Handout 7

More on tests for independence

The following example is taken from your textbook 'An Introduction to Categorical Data Analysis'. It is a 2×3 table of counts on gender and party identification.

Example 1 (Table in the textbook).

The following table lists the counts as well as the estimated expected counts (under independence) and standardized residuals in the brackets. The standardized residuals are defined as

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - n_{i+}/n)(1 - n_{+j}/n)}}, \text{ where } \hat{\mu}_{ij} = n_{i+}n_{+j}/n.$$

Table 1

	Party Identification		
Gender	Democrat	Independent	Republican
Female	762 (703.7, 4.50)	327 (319.6, 0.70)	468 (533.7, -5.32)
Male	484 (542.3, -4.50)	239 (246.4, -0.70)	477 (411.3, 5.32)

When testing for independence of gender and party identification, the values of Pearson chi-square and the LR statistics are $X^2 = 30.1$ and $G^2 = 30.0$, with df= $(I-1)(J-1) = (2-1)(3-1) = 2$. The p-value for both the statistics are smaller than 0.0001. Thus the null hypothesis of independence can be rejected, ie, gender and party identification seems to related.

Let us take at the standardized residuals. For Democrats and Republicans, they are all larger than 4.5 in magnitude. For instance, there are far more democrats among women in the sample than would be expected under the hypothesis of independence. Similarly, there are more republicans among men in the sample than would be expected under the hypothesis of independence. Let us look at this table more closely.

Three 2×2 sub-tables for race and party identification.

We now look at three 2×2 sub-tables obtained from the table in Example 1. In each case, we test for independence, obtain the p-value, and write down the estimate of the appropriate odds ratio.

Table 2

	Party Identification	
Gender	Democrat	Independent
Female	762	327
Male	484	239

Here df=1, $X^2 = 1.856$ and $G^2 = 1.850$ (p-value=0.174). The odds ratio of being a democrat when comparing female to male is $\hat{\theta} = 1.151$, which is substantially higher than 1.

Table 3

	Party Identification	
Gender	Democrat	Republican
Female	762	468
Male	484	477

Here, df=1, $X^2 = 29.529$ and $G^2 = 29.526$ (p-value<0.0001) with df=1. It seems that the null hypothesis of independence cannot be rejected. The odds ratio for being a democrat when comparing females to males is 1.605

Table 4

	Party Identification	
Gender	Independent	Republican
Female	327	468
Male	239	477

Here, df=1, $X^2 = 9.661$and $G^2 = 9.689$ (p-value=0.002). The odds ratio for independent when comparing females to males is 1.395

Partitioning of G^2.

Table 2 indicates that we can merge Democrats and Independents since the proportions of democrats among females seem to be nearly the same as the proportion of democrats among the males. Thus we can merge the first two columns of Table 1 and create a new 2×2 table with Democrats+Independents and Republicans.

Table 5

	Party Identification	
Gender	Democrat+Independent	Republican
Female	1089	468
Male	723	477

Here, df=1, $X^2 = 28.256$ and $G^2 = 28.164$ (p-value<0.0001). The p-value for G^2 . The odds ratio for democrat+Independent which compares females to males is1.536.

Note that G^2 for Table 1 is the sum of G^2 values from Tables 2 and 5. This is an example of partitioning of G^2.

Table 2 contains the first two columns of Table 1, and Table 5 is obtained from Table 1 by merging the first two columns.

Note that X^2 of Table 1 is not the sum of X^2 values of Tables 2 and 5.

Remark 1

For any $I \times J$ table, one can partition G^2 as sum of $G^2$of 2×2 tables. There are explicit rules for such partitioning, and they can be found in the book 'Categorical Data Analysis' by Agresti. It should be noted that X^2 for the $I \times J$ table is not equal to the sum of partitioned X^2 of 2×2 tables. Still it may be worthwhile to examine the values of X^2 for the 2×2 tables.

Tests for independence with ordinal data.

This data is taken from the textbook, and it is a 5×2 contingency table. The independent variable (X) is alcohol consumption (average number of drinks per day), and the dependent variable (Y) is infant malfunction.

Alcohol consumption	Malfunction		Total	Percentage Present	Standardized Residual
	Absent	Present			
0	17066	48	17114	0.28	-0.18
< 1	14464	38	14502	0.26	-0.71
$1-2$	788	5	793	0.63	1.84
$3-5$	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

Test of independence: df=4, $X^2 = 12.1$ (p-value=0.02), $G^2 = 6.2$ (p-value=0.19).

