

STA4030: Categorical Data Analysis

Two-way Tables: Tests for Independence and Homogeneity

Instructor: Bojun Lu

School of Science and Engineering
CUHK(SZ)

September 22, 2020

Agenda

- 1 3.1 Two-way Tables
- 2 3.2 Sampling Models
- 3 3.3 Test of Independence and Test of Homogeneity
- 4 3.4 Fisher's Exact Test

3.1 Two-way Tables

3.1.1 Examples

“Having been given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things”

M.H. Doolittle, (1887)

3.1 Two-way Tables

(From Exercise 2.4 in Agresti (2013)) All the auto accidents in Florida in 2008 were classified according to whether they were fatal and whether the person involved was wearing a seat-belt.

Seat-Belt Use	Injury	
	Fatal	Non-fatal
No	1085	55623
Yes	703	441239

Source: Florida Department of Highway Safety and Motor Vehicles

3.1 Two-way Tables

Five hundred residents of Massachusetts are asked their political affiliation and their opinion on some political statement. The results are

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

3.1 Two-way Tables

The Physician's Health Study was a 5-year randomized, blind study to see if taking aspirin regularly reduces the mortality from heart disease. Every other day, physicians taking part in the study either took one aspirin or one placebo. The results were

	Myocardial Infarction			Total
	Fatal Attack	Nonfatal Attack	No Attack	
Placebo	18	171	10845	11034
Aspirin	5	99	10933	11037

Source: Preliminary Report: Findings from the aspirin component of the ongoing Physician's Health Study. N. Engl. J. Med. 318:262-264, 1988.

3.1 Two-way Tables

A British woman claims to be able to tell if a cup of tea with milk has had the milk poured first or the tea poured first. To test this, she is presented with 8 cups of tea, with 4 cups of each type in random order. She tastes each cup and guesses what was poured first for each. The results were

Poured First \ Guess Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

3.1 Two-way Tables

3.1.2 Response and explanatory variables

Two-way tables involve two categorical variables, X with r categories and Y with c . Therefore, there are rc possible combinations. (Traditionally, X is the row variable and Y the column variable.)

Sometimes, both X and Y will be *response* variables, in which case it makes sense to talk about their joint distribution.

On other occasions, Y will be the response variable and X will be the *explanatory* variable. In this case, it does not make sense to talk about the joint distribution of X and Y . Instead, we focus on the conditional distribution of Y given X . That is, how does Y change as X changes.

3.1 Two-way Tables

Consider this example: we sample 100 students at random outside University station and categorise them by their gender and whether they like beer. The results are

	Likes Beer	Does Not Like Beer
Male	45	5
Female	15	35

By chance, we have observed equal numbers of male and female students. For this study, it makes sense to talk about the joint distribution of the counts in the four cells. Three of entries are free and the last must be 100 minus the sum of the rest. It is essentially and one-way table, written in two dimensions.

3.1 Two-way Tables

Now say we sample 50 female students and 50 male students outside University station and ask them if they like beer. The results are

	Likes Beer	Does Not Like Beer
Male	45	5
Female	15	35

In this study, the entries for the table are not as free. In each row, one entry determines the entire row. A joint distribution of the four counts doesn't make sense here. It makes sense to focus on the distribution of each row. It is essentially two one-way tables put together.

3.1 Two-way Tables

3.1.3 Notation

n : total number of observations (sample size).

n_{ij} : number of observations in row i and column j .

$p_{ij} = n_{ij}/n$: proportion of the total sample falling in the (i,j) -th cell. $\sum_i \sum_j p_{ij} = 1$.

Sample joint distribution: the set $\{p_{ij}\}$.

Sample marginal distribution: the set $\{p_{i+}\}$ and the set $\{p_{+j}\}$.

Sample conditional distribution: the set $\{p_{j(i)}\}$ or the set $\{p_{i(j)}\}$.

3.1 Two-way Tables

Let X and Y denote two categorical variables with r and c categories, respectively. Classifications of subjects on both variables have $r \times c$ possible combinations.

