Chi-Square Tests

Nathaniel E. Helwig

University of Minnesota

1 Introduction

In the previous chapter, we looked at inferential methods for a single proportion or for the difference between two proportions. In this chapter, we will extend these ideas to look more generally at contingency table analysis. All of these methods are a form of "categorical data analysis", which involves statistical inference for nominal (or categorial) variables. For the proportion tests, the categorical variable being analyzed had two levels: 0 (failure) or 1 (success). In the extensions discussed in this chapter, we will cover methods for testing hypotheses about distributions of categorical variables with two or more levels, as well as methods for testing the independence between categorical variables.

2 Goodness of Fit

Suppose that X is a categorical (i.e., nominal) variable that has J possible realizations, and denote the J realizations of X using the labels $0, \ldots, J-1$, so that $X \in \{0, \ldots, J-1\}$. Furthermore, let's suppose that

$$P(X = j) = \pi_j$$

where π_j is the probability that X is equal to j for $j = 0, \ldots, J-1$. Assume that the probabilities satisfy $\sum_{j=0}^{J-1} \pi_j = 1$, so that $\{\pi_j\}_{j=0}^{J-1}$ defines a valid probability mass function for the random variable X. Note that if $J = 2$, then this is simply the Bernoulli distribution. For $J > 2$, this is a multinomial distribution with $N = 1$ trial, which is a generalization of the Bernoulli distribution for simple experiments that have $J > 2$ possible outcomes.

Suppose that we have an independent and identically distributed sample $x_1, \ldots, x_n$ where $x_i \overset{\text{iid}}{\sim} F$ with F denoting the probability mass function defined by $\{\pi_j\}_{j=0}^{J-1}$. Furthermore, suppose that we want to test a null hypothesis of the form $H_0 : \pi_j = \pi_{j0}\ \forall j$ versus the alternative hypothesis $H_1 : (\exists j)(\pi_j \neq \pi_{j0})$. Note that the symbol $\forall$ should be read as "for all" and the symbol $\exists$ should be read as "there exists". Thus, the null hypothesis states that the probability for the j-th category is equal to π_{j0} for all $j = 0, \ldots, J-1$, and the alternative hypothesis states that there exists at least one category where the probability for the j-th category is not equal to π_{j0}.

Definition. Given an independent and identically distributed sample of n observations of the random variable X, the observed frequency for the j-th category is given by

$$f_j = \sum_{i=1}^{n} I(x_i = j)$$

for $j = 0, \ldots, J-1$, which is the number of observed x_i that belong to the j-th category.

Definition. Given an independent and identically distributed sample of n observations of the random variable X, the expected frequency for the j-th category is given by

$$m_j = n\pi_j$$

for $j = 0, \ldots, J-1$, which is the sample size n multiplied by the probability π_j.

To test $H_0 : \pi_j = \pi_{j0}\ \forall j$ versus $H_1 : (\exists j)(\pi_j \neq \pi_{j0})$, consider the test statistic

$$X^2 = \sum_{j=0}^{J-1} \frac{(f_j - m_{j0})^2}{m_{j0}}$$

where $m_{j0} = n\pi_{j0}$ is the expected frequency assuming that H_0 is true. Note that if the observed frequencies are far from what would be expected assuming that H_0 is true, then the value of X^2 would be relatively large. But how large is large enough to reject H_0? That is a somewhat difficult question to answer... As the sample size $n \to \infty$, the X^2 test statistic approaches a χ^2 distribution with $J-1$ degrees of freedom. However, in practice, it is difficult to know when n is large enough for the chi-square approximation to be reasonable.

***Example* 1.** Suppose a researcher is interested in studying the prevalence of vaping among college students in the United States. The researcher asked a random sample of $n = 1000$ college students the typical number of pods they vape per day, and finds the following:

# Pods	Observed Frequency	Expected Frequency
0	780	830
0-1	140	110
1-2	60	50
> 2	20	10

Note that the expected frequencies are based on last year's data, which found the following probabilities for each category: $\pi_{00} = 0.83$, $\pi_{10} = 0.11$, $\pi_{20} = 0.05$, and $\pi_{30} = 0.01$. To test the null hypothesis $H_0 : \pi_j = \pi_{j0}\ \forall j$ versus the alternative hypothesis $H_1 : (\exists j)(\pi_j \neq \pi_{j0})$, the X^2 test statistic can be calculated, such as

$$X^2 = \frac{(780-830)^2}{830} + \frac{(140-110)^2}{110} + \frac{(60-50)^2}{50} + \frac{(20-10)^2}{10} = 23.19387$$

