The Analysis of Categorical Data

Zhaoxia Yu

May 13, 2009

1 Introduction

In the first lecture, we collected a data set with many variables: some are continuous, such as weight; some are categorical, such as gender. In the past six weeks, we have learned several tools for analyzing data: one-sample t-test, two-sample t-test, ANOVA, and linear regression. Check the following questions and think about what test(s) you should consider:

- Father's body mass index is greater than mother's body mass index (paired t-test, one-sample t-test)
- Body mass index is different between men and women (two-sample t-test, one-way ANOVA)
- The effect of gender and smoking on weight (two-way ANOVA)
- Kid's height is linearly associated with mother's height (linear regression with one covariate)
- Kid's height has a linear relationship with parents' (linear regression with two covariates)
- Smoking and gender?

So far we have only been concerned with continuous responses. In practice, it is also interesting to ask the relationship between two discrete variables. For example, is blood type related to a certain disease? In the following two weeks, we will focus on the analysis of data in the form of counts in various categories, especially the two-way tables.

Definition: A **categorical variable** has a measurement scale consisting of a set of categories. For instance, blood type is often measures as A, B, AB, or O. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant.

Categorical variables have two primary types of scales:

Definition: Variables having categories without a natural ordering are called **nominal**. E.g., blood type (A, B, AB, O).

Definition: Categorical variables with ordering are called **oridal**. E.g., socioeconomic class (upper, middle, low), patient condition (good, fair, serious, critical).

Variables concerned here are nominal variables.

1.1 Brief Review

- Multinomial distribution

  Suppose that each of n independent, identical trials can have outcome in any of c categories. Let N_i denote the number of trials having outcome in category i. The counts $(N_1, N_2, \cdots, N_c)$ have the multinomial distribution, i.e.,

  $$(N_1, N_2, \cdots, N_c) \sim Multinomial(, (\pi_1, \pi_2, \cdots, \pi_c))$$

  where π_i denote the probability of outcome in category i for each trial.

  The multinomial probability mass function is

  $$p(N_1 = n_1, N_2 = n_2, \cdots, N_c = n_c) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

  where

  $$\begin{aligned} \sum_{i=1}^{c} N_i &= n \\ \sum_{i=1}^{c} \pi_i &= 1 \\ E(N_i) &= n\pi_i \\ Var(N_i) &= n\pi_i(1-\pi_i) \\ cov(N_i, N_j) &= -n\pi_i\pi_j, i \neq j \end{aligned}$$

  Marginal distribution: $N_i \sim Binomial(, \pi_i)$.

  Conditional distribution:

  $$(N_1, \cdots, N_{c-1}) | N_c = n_c \sim Multinomial(n - n_c, \frac{\pi_1}{1-\pi_c}, \cdots, \frac{\pi_{c-1}}{1-\pi_c})$$

- Binomial distribution

  When $c = 2$, the multinomial distribution is the binomial distribuiton. Let π be the success rate and Y be the number of successes from n independent trials, then

  $$p(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

  and

  $$\mu = E(Y) = n\pi \ \sigma^2 = Var(Y) = n\pi(1-\pi)$$

  .

- Hypergeometric distribution

  Suppose that an urn contains n balls, of which r are read $n-r$ are white. Let X denote the number of red balls drawn when taking m balls without replacement. Then

$$Pr(X=k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$$

2 Fisher's Exact Test

Consider two categorical variables, each has two levels. We can summerize data using a 2-by-2 contigency table. When the sample size is small, the asymptotic test, such as the chi-squared test, is not an appropriate method of analysis if minimjum expected coutns are small. For example, n is less than 20, or if one of the expected counts is less than 5, the chi-squared test should be avoided.

A test that may be used when the size requirements of chi-squared test are not met was proposed in the mid-1930s by Fisher and others. The test has come to known as the Fisher exact test. It is called exact because, if desired, it permits us to calculate the exact probability of obtaining the observed results or results that are more extreme.

2.1 Data arragement

When we use the Fisher exact test, we arrange the data in the form of a 2-by-2 table:

	B	$\bar{B}$	Total
A	N_{11}	N_{12}	$n_{1\cdot}$
$\bar{A}$	N_{21}	N_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

2.2 Assumptions for Fisher's exact test

- Independent observations
- Fixed marginal counts
- For each characteristic (categorical variable), each observation can be categorized as one of the two mutually exclusive types.

