Class Notes: STAT 501

Nonparametrics & Log-Linear Models

$I \times J$ Tables

$2 \times 2 \times k$ Tables

$I \times J \times K$ Tables

Da Kuang
University of Pennsylvania

Contents

1 Basic Combinatorics

1.1 Permutation

How many possible combinations are there for the computer login password if it must consist of a letter followed by a number?

If a job consist of k separate tasks, the i-th of which can be done in n_i ways, $i = 1, \cdots, k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways.

Definition 1.1 (Factorial): The factorial of a positive integer n is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

Also, we define, $0! = 1$.

Definition 1.2 (Permutation): A permutation of a set of objects is an ordered arrangement of the objects.

1.2 Combination

Order is not always meaningful, for instance the order of a hand of poker cards is actually does not matter.

Definition 1.3 (Combination): We call a collection of r unordered elements a combination of size r. In general, the number of combinations of size r from a group of $n(n \geq r \geq 0)$ objects is $\binom{n}{r}$.

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} = \frac{n \times (n-1) \times (n-r+1)}{r!}.$$

1.3 Multinomial Coefficients

A set of n distinct items is to be divided into I distinct groups of respective size n_1, $n_2, \ldots, n_I$, where $\sum_{i=1}^{I} n_i = n$. How many different divisions are possible?

There are the following possible divisions.

$$
\begin{aligned}
&\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots n_{I-1}}{n_I} \\
=&\frac{n!}{(n-n_1)!n_1!}\frac{(n-n_1)!}{(n-n_1-n_2)!n_2!}\cdots\frac{(n-n_1-n_2-\cdots-n_{I-1})!}{0!n_I!} \\
=&\frac{n!}{n_1!\cdots n_I!}
\end{aligned}
$$

Definition 1.4 (Multinomial Coefficient): We define the multinomial coefficient as

$$\binom{n}{n_1 n_2 \cdots n_I} \equiv \frac{n!}{n_1!\cdots n_I!}$$

1.4 Multinomial Distribution

Suppose we have n independent trials,

- each trial can result in one of I types of outcomes;
- on each trial the probabilities of the I outcomes are respectively $p_1, p_2, \cdots, p_I$;

Let X_i be the total number of outcomes of type I in the n trials. Any particular sequence of trials giving rise to $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_I = x_I$ occurs with probability $p_1^{x_1} p_2^{x_2} \cdots p_I^{x_I}$.

Recall that there are $\binom{n}{x_1 x_2 \cdots x_I}$ such sequences. Therefore the probability for a certain sequence happens is that

$$p(x_1, \cdots, x_I) = \binom{n}{x_1 x_2 \cdots x_I} p_1^{x_1} p_2^{x_2} \cdots p_I^{x_I}$$

The marginal distribution of X_1 is

$$\sum_{x_2,\cdots,x_I} \binom{n}{x_1 x_2 \cdots x_I} p_1^{x_1} p_2^{x_2} \cdots p_I^{x_I}$$

Also, binomial distribution can be derived from multinomial distribution.

2 $I \times J$ Table

Suppose we have two categorical variables, X and Y. The number of categories of X is I and the number of categories of Y is J.

Definition 2.1 (Contingency Table): A rectangular table having I rows for the categories of X and J columns for the categories of Y has cells that display the IJ possible combinations of outcomes.

Sometimes, it is also called frequency table or cross-classification table.

Suppose the total number of observations $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$.

	Y				
X	1	2	$\cdots$	J	Total
1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_I
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	n

Figure 1: Example of $I \times J$ Contingency Table

3 Study Designs

3.1 Design 1: One Multinomial Sampling with n Fixed

If the total sample size of observations n is fixed in advance (e.g., in a cross-sectional study), then a multinomial sampling model might be used where cell counts are treated as multinomial random variables with index n and probabilities π_{ij}.

3.2 Design 2: I Multinomial Distributions

If the row totals n_i are fixed in advance, the counts n_{ij} have a multinomial distribution with index n_i. So there are I multinomial distributions.

Sometimes n_i is not fixed in advance. The row variable is an explanatory variable and the column variable is a response variable. $\mathbf{P}(Y = j|X = i)$.

3.3 Design 3: The Lady Tasting Tea

If both row and column totals are fixed by design, then a hyper-geometric sampling distribution applies for the cell counts.

3.4 Design 4

If observations are to be collected over a certain period of time and cross-classified into one of the $I \times J$ categories, then a Poisson sampling model might be used where cell counts are treated as independent Poisson random variables with parameters μ_{ij}'s.