When the cells contain frequency counts of outcomes for a sample, the table is called a contingency table or a cross-classification table. An $r \times c$ contingency table is

$X \backslash Y$	1	2	\dots	c	Total
1	n_{11}	n_{12}	\dots	n_{1c}	n_{1+}
2	n_{21}	n_{22}	\dots	n_{2c}	n_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r	n_{r1}	n_{r2}	\dots	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	\dots	n_{+c}	n

3.1 Two-way Tables

Florida accident data: sample joint distribution:

	Fatal	Non-fatal	Total
No	$p_{11} = \frac{1085}{498650}$	$p_{12} = \frac{55623}{498650}$	$p_{1+} = p_{11} + p_{12} = \frac{56708}{498650}$
Yes	$p_{21} = \frac{703}{498650}$	$p_{22} = \frac{441239}{498650}$	$p_{2+} = p_{21} + p_{22} = \frac{441942}{498650}$
Total	$p_{+1} = \frac{1788}{498650}$	$p_{+2} = \frac{496862}{498650}$	1.0

Marginal distribution:

The set $\{p_{i+}\}$ for the row variable.

The set $\{p_{+j}\}$ for the column variable.

3.1 Two-way Tables

Physician's Health Study: sample conditional distribution:

Explanatory variable: Placebo or Aspirin ($i = 1, 2$)

Response variable: Heart Disease ($j = 1, 2, 3$)

Sample conditional distribution: $p_{j(i)}$

	Fatal	Non-fatal	None
Placebo	$p_{1(1)} = \frac{18}{11034}$	$p_{2(1)} = \frac{171}{11034}$	$p_{3(1)} = \frac{10845}{11034}$
Aspirin	$p_{1(2)} = \frac{5}{11037}$	$p_{2(2)} = \frac{99}{11037}$	$p_{3(2)} = \frac{10933}{11037}$

3.1 Two-way Tables

π_{ij} : probability of an observation falling in the (i, j) -th cell.

Population joint distribution: the set $\{\pi_{ij}\}$, $\sum_i \sum_j \pi_{ij} = 1$.

Population marginal distribution:

the set $\{\pi_{i+}\}$ for the row variable, $\sum_j \pi_{i+} = 1$.

the set $\{\pi_{+j}\}$ for the column variable, $\sum_i \pi_{+j} = 1$.

Population conditional distribution:

If Y is a response variable, and X is an explanatory variable,
 $\pi_{j(i)}$ denotes the probability of falling in level j of the response variable given level i of the explanatory variable,

$$\sum_{j=1}^c \pi_{j(i)} = 1, \quad \text{for } i = 1, \dots, r$$

3.2 Sampling Models

3.2.1 Poisson Sampling

As before with one-way tables, if the total sample size n is completely random and so are both margins of the table, it is common to treat each cell as an independent Poisson variable.

Each cell frequency n_{ij} has an independent Poisson distribution with mean μ_{ij} . The joint probability mass function for this model is

$$\prod_{i,j} \frac{\mu_{ij}^{n_{ij}} e^{-\mu_{ij}}}{n_{ij}!}$$

3.2 Sampling Models

Furthermore, if n is not of interest, data from Poisson Sampling may be treated as if it were from Multinomial Sampling, by conditioning on the value of n , because the inferences for π are still valid. (Recall Slide 18 of the “Preliminaries” notes and Question 4 of “Exercises for Week 1”.)

The Florida auto accident study is an example where Poisson sampling could be assumed. The time frame in which accidents were observed was fixed, but the number of accidents observed was not. If our focus is not the *number* of accidents but the *proportions* of accidents in the various categories, we could treat the sample as being from a multinomial model, with sample size $n = 498650$.

3.2 Sampling Models

3.2.2 Multinomial Sampling

If the total sample size n is fixed and each element of the sample is classified according to two categories X and Y , X with r categories and Y with c , the joint distribution of the n_{ij} is $Mult(n, \pi)$, with $\pi = (\pi_{11}, \pi_{12}, \dots, \pi_{1c}, \pi_{21}, \dots, \pi_{rc})$.

The probability mass function of the cell counts is

$$\frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}}.$$

The study of 500 residents of Massachusetts, categorised according to their political affiliation and opinion on some political statement is an example of multinomial sampling.