2.3 Hypothesis

Depending on study design, which will be discussed later, we can answer one of the two questions.

- H_0: The proportion of A is the same between subjects with B and $\bar{B}$ H_1: The proportion of A is not the same between subjects with B and $\bar{B}$

- H_0: The two categorical variables are associated with each other. H_1: The two categorical variables are not associated with each other.

The above two problems are two-sided. One can aslo ask one-sided problems.

2.4 Test statistic

The test statistic is N_{11}, the number in sample 2 with the characteristic of interest.

2.5 Distribution of the test statistic under the null hypothesis

Under the null, the two smaples are from the same populatio with $n_{..}$ balls of which $n_{1\cdot}$ has the characteristic of interest. Therefore,

$$Pr(N_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-n_{11}}}{\binom{n_{..}}{n_{\cdot 1}}} = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{..}}{n_{\cdot 1}}}$$

2.6 Decision Rule

2.7 Example

The purpose of a study was to evaluate whether a student has a web page or not depends the gender of the student. We have a random sample of 21 studens. The collected data are shown in the following 2-by-2 table:

	Male	Female
have	2	7
not have	8	4

N_{11}	Prob	
0	0.00019	(**)
1	0.00561	(**)
2	0.05052	(*)
3	0.18862	
4	0.33008	
5	0.28292	
6	0.11789	
7	0.02245	(**)
8	0.00168	(**)
9	0.00003	(**)

Two-sided p-value: 0.00019 + 0.00561 + 0.05062 + 0.02245 + 0.00168 + 0.00003 = 0.08050. Becuase the p-value is greater than 0.05, we fail to reject the null hypothesis at significance level

0.05. Our conclusion is that there is not enough to evidence to support the dependence between gender and web page.

You can do the calculation using R

```
> prob.observed=dhyper(2, 9, 12, 10)
> prob.observed
[1] 0.05052223
> prob.all=dhyper(0:9, 9, 12, 10)
> cbind(x=0:9,prob=prob.all)
      x         prob
 [1,] 0 1.871194e-04
 [2,] 1 5.613581e-03
 [3,] 2 5.052223e-02
 [4,] 3 1.886163e-01
 [5,] 4 3.300786e-01
 [6,] 5 2.829245e-01
 [7,] 6 1.178852e-01
 [8,] 7 2.245433e-02
 [9,] 8 1.684074e-03
[10,] 9 3.402171e-05
> p.value=sum(prob.all[prob.all <= prob.observed])
> p.value
[1] 0.08049536
```

There is an R function to perform Fisher's exact time. See the following

```
> fisher.test( matrix(c(2,8,7,4),2,2) )

        Fisher's Exact Test for Count Data

data:  matrix(c(2, 8, 7, 4), 2, 2)
p-value = 0.0805
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.01098553 1.36857463
sample estimates:
odds ratio
 0.1586641
```

2.8 Limitations

Some disadvantages of Fisher's exact test:

1. For large samples, calculation required for extrem tables is demanding.
2. The calculation for m-by-n tables is not straightforward.

An alternative is to consider asymptotic tests, such as Pearson's chi-squared test and large-sample likelihood ratio test. We will focus on Pearson's chi-squared test in this course. An example of large-sample likelihood ratio test is provided in homework 6.

3 Asymptotic Tests for Contigency Tables

3.1 The General Two-Way Contigency Table

Consider a two-way table with I rows and J columns:

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{I1}	n_{J2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

where

- n_{ij} is the observed count in row i and column j
- $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$ is the total number of observations in row i.
- $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$ is the total number of observations in column j.
- $n_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ is the total number of observations in the table.

Different questions can be asked based on an observed contigency table. For example, are the two factors independent? Are counts from different groups homogenous? The Pearson's chi-squared test can be used to address this type of questions.

3.2 The Pearson's Chi-Squared Test

The genernal Pearson's chi-squared statistic is defined as

$$X^2 = \sum_{i=1}^{c} \frac{(Obs_i - Exp_i)^2}{Exp_i}$$

where c is the total number of cells; Obs_i is the observed count for cell i. Exp_i is the expected count for cell i under a specific null distribution and it can be calculated based on the mle of model parameters under a null distribution.

