Class 27

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
Agenda:

Recap Chapter 10.5 and 10.6

Lecture Chapter 11.1-11.3
Recap Chapter 10.5 and 10.6
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions

That is where 1. and 2. in the orange box below come from

If independent samples of size \( n_1 \) and \( n_2 \) are drawn … with \( p_1 = P_1(\text{success}) \) and \( p_2 = P_2(\text{success}) \), then the sampling distribution of \( p'_1 - p'_2 \) has these properties:

1. mean \( \mu_{p'_1-p'_2} = p_1 - p_2 \)
2. standard error \( \sigma_{p'_1-p'_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \) \hspace{1cm} (10.10)
3. approximately normal dist if \( n_1 \) and \( n_2 \) are sufficiently large.
   - \( I \) \( n_1, n_2 > 20 \)
   - \( II \) \( n_1p_1, n_1q_1, n_2p_2, n_2q_2 > 5 \)
   - \( III \) sample < 10% of pop

Rowe, D.B.
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions

Confidence Interval Procedure

Assumptions for ... difference between two proportions

\( p_1 - p_2 \): The \( n_1 \) ... and \( n_2 \) random observations ... are selected independently from two populations that are not changing

Confidence Interval for the Difference between Two Proportions \( p_1 - p_2 \)

\[
(p_1' - p_2') - z(\alpha / 2) \sqrt{\frac{p_1'q_1'}{n_1} + \frac{p_2'q_2'}{n_2}} \quad \text{to} \quad (p_1' - p_2') + z(\alpha / 2) \sqrt{\frac{p_1'q_1'}{n_1} + \frac{p_2'q_2'}{n_2}}
\]

where \( p_1' = \frac{x_1}{n_1} \) and \( p_2' = \frac{x_2}{n_2} \).

(10.11)
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions

Confidence Interval Procedure

Example:
Construct a 99% CI for proportion of female A’s minus male A’s difference \( p_f - p_m \).

128 values

\( n_m = 43 \)
\( n_f = 85 \)

\( x_m = 18 \)
\( x_f = 52 \)

\( p_f' = \frac{x_f}{n_f} = \frac{52}{85} = .61 \)

\( p_m' = \frac{x_m}{n_m} = \frac{18}{43} = .42 \)

\( z(\alpha / 2) = 2.05 \)

\( (p_f' - p_m') \pm z(\alpha / 2) \sqrt{\frac{p_f'q_f'}{n_f} + \frac{p_m'q_m'}{n_m}} \)

\( (.61 - .42) \pm 2.05 \sqrt{\frac{(.61)(.39)}{85} + \frac{(.42)(.58)}{43}} \)

.001 to .379
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions

Hypothesis Testing Procedure

We can perform hypothesis tests on the proportion

\[ H_0: p_1 \geq p_2 \text{ vs. } H_a: p_1 < p_2 \]

\[ H_0: p_1 \leq p_2 \text{ vs. } H_a: p_1 > p_2 \]

\[ H_0: p_1 = p_2 \text{ vs. } H_a: p_1 \neq p_2 \]

when \( p_1 = p_2 = p \).

Test Statistic for the Difference between two Proportions - Population Proportions Known

\[
z^* = \frac{(p_1' - p_2') - (p_{10} - p_{20})}{\sqrt{pq\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}
\]

\[
p_1' = \frac{x_1}{n_1} \quad p_2' = \frac{x_2}{n_2}
\]

\[(10.12)\]
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions
Hypothesis Testing Procedure

Test Statistic for the Difference between two Proportions - Population Proportions UnKnown

\[ z^* = \frac{(p'_1 - p'_2) - (p_{10} - p_{20})}{\sqrt{p'_p q'_p \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}} \]

where we assume \( p_1 = p_2 \) and use pooled estimate of proportion

\[ p'_1 = \frac{x_1}{n_1}, \quad p'_2 = \frac{x_2}{n_2}, \quad p'_p = \frac{x_1 + x_2}{n_1 + n_2} \]
10: Inferences Involving Two Populations
10.5 Inference for Difference between Two Proportions

Hypothesis Testing Procedure

**Step 1**

\( H_0: \ p_s - p_c \leq 0 \) vs. \( H_a: \ p_s - p_c > 0 \)

**Step 2**

\[
 z^* = \frac{(p'_s - p'_c) - (p_{0s} - p_{0c})}{\sqrt{p'_p q'_p \left[ \frac{1}{n_s} + \frac{1}{n_c} \right]}}
\]

