**Section 3.2 – I×J contingency tables and inference procedures (continued)**

Independence

Consider the one multinomial distribution case again. Independence between X and Y results when the occurrence of X = i does not have an effect of the occurrence of Y = j for each i = 1, …, I and j = 1, …, J. Symbolically, independence exists when

πij = πi+π+j

for i = 1, …, I and j = 1, …, J. In words, this means that the probability of an item being in cell (i,j) only involves knowing the separate marginal probabilities for X and for Y; thus, X and Y are independent of each other.

Why is independence important?

Independence helps to simplify the understanding of probabilities within the contingency table! There are only I – 1 + J – 1 = I + J – 2 unknown probability parameters. Note that the “-1” parts occur due to the  and .

Without independence, there are IJ – 1 unknown probability parameters, where the “-1” part occurs due to the . We will examine how to perform a hypothesis test for independence shortly.

Question: In the previous R example with one multinomial distribution that had

π11 = 0.2, π21 = 0.3, π12 = 0.2, π22 = 0.1, π13 = 0.1, and π23 = 0.1

how would you simulate a sample under independence?

Consider the I multinomial distributions case again. Similar to Section 1.2, it is often interest to know if these conditional probabilities are equal across the rows of the table.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Y |  |
|  |  | 1 | 2  |   | J |  |
| X | 1 | 1|1 | 2|1 |  | J|1 | 1 |
| 2  | 1|2 | 2|2 |  | J|2 | 1 |
|   |  |  |  |  |  |
| I | 1|I | 2|I |  | J|I | 1 |

Thus, we want to know if πj|1 = = πj|I for j = 1, …, J. Note that this is mathematically equivalent to πij = πi+π+j for i = 1, …, I and j = 1, …, J!

From your first statistics course, you learned that . Thus,



under independence. Because each πj|i  is equal to π+j for each i, we have πj|1 = = πj|I.

Because of this equivalence, I will refer to πj|1 = = πj|I for j = 1, …, J as “independence” as well.

Question: In the previous R example with I multinomial distributions example that had

π1|1 = 0.4, π2|1 = 0.4, π3|1 = 0.2, π1|2 = 0.6, π2|2 = 0.2, and π3|2 = 0.2

how would you simulate a sample under independence?

Test for independence

The hypotheses are:

H0: πij = πi+π+j for i = 1,…,I and j = 1,…,J

Ha: Not all equal

Remember that a Pearson chi-square test statistic calculates



for every cell of a contingency table and sums these quantities. The Pearson chi-square test for independence then uses the statistic



Notes:

* The estimated expected cell count is  =  under independence.
* X2 is equivalent to the corresponding statistic used in Section 1.2 for the Pearson chi-square test for a 2×2 contingency table.
* If the null hypothesis is true, X2 has an approximate  distribution for a large sample.
* Reject the null hypothesis if .

The LRT statistic is formed the usual way with



The numerator of Λ uses  to estimate πij, and the denominator of Λ uses  to estimate πij. The transformed statistic simplifies to



where we use 0×log(0) = 0. The large-sample distribution is the same as for X2.

Degrees of freedom

Where do the degrees of freedom come from for a test of independence?

A general way to find degrees of freedom for a hypothesis test is to calculate:

(Number of free parameters under Ha)

– (Number of free parameters under H0)

Under the alternative hypothesis for the test of independence, we have IJ πij parameters with the restriction that . When independence is true, we need the I different πi+ and the J different π+j parameters to find πij with the restriction that  and . Thus, the overall degrees of freedom is

(IJ – 1) – (I + J – 2) = (I – 1)(J – 1)

“Large sample”

What is a large enough sample to have the  work well?

This is not an easy question to answer! A common recommendation is for ni+n+j/n > 1 or > 5 for all cells of the contingency table.

What if these recommendations are not satisfied?

A  distribution may not work!

How could this affect your hypothesis test decision?

When a distributional approximation is in doubt, there are a few things that can be done:

1. Use exact inference methods (future section)
2. Use Monte Carlo simulation (discussed soon in this section!)

Example: Fiber enriched crackers (Fiber.R, Fiber.csv)

Dietary fiber is a healthful compound that is found in many vegetables and grains. Heavily processed foods are low in fiber, so it is sometimes added to such foods to make them more nutritious. Unfortunately, high-fiber foods can also have the side-effect of causing digestive bloating in some people. To investigate properties of a new fiber-enriched cracker, a study was performed to investigate these side-effects.

The participants ate the crackers and then a meal. Shortly afterward, the participants were instructed to describe any bloating that they experienced. Below is the data:

|  |  |  |
| --- | --- | --- |
|  |  | Bloating severity |
|  |  | None | Low | Medium | High |
| Fiber source | None | 6 | 4 | 2 | 0 |
| Bran | 7 | 4 | 1 | 0 |
| Gum | 2 | 2 | 3 | 5 |
| Both | 2 | 5 | 3 | 2 |

The purpose of this data is to determine if the fiber source has an effect on the bloating severity.