Unclear if the null hypothesis of independence can be rejected. Some of the counts are small: reliability of chi-square tests questionable.

The test do not use the ordinal nature of the data.

Use of correlation assuming linear association.

Let $u_1 \leq \cdots \leq u_I$ be the scores for variable X (alcohol consumption), and $v_1 \leq \cdots \leq v_J$ be the scores for malfunction.

Here we take $u_1 = 0, u_2 = 1$, and $v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4.0$ and $v_5 = 7.0$.

Let $\bar{u}$ and $\bar{v}$ equal the sample means of the row and column scores: $\bar{u} = \sum u_i p_{i+}$ and $\bar{v} = \sum v_j p_{+j}$, where $p_{i+} = n_{i+}/n, p_{+j} = n_{+j}/n$.

The sample[correlation is

$$r = \frac{\sum_i \sum_j (u_i - \bar{u})(v_j - \bar{v}) p_{ij}}{\sqrt{\sum_i (u_i - \bar{u})^2 p_{i+} \ \sum_j (v_j - \bar{v})^2 p_{+j}}} = 0.0142, \text{.where } p_{ij} = n_{ij}/n.$$

The sample correlation r is an estimate of ρ, the population correlation between alcohol consumption and malfunction.

In order to test $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, we may use the statistic $M = (n-1)r^2$. Under H_0, $\sim \chi^2_1$.

Here $M = (n-1)r^2 = (32573)((0.0142)^2 = 6.6$, and the p-value=0.01, suggesting a nonzero linear association.

If we want to test $H_0 : \rho = 0$ vs $H_1 : \rho > 0$, we may use $M = \sqrt{n-1}\, r$. Under H_0, $\sim N(0,1)$. Here $M = 2.56$ with p-value=0.005. Thus we can reject H_0.

Remark 2.

(a) In the example above, the use of correlation as a measure of association and testing for zero correlation (independence) seems to have provided a more appropriate testing method than the usual chi-square tests.

(b) Sometimes the choice of scores for a ordinal categorical variable is unclear, and the value of r can change withe the choice of the scores. However, for a binary variable, for any choice of two numbers to describe it does not change the value of r.

(c) Correlation is a measure of linear association, and it may not be always appropriate as there may be nonlinear association.

(d) Model based methods (to be discussed later) are more appropriate than using a sample correlation for measuring association between malfunction and alcohol consumption.

Fisher's Exact Test for 2×2 Tables

When sample sizes are small, it is not appropriate to use chi-square tests. In a 2×2 table, one may use Fisher's exact test for testing independence even for small counts.

Given the row and column totals, n_{11} has a hypergeomteric distribution, ie,

$$p(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}},\ t \geq 0,$$

where $n_+ \geq t$, and $n_{+1} \geq n_{+1} - t$.

For a 2×2 table we may want to test

$H_0 : \theta = 1$ (independence) vs $H_1 : \theta > 1$ (dependence), where θ is the odds ratio.

Fisher's exact test calculates the p-value $\sum_{t \geq n_{11}} p(t)$ where the formula for $p(t)$ is given above and n_{11} is observed value.

Consider the following example.

Example 2. An English lady claimed that she could guess if or milk was pored first in a cup of tea. Eight cups of tea was prepared: 4 had milk first and 4 had tea first. Her guess and the actual was recorded.

		Lady's guess	
Actual	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Let θ be the odds ratio

$$\theta = \frac{\text{odds of guessing milk first}}{\text{odds of guessing tea first}}$$

If we want to test

$H_0 : \theta = 1$ aginst $H_1 : \theta > 1$.

The p-value is

$$p(3) + p(4) = \frac{\binom{4}{3}\binom{4}{4-3}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243.$$

This p-value is not small and thus we cannot reject H_0.

For the two-sided test $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$, the p-value is

$$\begin{aligned} & p(0) + p(1) + p(3) + p(4) \\ = & \frac{\binom{4}{0}\binom{4}{4-0}}{\binom{8}{4}} + \frac{\binom{4}{1}\binom{4}{4-1}}{\binom{8}{4}} + \frac{\binom{4}{3}\binom{4}{4-3}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} \\ = & 0.014 + 0.229 + 0.229 + 0.014 = 0.486. \end{aligned}$$

Once again we cannot reject H_0.

Remark 3.

(a) Fisher's exact test is sometimes criticized for being too conservative, ie, it tends to favor H_0 more than it should. However, there are many others who defend Fisher's exact test being appropriate

(b) Fisher's exact test is very computationally intensive for n not small.