Categorical Data Analysis
Lecture Note 3

Instructor: Seokho Lee

Hankuk University of Foreign Studies

3. Two-Way Contingency Tables

In this section we will examine association between two categorical variables.

Example: College students were classified according both to frequency of marijuana use and parental use of alcohol and psychoactive drugs.

		Level of Marijuana Use		
		Never	Occasional	Regular
Parental Use of Alcohol and Drugs	Neither	141	44	40
	One	68	44	51
	Both	17	11	19

We have two discrete variables, X and Y:

$X =$ parental use of alcohol and drugs

$Y =$ level of marijuana use

3.1. Probability Structure for Contingency Tables

3.1.1. Parameters:

$$p_{ij} = P(X = i, Y = j), \quad p_{i+} = \sum_{j=1}^{J} p_{ij}, \quad p_{+j} = \sum_{i=1}^{I} p_{ij}, \quad p_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$$

$\{p_{ij}\}$ gives the joint distribution of (X, Y)

$\{p_{i+}\}$ gives the marginal distribution of X

$\{p_{+j}\}$ gives the marginal distribution of Y

We obtain a table of cell probabilities:

$X \backslash Y$	1	2	$\cdots$	J	Total
1	p_{11}	p_{12}	$\cdots$	p_{1J}	p_{1+}
2	p_{21}	p_{22}	$\cdots$	p_{2J}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	p_{I+}
Total	p_{+1}	p_{+2}	$\cdots$	p_{+J}	$p_{++} = 1$

3.1.2. Data:

The sample data are the cell counts: n_{ij}.

We define row totals, column totals, and the grand total:

$$n_{i+} = \sum_{j=1}^{J} n_{ij}, \quad n_{+j} = \sum_{i=1}^{I} n_{ij}, \quad n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

$X \backslash Y$	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+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

The cell probabilities are estimated by the cell proportions:

$$p_{ij} = \frac{n_{ij}}{n}$$

3.1.3. Independence

We say that X and Y are statistically independent if

$$p_{ij} = p_{i+} \times p_{+j}, \qquad i = 1, \cdots, I, \quad j = 1, \cdots, J$$

Goal:

Use $\{n_{ij}\}$ to make inferences about $\{p_{ij}\}$, $\{p_{i+}\}$, or $\{p_{+j}\}$. This will tell us how X and Y are associated.

3.1.4. Poisson, Binomial and Multinomial Sampling

Consider a 2×2 table of gender by response to a prescription under investigation:

	Responder	Nonresponder	Total
Male	p_{11}, n_{11}	p_{12}, n_{12}	p_{1+}, n_{1+}
Female	p_{21}, n_{21}	p_{22}, n_{22}	p_{2+}, n_{2+}
Total	p_{+1}, n_{+1}	p_{+2}, n_{+2}	p_{++}, $n_{++} = n$

Parameters: $\{p_{ij},\ i = 1, 2,\ j = 1, 2\}$

Data: $\{n_{ij},\ i = 1, 2,\ j = 1, 2\}$

Notation: $\mu_{ij} = E(n_{ij})$

- Experiment 1: Collect $\{n_{ij}\}$ from all patients coming to the pharmacy for this prescription in the next 6 months.
  - n is random, as well as n_{ij}
  - each cell is an independent Poisson random variable: $n_{ij} \sim$ Poisson(μ_{ij})

- Experiment 2: Collect $\{n_{ij}\}$ from next 200 patients coming to the pharmacy for this prescription.
  - $n = 200$ is fixed
  - The four cell counts form a multinomial distribution with four categories:

$$\begin{pmatrix} n_{11} \\ n_{12} \\ n_{21} \\ n_{22} \end{pmatrix} \sim \text{Mult}\left(n, \begin{pmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{pmatrix} \right)$$

- Experiment 3: Collect $\{n_{ij}\}$ from next 100 male and 100 female patients coming to the pharmacy for this prescription.
  - n_{1+} and n_{2+} are fixed
  - This leads to two independent binomial samples:

$$n_{11} \sim \text{Binomial}\left(n_{1+}, \frac{p_{11}}{p_{1+}} \right), \qquad n_{21} \sim \text{Binomial}\left(n_{2+}, \frac{p_{21}}{p_{2+}} \right)$$

3.2. Comparing Proportion in 2×2 Tables

We will study methods of analyzing 2×2 tables to introduce methodology that can be extended to the analysis of $I \times J$ tables.