Before beginning the data analysis, below are a few notes:

* Notice the columns have ordinal levels. We will take this into account later in the chapter. It is instructive in a course setting to analyze the data first without taking the order into account. This will help us see the benefits later of using the ordering.
* I would expect that each person fits in one and only one cell of the table. Why would this be important to know?
* Given the layout of the data, it is likely that the sample size for each row was fixed. Thus, this would correspond to the I multinomial distribution setting.
* Fiber source could actually be analyzed as two separate explanatory variables – bran (“yes” or “no”) and gum (“yes” or “no”). I am going to analyze this data in a 4×4 contingency table format here as it was given in its original source. Please see my book for how this data can be analyzed as two separate explanatory variables through using regression models.

I read in the data from a comma-delimited file and then transform it to a contingency table format.

> diet <- read.csv(file = "C:\\data\\Fiber.csv", stringsAsFactors = TRUE)

> head(diet)

 fiber bloat count

1 bran high 0

2 gum high 5

3 both high 2

4 none high 0

5 bran medium 1

6 gum medium 3

> # Match order given in table

> diet$fiber <- factor(x = diet$fiber, levels = c("none", "bran", "gum", "both"))

> diet$bloat <- factor(x = diet$bloat, levels = c("none", "low", "medium", "high"))

> diet.table <- xtabs(formula = count ~ fiber + bloat, data = diet)

> diet.table

 bloat

fiber none low medium high

 none 6 4 2 0

 bran 7 4 1 0

 gum 2 2 3 5

 both 2 5 3 2

Pay special attention to my use of the factor() function. While it is not necessarily to use here, I use it to order the rows and columns as shown in the table.

I use three different functions next to test for independence. In practice, you only need to use one of these!

> ind.test <- chisq.test(x = diet.table, correct = FALSE)

> ind.test

 Pearson's Chi-squared test

data: diet.table

X-squared = 16.9427, df = 9, p-value = 0.04962

Warning message:

In chisq.test(diet.table, correct = FALSE) :

 Chi-squared approximation may be incorrect

> library(package = vcd)

> assocstats(x = diet.table)

 X^2 df P(> X^2)

Likelihood Ratio 18.880 9 0.026230

Pearson 16.943 9 0.049621

Phi-Coefficient : 0.594

Contingency Coeff.: 0.511

Cramer's V : 0.343

> class(diet.table)

[1] "xtabs" "table"

> summary(diet.table)

Call: xtabs(formula = count ~ fiber + bloat, data = diet2)

Number of cases in table: 48

Number of factors: 2

Test for independence of all factors:

 Chisq = 16.943, df = 9, p-value = 0.04962

 Chi-squared approximation may be incorrect

> qchisq(p = 0.95, df = 9)

[1] 16.91898

In summary,

* X2 = 16.94
* -2log(Λ) = 18.88
* 
* P-value using X2 is P(A > 16.94) = 0.0496 where A has a  distribution
* P-value using -2log(Λ) is P(A > 18.88) = 0.0262 where A has a  distribution
* Because the p-value is small, but not extremely so, we would say there is marginal evidence against independence. Thus, there is marginal evidence of dependence.
* Thus, there is marginal evidence that bloating severity is dependent on the fiber source.

Can we trust the  approximation? Below are the expected cell counts:

> ind.test$expected

 bloat

fiber none low medium high

 none 4.25 3.75 2.25 1.75

 bran 4.25 3.75 2.25 1.75

 gum 4.25 3.75 2.25 1.75

 both 4.25 3.75 2.25 1.75

These only partially satisfy the recommendations given earlier.

Alternatives to using a  approximation

I will focus here on using Monte Carlo simulation. Closely related methods are Fisher’s exact and permutation tests. These tests are discussed in another chapter.

Below is a summary of the Monte Carlo simulation process:

1. Simulate a large number of contingency tables of size n assuming the null hypothesis is true (i.e., **set** πij =  for one multinomial or **set** πj|i =  for I multinomials); let B be the number of contingency tables.
2. Calculate the X2 (or -2log(Λ)) statistic for each simulated contingency table; call these values ’s to differentiate them from the original observed X2 value
3. Plot a histogram  and overlay the 
4. Calculate ; this is the p-value for a hypothesis test

Notes:

* Step 3 helps to visualize if the original  distribution approximation was appropriate.
* Some simulated contingency tables may have less than I rows or J columns. These contingency tables should be excluded. Why?
* If there are a lot of contingency tables excluded for reasons given in the previous bullet, one should use Fisher’s exact or permutation tests instead.

Example: Fiber enriched crackers (Fiber.R, Fiber.csv)

Please see the supplemental notes file for the computational details.

After recording video: See the practice problems for the details

Below is a histogram with a  approximation overlaid. Also, I have included a plot of the CDF for a  along with an empirical CDF of the  values.

For a sample of observations , the empirical CDF at w is the proportion of observations at or below w: .



We see that the chi-square distribution approximation is quite good.

The statistic calculated on the observed data is X2 = 16.94. The p-value using Monte Carlo simulation is 0.0446. When using a  approximation previously, we had a p-value of 0.0496. Through using both inference methods, there is moderate evidence against independence.

If independence is rejected, we would like to determine why it is rejected. For example, perhaps only particular combinations of X and Y are causing the dependence. Also, we would like to determine how much dependence exists. There are a number of ways to examine a contingency table further to understand the dependence. My preference is to use statistical models instead for this purpose, while even using these models to help test for independence. The remainder of this chapter describe these models.