and comparing this to a χ^2 distribution with three degrees of freedom we find a p-value of

$$p = P(\chi^2_3 > 23.19387) = 0.00003679 = 3.679 \times 10^{-5}$$

Thus, we have reason to suspect that the prevalence of vaping this year is significantly different from last year. If we wanted to do this example in R to confirm our result, we can use the `chisq.test` function

```
> f <- c(780, 140, 60, 20)
> pi0 <- c(0.83, 0.11, 0.05, 0.01)
> chisq.test(f, p = pi0)

Chi-squared test for given probabilities

data:  f
X-squared = 23.194, df = 3, p-value = 3.679e-05
```

3 Tests of Association

Suppose that A and B are both categorical (i.e., nominal) variables with $A \in \{1, \ldots, J\}$ and $B \in \{1, \ldots, K\}$. Furthermore, suppose that $P(A = j) = \pi_{j\cdot}$ for $j = 1, \ldots, J$ and $P(B = k) = \pi_{\cdot k}$ for $k = 1, \ldots, K$ with the constraints that $\sum_{j=1}^{J} \pi_{j\cdot} = 1$ and $\sum_{k=1}^{K} \pi_{\cdot k} = 1$. In addition, suppose that the probability of observing the joint event is given by

$$P(A = j \cap B = k) = \pi_{jk}$$

where the joint probabilities satisfy $\sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} = 1$.

Definition. Given an iid sample of n observations $(a_i, b_i) \overset{\text{iid}}{\sim} F$ from the joint probability distribution F, the $J \times K$ table that contains the observed frequency for each combination of A and B is referred to as a contingency table, which is also known as a cross tabulation.

	$B = 1$	$B = 2$	$\cdots$	$B = k$	$\cdots$	$B = K$	Row Totals
$A = 1$	f_{11}	f_{12}	$\cdots$	f_{1k}	$\cdots$	f_{1K}	$f_{1\cdot}$
$A = 2$	f_{21}	f_{22}	$\cdots$	f_{2k}	$\cdots$	f_{2K}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$A = j$	f_{j1}	f_{j2}	$\cdots$	f_{jk}	$\cdots$	f_{jK}	$f_{j\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$A = J$	f_{J1}	f_{J2}	$\cdots$	f_{Jk}	$\cdots$	f_{JK}	$f_{J\cdot}$
Column Totals	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot k}$	$\cdots$	$f_{\cdot K}$	$f_{\cdot\cdot} = n$

In the above table, f_{jk} denotes the number of observations that are cross-classified in category j of variable A and category k of variable B, i.e.,

$$f_{jk} = \sum_{i=1}^{n} I(a_i = j) I(b_i = k)$$

for all $j, k \in \{1, \ldots, J\} \times \{1, \ldots, K\}$. The row and column totals are the marginal observed frequencies for variables A and B, which are defined as

$$f_{j\cdot} = \sum_{i=1}^{n} I(a_i = j) = \sum_{k=1}^{K} f_{jk} \qquad \text{and} \qquad f_{\cdot k} = \sum_{i=1}^{n} I(b_i = k) = \sum_{j=1}^{J} f_{jk}$$

for all $j \in \{1, \ldots, J\}$ and all $k \in \{1, \ldots, K\}$.

To test if A and B are independent of one another, we want to test the null hypothesis $H_0 : \pi_{jk} = \pi_{j.}\pi_{.k}\ \forall j, k$ versus the alternative hypothesis $H_1 : (\exists j, k)(\pi_{jk} \neq \pi_{j.}\pi_{.k})$. To test this null hypothesis, we can use a similar idea as was used for the goodness of fit test in the previous section. Assuming that the null hypothesis H_0 is true, the expected number of observations cross-classified in cell (j, k) of the contingency table would be

$$m_{jk} = n\pi_{j.}\pi_{.k}$$

which is the sample size n multiplied by the marginal probabilities for events A and B. In practice, these marginal probabilities are almost always unknown. However, we can use the sample estimates of these marginal probabilities to estimate the expected number of observations in cell (j, k), which has the form

$$\hat{m}_{jk} = n\hat{\pi}_{j.}\hat{\pi}_{.k}$$

where $\hat{\pi}_{j.} = f_{j.}/n$ and $\hat{\pi}_{.k} = f_{.k}/n$ are the sample estimates of the marginal probabilities.