\( \alpha = .05 \)

**Step 3**

\[
 z^* = \frac{(.10 - .04) - (0)}{\sqrt{(.07)(.93) \left[ \frac{1}{150} + \frac{1}{150} \right]}} = 2.04
\]

**Cellular Phone Sample Information**

<table>
<thead>
<tr>
<th>Product</th>
<th>Number Defective</th>
<th>Number Checked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salesperson’s</td>
<td>( x_s = 15 )</td>
<td>( n_s = 150 )</td>
</tr>
<tr>
<td>Competitor’s</td>
<td>( x_c = 6 )</td>
<td>( n_c = 150 )</td>
</tr>
</tbody>
</table>

\[
 p'_s = \frac{x_s}{n_s} = \frac{15}{150} \quad p'_c = \frac{x_c}{n_c} = \frac{6}{150}
\]

\[
 p'_p = \frac{x_s + x_c}{n_s + n_c} = \frac{15 + 6}{150 + 150}
\]

**Step 4**

\( z(\alpha) = 1.65 \)

**Step 5**

Reject \( H_0 \) if \( z^* < .05 \)

\( .02 < \text{p-value} < .023 \) or \( 2.04 > 1.65 \)

Figure from Johnson & Kuby, 2008.
We can perform hypothesis tests on two variances

\[ H_0 : \sigma_1^2 \geq \sigma_2^2 \quad \text{vs.} \quad H_a : \sigma_1^2 < \sigma_2^2 \]

\[ H_0 : \sigma_1^2 \leq \sigma_2^2 \quad \text{vs.} \quad H_a : \sigma_1^2 > \sigma_2^2 \]

\[ H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_a : \sigma_1^2 \neq \sigma_2^2 \]

**Assumptions:** Independent samples from normal distribution

**Test Statistic for Equality of Variances**

\[ F^* = \frac{s_n^2}{s_d^2} \]

with \( df_n = n_n - 1 \) and \( df_d = n_d - 1 \).

(10.16)

Will also need critical values.

\[ P \left( F > F(df_n, df_d, \alpha) \right) = \alpha \]

Use new table to find areas for new statistic. Table 9, Appendix B, Page 668
10: Inferences Involving Two Pops.
10.6 Inference Ratio of Two Variances

Example: Find $F(5, 8, 0.05)$.

$df_n = n_n - 1$  
$df_d = n_d - 1$

Table 9, Appendix B, Page 668.

<table>
<thead>
<tr>
<th>$df_d$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>18.5</td>
<td>19.0</td>
<td>19.2</td>
<td>19.2</td>
<td>19.3</td>
<td>19.3</td>
<td>19.4</td>
<td>19.4</td>
<td>19.4</td>
<td>19.4</td>
</tr>
<tr>
<td>3</td>
<td>10.1</td>
<td>9.55</td>
<td>9.28</td>
<td>9.12</td>
<td>9.01</td>
<td>8.94</td>
<td>8.89</td>
<td>8.85</td>
<td>8.81</td>
<td>8.79</td>
</tr>
<tr>
<td>4</td>
<td>7.71</td>
<td>6.94</td>
<td>6.59</td>
<td>6.39</td>
<td>6.26</td>
<td>6.16</td>
<td>6.09</td>
<td>6.04</td>
<td>6.00</td>
<td>5.96</td>
</tr>
<tr>
<td>5</td>
<td>6.61</td>
<td>5.79</td>
<td>5.41</td>
<td>5.19</td>
<td>5.05</td>
<td>4.95</td>
<td>4.88</td>
<td>4.82</td>
<td>4.77</td>
<td>4.74</td>
</tr>
<tr>
<td>6</td>
<td>5.99</td>
<td>5.14</td>
<td>4.76</td>
<td>4.53</td>
<td>4.39</td>
<td>4.28</td>
<td>4.21</td>
<td>4.15</td>
<td>4.10</td>
<td>4.06</td>
</tr>
<tr>
<td>7</td>
<td>5.59</td>
<td>4.74</td>
<td>4.35</td>
<td>4.12</td>
<td>3.97</td>
<td>3.87</td>
<td>3.79</td>
<td>3.73</td>
<td>3.68</td>
<td>3.64</td>
</tr>
<tr>
<td>8</td>
<td>5.32</td>
<td>4.46</td>
<td>4.07</td>
<td>3.84</td>
<td>3.69</td>
<td>3.58</td>
<td>3.50</td>
<td>3.44</td>
<td>3.39</td>
<td>3.35</td>
</tr>
<tr>
<td>9</td>
<td>5.12</td>
<td>4.26</td>
<td>3.86</td>
<td>3.63</td>
<td>3.48</td>
<td>3.37</td>
<td>3.29</td>
<td>3.23</td>
<td>3.18</td>
<td>3.14</td>
</tr>
<tr>
<td>10</td>
<td>4.96</td>
<td>4.10</td>
<td>3.71</td>
<td>3.48</td>
<td>3.33</td>
<td>3.22</td>
<td>3.14</td>
<td>3.07</td>
<td>3.02</td>
<td>2.98</td>
</tr>
</tbody>
</table>