3.2.1. Difference of Two Proportions

In a two-way table, we consider the rows as the two groups and the columns as the binary categories. For row 1 we let p_1 be the probability of success and the row 2 we let p_2 be the probability of success. Such data often result from *independent binomial experiments*. (Note that we are dealing with conditional probabilities.)

Experiment 1: Observe n_{11} successes in $n_1 = n_{1+}$ trials with probability of success p_1.

Experiment 2: Observe n_{21} successes in $n_2 = n_{2+}$ trials with probability of success p_2.

We are interested in inference concerning the difference in probabilities of success, $p_1 - p_2$.
We will use the statistic:

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 &= \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2} \\ E[\hat{p}_1 - \hat{p}_2] &= p_1 - p_2 \\ Var(\hat{p}_1 - \hat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

A 95% confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\widehat{s.e.}(\hat{p}_1 - \hat{p}_2)$$

where

$$\widehat{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example: An antibiotic for pneumonia was injected into 100 patients with kidney malfunctions and into 100 normal patients. Some allergic reactions developed in 38 uremic patients and 21 normal patients.

$$\text{Uremic:} \quad n_1 = 100 \quad n_{11} = 38 \quad \hat{p}_1 = 0.38$$

$$\text{Normal:} \quad n_2 = 100 \quad n_{21} = 21 \quad \hat{p}_2 = 0.21$$

$$\widehat{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(0.38)(0.62)}{100} + \frac{(0.21)(0.79)}{100}} = 0.0634$$

$$0.38 - 0.21 \pm (1.96)(0.0634)$$

$$0.17 \pm 0.124 \quad \text{or} \quad (0.058,\ 0.294)$$

Example: In 1954, 401,974 children were participants in a large clinical trial for polio vaccine. Of these 201,229 were given the vaccine, and 200,745 were given a placebo. 110 of the children who received a placebo got polio and 33 of those given the vaccine got polio. Was the vaccine effective?

$$\text{Vaccine:} \quad n_1 = 201229 \quad n_{11} = 33 \quad \hat{p}_1 = .000164$$

$$\text{Placebo:} \quad n_2 = 200745 \quad n_{21} = 110 \quad \hat{p}_2 = .000548$$

$$\widehat{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(.000164)(.999836)}{201229} + \frac{(.000548)(.999452)}{200745}} = 0.0000595$$

$$.000164 - .000548 \pm (1.96)(.0000595)$$

$$-0.000384 \pm 0.000117 \quad \text{or} \quad (-.0005,\ -.0003)$$

Remark: We could test the hypothesis $H_0 : p_1 - p_2 = 0$ by determining whether 0 is in the confidence interval for $p_1 - p_2$. Alternatively, we could use the test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{whether} \quad \bar{p} = \frac{n_{11} + n_{21}}{n_1 + n_2} = \frac{n_{+1}}{n_{++}}$$

3.2.2. Relative Risk

A difference between two proportions of a given size is more important when the proportions are near 0 or 1 than in the middle of the range.

Example:

$$\begin{cases} p_1 = 0.460 \\ p_2 = 0.452 \end{cases} \quad \text{and} \quad \begin{cases} p_1 = 0.010 \\ p_2 = 0.002 \end{cases}$$

In both cases the difference is 0.008, but the second difference seems more noteworthy.

The **relative risk** is the ratio of the success probabilities:

$$RR = \frac{p_1}{p_2}$$

Example: $\frac{0.460}{0.452} = 1.018$ and $\frac{0.010}{0.002} = 5$.

Example: The sample relative risk in an antibiotic example is

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.38}{0.21} = 1.8095$$

Using PROC FREQ in SAS, a 95% confidence interval for the relative risk is (1.1478, 2.8526).

Example: The sample relative risk in an vaccine example is

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.000164}{0.000548} = 0.2993$$

Using PROC FREQ in SAS, a 95% confidence interval for the relative risk is (0.2028, 0.4416).

