Statistics 3858 : Contingency Tables

1 Introduction

Before proceeding with this topic the student should review generalized likelihood ratios $\Lambda(X)$ for multinomial distributions, its relation to Pearson's chi-squared statistic, and the Theorem about the chi-square limit distribution for $-2\log(\Lambda(X))$

Contingency table data are counts for categorical outcomes and look to be of the form

	columns				
row	1	2	...	J	
1	$n_{1,1}$	$n_{1,2}$		$n_{1,J}$	$n_{1\cdot}$
2	$n_{2,1}$	$n_{2,2}$		$n_{2,J}$	$n_{2\cdot}$
$\vdots$					
I	$n_{I,1}$	$n_{I,2}$		$n_{I,J}$	$n_{I\cdot}$
	$n_{\cdot,1}$	$n_{\cdot,2}$		$n_{\cdot,J}$	$n_{\cdot\cdot}$

This table is of J columns and I rows, which we refer to as a I by J contingency table. The count for the cell (i, j) is $n_{i,j}$. The row totals, the last column in the table, is not part of the data, but is calculated as the row sum over the column categories $1, 2, \ldots, J$ and is denoted as

$$n_{i\cdot} = n_{i,1} + n_{i,2} + \ldots + n_{i,J} = \sum_{j=1}^{J} n_{i,j}$$

The notation of the subscript $i, \cdot$ is convenient to tell us we are summing over the second index. Similarly the column totals are

$$n_{\cdot j} = n_{1,j} + n_{2,j} + \ldots + n_{I,j} = \sum_{i=1}^{I} n_{i,j}$$

Finally the total number of observations or counts is the sum of the row and column counts and is

$$n_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j} \ .$$

A random contingency table will have random counts $N_{i,j}$ in the various categories.

2 Data Generating Mechanisms

There are different sampling mechanisms that can generate random contingency tables. In our statistical inference language we have to recognize there are different statistical models, and hence different parameter spaces. When we study hypothesis testing in these models, the hypotheses depend very much on the statistical model and parameter space. There are two main mechanisms for generating random data, that is two main statistical models.

Mechanism 1

Data is collected from J experiments. For each experiment there n_j trials, and each trial falls into categories 1, 2 , ..., I. This is the type of data obtained from the Jane Austin example. For experiment j the data is

$$\mathbf{N}_j = (N_{1,j}, N_{2,j}, \ldots, N_{I,j})$$

which has a multinomial distribution with $n_{\cdot,j}$ trials and parameter

$$\mathbf{p}_{\cdot,j} = (\pi_{1,j}, \pi_{2,j}, \ldots, \pi_{I,j})$$

which belongs to the simplex of order I, say $\mathcal{S}_I$. For the experiment the parameter space is

$$\Theta = \mathcal{S}_I \times \ldots \times \mathcal{S}_I$$

This is a parameter space of dimension $J(I-1)$.

Mechanism 2

Each trial has an outcome of one of IJ categories, that is all possible (i,j) categories. The experiment consists of $n = n_{\cdot,\cdot}$ independent multinomial single trials. The random data is

$$\mathbf{N} = (N_{i,j})_{i=1,\ldots,I, j=1,\ldots,J}$$

which has a multinomial distribution with parameter n and

$$\mathbf{p} = (p_{1,1}, \ldots, p_{1,J}, p_{2,1}, \ldots, p_{I,J})$$

which belongs to the simplex of order IJ, that is

$$\Theta = S_{IJ} \ .$$

This parameter space has dimension $IJ - 1$.

The notation for a vector

$$\mathbf{a} = (a_{i,j})_{i=1,\ldots,I,j=1,\ldots,J}$$

is a convenient notation for the more cumbersome notation

$$\mathbf{a} = (a_{1,1}, \ldots, a_{1,J}, a_{2,1}, \ldots, a_{I,J})$$

and means that we are writing the components in this specific order. Actually it does not matter on the specific order as long as we are consistent in our writing, so in particular we use the ordering given here. This notation is used in the remainder of the handout where appropriate.

Notice the two sampling mechanisms and their models are quite different.