3.2 Sampling Models

3.2.3 Product Multinomial Sampling

Often, observations on a response Y occur separately at each setting of X . Treating row totals as fixed, and using the notation n_{i+} , suppose that n_{i+} observations on Y at setting i of X are independent, each with probability distribution $\{\pi_{1(i)}, \dots, \pi_{c(i)}\}$. The counts $\{n_{ij}, j = 1, \dots, c\}$ satisfy $\sum_j n_{ij} = n_{i+}$, and have the multinomial form:

$Y X = i$	1	2	\dots	c	Total
n_{ij}	n_{i1}	n_{i2}	\dots	n_{ic}	n_{i+}
$\pi_{j(i)}$	$\pi_{1(i)}$	$\pi_{2(i)}$	\dots	$\pi_{c(i)}$	π_{i+}

$$, \quad \frac{n_{i+}!}{\prod_{j=1}^c n_{ij}!} \prod_{j=1}^c \pi_{j(i)}^{n_{ij}}$$

Product Multinomial Sampling is usually the sampling scheme when one of the variables is the response and the other is the explanatory variable.

3.2 Sampling Models

When samples at different settings of X are independent, the joint probability function for the entire data set is the product of the above multinomial functions from the various settings, i.e.

$$\prod_{i=1}^r \left(\frac{n_{i+}!}{\prod_{j=1}^c n_{ij}!} \prod_{j=1}^c \pi_{j(i)}^{n_{ij}} \right)$$

This sampling scheme is also called independent multinomial sampling.

The Physician's Health Study appears to be an example of independent multinomial sampling - the size of the Placebo and Aspirin groups were chosen to be very similar.

3.2 Sampling Models

3.2.4 Hypergeometric Sampling

Occasionally, we encounter studies in which both marginal totals (row and column) of the contingency table are fixed by design. Consider the case where $r = c = 2$. Then a single entry in the two-way table determines all the rest.

The situation is analogous to the following. We have N balls, of which K are red and the rest are blue. We draw a sample of n balls *without replacement*. What is the probability that our sample contains k red balls?

3.2 Sampling Models

For our sample of n balls to have k reds, we must choose k reds from the K available. Likewise, we must choose $n - k$ blues from the $N - K$ available. Moreover, there are $\binom{N}{n}$ possible sample of size n from the population of N . The probability of us choosing k reds in a sample of n is therefore

$$p(k; N, K, n) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \text{ for } k = 0, 1, \dots, n.$$

This is the hypergeometric distribution. (When the table is larger than 2×2 , we have the multivariate hypergeometric distribution.)

3.2 Sampling Models

The “Lady Tasting Tea” study is an (very famous) example of hypergeometric sampling. Say a success (picking a red ball) is the lady correctly identifying a cup of tea with milk poured first. Then

N = # cups of tea = 8;

K = # cups of tea with milk poured first = 4;

n = # sample size = 4;

k = # correct guesses by the lady, $k \in \{0, 1, 2, 3, 4\}$.

The two-way table for this study has the hypergeometric distribution if the lady is essentially guessing (drawing a random sample) whether the cups of tea have had milk poured first or not. If she genuinely has skill, the distribution will be different.

3.3 Test of Independence and Test of Homogeneity

3.3.1 Definitions

A principal aim of many studies is to compare conditional distributions of Y at various levels of explanatory variables.

When *both variables are response* variables, descriptions of the association can use their joint distribution, marginal distributions, and conditional distribution of Y given X , or conditional distribution of X given Y .

Independence, therefore, would mean that for a member of the population $P(X = i, Y = j) = P(X = i)P(Y = j)$, or in terms of the population parameters

$$\pi_{ij} = \pi_{i+}\pi_{+j} \text{ for } i = 1, \dots, r, j = 1, \dots, c$$

3.3 Test of Independence and Test of Homogeneity

When Y is a response and X is an explanatory variable (for example, in a product multinomial design), we are cautious about using a joint distribution π_{ij} to define “independence”.

