Chapter 2. Two-Way Contingency Tables

- **Example 2.1 (Introductory example)**: The following table cross classifies a sample of Americans according to their gender and their opinion about an afterlife. For the females in the sample, e.g. 435 said they believed in an afterlife and 147 said they did not or were undecided.

	Belief in Afterlife	
Gender	Yes	No or Undecided
Females	435	147
Males	375	134

2.1 Notation and Definition

- Questions:
  1. Whether an association exists between gender and belief in an afterlife. Is one sex more likely than the other to believe in an afterlife, or is belief in an afterlife independent of gender?
  2. How to describe and find the association?

This chapter answers above questions, and describes how to measure the association between two categorical variables.

2.1.1 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 $\{p_{i(j)}\}$.

2.1.2 Contingency table

Let X and Y denote two categorical response 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	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

2.1.3 Examples of contingency table

Example 2.2: Death Penalty Verdict by Defendant's Race

Defendant's Race	Death Penalty Yes	Death Penalty No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

2.1 Notation and Definition

Sample joint distribution:

	Yes	No	Total
White	$p_{11} = \frac{n_{11}}{n} = \frac{19}{326}$	$p_{12} = \frac{141}{326}$	$p_{1+} = p_{11} + p_{12} = \frac{160}{326}$
Black	$p_{21} = \frac{17}{326}$	$p_{22} = \frac{149}{326}$	$p_{2+} = p_{21} + p_{22} = \frac{166}{326}$
Total	$p_{+1} = \frac{36}{326}$	$p_{+2} = \frac{290}{326}$	1.0

Marginal distribution:

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

- **Example 2.3 (Example 2.2 continued):**

Defendant's Race	Death Penalty Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Explanatory variable: Race ($i = 1, 2$)
Response variable: Death Penalty ($j = 1, 2$)
Sample conditional distribution: $p_{j(i)}$

	Yes	No	
White	$p_{1(1)} = \frac{19}{160} = .119$	$p_{2(1)} = \frac{141}{160} = .881$	$\sum_j p_{j(1)} = 1$
Black	$p_{1(2)} = \frac{17}{166} = .102$	$p_{2(2)} = \frac{149}{166} = .898$	$\sum_j p_{j(2)} = 1$

- **Examples of multi-way contingency table:**

  Three-way $(2 \times 2 \times 2)$ table:

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

where Y = response (success, failure),
X = drug treatment (A, B),
Z = clinic (1, 2)

2.1 Notation and Definition

Three-way ($2 \times 2 \times 2$) table:

In a survey study, 2,276 students are asked whether they had ever used alcohol (A), cigarettes (C), or marijuana (M) in their final year of high school in a nonurban area near Dayton, Ohio.

Alcohol Use	Cigarette Use	Marijuana Use: Yes	Marijuana Use: No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

2.1 Notation and Definition

Four-way $(2 \times 2 \times 2 \times 2)$ table. The table refers to observations of 68,694 passengers in autos and light trucks involved in accidents in the state of Maine in 1991. The table classifies passengers by gender (G), location of accident (L), seat-belt use (S), and injury (I).

			Injury	
Gender	Location	Seat Belt	No	Yes
Female	Urban	No	7287	996
		Yes	11587	759
	Rural	No	3246	973
		Yes	6134	757
Male	Urban	No	10381	812
		Yes	10969	380
	Rural	No	6123	1084
		Yes	6693	513

2.1.4 Population analogs:

π_{ij}: probability of an observation falls 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, $\sum_i \pi_{i+} = 1$.
the set $\{\pi_{+j}\}$ for the column variable, $\sum_j \pi_{+j} = 1$.

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,

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

2.1 Notation and Definition

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.

The conditional distribution of Y given X relates to the joint distribution by

$$\pi_{j(i)} = \pi_{ij}/\pi_{i+} \text{ for } i = 1, \ldots, r,\ j = 1, \ldots, c.$$

2.1 Notation and Definition

Two categorical response variables are defined to be independent if

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

When X and Y are independent,

$$\pi_{j(i)} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j}, \text{ for } i = 1, \ldots, r.$$

When Y is a response variable and X is an explanatory variable,

$$\pi_{j(i)} = \pi_{+j}, \quad i = 1, \ldots, r, \; j = 1, \ldots, c$$

is a more natural way to define independence than $\pi_{ij} = \pi_{i+}\pi_{+j}$.