Under a specific null hypothesis, the chi-squared statistic follows a chi-squared distribution asymptotically. The degrees of freedom of a X^2 are the number of independent parameters in the full model minus that in the reduced model. For the two-way contigency table with I rows and J columns, the chi-squared statistic can be written as

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

of which the degrees of freedom depends on the null hypothesis.

3.2.1 The Theorectical Justification of the Pearson's Chi-Squared Statistic

For a multinomial sample $(n_1, n_2, \cdots, n_c)$ of size n, the marginal distribution of n_i is the $Binomial(n, \pi)$ distriubiton. For large n, by the normal approximation to the binomial, n_i has approximate normal distribution. By the central limit theorem,

$$\hat{\pi} = (n_1/n, \cdots, n_{c-1}/n)^T$$

has an approximate multivariate normal distribution. Let Σ_0 denote the null covariance matrix of $\sqrt{n}\hat{\pi}$, and let $\pi_0 = (\pi_{10}, \pi_{20}, \cdots, \pi_{c-1,0})^T$. Under H_0,

$$\sqrt{n}(\hat{\pi} - \pi_0) \to N(0, \Sigma_0)$$

Therefore, the quadratic form

$$n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1} (\hat{\pi} - \pi_0) \to \chi^2_{c-1}$$

The covariance matrix of $\sqrt{n}\hat{\pi}$ has elements

$$\begin{aligned} \sigma_{jk} &= -\pi_j \pi_k \text{ if } j \neq k \\ \sigma_j &= \pi_j(1 - \pi_j) \text{ if } j = k \end{aligned}$$

The matrix Σ_0^{-1} has (j,k)th element $1/\pi_{c0}$ when $j \neq k$ and $(1/\pi_{j0} + 1/\pi_{c0})$ when $j = k$. It can be shown that the quadratic form is identical to the Pearson's Chi-squared statistic.

Below is an argument used by R. A. Fisher (1922). Suppose $(n_1, n_2, \cdots, n_c)$ are independent Poisson r.v.s with means $(\mu_1, \mu_2, \cdots, \mu_c)$. For large $\{\mu_j\}$, the standardized values $\{z_i = (n_i - \mu_i)/\sqrt{\mu_i}\}$ have approximate standard normal distributions. Thuse $X^2 = \sum_{i=1}^{c} z_i^2$ has an approximate chi-squared distribution with c degrees of freedom. Adding the linear constraint $\sum_{i=1}^{n}(n_i - \mu_i) = 0$, we lose a degree of freedom.

3.3 Test for Independence and Homogeneity

The following are from

`http://inspire.stat.ucla.edu/unit_13`

- The "test of homogeneity" is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable. For example, do people of different races have the same proportion of smokers to non-smokers, or do different education levels have different proportions of Democrats, Republicans, and Independent. The test of homogeneity expands on the two-proportion z-test. The two proportion z-test is used when the response variable has only two categories as outcomes and we are comparing two groups. The homogeneity test is used if the response variable has several outcome categories, and we wish to compare two or more groups.

- The "test of independence" is a way of determining whether two categorical variables are associated with one another in the population, like race and smoking, or education level and political affiliation. In the probability unit we looked at this question without paying attention to the variability of our sample. Now we will have a method for deciding whether our observed P(A—B) is "too far" from our observed P(A) to conclude independence.

- If you're thinking, "homogeneity and independence sound the same!", you're nearly right. The difference is a matter of design. In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit. In the test of homogeneity, the data are collected by randomly sampling from each sub-group separately. (Say, 100 blacks, 100 whites, 100 American Indians, and so on.) The null hypothesis is that each sub-group shares the same distribution of another categorical variable. (Say, "chain smoker", "occasional smoker", "non-smoker".) The difference between these two tests is subtle yet important.

3.4 The Chi-Squared Test of Independence

Suppose that 200 students are selected at random from the entire enrollment at a large university, and each student in the sample is classified both according to his/her major and according to his/her preference for either of two candidates A and B in a forthcoming election. Suppose that the results are as presented in the following table:

	Biology	Engineering and Science	Social Science	Other	Totas
A	24	24	17	27	92
B	23	14	8	19	64
Undecided	12	10	13	9	44
Totals	59	48	38	55	200

Keep this table

In this table, each observation is classified in two ways. Such a table is called a two-way contigency table. One interesting question is whether the two classifications are independent. In other words, we may want to test the hypothesis that the major in which a student enrolled is independent of the candidate he/she prefers. Suppose a randomly selected student is of biology major, then the probability he/she choose candidate A is indepdent of the information of his/her major.