Degrees of Freedom for Numerator $df_n$

$\alpha = 0.05$

Figures from Johnson & Kuby, 2008.
10: Inferences Involving Two Populations

10.6 Inference for Ratio of Two Variances Two Ind. Samples

Hypothesis Testing Procedure

**One tailed tests:** Arrange $H_0$ & $H_a$ so $H_a$ is always “greater than”

$H_0: \sigma_1^2 \geq \sigma_2^2$ vs. $H_a: \sigma_1^2 < \sigma_2^2$ \quad \rightarrow \quad H_0: \frac{\sigma_2^2}{\sigma_1^2} \leq 1$ vs. $H_a: \frac{\sigma_2^2}{\sigma_1^2} > 1 \quad F^* = \frac{s_2^2}{s_1^2}$

$H_0: \sigma_1^2 \leq \sigma_2^2$ vs. $H_a: \sigma_1^2 > \sigma_2^2 \quad H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq 1$ vs. $H_a: \frac{\sigma_1^2}{\sigma_2^2} > 1 \quad F^* = \frac{s_1^2}{s_2^2}$

Reject $H_0$ if $F^* = \frac{s_n^2}{s_d^2} > F(df_n, df_d, \alpha)$.

**Two tailed tests:** put larger sample variance $s^2$ in numerator

$H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_a: \sigma_1^2 \neq \sigma_2^2$ \quad \rightarrow \quad H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs. $H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

$\sigma_n^2 = \sigma_1^2$ if $s_1^2 > s_2^2$, $\sigma_n^2 = \sigma_2^2$ if $s_2^2 > s_1^2$

Reject $H_0$ if $F^* = \frac{s_n^2}{s_d^2} > F(df_n, df_d, \alpha/2)$. 

Rowe, D.B.
10: Inferences Involving Two Populations

10.6 Inference for Ratio of Two Variances Two Ind. Samples

Hypothesis Testing Procedure

Step 1

\( H_0: \sigma_m^2 \geq \sigma_f^2 \) vs. \( H_a: \sigma_m^2 < \sigma_f^2 \)
\( H_0: \sigma_m^2 \leq \sigma_f^2 \) vs. \( H_a: \sigma_m^2 > \sigma_f^2 \)
\( H_0: \sigma_m^2 / \sigma_f^2 \leq 1 \) vs. \( H_a: \sigma_m^2 / \sigma_f^2 > 1 \)

Step 2

\[ F^* = \frac{s_m^2}{s_f^2} \]
\( df_m = 37 \)
\( df_f = 51 \)
\( \alpha = .01 \)

Step 3

\[ F^* = 15.8 / 9.2 = 1.72 \]

Step 4

\[ F(37, 51, .01) = 2.02 \]

Step 5

Fail to Reject \( H_0 \) if \( F^* < 2.02 \)
Chapter 10: Inferences Involving Two Populations

Questions?

Lecture Chapter 11.1-11.3
Chapter 11: Applications of Chi-Square

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
Data: The set of values collected from the variable from each of the elements that belong to the sample.
11: Applications of Chi-Square
11.1 Cooling a Great Hot Taste

Quite often we have qualitative data in categories.