3.3. Odds and Odds Ratio

We define the odds for success for each row:

$$\text{odds}_1 = \frac{p_1}{1-p_1} \qquad \text{odds}_2 = \frac{p_2}{1-p_2}$$

Example: $p_1 = 0.8 \rightarrow \text{odds}_1 = \frac{0.8}{0.2} = 4$

- odds ≥ 0
- If $p_1 > 1 - p_1$, then $\text{odds}_1 > 1$
- $p_1 = \frac{\text{odds}_1}{1+\text{odds}_1}$
- If $p_1 = p_2$, then $\text{odds}_1 = \text{odds}_2$

Example: Antibiotic data

The odds for the uremic group is

$$\text{odds}_1 = \frac{0.38}{0.62} = 0.6129$$

The odds for the normal group is

$$\text{odds}_2 = \frac{0.21}{0.79} = 0.2658$$

Example: Vaccine data

The odds for the vaccine group is

$$\text{odds}_1 = \frac{.000164}{.999836} = 0.000164$$

The odds for the non-vaccine group is

$$\text{odds}_2 = \frac{.000548}{.999458} = 0.000548$$

The **odds ratio** for group 1 relative to group 2 is defined as

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

Note: The odds ratio is sometimes known as the *cross-product ratio*. Consider the two way table:

	S	F
1	p_1	$1-p_1$
2	p_2	$1-p_2$

The odds ratio is the ratio of the product of the diagonal elements to the product of the off-diagonal elements.

Properties of the Odds Ratio

- $\theta \geq 0$
- $\theta = 1 \iff p_1 = p_2$
- $\theta > 1 \iff p_1 > p_2$
- $\theta < 1 \iff p_1 < p_2$
- θ having substantially larger or smaller than 1 indicates a stronger association (or we can say that two groups have very different proportions of events of interest)

- If $\theta = 1$, then $\log \theta = 0$
- Let $\theta' =$ odds ratio for group 2 relative to group 1. Then

$$\theta' = \frac{p_2(1-p_1)}{p_1(1-p_2)} = \frac{1}{\theta}$$

and

$$\log(\theta') = \log\left(\frac{1}{\theta}\right) = -\log\theta$$

Sample Odds Ratio:

$$\hat{\theta} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)} = \frac{\frac{n_{11}}{n_{1+}}\left(1-\frac{n_{21}}{n_{2+}}\right)}{\frac{n_{21}}{n_{2+}}\left(1-\frac{n_{11}}{n_{1+}}\right)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Example: Antibiotic Example:

$$\hat{\theta} = \frac{38 \times 79}{62 \times 21} = 2.3057$$

Example: Vaccine Example:

$$\hat{\theta} = \frac{32 \times 200635}{110 \times 201196} = 0.2901$$

3.3.1. Inference for Odds Ratios and Log Odds Ratios

The sampling distribution of $\hat{\theta}$ can be highly skewed. The sampling distribution of $\log \hat{\theta}$ is better behaved; that is, we can use a normal approximation.

- First find a confidence interval for $\log \theta$.
- Exponentiate the endpoints to form a confidence interval for θ.

The *asymptotic standard error* of the sampling distribution of $\log \hat{\theta}$ is

$$\widehat{s.e.}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

We approximate the distribution of $\log \hat{\theta}$ using a normal distribution with mean $\log \theta$ and standard deviation given by the above $\widehat{s.e.}(\log \hat{\theta})$.
A large sample level $1 - \alpha$ confidence interval for $\log \theta$ is given by

$$\log \hat{\theta} \pm z_{\alpha/2} \widehat{s.e.}(\log \hat{\theta}).$$

After exponentiating the endpoints, the large sample confidence interval for θ is

$$\left(\hat{\theta} / \exp\{z_{\alpha/2} \widehat{s.e.}(\log \hat{\theta})\},\ \hat{\theta} \times \exp\{z_{\alpha/2} \widehat{s.e.}(\log \hat{\theta})\} \right)$$

Example: Antibiotic

$$\widehat{s.e.}(\log\hat{\theta}) = \sqrt{\frac{1}{28} + \frac{1}{62} + \frac{1}{21} + \frac{1}{79}} = 0.3348$$

$$\log\hat{\theta} \pm z_{\alpha/2}\widehat{s.e.}(\log\hat{\theta}) = \log(2.3057) \pm (1.96)(0.3348) = 0.8354 \pm 0.6562$$

The 95% confidence interval for $\log\theta$ is (0.1792, 1.4916). The 95% C.I. for θ is (1.1963, 4.4442).