What does “independence” mean when we have Y as a response and X an explanatory variable? We basically want for Y not to change behaviour when we change the category of X . That is, as we move rows, the *conditional* distribution of Y does not change. In terms of population parameters, this means

$$\pi_{j(1)} = \pi_{j(2)} = \cdots = \pi_{j(r)} \text{ for } j = 1, \dots, c.$$

3.3 Test of Independence and Test of Homogeneity

3.3.2 Test of Independence

Observed frequencies:

$X \backslash Y$	1	2	\dots	c	Total
1	n_{11}	n_{12}	\dots	n_{1c}	n_{1+}
2	n_{21}	n_{22}	\dots	n_{2c}	n_{2+}
\vdots	\vdots	\vdots		\vdots	\vdots
r	n_{r1}	n_{r2}	\dots	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	\dots	n_{+c}	n

Sampling model: Multinomial model with size n and $r \times c$ categories; X and Y are responses.

3.3 Test of Independence and Test of Homogeneity

$X \backslash Y$	1	2	\dots	c	Total
1	π_{11}	π_{12}	\dots	π_{1c}	π_{1+}
2	π_{21}	π_{22}	\dots	π_{2c}	π_{2+}
\vdots	\vdots	\vdots		\vdots	\vdots
r	π_{r1}	π_{r2}	\dots	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	\dots	π_{+c}	1

$$\sum_i \sum_j \pi_{ij} = \sum_i \pi_{i+} = \sum_j \pi_{+j} = 1.$$

3.3 Test of Independence and Test of Homogeneity

We wish to test if the data are independent or not. We can frame this hypothesis in terms of the parameters:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \text{ for } i = 1, \dots, r, j = 1, \dots, c,$$

that is, the parameters π_{ij} can be completely described by parameters π_{i+} , $i = 1, \dots, r$ and π_{+j} , $j = 1, \dots, c$, which are constrained by $\sum_i \pi_{i+} = 1$ and $\sum_j \pi_{+j} = 1$.

As such, testing this null hypothesis of independence is a matter of testing a model of the multinomial parameters. Therefore, both the Likelihood Ratio Test and Pearson's Chi-squared test can be used.

3.3 Test of Independence and Test of Homogeneity

We find MLEs of π_{i+} and π_{+j} under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$.

Under H_0 , the log-likelihood function is

$$\begin{aligned} l(\pi) &= \sum_i \sum_j n_{ij} \log \pi_{ij} = \sum_i \sum_j n_{ij} (\log \pi_{i+} + \log \pi_{+j}) \\ &= \sum_i \left(\sum_j n_{ij} \right) \log \pi_{i+} + \sum_j \left(\sum_i n_{ij} \right) \log \pi_{+j} \\ &= \sum_i n_{i+} \log \pi_{i+} + \sum_j n_{+j} \log \pi_{+j}. \end{aligned}$$

Since

$$\sum_i \pi_{i+} = 1 \Rightarrow \frac{\partial \log \pi_{r+}}{\partial \pi_{i+}} = -1 \text{ and } \frac{\partial \log \pi_{r+}}{\partial \pi_{i+}} = -\frac{1}{\pi_{r+}}, \quad i = 1, \dots, r-1,$$

$$\sum_j \pi_{+j} = 1 \Rightarrow \frac{\partial \log \pi_{+c}}{\partial \pi_{+j}} = -1 \text{ and } \frac{\partial \log \pi_{+c}}{\partial \pi_{+j}} = -\frac{1}{\pi_{+c}}, \quad j = 1, \dots, c-1,$$

3.3 Test of Independence and Test of Homogeneity

then

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_{i+}} = \frac{n_{i+}}{\pi_{i+}} - \frac{n_{r+}}{\pi_{r+}} = 0 \Rightarrow \pi_{i+} = \frac{n_{i+}}{n_{r+}} \pi_{r+},$$

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_{+j}} = \frac{n_{+j}}{\pi_{+j}} - \frac{n_{+c}}{\pi_{+c}} = 0 \Rightarrow \pi_{+j} = \frac{n_{+j}}{n_{+c}} \pi_{+c}.$$

Summing both sides of $\pi_{i+} = (n_{i+}/n_{r+})\pi_{r+}$, we have

$$1 = \sum_i \pi_{i+} = \sum_i \frac{n_{i+}}{n_{r+}} \pi_{r+} = \frac{n}{n_{r+}} \pi_{r+} \Rightarrow \hat{\pi}_{r+} = \frac{n_{r+}}{n}.$$