In general, an I-by-J table can be defined in the parameter space by specifying the cell probabilities:

π_{11}	π_{12}	$\cdots$	π_{1J}	$\pi_{1\cdot}$
π_{21}	π_{22}	$\cdots$	π_{2J}	$\pi_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
π_{I1}	π_{I2}	$\cdots$	π_{IJ}	$\pi_{I\cdot}$
$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\cdots$	$\pi_{\cdot J}$	1

where

- π_{ij} is the probability of being in row i and column j.
- $\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij}$ is the probability of being in row i (marignal probability).
- $\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij}$ is the probability of being in column j (marginal probability).
- $\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = \sum_{i=1}^{I} \pi_{i\cdot} = \sum_{j=1}^{J} \pi_{\cdot j} = 1$.

Suppose that a random sample of $n_{\cdot\cdot}$ subjects is taken from the given population. Let N_{ij} denote the number of subjects who are classified in the ith row and the jth column, $i = 1, \cdots, I$; $j = 1, \cdots, J$. Similarly to the margianl probabilities, we also define marginal counts $N_{i\cdot}$ and $N_{\cdot j}$. Below shows an observed two-way table with I rows and J columns,

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{I1}	n_{J2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

3.5 Derivation of the mles

Under no constraints, we can think the counts N_{ij} follows a multinomial distribution with cell probabilities π_{ij}, $i = 1, \cdots, I$; $j = 1, \cdots, J$. Therefore, the likelihood is

$$L_1 = L(\pi_{11}, \cdots, \pi_{IJ}) = \binom{n_{\cdot\cdot}}{n_{11}, \cdots, n_{IJ}} \pi_{11}^{n_{11}} \cdots \pi_{IJ}^{n_{IJ}} \propto \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}}$$

Under the assumption of independence, we assume whether a subject belongs to a row is independent of which column it belongs to. In statistics, the null hypothesis is

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \text{ for } i = 1, \cdots, I,\ j = 1, \cdots, J$$

Under the null hypothesis of independence,

$$L_0 = L(\pi_{1\cdot}, \cdots, \pi_{I\cdot}, \pi_{\cdot 1}, \cdots, \pi_{\cdot J}) \propto \prod_{i=1}^{I} \pi_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{J} \pi_{j\cdot}^{n_{\cdot j}}$$

Recall that $\sum_{i=1}^{I} \pi_{i\cdot} = \sum_{j=1}^{J} \pi_{\cdot j} = 1$, we have

$$\begin{aligned} l(\pi_{11}, \cdots, \pi_{IJ}) &= logL(\pi_{11}, \cdots, \pi_{IJ}) = \sum_{i=1}^{I} n_{i\cdot} log\pi_{i\cdot} + \sum_{j=1}^{J} n_{\cdot j} log\pi_{\cdot j} + c \\ &= \sum_{i=1}^{I-1} n_{i\cdot} log\pi_{i\cdot} + n_{I\cdot} log(1 - \sum_{i=1}^{I-1} \pi_{i\cdot}) + \sum_{j=1}^{J-1} n_{\cdot j} log\pi_{\cdot j} + n_{\cdot J} log(1 - \sum_{j=1}^{J-1} \pi_{\cdot j}) \end{aligned}$$

The mles of $\pi_{i\cdot}$ are solutions to

$$\frac{\partial l}{\partial \pi_i} = \frac{n_{i\cdot}}{\pi_{i\cdot}} - \frac{n_{I\cdot}}{1 - \sum_{i=1}^{I-1} \pi_{i\cdot}} = \frac{n_{i\cdot}}{\pi_{i\cdot}} - \frac{n_{I\cdot}}{\pi_{I\cdot}} = 0 \ , \ i = 1, \cdots, I-1$$

which implies $\hat{\pi}_{i\cdot} = \hat{\pi}_{I\cdot} n_{i\cdot}/n_{I\cdot}$. Because π_i are probabilities of a multinomial distribution,

$$1 = \sum_{i=1}^{I} \hat{\pi}_{i\cdot} = \sum_{i=1}^{I} \frac{\hat{\pi}_{I\cdot} n_{i\cdot}}{n_{I\cdot}} = \frac{\hat{\pi}_{I\cdot} n_{\cdot\cdot}}{n_{I\cdot}}$$

giving $\hat{\pi}_{I\cdot} = \frac{n_{I\cdot}}{n_{\cdot\cdot}}$. Substituting it back leads to

$$\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \text{ , for } i = 1, \cdots, I$$

