Contingency Table Analysis

Cynthia Chen
Kelly Sun
Overview

- Basics of Contingency Tables
- Test of Independence and Homogeneity
  - Hypotheses, Chi-square Test Statistics
  - Examples
  - Comparisons
- Fisher’s Exact Test
- Take Home Messages
### Notation

- **O**: the observed frequency of an outcome
- **E**: the expected frequency of an outcome
- **k**: the number of different categories or outcomes
- **n**: the total number of trials
- **r**: the number of rows
- **c**: the number of columns
- **p**: the probability of an outcome
Contingency Table

- first used by the statistician Karl Pearson in 1904
- a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables
- often used to analyze and record the relationship between two or more discrete variables (i.e. binary, categorical variables)
- two common sampling models used for contingency tables
  - the Multinomial sampling model
  - the Poisson sampling model
Contingency Table

- The Multinomial sampling model
  - assumes that a fixed number of individuals from a random sample are classified with fixed probabilities of being assigned to the different “cells”.

- The Poisson sampling model
  - assumes that different “cells” in the table entries have independent Poisson distributions
Multinomial Distribution

- Multinomial experiment
  - Experiment consists of \( n \) identical and independent trials
  - Each trial results in one of \( K \) outcomes
  - Let \( p_i \) be the probability of outcome \( i \)
    1. Each \( p_i \) remains constant for each experiment
    2. \[ \sum_{i=1}^{K} p_i = 1 \]

- The pmf for \( k \) outcomes is

\[
P( n_1, n_2, ..., n_k ) = \frac{n!}{n_1!n_2!...n_k!} \prod_{i=1}^{k} p_i^{n_i}
\]

Notes:

\[
\sum_{i=1}^{k} p_i = 1 \quad \sum_{i=1}^{k} n_i = n \quad E(n_i) = n \times p_i
\]
Contingency Table \((r \times c)\)

- Two cross-classified qualitative variables \(A\) and \(B\)
  - \(A\) has \(r\) categories, denoted by \(i = 1, 2, \ldots, r\)
  - \(B\) has \(c\) categories, denoted by \(j = 1, 2, \ldots, c\)

- The observed outcomes \(\{n_{ij}\}\) in a contingency table with \(r\) rows and \(c\) columns are jointly modeled with a multinomial distribution with parameters \(\{p_{ij}\}\)
Contingency Table (r x c)

- General form of a two-dimensional r x c contingency table

<table>
<thead>
<tr>
<th>Rows (Variable A)</th>
<th>Columns (Variable B)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>N</td>
</tr>
</tbody>
</table>

\[
n_{i+} = \sum_{i=1}^{r} n_{ij} \quad n_{+j} = \sum_{j=1}^{c} n_{ij} \quad N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}
\]
Contingency Tables: Null Model

- Under the null hypothesis of independence, $p_{ij} = p_i \times p_j$ for all $i$ and $j$ where there are $r - 1$ free parameters for the row factor and $c - 1$ free parameters for the column factor, for a total of $k_0 = r + c - 2$.
- The maximum likelihood estimates are
  \[
  \hat{p}_i = \frac{\text{sum of observations in row } i}{N}
  \]
  for the row probabilities and
  \[
  \hat{p}_j = \frac{\text{sum of observations in column } j}{N}
  \]
  for the column probabilities.
Contingency Tables: Null Model

- The maximum log-likelihood is

\[
\ln L_0 = \ln \left( \frac{n}{n_{11}, \ldots, n_{rc}} \right) + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \ln \left( \frac{E_{ij}}{N} \right)
\]

- The maximum likelihood estimate for \( p_{ij} \) is

\[
\hat{p}_{ij} = \hat{p}_i \cdot \hat{p}_j = \frac{E_{ij}}{N}
\]
Contingency Tables: Alternative Model

- Under the alternative hypothesis of no independence, the only restriction on the probabilities is that they sum to one, so there are \( k_1 = rc - 1 \) free parameters.
- The maximum log-likelihood is

\[
\ln L_1 = \ln \left( \frac{n}{n_{i1}, \ldots, n_{rc}} \right) + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \ln \left( \frac{n_{ij}}{N} \right)
\]