For mechanism 1, the numbers $n_{\cdot,j}$ are typically not random or are set by the mechanism generating the data - ie the experimenter, but are the number of trials for experiment j. For mechanism 1 $N_{i,j}$ are random, and for a given j we have $\sum_{i=1}^{I} N_{i,j} = N_{\cdot j}$, due to a basic property of the multinomial distribution. For mechanism 2, the numbers $N_{\cdot,j}$ are random, and in fact have a multinomial distribution with probability vector

$$\pi_{\text{Col}} = (p_{\cdot,1}, p_{\cdot,2}, \ldots, p_{\cdot,J})$$

Similarly for mechanism 2 the row totals or row counts are random with probability vector

$$\pi_{\text{Row}} = (p_{1,\cdot}, p_{2,\cdot}, \ldots, p_{I,\cdot})$$

For mechanism 1 it makes sense to consider a hypothesis such as : The J probability vectors $\mathbf{p}_{\cdot,j}$ are all equal. This question does not make sense for mechanism 2.

For mechanism 2 we could consider the two marginal distributions for Row and Column, and the two marginal random vectors, corresponding to Row and Column. Each individual trial falls into one row and column, and so may be thought of as a bivariate random variable (R, C). We could then consider a hypothesis that these two random variables R and C are independent. This question is sensible for mechanism 2, but not for mechanism 1.

These are not the only types of hypotheses that one might test for mechanisms 1 and 2, but are the most basic and common null hypotheses of interest. We will not study other hypothesis for contingency tables in this course.

Mechanism 1 is well suited for studying categorical data from J different experiments.

Mechanism 2 is well suited to study an experiment is which each individual has categorical data on two variables, and each of these variables falls into I and J categories respectively. For example we may take measurements on patients with a particular type of disease and study two physiological variables, such as blood type and diet type, or those given some treatment (cancer treatment) and measure tumor

size (small, medium or large) after 1 month and gender type (male or female). It is interesting to know if the pair of variables are independent. If so, then knowing gender will not help to predict if after treatment the tumor is large or small. In an insurance setting one may look at some classes of accidents and consider similar questions.

Contingency tables may be of more than 2 dimensions, but we only consider them of dimension 2 or sometimes 3.

The method used in the hypothesis testing is GLR, generalized likelihood ratio. In the multinomial case recall that $-2\log(\Lambda(\mathbf{N}))$ is approximately equal to Pearson's chi-square statistic; this was obtained by a Taylor's formula approximation. Also $-2\log(\Lambda(\mathbf{N}))$ converges in distribution as the sample size $n = n_{\cdot\cdot} \to \infty$, under the null hypothesis assumption, to a $\chi^2_{(df)}$ distribution and the degrees of freedom is df. The student should review how this is calculated. For sampling mechanism 1 we also need to think about what does it mean for the sample size tending to infinity.

Student's should also recall one of the calculations about the GLR and its relation to the Pearson's chi-square statistic. Specifically we found, with the notation changed for our cells indexed by i, j instead of j,

$$\begin{aligned} -2\log(\Lambda(\mathbf{N})) &\approx -2\sum_{i=1}^{I}\sum_{j=1}^{J} N_{i,j}\log(\hat{p}^{(0)}_{i,j})/\hat{p}_{i,j}) \\ &= \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{\left(N_{i,j} - \frac{N_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}\right)^2}{\hat{E}_{i,j}} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{\left(N_{i,j} - \hat{E}_{i,j}\right)^2}{\hat{E}_{i,j}} \\ &\equiv X^2 \end{aligned}$$

where $\hat{p}^{(0)}_{i,j}$ is the MLE calculated under the null hypothesis assumption, which differs for mechanism 1 and 2, and $\hat{E}_{i,j}$ is the expected counts for cell i, j, which also differ for mechanism 1 and 2.

It is an algebraic artifact that the final formula for Pearson's chi-square X^2 happens to be the same algebraic formula in both mechanism 1 and 2. This will happen due to the different expressions for the MLE under the null hypothesis and an *algebraic conincidence* or *cancellation*.

3 Test of Homogeneity or Equal Population Probability Vectors

In this problem the data is obtained in a form of mechanism 1. There are J populations with probability vectors

$$\mathbf{p}_{\cdot,j} = (\pi_{1,j}, \pi_{2,j}, \ldots, \pi_{I,j})$$

which belongs to the simplex of order I, say $\mathcal{S}_I$. For the experiment the parameter space is

$$\Theta = \mathcal{S}_I \times \ldots \times \mathcal{S}_I \ .$$

The null hypothesis of interest is

$$H_0 : \mathbf{p}_{\cdot,j} \text{ are all equal } (= \pi), j = 1, \ldots, J$$

The vector π is the common value of these vectors. This null hypothesis is also referred to as homogeneity of these probability vectors, or equality of these probability vectors. The notation means for the vectors, of length I, the corresponding components are equal, that is for each component index i we have

$$\pi_{i1} = \pi_{i2} = \ldots = \pi_{iJ}$$

or equivalently

$$\pi_{i1} = \pi_{i2} = \ldots = \pi_{iJ} \text{ for all } i = 1, 2, \ldots, I \ .$$

The vector notation is more clear. The Rice textbook uses the component notation.