4 Divide into The Fist Two Designs

4.1 One Multinomial Sampling with n Fixed

The contingency Table for this design is as follows

	Y				
X	1	2	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_1
2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_I
Total	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.J}$	1

Figure 2: Example of Contingency Table

Here $\pi_{ij} = \mathbf{P}(X = i, Y = j)$ is the probability that (X, Y) falls in the ij-th cell. The combinations of π_{ij}'s form the joint distribution of X and Y.

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

The marginal distribution are the row and column totals of the joint probabilities, i.e. $\pi_{i.}$'s and $\pi_{.j}$'s.

$$\mathbf{P}(X = i) = \pi_{i.}, \mathbf{P}(Y = j) = \pi_{.j}$$

Given a sequence $z_{11}, \cdots, z_{IJ}$, its coresponding probability is

$$\mathbf{P}(n_{11} = z_{11}, \cdots, n_{ij} = z_{IJ}) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!}\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{n_{ij}}.$$

The estimator of the cell probability is $\widehat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n}$, and $\widehat{\pi}_i = \frac{n_i}{n}$.

4.2 Independence

In this step-up, we are usually interested in the dependency between X and Y. When X does not have an effect on the probabilities for the outcomes of Y, we say that Y is independent of X. When they are independent,

$$\mathbf{P}(X = i, Y = j) = \mathbf{P}(X = i)\mathbf{P}(Y = j)$$

$$\pi_{ij} = \pi_i \pi_{.j}$$

Independence simplifies the probability structure within a contingency table by reducing the number of unknown parameters from IJ to $I - 1 + J - 1 = I + J - 2$ marginal probabilities.

4.3 I Multinomial Distributions

The contingency Table for this design looks the same

	Y				
X	1	2	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_1
2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_I
Total	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.J}$	1

Figure 3: Example of Contingency Table

We have a separate J-category multinomial distribution in each of the I groups. Define $\pi_{j|i} = \mathbf{P}(Y = j|X = i)$ as the probability of observing response category j given that a unit is from group i.

$$\sum_{j=1}^{J} \pi_{j|i} = 1, \text{ for each } i = 1, \cdots, I.$$

The probability of observing a sequence $z_{i1}, \cdots, z_{iJ}$ from group i is

$$\mathbf{P}(n_{i1} = z_{i1}, \cdots, n_{iJ} = z_{iJ} | n_i = \frac{n!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}$$

The full model for the contingency table is the Product Multinomial Model

$$\prod_{i=1}^{I} \frac{n!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}$$

4.4 Independence

Independence of X and Y in the context of a product multinomial model means that the conditional probabilities for each Y are equal across the rows of the table.

For each $j = 1, \cdots J$, we have

- $\mathbf{P}(Y = j | X = 1) = \cdots = \mathbf{P}(Y = j | X = I) = \mathbf{P}(Y = j)$.
- $\pi_{j|1} = \cdots = \pi_{j|I} = \pi_{.j}$
- $\widehat{\pi}_{j|i} = \frac{n_{ij}}{n_i} = \frac{n_{ij}}{n} / \frac{n_i}{n} = \frac{\widehat{pi}_{ij}}{\widehat{\pi}_i}$.

Note that this condition is mathematically equivalent to independence as defined for the one multinomial model.

Proof. Because X is not random in the product multinomial model, we define π_i to be the fixed proportion of the total sample that is taken from group i.

Then $\pi_{j|i} = \frac{\pi_{ij}}{\pi_i} = \pi_{.j}$ together imply that $\pi_{ij} = \pi_i \pi_{.j}$ ■

4.5 Summary

- Parameter estimates from the one and product multinomial models are the same.
- The definitions of independence in the two models are equivalent.
- The two models lead to exactly the same conditional distributions for Y given $X = i$.
- As a consequence, analyses conducted based on each model generally yield the same results.
- Therefore, when developing tests for independence and other analyses on contingency tables, we assume whichever model for the table is most convenient.

5 Chi-squared Test of Independence

- Null Hypothesis: X and Y are independent.
- Test Statistic:

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_i n_{.j}}{n})^2}{\frac{n_i n_{.j}}{n}}$$

- Reject the null hypothesis if

$$T > \chi^2_{(I-1)(J-1),1-\alpha}$$

5.1 Example

Scarlet fever is a childhood infection that among other symptoms gives rise to severe irritation of the nose, throat and ears.

In a study, six districts A to F were chosen. In each district, patients were located, and parents were asked to state the site at which they thought their child's irritation was worst.