Similarly,

$$\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}} \ , \ j = 1, \cdots, J$$

Therefore, under the null hypothesis of independence,

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

The Chi-square test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}}$$

When the null hypothesis is true, the test statistic X^2 has an approximate chi-squared distribution with df degrees of freedom. Here df equals the number of uniques parmaters under the full (unconstrained) model minus the number of unique parameters under the null (reduced, constrained) model.

Under the full model, we have IJ-1 unique probabilities, as the IJ probabilities add up to 1.

Under the reduced model, we have I-1 unique parameters for row marginal probabilities. This is because the I row marginal probabilities add up to 1. Similary, we have J-1 unique parameters for column marginal probabilities. In summary, there are I+J-2 unique parameters in the reduced model.

The difference is (IJ-1)-(I+J-2)=IJ-I-J+1=(I-1)(J-1).

Return to the example, the expected cell counts are:

	Biology	Engineering and Science	Social Science	Other	Totas
A	27.14	22.08	17.48	25.30	92
B	18.88	15.36	12.16	17.60	64
Undecided	12.98	10.56	8.36	12.10	44
Totals	59	48	38	55	200

The corresponding chi-square statistic is 6.68. Since $I = 3$ and $J = 4$, we compare the observed chi-square statistic to the chi-square distribution with $(3-1)\times(4-1) = 6$ degrees of freedom. Since the upper 5% point of χ_6 is 12.59, we do not reject the null hypothesis at level 0.05. So there is not enough evidence to support the dependence between major and candidate preference. We can also calculate the p-value: 1-pchisq(6.68,6)=0.35, from which we can draw the same conclusion.

3.6 The Chi-Square Test of Homogeneity

Materials for this section can be found in 9.4 of DeGroot.

3.6.1 Introduction

Back to the relationship between major and preference of candidate problem we discussed in the previous lecture. We now assume (pretend) that the data were not from a random sample of the population (the enrolments in a large university). They were obtained in the following way:

First, 59 students are selected at random from all enrolled students in the Biology major. Each of the 59 students is classified based on his/her prefence of candidates.

Second, 48 student are selected at random from students in the Engineering and Science major. Each of them is clasified according to his/her preference.

Third, 38 stduents are selected at random from students in the Social Science major. Each of them is classified according to his/her preference.

Last, 55 students are selected at random from students in other majors. Each of them is classified according to his/her preference.

Put preferences as rows and majors as columns, we have the following two-way table (it is the same as the one we saw last time!). We are interested in testing the hypothesis that, in all four populations, the same proportion of students prefers candidate A, the same proportion prefers candidate B, and the same proportion is undecided.

	Biology	Engineering and Science	Social Science	Other	Totas
A	24	24	17	27	92
B	23	14	8	19	64
Undecided	12	10	13	9	44
Totals	59	48	38	55	200

Suppose we have observations from J multinomial distriubtions. Each multinomial distribution consists of I categories. Our purpose is to test whether the J multinomial distributions are the same or not. In other words, we want to know whether or not the observations are from a same multinomial distriubtion.

The observed counts can also be arranged using the following two-way table with I rows and J columns,

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{I1}	n_{J2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

An I-by-J table can be defined in the parameter space by specifying the cell probabilities:

Pop_1	Pop_2	$\cdots$	Pop_J
p_{11}	p_{12}	$\cdots$	p_{1J}
p_{21}	p_{22}	$\cdots$	p_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_{I1}	p_{I2}	$\cdots$	p_{IJ}
1	1	$\cdots$	1

where

- p_{ij} is the probability that an observation chosen at random from the jth population will be of type i.
- $\sum_{i=1}^{I} p_{ij} = 1$, for $j = 1, \cdots, J$.