We might also write

$$\mathbf{p}^{(j)} = (\pi_{1,j}, \pi_{2,j}, \ldots, \pi_{I,j})$$

or

$$\pi^{(j)} = (\pi_{1,j}, \pi_{2,j}, \ldots, \pi_{I,j})$$

as another notation or way of writing $\mathbf{p}_{\cdot,j}$.

The set Θ_0 is

$$\Theta_0 = \{(\mathbf{p}_{\cdot,1}, \ldots, \mathbf{p}_{\cdot,J}) \in \Theta \ : \ \mathbf{p}_{\cdot,j} = \mathbf{p} = (\pi_1, \ldots, \pi_I) \text{ for some } \mathbf{p} \in \mathcal{S}_I\}$$

We have

$$\text{Dim}(\Theta_0) = I - 1 \ .$$

The alternative is the general alternative

$$H_A : \mathbf{p}_{\cdot,j} \in \Theta_0^c$$

Thus the limiting χ^2 distribution has degrees of freedom

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

The student should review the multinomial MLE handout. Following this method the student should verify that the argmax for the denominator in the GLR is

$$\hat{\pi}_{i,j} = \frac{n_{i,j}}{n_{\cdot j}}$$

Next we need to calculate $\hat{\mathbf{p}}_0$, the MLE under the null hypothesis. The MLE for this case is

$$\hat{\pi}_n^{(0)} = (\hat{\pi}_{1,n}, \ldots, \hat{\pi}_{1,n}) \left(\frac{n_{1\cdot}}{n_{\cdot\cdot}}, \ldots, \frac{n_{I\cdot}}{n_{\cdot\cdot}} \right)$$

For convenience we are using the notation superscript (0) to denote the MLE under the null hypothesis assumption.

For the student : Write out the log likelihood, under the null hypothesis, and derive the MLE.

We now calculate the Pearson's chi-square formula. Notice the estimate expected counts for cell i, j is

$$\hat{E}_{i,j} = n_{\cdot j} \frac{n_{i\cdot}}{n_{\cdot\cdot}} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \ .$$

Recall mechanism 1 and that the random variable in this experiment is $\mathbf{N}_j$, whereas the trials in a given experiment is $n_{\cdot,j}$. Pearson's chi-square is then

$$\begin{aligned} X^2 &= \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left(O_{i,j} - \hat{E}_{i,j} \right)^2}{\hat{E}_{i,j}} \\ &= \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left(N_{i,j} - \frac{N_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \right)^2}{\hat{E}_{i,j}} \end{aligned}$$

The degrees of freedom for the limiting chi-square distribution are $J(I-1) - (I-1) = (J-1)(I-1)$. Thus the Pearson's chi-square r.v. $X^2 \to \chi^2_{((J-1)(I-1))}$ in distribution as $n = n_{\cdot\cdot} \to \infty$.

See Jane Austin examples using the data from the text.

4 Test of Independence for Contingency Table Data

For this problem the experiment has data collected by mechanism 2, that is each random variable is an observation from a multinomial trial with IJ categories (or classes). Here the parameter space is

$$\Theta = \mathcal{S}_{IJ}$$

Since each random observation falls into a give row and column, we may also think of it as a bivariate random variable. That is an observable random variable, say Y is also

$$Y \equiv (R, C)$$

where R and C are the random row and column for observation Y.

The null hypothesis is that the random variables R and C are independent. If the marginal distributions of these random variables are

$$\begin{aligned} \pi_{\text{Row}} &= (\pi_{1\cdot}, \pi_{2\cdot}, \ldots, \pi_{I\cdot}) \\ \pi_{\text{Col}} &= (\pi_{\cdot 1}, \pi_{\cdot 2}, \ldots, \pi_{\cdot J}) \end{aligned}$$

then the null hypothesis is that for each i, j

$$p_{i,j} = \pi_{i\cdot}\pi_{\cdot j} \ .$$

The null hypothesis does not specify what are these marginal probability vectors, so this is a composite null hypothesis.