- The maximum likelihood estimates are

\[
p_{ij} = \frac{n_{ij}}{N}
\]
Contingency Table (2 x 2)

- The simplest contingency table is the 2x2 table.
- Here the cells are formed by cross-classification of 2 variables.
- 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Column Levels</th>
<th>Row Levels</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(n_{11})</td>
<td>(n_{12})</td>
</tr>
<tr>
<td>2</td>
<td>(n_{21})</td>
<td>(n_{22})</td>
</tr>
<tr>
<td>Total</td>
<td>(n_{+1})</td>
<td>(n_{+2})</td>
</tr>
</tbody>
</table>
Test of Independence

- **Definition:**
  there is no association between the row variable and the column variable in a contingency table

- **Null hypothesis:**
  The row and column variables are independent

- **Requirements:**
  - The sample data are randomly selected
  - represented as frequency counts in a two-way table
  - For every cell, the expected frequency E is at least 5
Test of Independence

- Test Statistic:
  \[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

- Degree of freedom:
  \[ df = (r-1)(c-1) \]
**Expected Frequency :** $E$

\[ E = \frac{\text{grand total} \cdot \text{row total}}{\text{grand total}} \cdot \frac{\text{column total}}{\text{grand total}} \]

\[ E = \frac{(\text{row total}) (\text{column total})}{(\text{grand total})} \]

- **Grand total:**
  Total number of all observed frequencies in the table.
**Example: Study of Motorcycle Drivers**

<table>
<thead>
<tr>
<th></th>
<th>Controls (not injured)</th>
<th>Cases (injured or killed)</th>
<th>Column Totals</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>491</td>
<td>213</td>
<td>704</td>
<td>899</td>
</tr>
<tr>
<td>White</td>
<td>377</td>
<td>112</td>
<td>489</td>
<td>333</td>
</tr>
<tr>
<td>Yellow/Orange</td>
<td>31</td>
<td>8</td>
<td>39</td>
<td>1232</td>
</tr>
</tbody>
</table>

For the upper left hand cell:

\[ E = \frac{(899)(704)}{1232} = 513.714 \]
Example: Study of Motorcycle Drivers

- Calculate expected for all cells:

<table>
<thead>
<tr>
<th></th>
<th>Black</th>
<th>White</th>
<th>Yellow/Orange</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls (not injured)</td>
<td>491</td>
<td>377</td>
<td>31</td>
<td>899</td>
</tr>
<tr>
<td>Expected</td>
<td>513.714</td>
<td>356.827</td>
<td>28.459</td>
<td></td>
</tr>
<tr>
<td>Cases (injured or killed)</td>
<td>213</td>
<td>112</td>
<td>8</td>
<td>333</td>
</tr>
<tr>
<td>Expected</td>
<td>190.286</td>
<td>132.173</td>
<td>10.541</td>
<td></td>
</tr>
<tr>
<td>Column Totals</td>
<td>704</td>
<td>489</td>
<td>39</td>
<td>1232</td>
</tr>
</tbody>
</table>
Example: Study of Motorcycle Drivers

- Using a 0.05 significance level, test the claim group (control or case) is independent of the helmet color.
- $H_0$: Whether a subject is in the control group or case group is independent of the helmet color.
- $H_1$: The group and helmet color are dependent.

\[
\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(491 - 513.714)^2}{513.714} + ... + \frac{(8 - 10.541)^2}{10.541}
\]

\[
\chi^2 = 8.775
\]
Example: Study of Motorcycle Drivers

- Degree of freedom: \( df = (2-1)(3-1)=2 \)
- The critical value: \( \chi^2_{0.05,2} = 5.991 \).
- We reject the null hypothesis. It appears there is an association between helmet color and motorcycle safety.
Test of Homogeneity

- **Definition:**
  we test the claim that different populations have the same proportions of same characteristics

