Lecture 16
Outline

Contingency Tables
   Chi-Squared Tests

Paired Samples

Measuring Agreement

Diagnostic Tests

Simpson’s Paradox

Brief Review for Exam 2
Part I

Contingency Tables and More on Paired Samples
Generalizing the Chi-Square Statistic

- For a table with \( r \) rows and \( c \) columns (an \( r \times c \) table, or a contingency table), the null hypothesis is that the row and column variables are independent.

- Chi-squared statistic:

\[
X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.
\]

- The expected count in cell \( i \)th row and \( j \)th column:

\[
\text{expected count in cell } ij = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{grand total}}.
\]
Chi-Squared Test

- $X^2$ has a Chi-squared distribution with $(r - 1) \times (c - 1)$ degrees of freedom, i.e., a $\chi^2_{(r-1) \times (c-1)}$.

- Note in 2 $\times$ 2 case, $(r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$, as expected.

- $p$-value computation: $p = P_{\chi^2_{(r-1) \times (c-1)}}(X^2 > \chi^2)$. As before, this test is always two-sided.
In this study, 144 acute MI patients were age- and gender-matched with 144 individuals free of heart diseases.

<table>
<thead>
<tr>
<th>Diabetes</th>
<th>MI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>46</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>98</td>
<td>119</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>144</td>
</tr>
</tbody>
</table>

\[ OR = \frac{37}{16} \]

<table>
<thead>
<tr>
<th>Diabetes</th>
<th>No MI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>Diabetes</td>
<td>No Diabetes</td>
</tr>
<tr>
<td>Diabetes</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>16</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>119</td>
</tr>
</tbody>
</table>
Part II

Measuring Agreement and Diagnostic Tests
Survey Reliability

Amount of beef consumption reported by 537 female American nurses by two different surveys:

<table>
<thead>
<tr>
<th>Survey 1</th>
<th>Survey 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 1 servings/week</td>
<td>136, 92</td>
<td>228</td>
</tr>
<tr>
<td>&gt; 1 servings/week</td>
<td>69, 240</td>
<td>309</td>
</tr>
<tr>
<td>Total</td>
<td>205, 332</td>
<td>537</td>
</tr>
</tbody>
</table>

Question

How to measure (and test) the *agreement* between the two surveys (hence the reliability of the survey data)?
**Definition**

The Kappa ($\kappa$) statistic is used to measure the reproducibility/agreements between surveys, where

$$\kappa = \frac{p_o - p_e}{1 - p_e}.$$ 

Here, $p_o$ is the **observed** frequency of concordance (agreement) between the surveys, and $p_e$ is the **expected** probability of concordance between the surveys.
Example

- The number of concordant pairs is $136 + 240 = 376$, so $p_o = \frac{376}{537} = 0.7$.
- The expected probability of concordance is:

\[
p_e = \Pr(\leq 1 \text{ in survey 1}) \Pr(\leq 1 \text{ in survey 2}) + \\
\Pr(> 1 \text{ in survey 1}) \Pr(> 1 \text{ in survey 2})
\]

\[
= \frac{228 \times 205}{537^2} + \frac{309 \times 332}{537^2}
\]

\[
= \frac{228 \times 205 + 309 \times 332}{537^2}
\]

\[
= 0.52
\]
Evaluation of Kappa

\[ \kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{(0.7 - 0.52)}{(1 - 0.52)} = 0.37. \]

- If there is perfect agreement, \( p_o = 1 \) so \( \kappa = \frac{1 - p_e}{1 - p_e} = 1. \)

- General guidelines:
  \[
  \begin{align*}
  \kappa &> 0.75 \quad \text{excellent agreement} \\
  0.30 &< \kappa \leq 0.75 \quad \text{fair to good agreement} \\
  \kappa &\leq 0.30 \quad \text{poor agreement}
  \end{align*}
  \]

- SAS can also test for statistical significance of any \( \kappa \) value
  \[ \implies \text{check to see if } \kappa \text{'s 95\% CI excludes 0!} \]
data survey;
  input survey1 survey2 count;
cards;
  1  1  136
  1  2  92
  2  1  69
  2  2  240;
;
proc freq data = survey;
  weight count;
  tables survey1 * survey2 / agree;
  exact agree;
run;
McNemar’s Test

Statistic (S) 3.2857
DF 1
Asymptotic Pr > S 0.0699
Exact Pr >= S 0.0826
# Simple Kappa Coefficient

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa (K)</td>
<td>0.3782</td>
</tr>
<tr>
<td>ASE</td>
<td>0.0404</td>
</tr>
<tr>
<td>95% Lower Conf Limit</td>
<td>0.2989</td>
</tr>
<tr>
<td>95% Upper Conf Limit</td>
<td>0.4575</td>
</tr>
</tbody>
</table>

**Test of H0: Kappa = 0**

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE under H0</td>
<td>0.0430</td>
</tr>
<tr>
<td>Z</td>
<td>8.7987</td>
</tr>
<tr>
<td>One-sided Pr &gt; Z</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Two-sided Pr &gt;</td>
<td>Z</td>
</tr>
</tbody>
</table>

**Exact Test**

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-sided Pr &gt;= K</td>
<td>9.674E-19</td>
</tr>
<tr>
<td>Two-sided Pr &gt;=</td>
<td>K</td>
</tr>
</tbody>
</table>
Consider a screening test for some disease:

- **Sensitivity** is the probability of a positive test given the disease is present (the analogue of “power”)

- **Specificity** is the probability of a negative test given the disease is not present.