Formally

$$H_0 : \mathbf{p} = (p_{i,j})_{i=1\ldots I, j=1\ldots J} = (\pi_{i\cdot}\pi_{\cdot j})\, i = 1\ldots I, j = 1\ldots J \text{ for some } \pi_{\text{Row}} \in \mathcal{S}_I, \pi_{\text{Col}} \in \mathcal{S}_J$$

or equivalently

$$H_0 : \mathbf{p} \in \Theta_0$$

where

$$\Theta_0 = \left\{ (p_{i,j})_{i=1\ldots I, j=1\ldots J} = (\pi_{i\cdot}\pi_{\cdot j})\, i = 1\ldots I, j = 1\ldots J \text{ for some } \pi_{\text{Row}} \in \mathcal{S}_I, \pi_{\text{Col}} \in \mathcal{S}_J \right\}$$

The dimension of Θ_0 is $(I-1)(J-1)$.

MLE under the model and null hypothesis.

In this case

$$p_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

and the log likelihood is

$$\log L(\mathbf{p}) = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{i,j}\left(\log(\pi_{i\cdot}) + \log(\pi_{\cdot j})\right)$$

It is easiest to maximize using the Lagrange multiplier method since there are two linear constraints. The student should find the MLE using this method and show we obtain

$$\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \ , \ \hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}} \ .$$

Thus under the null hypothesis assumption we have

$$\hat{E}_{ij} = n_{\cdot\cdot}\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

Completing the calculation we find the Pearson's chi-square formula (in terms of the r.v. $\mathbf{N}$) is

$$\begin{aligned} X^2 &= \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(O_{i,j} - \hat{E}_{i,j}\right)^2}{\hat{E}_{i,j}} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(N_{i,j} - \frac{N_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}\right)^2}{\hat{E}_{i,j}} \end{aligned}$$

The degrees of freedom are $IJ - 1 - (I-1) - (J-1) = IJ - I - J + 1 = (I-1)(J-1)$, and by Theorem 9.4A $X^2 \to \chi^2_{(I-1)(J-1)}$ in distribution as $n_{\cdot\cdot} \to \infty$.

This is the same test statistic and same limit distribution as in sampling mechanism 1, even though the parameter space, sample space, null hypothesis and alternative hypothesis are different in this mechanism 2 versus mechanism 1.

5 Other Hypotheses and Contingency Tables

As the Rice text mentions there are other null hypotheses of interest. These are often special cases of interest in either biostatistics or sometimes social sciences. They will also apply to insurance data. In these various applications the sampling mechanism is mechanism 2.

The applications are often for a 2 by 2 contingency table. Since the sampling mechanism is mechanism 2, the data is of the type (R, C), that is each individual sampled corresponds to a bivariate observation, indicating the *row* and *column* random variable.

5.1 Odds Ratio

Rice uses the notation of rows and columns being numbered 0 and 1.

The null hypothesis is

$$H_0 : \text{Odds}(C = 1|R = 0) = \text{Odds}(C = 1|R = 0)$$

where for an event A we define

$$\text{Odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)} \ .$$

Therefore

$$\text{Odds}(C = 1|R = 0) = \frac{P(C = 1|R = 0)}{1 - P(C = 1|R = 0)} = \frac{\pi_{01}}{\pi_{00}}$$

and

$$\text{Odds}(C = 1|R = 1) = \frac{P(C = 1|R = 1)}{1 - P(C = 1|R = 1)} = \frac{\pi_{11}}{\pi_{10}}$$

Thus H_0 is written as

$$H_0 : \frac{\pi_{01}}{\pi_{00}} = \frac{\pi_{11}}{\pi_{10}} \ .$$

There are some different methods in the way data is obtained, the so called *restrospective* and *prospective* methods.

We will not discuss this method further in the course.

5.2 Matched Pairs

This is also used in biostatistics. It is a method specialized to this type of data problem and can be used in other areas such as social science. The distinction with the general mechanism 2 is what is the population from which the experimenter draws the sample. The data is chosen from a population of pairs of individuals, for example (patients , siblings), which is then one sampling unit. A given sampling unit then has recorded random observations (R, C).

The null hypothesis of interest is that the marginal distribution of R is the same as the marginal distribution of C Since there are two rows and columns R and C can takes values 1 or 2. The null hypothesis is then

$$H_0 : P(R = 2) = P(C = 2) \ .$$

Further simplification can be made.

We will not discuss this method further in the course.