2.2.1 Three possible sampling models:

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

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

2. **Multinomial model:**
   The complete table of count $n_{ij}, i = 1, \ldots, r, j = 1, \ldots, c$ have a multinomial distribution with sample size n and probability $\pi_{ij} > 0$. The probability function for this model is

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

3. **Product (independent) multinomial model:**

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 $\{\pi_{1(i)}, \ldots, \pi_{c(i)}\}$. The counts $\{n_{ij}, j = 1, \ldots, c\}$ satisfy $\sum_j n_{ij} = n_{i+}$, and have the multinomial form:

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

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

2.2 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 *independent multinomial sampling*, also called *product multinomial sampling*.

2.3.1 Test of Independence

Observed frequencies:

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

Sampling model: Multinomial model with size n and $r \times c$ categories

2.3 Test of Independence and Test of Homogeneity

Population distribution:

$X \backslash Y$	1	2	$\cdots$	c	Total
1	π_{11}	π_{12}	$\cdots$	π_{1c}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2c}	π_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	π_{r1}	π_{r2}	$\cdots$	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+c}	1.00

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

Null hypothesis of independence:

$$H_0 : \pi_{ij} = \Pr(X = i, Y = j) = \Pr(X = i)\Pr(Y = j) = \pi_{i+}\pi_{+j},$$
$$\forall i = 1, \ldots, r,\ j = 1, \ldots, c.$$

Test statistic: Pearson's χ^2 (Case B)

$$X^2 = \sum_{\text{cell}} \frac{(O-E)^2}{E} \quad = \quad \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n\pi_{ij})^2}{n\pi_{ij}}$$

$$\Big\downarrow \text{ Under } H_0$$

$$= \quad \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n\pi_{i+}\pi_{+j})^2}{n\pi_{i+}\pi_{+j}}.$$

Need to estimate π_{i+}, $i = 1, \ldots, r$ and π_{+j}, $j = 1, \ldots, c$ under the constraints $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$.

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

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

Since

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

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

then

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

Summating 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,

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

Similarly,

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

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

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

Thus, the estimated expected frequencies are

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

Pearson's χ^2 test statistic:

$$X^2 = \sum_{\text{cell}} \frac{(O - \hat{E})^2}{\hat{E}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}.$$

When n is large, X^2 has a chi-square distribution with

$$\begin{aligned} df &= \text{No. cells} - 1 - \text{No. independent parameter estimated} \\ &= rc - 1 - (r - 1 + c - 1) = (r-1)(c-1). \end{aligned}$$

Reject H_0 if observed $X^2 \geq$ tabled chi-square value.

- **Example 2.4: Contingency table for political affiliation and opinion**

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

H_0: The pattern of opinion is independent of political affiliation.
Test statistic:

$$X^2 = \sum_{\text{cell}} \frac{(O - \widehat{E})^2}{\widehat{E}} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}} \sim \chi_2^2.$$

e.g. $\hat{E}_{11} = 285 \times 202/500$, $\hat{E}_{23} = 215 \times 148/500$.
$X^2 = 22.152$, P-value = 0.0000, $\text{Reject } H_0$.

2.3.2 Test of homogeneity

(contingency table with one margin fixed)

The total for rows (or columns) are specified in advance. We are testing that the various columns (or rows) have the same proportions of individuals in the various categories.

Row variable: explanatory variable
Column variable: response variable

For instance,
Row variable: Sex (male and female)
Column variable: Alcoholism (alcoholic or nonalcoholic)

Sampling scheme: Sample with fixed numbers of males and females.

2.3 Test of Independence and Test of Homogeneity

Observed frequencies:

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

Sampling model: Product multinomial model

$n_{1+}, n_{2+}, \ldots, n_{r+}$ fixed in advance

2.3 Test of Independence and Test of Homogeneity

Population distribution:

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

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

Null hypothesis of homogeneity:

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

2.3 Test of Independence and Test of Homogeneity

Test statistic: Pearson's χ^2 (Case B)

$$X^2 = \sum_{\text{cell}} \frac{(O-E)^2}{E} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i+}\pi_{j(i)})^2}{n_{i+}\pi_{j(i)}}$$

$$\Big\downarrow \ \text{Under } H_0$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i+}\pi_{j})^2}{n_{i+}\pi_{j}}.$$

Need to estimate π_j under the constraints $\sum_j \pi_j = 1$.

It can be shown that

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

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

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

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

2.3 Test of Independence and Test of Homogeneity

Pearson's χ^2 test statistic:

$$X^2 = \sum_{\text{cell}} \frac{(O - \hat{E})^2}{\hat{E}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

When n is large, X^2 has a chi-square distribution with

$$\begin{aligned} df &= r(c-1) - \text{No. of parameter estimated} \\ &= r(c-1) - (c-1) \\ &= (r-1)(c-1) \end{aligned}$$

Reject H_0 if observed $X^2 \geq$ tabled chi-square value.