**Example:** Cooling mouth after hot spicy food.

<table>
<thead>
<tr>
<th>Method</th>
<th>Water</th>
<th>Milk</th>
<th>Soda</th>
<th>Beer</th>
<th>Bread</th>
<th>Other</th>
<th>Nothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>73</td>
<td>35</td>
<td>20</td>
<td>19</td>
<td>29</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>
11: Applications of Chi-Square

11.2 Chi-Square Statistic

**Example:** Cooling mouth after hot spicy food.

<table>
<thead>
<tr>
<th>Method</th>
<th>Water</th>
<th>Milk</th>
<th>Soda</th>
<th>Beer</th>
<th>Bread</th>
<th>Other</th>
<th>Nothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>73</td>
<td>35</td>
<td>20</td>
<td>19</td>
<td>29</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

**Data set up:** $k$ cells $C_1, \ldots, C_k$ that $n$ observations sorted into

- Observed frequencies in each cell $O_1, \ldots, O_k$, \hspace{1cm} $O_1 + \ldots + O_k = n$
- Expected frequencies in each cell $E_1, \ldots, E_k$, \hspace{1cm} $E_1 + \ldots + E_k = n$

<table>
<thead>
<tr>
<th>Cell</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>$C_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>$O_1$</td>
<td>$O_2$</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>$O_k$</td>
</tr>
<tr>
<td>Expected</td>
<td>$E_1$</td>
<td>$E_2$</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>$E_k$</td>
</tr>
</tbody>
</table>
11: Applications of Chi-Square

11.2 Chi-Square Statistic

Outline of Test Procedure

When we have observed cell frequencies \( O_1, \ldots, O_k \), we can test to see if they match with some expected cell frequencies \( E_1, \ldots, E_k \).

Test Statistic for Chi-Square

\[
\chi^2* = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad df = k - 1
\]  

If the \( O_i \)'s are different from \( E_i \)'s then \( \chi^2* \) is “large.”

Go through 5 hypothesis testing steps as before.

Figure from Johnson & Kuby, 2008.
11.3 Inferences Concerning Multinomial Experiments

Example: We work for Las Vegas Gaming Commission. We have received many complaints that dice at particular casino are loaded (weighted, not fair). We confiscate all dice and test one. We roll it $n=60$ times. We get following data.

<table>
<thead>
<tr>
<th>Cell, $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed, $O_i$</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Expected, $E_i$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Expected Value for Multinomial Experiment:

$$E_i = np_i$$  \hspace{1cm} (11.3)
11: Applications of Chi-Square
11.3 Inferences Concerning Multinomial Experiments

Example: We roll it $n=60$ times. We get following data.

<table>
<thead>
<tr>
<th>Cell, $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed, $O_i$</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Expected, $E_i$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

$E_i = 60(1/6)$

Is the die fair? Need to go through the hypothesis testing procedure to determine if it is fair.
11: Applications of Chi-Square

11.3 Inferences Concerning Multinomial Experiments

Example: Is the die fair?

\[ \chi^2* = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]

Computations for Calculating \( \chi^2 \)

<table>
<thead>
<tr>
<th>Number</th>
<th>Observed (( O ))</th>
<th>Expected (( E ))</th>
<th>( O - E )</th>
<th>( (O - E)^2 )</th>
<th>( \frac{(O - E)^2}{E} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>10</td>
<td>-3</td>
<td>9</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>10</td>
<td>-2</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Total 60 \( O_1 + \ldots + O_k = n \) 60 \( E_1 + \ldots + E_k = n \) \( \chi^2* = 2.2 \)

D of F for Mult: \( df = k - 1 \) (11.2)

Figure from Johnson & Kuby, 2008.

Rowe, D.B.
11: Applications of Chi-Square

11.3 Inference for Mean Difference Two Dependent Samples

Observed different than expected?

Step 1

\( H_0: \text{Die fair } p_i's = \frac{1}{6} \quad H_a: \text{Die not fair } p_i's \neq \frac{1}{6} \)

Step 2

\[ \chi^2* = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \alpha = .05 \quad df = k - 1 \]

Step 3

\[ \chi^2* = 2.2 \]

Step 4

\[ \chi^2(df, \alpha) = 11.1 \]

Step 5

\[ df = 5 \]

Since \( .05 < p-value = .82 \) or because \( \chi^2* < \chi^2(df, \alpha) \), fail to reject \( H_0 \)
3: Descriptive Analysis and Bivariate Data

3.2 Bivariate Data

**Bivariate data:** The values of two different variables that are obtained from the same population element.

**Cross-tabulation tables or contingency tables**

Sometimes called $r$ by $c$ ($r \times c$)
3: Descriptive Analysis and Bivariate Data

3.2 Bivariate Data: two qualitative

Example:
Construct a 2×3 table.