Example: Vaccine

$$\widehat{s.e.}(\log\hat{\theta}) = \sqrt{\frac{1}{33} + \frac{1}{201196} + \frac{1}{110} + \frac{1}{200635}} = 0.1985$$

$$\log\hat{\theta} \pm z_{\alpha/2}\widehat{s.e.}(\log\hat{\theta}) = \log(0.2992) \pm (1.96)(0.1985) = -1.2066 \pm 0.3891$$

The 95% confidence interval for $\log\theta$ is (-1.5957, -0.8176). The 95% confidence interval for θ is (0.2027, 0.4414).

3.3.2. Relationship Between Odds Ratio and Relative Risk

$$\text{Odds Ratio} = \hat{\theta} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = RR \times \frac{1-p_2}{1-p_1}$$

- If p_1 and p_2 are both close to zero, then $\hat{\theta} \approx RR$.
- For some data sets, due to their study designs, it is not possible to calculate RR. Then we use $\hat{\theta}$ to approximate RR. (We will see this issue later)

3.3.3. Interpretation of difference of proportions, RR and OR

Suppose we conclude $p_1 = p_2$ from inference using difference of proportions, relative risk and odds ratio. This implies that event of our interest occurs with the same probability across groups. In this sense, event variable and group variable are **not associated**, or **independent**.

- Difference of proportions: $p_1 - p_2 = 0$
- Relative risk: $p_1/p_2 = 1$
- Odds ratio: $p_1(1-p_2)/p_2(1-p_1) = 1$

All equations are determined under the statistical sense, not the deterministic way. (confidence intervals, statistical tests, etc.)

3.4. Types of Studies

1. Prospective Study – subjects followed through time
   - Cohort Study – individuals stratified according to some variable (perhaps choice)
   - Clinical Trial – individuals assigned at random to groups of interest

2. Retrospective Study – the past of current subjects is studied
   - Case-control Study – match subjects with the condition (cases) with others free of the condition (controls)
   - Cross-sectional Study – classify individuals simultaneously on group and condition

All the above are **observational studies** except for the clinical trials which is **experimental study**.

Examples: Cross-classification of Smoking Status and Myocardial Infarction (MI)

Ever Smoker	Myocardial Infarction	Controls
Yes	172	173
No	90	346

- First column refers to 262 women with acute MI
- Second column refers to sets of two patients with other acute conditions matched to each patient in column 1
- All patients were classified according to smoking status
- This is a *case-control study*
- We cannot estimate Pr(MI) or Pr(MI|Smoker) from this study
- We can estimate Pr(Smoker|MI).

Now we consider other study designs and how to compute odds ratio (OR) and relative risk (RR) of MI occurrence for group Smoker relative to group Nonsmoker.

1. **Cross-Sectional Study** – Classify a random sample of n patients according to MI and to smoker. We obtain the following table of expected frequencies:

	MI	C
Y	$n \times p_{Y,MI}$	$n \times p_{Y,C}$
N	$n \times p_{N,MI}$	$n \times p_{N,C}$

$$OR = \frac{p_{MI|Y}/p_{C|Y}}{p_{MI|N}/p_{C|N}} = \frac{p_{Y,MI} \times p_{N,C}}{p_{N,MI} \times p_{Y,C}}, \qquad RR = \frac{p_{MI|Y}}{p_{MI|N}} = \frac{p_{Y,MI}(p_{N,MI} + p_{N,C})}{p_{N,MI}(p_{Y,MI} + p_{Y,C})}$$

2. **Case-Control Study** – For each case, match a given number of control subjects. We can estimate $p_{Y|MI}$ and $p_{Y|C}$, but not $p_{MI|Y}$, $p_{MI|N}$. Thus, relative risk is not estimable.

	MI	C
Y	$n_{MI} \times p_{Y|MI}$	$n_C \times p_{Y|C}$
N	$n_{MI} \times p_{N|MI}$	$n_C \times p_{N|C}$

$$OR = \frac{p_{Y,MI} \times p_{N,C}}{p_{N,MI} \times p_{Y,C}} = \frac{\left(\frac{p_{Y,MI}}{p_{MI}}\right) \times \left(\frac{p_{N,C}}{p_C}\right)}{\left(\frac{p_{N,MI}}{p_{MI}}\right) \times \left(\frac{p_{Y,C}}{p_C}\right)} = \frac{p_{Y|MI} \times p_{N|C}}{p_{N|MI} \times p_{Y|C}}$$

3. **Clinical Trial** – Randomly assign n_Y smokers and n_N nonsmokers. Classify according to MI. We can estimate $p_{MI|Y}$ and $p_{MI|N}$, but not $p_{Y|MI}$, $p_{Y|C}$.

