # of rows and columns? This is because we know that $\sum_{i=1}^{I} \pi_{i+} = 1$ and $\sum_{j=1}^{J} \pi_{+j} = 1$. Therefore, if we know all but 1 of the marginal row probabilities, we can find the remaining 1 by subtraction. The same is true for the marginal column probabilities.

Therefore, under the null hypothesis, there are a total of (I-1) + (J-1) = I + J – 2 non-redundant parameters.

How many non-redundant parameters are there associated with the alternative hypothesis, which is

H_a: {X & Y are not independent}.

Under H_a, we are not imposing any restrictions on the IJ cell probabilities. However, we do know that

$$\sum_{j=1}^{J}\sum_{i=1}^{I} \pi_{ij} = 1.$$

Therefore, there are IJ-1 non-redundant parameters since any one of the cell probabilities can be obtained by substracting the sum of the remaining probabilities from 1.

Therefore, for the test of independence,

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

$$= IJ - 1 - I + 1 - J + 1$$

$$= IJ - I - J + 1$$

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

Therefore, to test the independence of X & Y, calculate either X^2 (χ^2 test) or G^2 (LR test) and use the $\chi^2(df)$ distribution to calculate an approximate p-value.

Example (Toxicology and Trauma Study)

Consider the following 2x3 table (SPSS output):

positive screen for cocaine * mechanism of injury crosstabulation

			mechanism of injury			Total
			blunt	burn	penetrat	
positive screen for cocaine	no	Count	626	21	73	720
		Expected Count	615.0	21.9	83.1	720.0
		% within positive screen for cocaine	86.9%	2.9%	10.1%	100.0%
		Adjusted Residual	3.3	-.5	-3.3	
	yes	Count	77	4	22	103
		Expected Count	88.0	3.1	11.9	103.0
		% within positive screen for cocaine	74.8%	3.9%	21.4%	100.0%
		Adjusted Residual	-3.3	.5	3.3	
Total		Count	703	25	95	823
		Expected Count	703.0	25.0	95.0	823.0
		% within positive screen for cocaine	85.4%	3.0%	11.5%	100.0%

This contingency table examines the association between a positive tox screen for cocaine and the mechanism of traumatic injury (blunt, burn, or penetrating). The observed cell counts, expected cell counts, and % breakdown by mechanism are shown in the table.

The expected cell frequency for cell (1,2), for example, is calculated using:

$$\hat{\mu}_{12} = \frac{n_{1+}n_{+2}}{n} = \frac{(720)(25)}{823} = 21.9.$$

Thus, the χ^2 test statistic is $x^2 = 11.672$ and the LR test statistic is $g^2 = 10.024$.

The degrees of freedom are df = (I-1)(J-1) = (2-1)(3-1) = 2.

Using a $\chi^2(2)$ distribution, we find the following approximate p-values:

$$\chi^2 \text{ test: p-value} \approx \Pr(X^2 \geq 11.672) = .003$$

$$G^2 \text{ test: p-value} \approx \Pr(G^2 \geq 10.024) = .007.$$

Both tests are statistically significant, indicating that a positive tox screen for cocaine is *not* independent of the mechanism of injury.

Generally speaking, the p-values for the χ^2 and LR tests will not be that different. However, if some of the cells in the I x J contingency table have small expected frequencies, the two tests can result in different conclusions. There *are* exact versions of the two tests, and these always yield the same p-values. We will discuss these exact tests in Section 2.6. They can be performed using either SAS or SPSS.

STATISTICAL COMPUTING NOTE: We can use SimCalc to perform the usual χ^2 and LR tests:

```
CONTINGENCY TABLE - TEST OF INDEPENDENCE

    626     21     73
     77      4     22

         Chi-square =  11.67   d.f. =  2    P = .0029
   Yates correction =  10.35                P = .0056
   Likelihood ratio =  10.02                P = .0067
```

For another example of the application of the χ^2 and LR tests, see "Gender Gap Example," Section 2.4.4, pp. 30-31 in our text.

Analysis of Residuals (Sec. 2.4.5)

The p-values for the χ^2 and LR tests only tell us whether we should reject the null hypothesis of independence or not. In order to get a better understanding of what is going on with the data, we should also examine the *residuals*, which measure the discrepancy between the observed cell frequencies and the expected cell frequencies.

(Recall from regression analysis that the *residual* for the i'th observation is the difference between the observed value y_i and the predicted value $\hat{y}_i$.

However, in analyzing contingency table data, it is not sufficient to examine the (raw) residuals $n_{ij} - \hat{\mu}_{ij}$. The raw residuals do not adjust for the fact that cells with larger frequencies will tend to have larger residuals, even if the null hypothesis is true.

Agresti recommends using the *adjusted residuals*

$$ar_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}.$$

For cell (1,3) in our toxicology and trauma example, this is

$$ar_{13} = \frac{73 - 83.1}{\sqrt{83.1(1-.875)(1-.115)}} = -3.3.$$

(See the SPSS output on p. 28 of these notes.)

When the null hypothesis of independence is true, each adjusted residual has an approximate N(0,1) distribution.. Therefore, any value greater than 2 in absolute value indicates an observation that is unlikely to occur if H_0 is true.

Looking at the residuals in the toxicology and trauma example, we see "large" adjusted residuals for both "blunt" trauma and "penetrating injury." (Both are ≈ 3.3 in absolute value.) This tells us that those patients who screened positive for cocaine differed from those who did not in terms of the likelihood of blunt injury (cocaine users less likely) and penetrating injury (cocaine users more likely).

The table of residuals on p. 28 of these notes indicates that there is only one non-redundant residual in each column of the table, i.e., the residual for positive cocaine screen is the negative of the one for negative cocaine screen. Note that the absolute values of the residuals for "blunt" trauma and "penetrating injury" agree to within one significant digit only by coincidence; there is no general relationship among the residuals within a row of a 2 x J contingency table.