### TABLE 3.1
**Genders and Majors of 30 College Students**

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>M</td>
<td>LA</td>
<td>Feeney</td>
<td>M</td>
<td>T</td>
<td>McGowan</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Argento</td>
<td>F</td>
<td>BA</td>
<td>Flanigan</td>
<td>M</td>
<td>LA</td>
<td>Mowers</td>
<td>F</td>
<td>BA</td>
</tr>
<tr>
<td>Baker</td>
<td>M</td>
<td>LA</td>
<td>Hedge</td>
<td>F</td>
<td>LA</td>
<td>Ornt</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>Bennett</td>
<td>F</td>
<td>LA</td>
<td>Holmes</td>
<td>M</td>
<td>T</td>
<td>Palmer</td>
<td>F</td>
<td>LA</td>
</tr>
<tr>
<td>Brand</td>
<td>M</td>
<td>T</td>
<td>Jepson</td>
<td>F</td>
<td>T</td>
<td>Pullen</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>Brock</td>
<td>M</td>
<td>BA</td>
<td>Kee</td>
<td>M</td>
<td>BA</td>
<td>Rattan</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Chun</td>
<td>F</td>
<td>LA</td>
<td>Kleeberg</td>
<td>M</td>
<td>LA</td>
<td>Sherman</td>
<td>F</td>
<td>LA</td>
</tr>
<tr>
<td>Crain</td>
<td>M</td>
<td>T</td>
<td>Light</td>
<td>M</td>
<td>BA</td>
<td>Small</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>Cross</td>
<td>F</td>
<td>BA</td>
<td>Linton</td>
<td>F</td>
<td>LA</td>
<td>Tate</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Ellis</td>
<td>F</td>
<td>BA</td>
<td>Lopez</td>
<td>M</td>
<td>T</td>
<td>Yamamoto</td>
<td>M</td>
<td>LA</td>
</tr>
</tbody>
</table>

### TABLE 3.2
**Cross-Tabulation of Gender and Major (tallied)**

<table>
<thead>
<tr>
<th>Gender</th>
<th>LA</th>
<th>BA</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>F</td>
<td>(6)</td>
<td>(4)</td>
<td>(2)</td>
</tr>
</tbody>
</table>

LA = liberal arts
BA = business admin
T = technology

Figures from Johnson & Kuby, 2008.
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables

Example:
Construct a 2×3 table.

Each in group of 300 students identified as male or female and asked if preferred classes in math-science, social science, or humanities.

<table>
<thead>
<tr>
<th></th>
<th>Math-Science (MS)</th>
<th>Social Science (SS)</th>
<th>Humanities (H)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (M)</td>
<td>37</td>
<td>41</td>
<td>44</td>
<td>122</td>
</tr>
<tr>
<td>Female (F)</td>
<td>35</td>
<td>72</td>
<td>71</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>113</td>
<td>115</td>
<td>300</td>
</tr>
</tbody>
</table>

Figure from Johnson & Kuby, 2008.
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables
Test of Independence

Is “Preference for math-science, social science, or humanities” … “independent of the gender of a college student?”

<table>
<thead>
<tr>
<th>TABLE 11.5</th>
<th>Sample Results for Gender and Subject Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Favorite Subject Area</td>
</tr>
<tr>
<td></td>
<td>Math-Science (MS)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male (M)</td>
<td>37</td>
</tr>
<tr>
<td>Female (F)</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
</tr>
</tbody>
</table>

Figure from Johnson & Kuby, 2008.
Is “Preference for math-science, social science, or humanities” … “independent of the gender of a college student?”

There is a Hypothesis test (of independence) to determine this.

Is Favorite Subject independent of Gender.

Similar to Example with the Die, now have rows and columns.

\[
\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

What are \(E_{ij}'s\)?

---

<table>
<thead>
<tr>
<th>Gender</th>
<th>Math-Science (MS)</th>
<th>Social Science (SS)</th>
<th>Humanities (H)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (M)</td>
<td>37</td>
<td>41</td>
<td>44</td>
<td>122</td>
</tr>
<tr>
<td>Female (F)</td>
<td>35</td>
<td>72</td>
<td>71</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>113</td>
<td>115</td>
<td>300</td>
</tr>
</tbody>
</table>

Figure from Johnson & Kuby, 2008.
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables

Test of Independence

\[ \chi^2 = \sum_{all cells} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]

D of F for Contingency Tables:

\[ df = (r - 1)(c - 1) \]  \hspace{1cm} (11.4)

Expected Frequencies for Contingency Tables

\[ E_{ij} = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{R_i C_j}{n} \]  \hspace{1cm} (11.5)

Where does this formula for \( E_i \)'s come from?
4: Probability
4.6 Independent Events

**Independent events:** Two events are independent if the occurrence or nonoccurrence of one gives us no information about the likeliness of occurrence for the other.