So,

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n_{r+}} \hat{\pi}_{r+} = \frac{n_{i+}}{n_{r+}} \times \frac{n_{r+}}{n} = \frac{n_{i+}}{n}, \quad i = 1, \dots, r.$$

Similarly,

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n}, \quad j = 1, \dots, c.$$

3.3 Test of Independence and Test of Homogeneity

Under H_0 , the MLEs of π_{i+} and π_{+j} are

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n}, \quad \hat{\pi}_{+j} = \frac{n_{+j}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

Thus, the estimated expected frequencies are

$$E_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}.$$

The LR test statistic is

$$G^2 = 2 \sum_{\text{cells}} O_{ij} \log \left(\frac{O_{ij}}{E_{ij}} \right) = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \log \left(\frac{n_{ij}n}{n_{i+}n_{+j}} \right)$$

and Pearson's Chi-squared test statistic is

$$X^2 = \sum_{\text{cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n} \right)^2}{\frac{n_{i+}n_{+j}}{n}}.$$

3.3 Test of Independence and Test of Homogeneity

Both G^2 and X^2 have the χ^2_{df} distribution when n is large, where $df = \# \text{ parameters estimated under } H_0 \cup H_1 - \# \text{ parameters estimated under } H_0$.

Under H_0 , we estimated $\pi_{1+}, \dots, \pi_{r+}$ and $\pi_{+1}, \dots, \pi_{+c}$, with the constraints $\sum_i \pi_{i+} = 1$ and $\sum_j \pi_{+j} = 1$. Under $H_0 \cup H_1$, we estimated $\pi_{ij}, i = 1, \dots, r, j = 1, \dots, c$ with the constraint $\sum_{i,j} \pi_{ij} = 1$. Therefore

$$df = rc - 1 - ((r - 1) + (c - 1)) = (r - 1)(c - 1).$$

The large sample approximation (i.e. the appeal to the Chi-squared distribution) requires that n is large. The rule of thumb is the same as for one-way tables: we can have $E_{ij} < 5$ for at most 20% of the cells, but none of the E_{ij} s should be smaller than 1.

3.3 Test of Independence and Test of Homogeneity

3.3.3 Example: Test of Independence

Recall the study of five hundred residents of Massachusetts and their political opinion and affiliation:

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

We test H_0 : the political affiliation and political opinion of a Massachusetts resident are independent. The test statistics are: $G^2 = 22.339$, $X^2 = 22.152$. For both, the P -value is less than 0.0001. We reject the null hypothesis that political affiliation and political opinion are independent. (Unsurprisingly.)

3.3 Test of Independence and Test of Homogeneity

3.3.4 Test of Homogeneity

Observed frequencies:

$X \backslash Y$	1	2	\dots	c	Total
1	n_{11}	n_{12}	\dots	n_{1c}	n_{1+}
2	n_{21}	n_{22}	\dots	n_{2c}	n_{2+}
\vdots	\vdots	\vdots		\vdots	\vdots
r	n_{r1}	n_{r2}	\dots	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	\dots	n_{+c}	n

Sampling model: Product multinomial model; $n_{1+}, n_{2+}, \dots, n_{r+}$ fixed in advance; Y is response, X is explanatory.

3.3 Test of Independence and Test of Homogeneity

Population distribution:

$X \backslash Y$	1	2	\dots	c	Total
1	$\pi_{1(1)}$	$\pi_{2(1)}$	\dots	$\pi_{c(1)}$	1.0
2	$\pi_{1(2)}$	$\pi_{2(2)}$	\dots	$\pi_{c(2)}$	1.0
\vdots	\vdots	\vdots		\vdots	\vdots
r	$\pi_{1(r)}$	$\pi_{2(r)}$	\dots	$\pi_{c(r)}$	1.0

Note: $\sum_{j=1}^c \pi_{j(i)} = 1, \forall i = 1, \dots, r$

3.3 Test of Independence and Test of Homogeneity

We wish to test if the data are independent or not, which in this sampling scheme means whether the conditional distribution of Y is the same for each level of X . We can frame this hypothesis in terms of the parameters:

$$H_0 : \pi_{j(1)} = \pi_{j(2)} = \cdots = \pi_{j(r)} = \pi_j, j = 1, \dots, c$$

that is, the conditional probabilities $\pi_{j(i)}$ can be completely described by parameters $\pi_j, j = 1, \dots, c$, constrained by $\sum_j \pi_j = 1$.