- **Positive predictive value** is the probability of the disease being present given the test is positive.

- **Negative predictive value** is the probability of the disease not being present given the test is negative.
In an ideal setting, $n$ random subjects are tested, and their true status (determined via some “gold standard,” e.g. biopsy) is also known:

\[
\begin{array}{c|cc}
\text{Test} & + & - \\
\hline
+ & a & b \\
- & c & d \\
\end{array}
\]

There are also cases where say $n_D$ diseased patients and $n_N$ non-diseased patients are tested, or, $n_+$ people with positive test results and $n_-$ with negative results are evaluated using the gold standard.

Be careful when trying to estimate proportions here (i.e., use the right sample!)
Part III

Data Aggregation and Simpson’s Paradox
Simpson’s Paradox

Definitions

- The association between $X$ and $Y$ for one fixed value of a third variable $Z$ is called the *conditional* association (say, given $Z = z$).
- For three variables $X$, $Y$, and $Z$, the association between $X$ and $Y$ *once we have summed over all levels of $Z$* is called the *marginal* association.
- The counterintuitive result that a *marginal* association can have a different direction from each *conditional* association is called *Simpson’s Paradox*.
- Can arise from *dramatically different population sizes* in the various categories defined by the conditioning $Z$ variable.
Example: Foreigners more likely to be insane?

Social class-stratified analysis of the prevalence of insanity in the foreign-born population relative to the native-born population: Massachusetts, 1854.

<table>
<thead>
<tr>
<th>Social Class</th>
<th>Birth Place</th>
<th>Insane</th>
<th>Not Insane</th>
<th>Percent Insane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pauper</td>
<td>Foreign</td>
<td>581</td>
<td>9090</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Native</td>
<td>941</td>
<td>12531</td>
<td>7.5</td>
</tr>
<tr>
<td>Non-Pauper</td>
<td>Foreign</td>
<td>44</td>
<td>220285</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Native</td>
<td>1066</td>
<td>880156</td>
<td>0.12</td>
</tr>
<tr>
<td>Total</td>
<td>Foreign</td>
<td>625</td>
<td>229375</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Native</td>
<td>2007</td>
<td>892687</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Here, $X =$ birth place, $Y =$ insanity status, and $Z =$ social class.
Note it “works” with odds ratios (OR’s) too:

- OR for Pauper table $= \frac{581(12531)}{941(9090)} = .85$
- OR for Non-Pauper table $= \frac{44(880156)}{1066(220285)} = .16$

BUT

- OR for total table $= \frac{625(892687)}{2007(229375)} = 1.21$

(Recall the “null” value for the OR is 1)
Simpson’s paradox can arise when we aggregate data (i.e., collapse over the variable Z) when Z is in fact an important predictor, or confounder. Other examples:

- **X = faculty salary, Y = gender, Z = scientific field**
  
  Women can do better in every subfield, yet still appear to have lower salaries overall due to their tendency to gravitate toward lower-paying fields (life and social sciences, public health) as opposed to physical sciences and engineering.

- **X = operation outcome, Y = hospital, Z = patient prognosis**
  
  Bill Clinton actually picked a “poor” hospital for his heart surgery; he knew that hospital only appeared poor because it attracted the most difficult cases!
Part IV

Brief Review for Exam 2
Exam 2 will cover the following sections in Moore and McCabe:
  - Secs 6.2-6.4
  - Secs 7.1-7.2
  - Secs 8.1-8.2

That is, it will cover Brad’s lectures 9-14. So just 6 lectures this time – albeit very full ones!

FYI, the Final Exam will cover Chapters 2, 9-11, and 14 in Moore and McCabe (i.e., the rest of the course). So this would be Brad’s lectures 15-23.
Confidence intervals for one-sample means: computation, sample size estimation, $Z$ versus $t$ distribution, use of $t$ when variance is unknown.

Departures from normality: transformations, bootstrapping!

Hypothesis testing: Basic definitions, null vs. alternative, accept/reject, accept really “fail to reject,” test statistics, $p$-values (one-sided vs. two-sided), equivalence of two-sided tests and CI’s.
Lectures 11-12

- **Hypothesis tests for one-sample means**: $Z$ versus $t$ tests, interpretations and misinterpretations of the $p$-value, critical value, rejection region

- **Paired two-sample inference**: To overcome lack of independence, reduce data to differences $D_i = X_i - Y_i$, and work as an appropriate one-sample problem

- **Independent two-sample inference**: two-sample $Z$ statistic (variances known and unequal), pooled $t$ statistic (variances unknown but assumed equal), two-sample $t$ statistic (variances unknown and assumed unequal), how to choose df and corresponding CIs in these settings

- **Error rates and power**: Type I error and $p$-values, Type II error and power, computation of power at a specific alternative
Lectures 13-14

▶ Sample size estimation: requirements, one- versus two-sample, \( n_1 = n_2 \) versus unequal

▶ One-sample binomial inference: binomial refresher, point estimate in one-sample case, CI using Wilson fix-up, CI for rare events (\( p \) close to 0 or 1), exact tests for \( p \) (using Table C), large-sample tests for \( p \) (using Tables A/D), choosing a sample size for a specified margin of error

▶ Conservativeness of tests: multiple testing, tradeoff of Type I error and power

▶ Two-sample binomial inference: 2 × 2 table format, design issues, Wilson CI, tests for equal proportions, relative risk