The jth column $(n_{1j}, n_{2j}, \cdots, n_{Ij})$ denotes the observations from the jth population with a multinomial distribution and parameters $((p_{1j}, p_{2j}, \cdots, p_{Ij}))$. Under the null hypothesis, the distributions for all the J columns are the same. The parameter table under the null hypothesis looks like:

Pop_1	Pop_2	$\cdots$	Pop_J
$p_{11} = p_1$	$p_{12} = p_1$	$\cdots$	$p_{1J} = p_1$
$p_{21} = p_2$	$p_{22} = p_2$	$\cdots$	$p_{2J} = p_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p_{I1} = p_I$	$p_{I2} = p_I$	$\cdots$	$p_{IJ} = p_J$
1	1	$\cdots$	1

The null hypothesis is

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{iJ} = p_i \text{ for } i = 1, \cdots, I$$

3.6.2 MLE under Homogeneity

.

Under $H_0 : p_{i1} = p_{i2} = \cdots = p_{iJ} = p_i$ for $i = 1, \cdots, I$ the likelihood is given by

$$\begin{aligned} L(p_1, \cdots, p_I) &= \prod_{j=1}^{J} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}} p_{1j}^{n_{1j}} p_{2j}^{n_{2j}} \cdots p_{Ij}^{n_{Ij}} \\ &= \prod_{j=1}^{J} \binom{n_{\cdot j}}{n_{1j} n_{2j} \cdots n_{Ij}} p_1^{n_{1j}} p_2^{n_{2j}} \cdots p_I^{n_{Ij}} \text{ under } H_0 \\ &\propto p_1^{n_{1\cdot}} p_2^{n_{2\cdot}} \cdots p_I^{n_{I\cdot}} \end{aligned}$$

With the constraint $\sum_{i=1}^{I} p_i = 1$, we have

$$L(p_1, \cdots, p_I) \propto p_1^{n_{1\cdot}} p_2^{n_{2\cdot}} \cdots (1 - \sum_{i=1}^{I-1} p_i)^{n_{I\cdot}}$$

The loglikelihood

$$\log L(p_1, \cdots, p_I) = c + n_{1\cdot} log(p_1) + \cdots + n_{(I-1)\cdot} log(p_{I-1}) + n_{I\cdot} log(1 - \sum_{i=1}^{I-1} \pi_i)$$

The mles are solutions to

$$\frac{\partial l}{\partial p_i} = \frac{n_{i\cdot}}{p_i} - \frac{nI\cdot}{1 - \sum_{i=1}^{I-1} p_i} = \frac{n_{1\cdot}}{p_i} - \frac{n_{I\cdot}}{p_I} = 0 \ , \ i = 1, \cdots, I$$

which implies $\hat{p}_i = \hat{p}_I n_{i\cdot}/n_{I\cdot}$. Because p_i are probabilities of a multinomial distribution,

$$1 = \sum_{i=1}^{I} \hat{p}_i = \sum_{i=1}^{I} \frac{\hat{p}_I n_{i\cdot}}{n_{I\cdot}} = \frac{\hat{p}_I n_{\cdot\cdot}}{n_{I\cdot}}$$

giving $\hat{p}_I = \frac{nI\cdot}{n_{\cdot\cdot}}$. Substituting it back leads to

$$\hat{p}_i = \frac{ni\cdot}{n\cdot\cdot} \ , \text{ for } i = 1, \cdots, I$$

Therefore, under the null hypothesis of homogeneity,

$$Exp_{ij} = n_{\cdot j} \hat{p}_i = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

The Chi-squared test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}}$$

Under the null hypothesis, X^2 follows $\chi^2_{(I-1)(I-1)}$. Here is the justification of the df. In the full parameter space, we use $(I-1)$ parameters for each population; since there are J populations, there are $(I-1)J$ unique parameters. In the reduced model (under H_0), there are $(I-1)$ unique parameters. df equals the difference, which is $(I-1)J - (I-1) = (I-1)(J-1)$.