- **Test Statistic:**
  \[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

- **Degree of freedom:**
  \[ df = (r-1)(c-1) \]
Example: Study of influence of gender

- Jail inmates can be classified into one of four crime categories

<table>
<thead>
<tr>
<th>Type of Crime</th>
<th>Male</th>
<th>Female</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent</td>
<td>117</td>
<td>66</td>
<td>183</td>
</tr>
<tr>
<td>Property</td>
<td>150</td>
<td>160</td>
<td>310</td>
</tr>
<tr>
<td>Drug</td>
<td>109</td>
<td>168</td>
<td>277</td>
</tr>
<tr>
<td>Public-Order</td>
<td>124</td>
<td>106</td>
<td>230</td>
</tr>
<tr>
<td>Column Totals</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>
Example: Study of influence of gender

- Calculate expected for all cells:

<table>
<thead>
<tr>
<th>Type of Crime</th>
<th>Male</th>
<th>Female</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent</td>
<td>117</td>
<td>117</td>
<td>183</td>
</tr>
<tr>
<td>Expected</td>
<td>91.5</td>
<td>91.5</td>
<td></td>
</tr>
<tr>
<td>Property</td>
<td>150</td>
<td>150</td>
<td>310</td>
</tr>
<tr>
<td>Expected</td>
<td>155</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>109</td>
<td>109</td>
<td>277</td>
</tr>
<tr>
<td>Expected</td>
<td>138.5</td>
<td>138.5</td>
<td></td>
</tr>
<tr>
<td>Public- Order</td>
<td>124</td>
<td>124</td>
<td>230</td>
</tr>
<tr>
<td>Expected</td>
<td>115</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>Column Totals</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>
Example: Study of influence of gender

- Using a 0.05 significance level, test the population proportions of the categories are same
- Ho: The population proportions of the categories are the same for male and female inmates.
- H1: At least one category differs.

\[ \chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(117-91.5)^2}{91.5} + \cdots + \frac{(106-115)^2}{115} \]

\[ \chi^2 = 28.5 \]
Example: Study of influence of gender

- Degree of freedom: \( df = (4-1)(2-1) = 3 \)
- The critical value: \( \chi^2_{0.05,3} = 7.82 \).
- We reject the null hypothesis and conclude that the proportions for one or more categories of crime differ for male and female inmates.
# Comparisons

<table>
<thead>
<tr>
<th>Independence</th>
<th>Homogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have two variables measured on one population. Want to know if they are somehow connected.</td>
<td>Have a single variable measured independently on two or more populations. Want to know if proportions are alike.</td>
</tr>
<tr>
<td>Q: Are the variables independent? Q: Is there an association between the two variables? Q: Is one variable dependent on the other?</td>
<td>Q: Are the groups homogenous (the same)? Q: Does there appear to be a difference between the groups?</td>
</tr>
</tbody>
</table>
Comparisons

- “Independence”, “No Association”, “Homogeneity of Proportions” are all the same null hypothesis.

- Thus, we use the same procedure (the chi-square test) to test hypotheses of “independence”, “no association”, “homogeneity of proportions” in an analysis of contingency table data.
When NOT to do a Chi-Squared Test

- Do not do a Chi-squared test for cell values that are not observed frequencies.

<table>
<thead>
<tr>
<th>sex</th>
<th>Voted in last election?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>women</td>
<td>35%</td>
<td>15%</td>
<td>50%</td>
</tr>
<tr>
<td>men</td>
<td>20%</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>total</td>
<td>55%</td>
<td>45%</td>
<td>100%</td>
</tr>
</tbody>
</table>

- The Problem: If you use percentages, you misstate the sample size as 100.
When NOT to do a Chi-Squared Test

- Do not do a Chi-squared test to find a difference in population proportions for dependent samples.