Given the estimates of the expected cell counts (assuming H_0 is true), we can use the chi-square test statistic

$$X^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(f_{jk} - \hat{m}_{jk})^2}{\hat{m}_{jk}}$$

to test whether or not the null hypothesis of independence is reasonable. Assuming that H_0 is true, the test statistic X^2 will follow a chi-square distribution with $(J-1)(K-1)$ degrees of freedom if n is large, i.e., as $n \to \infty$, we have that $X^2 \dot{\sim} \chi^2_{(J-1)(K-1)}$. Note that this is known as Pearson's chi-square test for association, given that Pearson (1900) is the statistician who originally determined that X^2 asymptotically follows a $\chi^2_{(J-1)(K-1)}$ distribution under the null hypothesis of independence between A and B. The chi-square approximation arises from the fact that, assuming H_0 is true, we have that

$$z_{jk} = \frac{f_{jk} - \hat{m}_{jk}}{\sqrt{\hat{m}_{jk}}} \xrightarrow{d} N(0, 1)$$

as the sample size $n \to \infty$. Again, in practice, it is difficult to know when n is large enough for the chi-square approximation to be reasonable.

***Example* 2.** The following contingency table is from Table 4 of Radelet and Pierce (1991), which cross-classifies individuals who received a death penalty sentence by race:

	Death Penalty		
Defendant	Yes	No	Total
White	53	430	483
Black	15	176	191
Total	68	606	674

Suppose that we want to test the null hypothesis that race and death penalty sentence are independent. The marginal probability estimates for the rows are

$$\hat{\pi}_{1\cdot} = 483/674 = 0.7166172 \qquad \text{and} \qquad \hat{\pi}_{2\cdot} = 191/674 = 0.2833828$$

and the marginal probability estimates for the columns are

$$\hat{\pi}_{\cdot 1} = 68/674 = 0.1008902 \qquad \text{and} \qquad \hat{\pi}_{\cdot 2} = 606/674 = 0.8991098$$

This implies that the null hypothesized (estimates of the) expected frequency for each cell is

	Death Penalty	
Defendant	Yes	No
White	48.72997	434.27
Black	19.27003	171.73

so the Pearson's chi-square test statistic is given by

$$\begin{aligned} X^2 &= \frac{(53-48.72997)^2}{48.72997} + \frac{(430-434.27)^2}{434.27} + \frac{(15-19.27003)^2}{19.27003} + \frac{(176-171.73)^2}{171.73} \\ &= 1.468519 \end{aligned}$$

Comparing this to a χ^2_1 distribution, the p-value for testing the null hypothesis is

$$p = P(\chi^2_1 > 1.468519) = 0.2255796$$

so we don't have sufficient evidence (using any rationale significance level) to reject the null hypothesis that race and death penalty sentence are independent.

We can confirm this result using the `chisq.test` function in R:

```
> xtab <- matrix(c(53, 15, 430, 176), 2, 2)
> colnames(xtab) <- c("Yes", "No")
> rownames(xtab) <- c("White", "Black")
> xtab
      Yes  No
White  53 430
Black  15 176
> chisq.test(xtab, correct = FALSE)

Pearson's Chi-squared test

data:  xtab
X-squared = 1.4685, df = 1, p-value = 0.2256
```

A careful reader may notice that the result of Pearson's chi-squared test is the exact same as the asymptotic test for the difference in proportions that was conducted in the previous chapter (via the `prop.test` function). For a 2×2 contingency table, it can be shown that the X^2 test statistic is identical to the Z^2 test statistic that was used for the asymptotic test of the difference between two proportions. This is because, for 2×2 contingency tables, testing the null hypothesis $H_0 : \pi_{jk} = \pi_{j\cdot}\pi_{\cdot k}\ \forall j, k$ is equivalent to testing the null hypothesis $H_0 : \pi_{1\cdot} = \pi_{2\cdot}$, i.e., the null hypothesis of independence is equivalent to assuming that the probability of success is the same for both levels of A. Note that if the probability of success is the same for both levels of A (i.e., if $\pi_{1\cdot} = \pi_{2\cdot}$), this implies that there is no association between the row variable (population) and the column variable (success/failure).

4 Conditional Association Tests

In the previous example, we looked at a two-way table (Defendant's Race by Death Penalty), however these data could actually be arranged into a three-way table where the third variable is the Race of the victim. Note that all of these defendants were on trial for committing multiple homicides, so we can look at how both the defendant's and victim's race affects the probability of receiving a death penalty sentence.

Definition. A three-way contingency table cross-classifies observations on three different categorical variables, such as the below table.