4 Matched-Pairs Designs

4.1 Introduction: Two Examples

4.1.1 Example 1 (Rice)

Johnson and Johnson (1971) selected 85 Hodgkin's patients who had a sibling of the same sex who was free of the disease and whose age was within 5 years of the patient's. These investigators presented the following table:

	Tonsillectomy	No Tonsillectomy
Hodgkin's	41	44
Control	33	52

They wanted to know whether the tonsil act as a protective barrier against Hodgkin's disease. The Pearson's chi-squared statistic was calculated: 1.53 (p-value 0.22), which is not significant and they they concluded that the tonsil is not a protector against Hodgkin's disease. Any problme with this? Shortly after their result was published, seveal letters to the editor pointed out that those investigators had made an error in their analysis by ignoring the pairings. The assumption behind the chi-squared test of homogeneity/independence is that independent multinomial samples are compared, and the samples in Johnson and Johnson were not, because siblings were paired.

4.1.2 Example 2 (DeGroot)

Suppose that 100 persons were selected at random in a certain city, and that each person was asked whether he/she thought the service provided by the fire department in the city was satisfactory. Shortly after this survey was carried out, a large fire occured in the city. Suppose that after this fire, the same 100 persons were again asked whether they thought that the service provided by the fire department was satisfactory. The results are presented in the table below:

	Satifactory	Unsatisfactor
Before the fire	80	20
After the fire	72	28

Suppose we want to know whether people's opion was changed after the fire, how should we analyze the data? You may want to consider a test of homogeneity using a chi-square test. You apply the chi-square test for homogeneity and obtain a chi-square statistic 1.75 and the corresponding p-value 0.19. However, it would not be appropriate to do so for this table because the observations taken before the fire and the observations taken after the fire are not independent. Although the total number of observations in the table is 200, only 100 independtly chosen persons were questioned in the surveys. It is reasonable to believe that a particular person's opinion before the fire and after the fire are dependent.

4.2 The Proper Way to Display Correlated Tables

To take the pairing/correlation nature of data into consideration, the data in the two examples should be displayed in a way that exhibits the pairing.

4.2.1 Example 1

		Sibling No Tonsillectomy	Sibling Tonsillectomy
Patient	No Tonsillectomy	37	7
Patient	Tonsillectomy	15	26

4.2.2 Example 2

		After the fire Satisfactory	Unsatisfactory
Before the fire	Satisfactory	70	10
Before the fire	Unsatisfactory	2	18

4.3 Data Analysis

With the approproiate presentation, the data are a sample of size n from a multinomial distribution with four cells. We can represent the probabilities in the tables as follows:

π_{11}	π_{12}	$\pi_{1\cdot}$
π_{21}	π_{22}	$\pi_{2\cdot}$
$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

The appropriate null hypothesis states that the probabilities of tonsillectomy and no tonsillectomy are the same for patients and siblings - that is, $\pi_{1\cdot} = \pi_{\cdot 1}$ and $\pi_{2\cdot} = \pi_{\cdot 2}$, or

$$\begin{aligned} \pi_{11} + \pi_{12} &= \pi_{11} + \pi_{21} \\ \pi_{12} + \pi_{22} &= \pi_{21} + \pi_{22} \end{aligned}$$

These equations simplify to

$$H_0 : \pi_{12} = \pi_{21}$$

Under the null hypothesis, the off-diagnoal probabilities are equal, and under the alternative hypothesis they are not. Under the nul, it can be shown that the mle's of the cell probabilities are

$$\begin{aligned} \hat{\pi}_{11} &= \frac{n_{11}}{n} \\ \hat{\pi}_{22} &= \frac{n_{22}}{n} \\ \hat{\pi}_{12} &= \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n} \end{aligned}$$

The test-statistic is

$$\begin{aligned} X^2 &= \frac{(n_{11}-n_{11})^2}{n_{11}} + \frac{(n_{12}-(n_{12}+n_{21})/2)^2}{(n_{11}+n_{21})/2} + \frac{(n_{21}-(n_{12}+n_{21})/2)^2}{(n_{11}+n_{21})/2} + \frac{(n_{22}-n_{22})^2}{n_{22}} \\ &= \frac{(n_{12}-(n_{12}+n_{21})/2)^2}{(n_{11}+n_{21})/2} + \frac{(n_{21}-(n_{12}+n_{21})/2)^2}{(n_{11}+n_{21})/2} + \frac{(n_{22}-n_{22})^2}{n_{22}} \\ &= \frac{(n_{12}-n_{21})^2}{n_{21}+n_{21}} \end{aligned}$$

Since there are three unique parameters in the full parameter space and there are two in the reduced, the null distribution of X^2 is χ^2_1.