Thus testing the null hypothesis is a matter of testing a model of the product multinomial parameters. Therefore, both the LR test and Pearson's Chi-squared test can be used. This test is called the Test of Homogeneity.

3.3 Test of Independence and Test of Homogeneity

We find MLEs of π_j under H_0 . Under H_0 , the log-likelihood function is

$$l(\boldsymbol{\pi}) = \sum_{i=1}^r \sum_{j=1}^c n_{ij} \log \pi_j = \sum_{j=1}^c \left(\sum_{i=1}^r n_{ij} \right) \log \pi_j = \sum_{j=1}^c n_{+j} \log \pi_j.$$

Since

$$\sum_{j=1}^c \pi_j = 1 \Rightarrow \frac{\partial \log \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c}, \text{ for } j = 1, \dots, c-1,$$

from

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = \mathbf{0} \Rightarrow \frac{n_{+j}}{\pi_j} - \frac{n_{+c}}{\pi_c} = 0, \quad \text{or} \quad \pi_j = \frac{n_{+j}}{n_{+c}} \pi_c.$$

3.3 Test of Independence and Test of Homogeneity

Summing both sides of the last equation, we have

$$1 = \sum_{j=1}^c \pi_j = \frac{\sum_{j=1}^c n_{+j}}{n_{+c}} \pi_c = \frac{n}{n_{+c}} \pi_c,$$

So,

$$\hat{\pi}_c = \frac{n_{+c}}{n}, \quad \text{and} \quad \hat{\pi}_j = \frac{n_{+j}}{n_{+c}} \hat{\pi}_c = \frac{n_{+j}}{n}.$$

Thus the estimated expected frequencies are

$$E_{ij} = n_{i+} \hat{\pi}_{j(i)} = n_{i+} \hat{\pi}_j = \frac{n_{i+} n_{+j}}{n}.$$

3.3 Test of Independence and Test of Homogeneity

We therefore have the rather remarkable result that no matter we are conducting a test of independence or a test of homogeneity, we calculate the LR test statistics with the same formula. The same is true, of course, for the Pearson's chi-squared test statistic.

Moreover, since under $H_0 \cup H_1$ there are $r \times (c - 1)$ parameters to estimate and under H_0 we estimated $c - 1$, the test statistics for the test of homogeneity have

$$df = r(c - 1) - (c - 1) = (r - 1)(c - 1)$$

degrees of freedom. Just like for the test of independence. So, when n is large, X^2 and G^2 follow the χ^2_{df} distribution. We use the same rule of thumb for deciding whether n is large enough.

3.3 Test of Independence and Test of Homogeneity

3.3.5 Example: Test of Homogeneity

Recall the Physician's Health Study:

	Myocardial Infarction			Total
	Fatal Attack	Nonfatal Attack	No Attack	
Placebo	18	171	10845	11034
Aspirin	5	99	10933	11037

We test H_0 : taking aspirin has no more affect on heart disease than taking a placebo. The test statistics are $G^2 = 27.589$ and $X^2 = 26.903$. For both, the P -value is less than 0.0001. We reject the null hypothesis that aspirin and a placebo have the same affect on heart disease.

3.4 Fisher's Exact Test

3.4.1 The Lady Tasting Tea

Recall the experiment with the lady who thought she could tell whether a cup of tea has had milk or tea poured first. The results were

Poured First \ Guess Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Since both marginal totals (rows and columns) are fixed, the LR and Pearson Chi-squared tests are not appropriate for this data. Moreover, the sample size is small, so an appeal to large sample approximations should not be made. We shall have to use the *exact* distribution instead.

3.4 Fisher's Exact Test

Fortunately we know what the exact distribution of the table is, under the hypothesis that the lady is guessing her answers - it is the hypergeometric.