	District						
	A	B	C	D	E	F	Total
Nose	1	1	0	1	8	0	11
Throat	0	1	1	1	0	1	4
Ears	1	0	0	0	7	1	9
Total	2	2	1	2	15	2	24

Figure 4: Example of Chi-square Test

```
A=c(1,0,1)
B=c(1,1,0)
C=c(0,1,0)
D=c(1,1,0)
E=c(8,0,7)
Ff=c(0,1,1)

da=cbind(A,B,C,D,E,Ff)
da
# A B C D E Ff
# [1,] 1 1 0 1 8 0
# [2,] 0 1 1 1 0 1
# [3,] 1 0 0 0 7 1
chisq.test(da)
# Pearson's Chi-squared test
#
# data: da
# X-squared = 14.96, df = 10, p-value = 0.1335
#
# Warning message:
# In chisq.test(da) : Chi-squared approximation may be incorrect
fisher.test(da)
# Fisher's Exact Test for Count Data
#
```

```
# data: da
# p-value = 0.02613
# alternative hypothesis: two.sided
```

Note that there is warning when running the `chisq.test`. It is because that the chi-square test is based on CLT but the sample size is not large enough for approximation.

To have enough sample size, for all cells of the contingency table, we have

$$\frac{n_i n_{.j}}{n} > 1 \text{ or } > 5.$$

We turn to Fisher Exact test to have better result. But note that the odd ratio is undefined for multinomial distribution.

6 $2 \times 2 \times k$ Table

6.1 Setup

The data consist of k strata, $i = 1, 2, \cdots, k$. Within each stratum, we have a 2×2 table.

	Success	Failure	Total
Treatment	$n_{11,i}$		$n_{1,i}$
Control	$n_{21,i}$		$n_{2,i}$
Total	$n_{.1,i}$	$n_{.2,i}$	n_i

Figure 5: One of the Stratum

The two rows of the 2×2 table in the i-th stratum are viewed as data from two independent binomial distributions.

	Success	Failure
Treatment	$\pi_{1,i}$	
Control	$\pi_{2,i}$	

Figure 6: Probabilities in One of the Stratum

6.2 Hypothesis Test

Null hypothesis is that within each straum, the success probabilities are equal.

$$H_0 : \pi_{1,i} = \pi_{2,i}, i = 1, \cdots, k.$$

Let θ denote the odds ratio for the i-th table.

$$\theta_i = \frac{\pi_{1,i}/(1-\pi_{1,i})}{\pi_{2,i}/(1-\pi_{2,i})}$$

Use odds ratio to represent the null hypothesis.

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = 1$$

Note that we are testing that there is a common odds ratio and it is equal to 1. Moreover, H_0 allows for the common success probabilities to differ from stratum to stratum.

Note that the alternative hypothesis must be consistent across the stratum. By consistent, I mean it is either

$$H_1 : \pi_{1,i} \geq \pi_{2,i} \text{ for all } i = 1, \cdots, k \text{ with at least one inqeuality.}$$

or

$$H_1 : \pi_{1,i} \leq \pi_{2,i} \text{ for all } i = 1, \cdots, k \text{ with at least one inqeuality.}$$

7 Mantel-Haenszel Chi-Squared Test

In the i-th table, given the marginal totals $n_{1,i}$, $n_{2,i}$, $n_{.1,i}$, $n_{.2,i}$ are fixed, the random variable $n_{11,i}$ has a hyper-geometric distribution.

$$\mathbf{P}(n_{11,i} = z) = \frac{\binom{n_{1,i}}{z}\binom{n_{2,i}}{n_{.1,i}-x}}{\binom{n_i}{n_{.1,i}}}.$$

Under the H_0, we have

$$\begin{aligned}\mathbf{E}(n_{11,i}) =& \frac{n_{1,i}n_{.1,i}}{n_i};\\ \mathbf{Var}(n_{11,i}) =& \frac{n_{1,i}n_{2,i}n_{.1,i}n_{.2,i}}{n_i^2(n_i-1)}\end{aligned}$$

Also under the H_0, we have the statistic MH for Mantel-Haenszel Chi-squared Test.

$$\mathrm{MH} = \frac{\sum_{i=1}^{i}(n_{11,i} - \mathbf{E}(n_{11,i}))}{\sqrt{\sum_{i=1}^{i}\mathbf{Var}(n_{11,i})}}$$

The rejected regions are as follows.