It is obvious that the contributions to the chi-squared statistic from n_{11} and n_{22} are zero. This is becuase the diagonal probabilities do not distinguish the null and alternative hypotheses. Under the alternative, there are three free parameters. Under the null, there is the additional constraint $\pi_{12} = \pi_{21}$, and there are two free paramters. Therefore, the chi-square statistic has one degree of freddom. This test is called **McNemar's test**.

For example 1, $X^2 = (15-7)^2/(15+7) = 2.91$; the corresponding p-value is 0.09. We fail to reject the null hypothesis. There is not evidence that tonsillectomy changes the risk of Hodgkin's disease.

For example 2, $X^2 = (10-2)^2/(10+2) = 5.33$; the corresponding p-value si 0.02. We reject the null hypothesis and conclude that the opinions of the residents changed by the fire.

5 Measurements of Association: Odds Ratio

5.1 Odds

If an event A has probability $P(A)$ of occuring, the **Odds** of A is the ratio of $P(A)$ to $P(A^c) = 1 - P(A)$:

$$odds(A) = \frac{P(A)}{1 - P(A)}$$

Based on this definition,

$$P(A) = \frac{odds(A)}{1 + Odds(A)}$$

5.2 Odds Ratio

Let X denote the event that an individual is exposed to a potentially harmful agent and D denote the event that the individual becomes diseased. We denote the complementary events as X^c and D^c. The odds of an individual contracting the disease given that he is exposed is

$$odds(D|X) = \frac{P(D|X)}{1 - P(D|X)}$$

and the odds of contracting the disease given that he is not exposed is

$$odds(D|X^c) = \frac{P(D|\bar{X})}{1 - P(D|X^c)}$$

The **odds ratio**

$$\delta = \frac{odds(D|X)}{odds(D|X^c)}$$

is a measure of the influence of exposure on subsequent disease. How can the odds and odds ratio be estimated? Suppose we have a sample from a population with joint and marginal probabilities defined as in the following table:

	D^c	D	
X^c	π_{00}	π_{01}	$\pi_{0\cdot}$
X	π_{10}	π_{11}	$\pi_{1\cdot}$
	$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	1

With this notation,

$$\begin{aligned} P(D|X) &= \frac{\pi_{11}}{\pi_{10} + \pi_{11}} \\ P(D|X^c) &= \frac{\pi_{01}}{\pi_{00} + \pi_{01}} \end{aligned}$$

so that

$$\begin{aligned} odds(D|X) &= \frac{\pi_{11}}{\pi_{10}} \\ odds(D|X^c) &= \frac{\pi_{01}}{\pi_{00}} \end{aligned}$$

and the odds ratio is

$$\delta = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

With a sample from a population, we can esitmate π_{ij} by $\hat{\pi}_{ij}$, and δ can be estimated by

$$\hat{\delta} = \frac{\hat{\pi}_{11}\hat{\pi}_{00}}{\hat{\pi}_{01}\hat{\pi}_{10}}$$

So fare we have assumed that data were obtained as a random sample from a population. In practice, we may consider more efficient methods of sampling, such as prospective and retrospective.

In a prospective study, a fixed number of exposed and nonexposed individuals are sampled, and the incidences of diseases in those two groups are compared. In this situation, we can estimate $P(D|X)$ and $P(D|X^c)$, which allow use to estimate the odds ratio.

In a retrospective study, a fixed number of diseased and undiseased individuals are sampled and the incidences of exposure in the two groups are compared. In this situation, we can estimate $P(X|D)$, thus $odds(X|D)$ and $P(X|D^c)$, thus $odd(X|D^c)$. It can shown that

$$\delta = \frac{odds(D|X)}{odds(D|X^c)} = \frac{odds(X|D)}{odds(X|D^c)}$$

Therefore, we are still be able to estimate δ.

In summary, if we have the counts from based on any of the three studies

	D^c	D
X^c	n_{00}	n_{01}
X	n_{10}	n_{11}
	$n_{\cdot 0}$	$n_{\cdot 1}$

The estimate of the odds ratio is

$$\hat{\delta} = \frac{n_{00}n_{11}}{n_{01}n_{10}}$$