		Death Penalty		
Victim	Defendant	Yes	No	Total
White	White	53	414	467
	Black	11	37	48
Black	White	0	16	16
	Black	4	139	143
Total	White	53	430	483
	Black	15	176	191

In this example, the three variables are A = Defendant's Race, B = Death Penalty Verdict, and C = Victim's Race. Note that the previous two-way table that we looked at was a "marginal table", which is formed by aggregating the three-way table across the levels of one variable (in this case Victim's Race). We could also look at the two relevant "partial tables", which are two-way slices of the three-way table.

White Victim	Death Penalty		
Defendant	Yes	No	Total
White	53	414	467
Black	11	37	48
Total	64	451	515

Black Victim	Death Penalty		
Defendant	Yes	No	Total
White	0	16	16
Black	4	139	143
Total	4	155	159

Definition. Two variables A and B are said to be marginally independent if they are independent after aggregating across the levels of a third variable C. In contrast, two variables A and B are said to be conditionally independent given C if they are independent at all levels of the third variable C.

Note that marginal independence does not imply conditional independence. Likewise, conditional independence does not imply marginal independence. It is possible to have the marginal and conditional associations be significant in the opposite direction—which is an example of Simpson's paradox. Consequently, it is important to decide which variable(s) should be conditioned on when exploring associations in contingency tables.

***Example* 3.** In our previous example, we tested the null hypothesis that A (Defendant's Race) and B (Death Sentence Verdict) are marginally independent after collapsing across both of the levels of C (Victim's Race). In this example, we will explore the conditional independence between A and B given C, which involves testing the null hypothesis of independence separately for each level of Victim's Race, i.e., $H_0 : \pi_{jk(\ell)} = \pi_{j\cdot(\ell)}\pi_{\cdot k(\ell)}\ \forall j, k, \ell$ versus $H_1 : (\exists j, k, \ell)(\pi_{jk(\ell)} \neq \pi_{j\cdot(\ell)}\pi_{\cdot k(\ell)})$ where $\ell \in \{1, 2\}$ denotes the Victim's Race.

When the victim is white, the estimated probabilities of receiving the death penalty are

$$\hat{\pi}_{1\cdot(1)} = 53/467 = 0.1134904 \qquad \text{and} \qquad \hat{\pi}_{2\cdot(1)} = 11/48 = 0.2291667$$

Note that the probability of receiving the death penalty is twice as high for black defendants when the victim is white. To test if this difference is significant, we can use the same approach as before. The observed p-value is $p = 0.0207$, so there is some evidence to suggest that the defendant's race and the death penalty are dependent when the victim is white.

```
> white.victim <- matrix(c(53, 11, 414, 37), 2, 2)
> colnames(white.victim) <- c("Yes", "No")
> rownames(white.victim) <- c("White", "Black")
> chisq.test(white.victim, correct = FALSE)

Pearson's Chi-squared test

data:  white.victim
X-squared = 5.3518, df = 1, p-value = 0.0207
```

When the victim is black, the estimated probabilities of receiving the death penalty are

$$\hat{\pi}_{1\cdot(2)} = 0/16 = 0 \qquad \text{and} \qquad \hat{\pi}_{2\cdot(2)} = 4/143 = 0.02797203$$

and the p-value for testing the null hypothesis of independence is $p = 0.498$. However, this p-value is questionable given that we have a cell of the table with no observations—this is why R's `chisq.test` function outputs a warning message to us. Note that Pearson's chi-square test doesn't perform well when the observed frequency in any cell is too small (e.g., < 5 or < 10), so the large sample approximation should not be trusted in this case.

```
> black.victim <- matrix(c(0, 4, 16, 139), 2, 2)
> colnames(black.victim) <- c("Yes", "No")
> rownames(black.victim) <- c("White", "Black")
> chisq.test(black.victim, correct = FALSE)

Pearson's Chi-squared test

data:  black.victim
X-squared = 0.4591, df = 1, p-value = 0.498

Warning message:
In chisq.test(black.victim, correct = FALSE) :
  Chi-squared approximation may be incorrect
```

Regardless of whether we can trust the results for the black victims, it seems apparent that the Defendant's Race and the Death Penalty Sentence are **not** conditionally independent given the Victim's Race. We have evidence to suggest that black defendants are more likely than white defendants to receive the death penalty if the homicide victim is white.

References

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine 50*(302), 157–175.

Radelet, M. L. and G. L. Pierce (1991). Choosing those who will die: Race and the death penalty in Florida. *Florida Law Review 43*(1), 1–34.