Treating the $(1, 1)$ th entry as the variable, we calculate that the probability of observing this table under “ H_0 : the lady has no special tea-tasting ability” is

$$P(n_{11} = 3) = \frac{\binom{4}{3} \binom{4}{1}}{\binom{8}{4}} = 0.229$$

The null hypothesis is the same as “The lady’s identification of a cup has had milk poured first is independent of whether the cup has had milk poured first”.

3.4 Fisher's Exact Test

Recall what a P -value is: a P -value is the probability of observing results as extreme or more extreme than what was actually observed.

The only other more extreme result would have been if the lady correctly identifies all 4 cups with milk poured first. Under H_0 , that could occur with probability

$$P(n_{11} = 4) = \frac{\binom{4}{4} \binom{4}{0}}{\binom{8}{4}} = 0.014$$

hence the P -value is $0.229 + 0.014 = 0.243$ and we are unable to reject H_0 .

3.4 Fisher's Exact Test

3.4.2 A test for small sample sizes

It is rare to see a study design which fixes the totals of rows *and* columns. However, it is common to see studies with small sample sizes.

In such situations, it is common to use an exact test like Fisher's. By conditioning on the margin totals, it is possible to calculate P -values and test independence of X and Y .

However, note that these tests consider a small number of possible realizations (only 5 possible outcomes for the lady). As a result, the P -values can be conservative, i.e. larger than they actually are. For example, if we chose our significance level to $\alpha = 0.01$, under what realization could we reject H_0 in favour of the lady?

3.4 Fisher's Exact Test

3.4.3 Example: a test for small sample sizes

A study is made to compare the effectiveness of two treatments in curing a rare blood disease. A total of 15 patients afflicted with approximately the same severity of this rare disease are available subjects for the study. Of these, 7 are randomly chosen to receive treatment 1; the remaining 8 are to receive treatment 2. The observed results are

	cured	uncured	Total
Treatment 1	4	3	7
Treatment 2	1	7	8
Total	5	10	15

3.4 Fisher's Exact Test

H_0 : no treatment difference

H_1 : treatment 1 is more effective

Now, under H_0 , it is a randomization model. That is, we consider the 5 cured and 10 uncured persons who are randomly divided into groups of 7 and 8.

There are many possible configurations under the randomization model.

Probability of the current configuration:

$$P(n_{11} = 4) = \frac{\binom{5}{4} \binom{10}{3}}{\binom{15}{7}} = \frac{5 \times 120}{6435} = 0.093.$$

3.4 Fisher's Exact Test

More extreme configuration that favor H_1 :

	cured	uncured	Total
Treatment 1	5	2	7
Treatment 2	0	8	8
Total	5	10	15

Probability of this configuration:

$$P(n_{11} = 5) = \frac{\binom{5}{5} \binom{10}{2}}{\binom{15}{7}} = 0.007.$$

Significant probability = $0.093 + 0.007 = 0.1$.

\therefore Do not reject H_0 at $\alpha = 0.05$

3.4 Fisher's Exact Test

If H_1 : there is some treatment difference (Two-Tail test)

Need to search more extreme configuration in another direction.

Let $n_{11} = x$, we then have

	cured	uncured	Total
Treatment 1	x	$7-x$	7
Treatment 2	$5-x$	$3+x$	8
Total	5	10	15

Find x such that

$$\left| \frac{x}{7} - \frac{5-x}{8} \right| \geq \left| \frac{4}{7} - \frac{1}{8} \right|,$$

$$\text{or} \quad |3x - 7| \geq 5.$$

$$\therefore x = 0 \text{ or } x = 4 \text{ or } x = 5.$$

3.4 Fisher's Exact Test

The configuration for $x = 0$:

	cured	uncured	Total
Treatment 1	0	7	7
Treatment 2	5	3	8
Total	5	10	15

Probability of this configuration:

$$P(n_{11} = 0) = \frac{\binom{5}{0} \binom{10}{7}}{\binom{15}{7}} = 0.019.$$

Significant probability = $0.093 + 0.007 + 0.019 = 0.119$.

\therefore Do not reject H_0 at $\alpha = 0.05$.