H_1: $\pi_{1,i} \geq \pi_{2,i}$ for all $i = 1, \dots, k$ with at least one inequality.

- Reject H_0 if $MH \geq z_\alpha$.

H_1: $\pi_{1,i} \leq \pi_{2,i}$ for all $i = 1, \dots, k$ with at least one inequality.

- Reject H_0 if $MH \leq -z_\alpha$.

H_1: $\pi_{1,i} \geq \pi_{2,i}$ for all $i = 1, \dots, k$ or $\pi_{1,i} \leq \pi_{2,i}$ for all $i = 1, \dots, k$ with at least one inequality.

- Reject H_0 if $(MH)^2 \geq \chi^2_{\alpha,1}$.

Figure 7: Rejected Region

7.1 Example 1

```
## Penicillin and Rabbits
## Investigation of the effectiveness of immediately injected or 1.5
## hours delayed penicillin in protecting rabbits against a lethal
## injection with beta-hemolytic streptococci.

Rabbits <-
  array(c(0, 0, 6, 5,
          3, 0, 3, 6,
          6, 2, 0, 4,
          5, 6, 1, 0,
          2, 5, 0, 0),
        dim = c(2, 2, 5),
        dimnames = list(
          Delay = c("None", "1.5h"),
          Response = c("Cured", "Died"),
          Penicillin.Level = c("1/8", "1/4", "1/2", "1", "4")))
Rabbits
# , , Penicillin.Level = 1/8
#
# Response
# Delay Cured Died
# None 0 6
# 1.5h 0 5
#
# , , Penicillin.Level = 1/4
#
# Response
# Delay Cured Died
# None 3 3
# 1.5h 0 6
```

```
#
# , , Penicillin.Level = 1/2
#
# Response
# Delay Cured Died
# None 6 0
# 1.5h 2 4
#
# , , Penicillin.Level = 1
#
# Response
# Delay Cured Died
# None 5 1
# 1.5h 6 0
#
# , , Penicillin.Level = 4
#
# Response
# Delay Cured Died
# None 2 0
# 1.5h 5 0

## Classical Mantel-Haenszel test
mantelhaen.test(Rabbits)
# Mantel-Haenszel chi-squared test with continuity correction
#
# data: Rabbits
# Mantel-Haenszel X-squared = 3.9286, df = 1, p-value = 0.04747
# alternative hypothesis: true common odds ratio is not equal to 1
# 95 percent confidence interval:
# 1.026713 47.725133
# sample estimates:
```

```
# common odds ratio
# 7
```

7.2 Example 2

Transform the data if it is not categorical.

```
Satisfaction <-
  as.table(array(c(1, 2, 0, 0, 3, 3, 1, 2,
                   11, 17, 8, 4, 2, 3, 5, 2,
                   1, 0, 0, 0, 1, 3, 0, 1,
                   2, 5, 7, 9, 1, 1, 3, 6),
                 dim = c(4, 4, 2),
                 dimnames =
                   list(Income =
                          c("<5000", "5000-15000",
                            "15000-25000", ">25000"),
                        "Job Satisfaction" =
                          c("Very_D", "A Little_S", "Moderately_S", "Very_S"
                            ),
                        Gender = c("Female", "Male"))))
Satisfaction
# , , Gender = Female
#
# Job Satisfaction
# Income Very_D A Little_S Moderately_S Very_S
# <5000 1 3 11 2
# 5000-15000 2 3 17 3
# 15000-25000 0 1 8 5
# >25000 0 2 4 2
#
# , , Gender = Male
```

```
#
# Job Satisfaction
# Income Very_D A Little_S Moderately_S Very_S
# <5000 1 1 2 1
# 5000-15000 0 3 5 1
# 15000-25000 0 0 7 3
# >25000 0 1 9 6
## (Satisfaction categories abbreviated for convenience.)
ftable(. ~ Gender + Income, Satisfaction)
# Job Satisfaction Very_D A Little_S Moderately_S Very_S
# Gender Income
# Female <5000 1 3 11 2
# 5000-15000 2 3 17 3
# 15000-25000 0 1 8 5
# >25000 0 2 4 2
# Male <5000 1 1 2 1
# 5000-15000 0 3 5 1
# 15000-25000 0 0 7 3
# >25000 0 1 9 6
## Table 7.8 in Agresti (2002), p. 288.
mantelhaen.test(Satisfaction)

# Cochran-Mantel-Haenszel test
#
# data: Satisfaction
# Cochran-Mantel-Haenszel M^2 = 10.2, df = 9, p-value = 0.3345
```