	MI	C
Y	$n_Y \times p_{MI\|Y}$	$n_Y \times p_{C\|Y}$
N	$n_N \times p_{MI\|N}$	$n_N \times p_{C\|N}$

$$OR = \frac{p_{Y,MI} \times p_{N,C}}{p_{N,MI} \times p_{Y,C}} = \frac{\left(\frac{p_{Y,MI}}{p_Y}\right) \times \left(\frac{p_{N,C}}{p_N}\right)}{\left(\frac{p_{N,MI}}{p_N}\right) \times \left(\frac{p_{Y,C}}{p_Y}\right)} = \frac{p_{MI|Y} \times p_{C|N}}{p_{MI|N} \times p_{C|Y}}, \qquad RR = \frac{p_{MI|Y}}{p_{MI|N}}$$

3.5. Chi-Squared Tests of Independence

We consider tests of the null hypothesis (H_0) that the cell probabilities equal certain fixed values $\{p_{ij}\}$. When H_0 is true, the expected frequencies are

$$e_{ij} = np_{ij} = E(n_{ij}).$$

The test statistics measure the closeness of the observed cell frequencies $\{n_{ij}\}$ to $\{e_{ij}\}$.

1. Pearson's Chi-Squared Statistic

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

2. Likelihood Ratio Statistic

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right)$$

- Both statistics have an approximate chi-squared distribution when H_0 is true.
- We typically require that all $e_{ij} \geq 5$.
- Both test statistics equal zero if all $n_{ij} = e_{ij}$
- We reject H_0 for large values of the test statistic.
- Usually the two statistics have similar numerical values and will lead to the same conclusion. (But not always!!)
- Both statistics stay the same if the rows or columns are reordered.

3.5.1. Tests of Independence

We now consider testing whether X and Y are independent (or not associated):

$$H_0 \ : \ p_{ij} = p_{i+}p_{+j}, \quad \text{all } i, j$$

Under H_0, $p_{ij} = p_{i+}p_{+j}$ and $e_{ij} = E(n_{ij}) = np_{ij} = np_{i+}p_{+j}$. This is estimated by

$$e_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

For testing independence in an $I \times J$ table, Pearson's and the LR statistics are

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad \text{and} \quad G^2 = 2\sum_{i=1}^{J}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{e_{ij}}\right)$$

The degrees of freedom equal $(I-1)(J-1)$.

In general the degrees of freedom of the test statistic equals the difference in the numbers of parameters between the null and alternative hypotheses.

Null hypothesis: $I-1$ row probabilities and $J-1$ column probabilities for a total of $I+J-2$ parameters.

Alternative hypothesis: $IJ-1$ cell probabilities

$$df = (IJ-1) - (I+J-2) = IJ - I - J - 1 = (I-1)(J-1)$$

3.6. Analysis of $I \times J$ Tables

The methods that we studied for 2×2 tables form the basis for the study of $I \times J$ tables. We can use the Pearson statistic or likelihood ratio statistic for independence of row categories and column categories.

3.6.1. Odds Ratios in $I \times J$ Tables

Odds ratios can be formed for each pair of rows with each pair of columns resulting in $\binom{I}{2} \times \binom{J}{2}$ possible odds ratios. Often the *local odds ratios* between adjacent rows and columns are computed.

Example: Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack (1)	Nonfatal Attack (2)	No Attack (3)
Placebo	18	171	10845
Aspirin	5	99	10933

The local sample odds ratios are

$$\hat{\theta}_{12} = \frac{18 \times 99}{5 \times 171} = 2.0842 \qquad \hat{\theta}_{23} = \frac{171 \times 10933}{99 \times 10845} = 1.7413$$

The other odds ratio between column 1 and 3 can be computed by taking the product,

$$\hat{\theta}_{13} = \hat{\theta}_{12} \times \hat{\theta}_{23} = 2.0842 \times 1.7413 = 3.6293$$

3.6.2. Residual Analysis for Tests of Independence

We often wish to follow up a test of independence that results in a significant association by studying the nature of the association. One approach is to examine the observed and estimated expected frequencies of the various cells. Since larger differences tend to occur in cells with larger expected frequencies, it is useful to use adjusted residuals rather than the raw difference.