<table>
<thead>
<tr>
<th>Before speech:</th>
<th>Number supporting death penalty:</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>After hearing speech:</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>40</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>120</td>
<td>80</td>
<td>200</td>
</tr>
</tbody>
</table>

- **The Problem**: You want to know if the speech changed people’s opinions. A $\chi^2$ test would tell you if opinions after the speech depend on opinions before the speech.
When NOT to do a Chi-Squared Test

- Do not do a Chi-squared test when the expected value of a cell is less than 5.

<table>
<thead>
<tr>
<th>age</th>
<th>Party identification</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Democrat</td>
<td>Independ.</td>
<td>Republican</td>
<td>Total</td>
</tr>
<tr>
<td>&lt;65</td>
<td>42 (40)</td>
<td>5 (8)</td>
<td>33 (32)</td>
<td>80</td>
</tr>
<tr>
<td>65</td>
<td>8 (10)</td>
<td>5 (2)</td>
<td>7 (8)</td>
<td>20</td>
</tr>
<tr>
<td>total</td>
<td>50</td>
<td>10</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

- The Problem: The total $\chi^2$ is 6.28, so $p<.05$, but $((O-E)^2)/E = ((5-2)^2)/2 = 4.5$ of the total comes from one cell with the expected value of 2.

- *It is okay to do a Chi-squared test if a cell has an expected value above 5 and an observed value below 5!*
Fisher’s Exact Test

- often used for a 2×2 contingency table with one or more expected frequencies that are below 5
- based on exact probabilities from a specific distribution (the hypergeometric distribution).
- computes the probability, given the observed marginal frequencies, of obtaining exactly the frequencies observed and any configuration more extreme
- has no formal test statistic and no critical value, and it only gives you a p-value.
- also work on large samples, but sometimes will bog down the computer with lengthy computations. (This is especially likely to happen when the tables are 5×4 or larger).
Chi-Square & Fisher’s Tests in R

2 x 3 Contingency Table:

<table>
<thead>
<tr>
<th>age</th>
<th>Party identification</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Democrat</td>
<td>Independ.</td>
<td>Republican</td>
<td>Total</td>
</tr>
<tr>
<td>&lt;65</td>
<td>42 (40)</td>
<td>5 (8)</td>
<td>33 (32)</td>
<td>80</td>
</tr>
<tr>
<td>65</td>
<td>8 (10)</td>
<td>5 (2)</td>
<td>7 (8)</td>
<td>20</td>
</tr>
<tr>
<td>total</td>
<td>50</td>
<td>10</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

R code:
- Contingency<-matrix(c(42,8,5,5,33,7), nrow=2)
- chisq.test(Contingency)
- fisher.test(Contingency)
Output in R

```
> Contingency
   [,1] [,2] [,3]
[1,]  42   5  33
[2,]   8   5   7

> chisq.test(Contingency)

Pearson's Chi-squared test

data:  Contingency
X-squared = 6.2812, df = 2, p-value = 0.04326

> fisher.test(Contingency)

Fisher's Exact Test for Count Data

data:  Contingency
p-value = 0.06134
alternative hypothesis: two.sided
```
Fisher’s Exact Test

- The exact test was developed for the case of fixed marginals.
- The probability (p-value) computed by the Fisher’s test is exact (unlike the chi-square test, which relies on approximations).
- However, this setup is unrealistic for most studies – even if we know how many samples we will get in each group, we generally cannot fix in advance both margins.
Collapsing Tables

- combine adjacent rows and/or columns to attain required minimum expected cell frequencies.
  - may increase or reduce interpretability
  - may create or destroy structure in the table
- disadvantages of this approach
  - a loss of degrees of freedom
  - the resulting test statistic is less powerful
- no clear guidelines
  - avoid simply trying to identify the combination of cells that produces a “significant” result
Take Home Messages

- Contingency tables arrange categorical data in a table with at least two rows and at least two columns.
- Test of Independence tests the claim that the row and column variables are independent of each other.
- Test of Homogeneity tests the claim that different populations have the same proportion of some characteristics.
- The chi-square tests are easy to do for larger contingency tables. When the expected cell counts are less than 5, it is better to use the Fisher’s Exact Test.
Thank you!