Two commonly used cell residuals that are the Pearson's residual

$$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

and the adjusted residual

$$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}(1 - p_{i+})(1 - p_{+j})}}.$$

When the null hypothesis of independence is true, the adjusted residuals have approximately a standard normal distribution. Large values of these residuals indicate the cells that are discrepant with independence.

In many tables, there will be a general interaction between the variables that cause us to reject independence. In other tables, the assumption of independence may be reasonable except for a few unusual cells. In this case we can use the residuals to identify these *outlier cells*.

Example: College students were classified according both to frequency of marijuana use and parental use of alcohol and psychoactive drugs.

The SAS output for the test of independence follows:

```
Chi-Square                      4      24.4171    <.0001
Likelihood Ratio Chi-Square     4      24.3571    <.0001
```

We conclude that there is strong evidence of association between parental use of alcohol and student marijuana use. The standardized residuals are in the following table:

		Level of Marijuana Use		
		Never	Occasional	Regular
Parental Use of Alcohol and Drugs	Neither	4.63	−1.64	−3.73
	One	−3.31	1.63	2.23
	Both	−2.29	0.11	2.53

3.7. Exact Tests for Small Samples

3.7.1. Fisher's Exact Test

- Consider first the 2×2 table. We will consider the exact distribution of counts for the sets of tables have the same row and column totals as the observed data.
- Under Poisson, independent binomial, or multinomial sampling for the cell counts, the distribution that applies to the set of tables with fixed row and column totals is the **hypergeometric distribution**.
- H_0: independence of two categorical variables.

Review of the Hypergeometric Distribution

Consider a population N objects of which M are labelled as defective. Take a sample of size n without replacement from the population. Define the random variable X to be the number of defective items in the sample. The random variable X has probability mass function

$$f(x) = \Pr(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where $\max(0, n - N + M) \leq x \leq \min(n, M)$.

The mean and variance of a hypergeometric random variable are

$$E[X] = \mu = n\frac{M}{N} = n\pi \quad \text{where } \pi = \frac{M}{N}$$

$$Var(X) = \sigma^2 = \left(\frac{N-n}{N-1}\right) n\frac{M}{N}\left(1 - \frac{M}{N}\right) = \left(\frac{N-n}{N-1}\right) n\pi\,(1 - \pi)$$

- For 2×2 tables, we can express probabilities for the 4 cell counts in terms of n_{11} only.

$X \backslash Y$	1	2	Total
1	n_{11}		n_{1+}
2			n_{2+}
Total	n_{+1}	n_{+2}	n

- To test H_0, the p-value is the sum of the hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcomes.
- When H_0 holds, the probability of a particular value for n_{11} is

$$\Pr(n_{11} = x) = \frac{\binom{n_{1+}}{x}\binom{n_{2+}}{n_{+1}-x}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

- The mean and variance of the cell counts are

$$E[n_{ij}|H_0] = \frac{n_{i+}n_{+j}}{n} \quad \text{and} \quad Var(n_{ij}|H_0) = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)}$$

3.7.2. Fisher's Tea Taster

A colleague of R. A. Fisher claimed she could determine if milk were added to the cup of tea first. Fisher designed an experiment where she tasted eight cups of tea. Four of these had milk added first, and four had tea added first. The results of the experiment were:

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

In this experiment, both column and row totals are fixed since she knew that 4 cups had milk added first. The distribution of n_{11} is hypergeometric with potential values $\{0, 1, 2, 3, 4\}$.

The possible tables are

4	0	3	1	2	2	1	3	0	4
0	4	1	3	2	2	3	1	4	0

Use the hypergeometric distribution to find probabilities. For example,

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

Hypergeometric Distribution for Example

n_{11}	Probability	p-value	X^2
0	.014	1.000	8.0
1	.229	.986	2.0
2	.514	.757	0.0
3	.229	.243	2.0
4	.014	.014	8.0

The table gives p-values for testing $H_0 : \theta = 1$ versus $H_a : \theta > 1$. The only table more extreme than the one observed is the last one. Thus, the p-value is

$$p\text{-value} = P(3) + P(4) = .229 + .014 = .243$$

Two-Sided Alternative: Consider testing $H_0 : \theta = 1$ versus $H_a : \theta \neq 1$. The exact p-value is defined as the sum of probabilities of tables no more likely than the observed table.

Example: Fisher's Tea Taster

$$p\text{-value} = P(0) + P(1) + P(3) + P(4) = .486$$