In algebra:

\[ P(A) = P(A \mid B) = P(A \mid \text{not } B) \]

In words:

1. Prob of \( A \) unaffected by knowledge that \( B \)
   has occurred, not occurred, or no knowledge.
2. …
3. …
4: Probability
4.6 Independent Events

Two events $A$ and $B$ are independent if the probability of one is not “influenced” by the occurrence or nonoccurrence of the other.

Two Events $A$ and $B$ are independent if:

1. $P(A) = P(A|B)$
2. $P(B) = P(B|A)$
3. $P(A \text{ and } B) = P(A) \cdot P(B)$

Examples:
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables
Test of Independence

Where does this formula for $E_{ij}$’s come from?

\[ E_{ij} = \frac{R_i C_j}{n} \]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Math-Science (MS)</th>
<th>Social Science (SS)</th>
<th>Humanities (H)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (M)</td>
<td>37</td>
<td>41</td>
<td>44</td>
<td>122</td>
</tr>
<tr>
<td>Female (F)</td>
<td>35</td>
<td>72</td>
<td>71</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>113</td>
<td>115</td>
<td>300</td>
</tr>
</tbody>
</table>

If Favorite Subject (column variable) is independent of Gender (row variable), then

\[ P(MS \mid M) = P(MS \mid F) = P(MS) \]

\[ P(A) = P(A \mid B) \]

\[ P(A \text{ and } B) = P(A) \cdot P(B) \]

Figure from Johnson & Kuby, 2008.
11: Applications of Chi-Square

11.4 Inferences Concerning Contingency Tables

Test of Independence

Where does this formula for $E_{ij}$’s come from?

<table>
<thead>
<tr>
<th></th>
<th>MS</th>
<th>SS</th>
<th>H</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>29.28</td>
<td>45.95</td>
<td>46.77</td>
<td>122.00</td>
</tr>
<tr>
<td>Female</td>
<td>42.72</td>
<td>67.05</td>
<td>68.23</td>
<td>178.00</td>
</tr>
<tr>
<td>Total</td>
<td>72.00</td>
<td>113.00</td>
<td>115.00</td>
<td>300.00</td>
</tr>
</tbody>
</table>

If Favorite Subject is independent of Gender, then

$$P(M \ and \ MS) = P(M)P(MS) = (122/300)(72/300)$$

$$E(M \ and \ MS) = nP(M)P(MS) = 300(122/300)(72/300)$$

$$E(M \ and \ MS) = 122 \times 72 / 300$$

Figure from Johnson & Kuby, 2008.
11: Applications of Chi-Square

11.4 Inferences Concerning Contingency Tables

Test of Independence

Where does this formula for $E_{ij}$’s come from?

$$E_{ij} = \frac{R_i C_j}{n}$$

<table>
<thead>
<tr>
<th>Gender</th>
<th>MS</th>
<th>SS</th>
<th>H</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>37 (29.28)</td>
<td>41 (45.95)</td>
<td>44 (46.77)</td>
<td>122</td>
</tr>
<tr>
<td>Female</td>
<td>35 (42.72)</td>
<td>72 (67.05)</td>
<td>71 (68.23)</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>113</td>
<td>115</td>
<td>300</td>
</tr>
</tbody>
</table>

If Favorite Subject is independent of Gender, then

$$\chi^2* = \sum_{all\ cells} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right) < \chi^2(2, 0.05)$$

$\chi^2* = 4.604 < \chi^2(2, 0.05) = 5.99$

$\alpha = 0.05$

$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1)$$

Figure from Johnson & Kuby, 2008.
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables
Test of Independence

$E_{ij} = \frac{R_i C_j}{n}$

Expected Frequencies for an $r \times c$ Contingency Table

<table>
<thead>
<tr>
<th>Row</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$j^{th}$ column</th>
<th>...</th>
<th>$c$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{R_1 \times C_1}{n}$</td>
<td>$\frac{R_1 \times C_2}{n}$</td>
<td>...</td>
<td>$\frac{R_1 \times C_j}{n}$</td>
<td>...</td>
<td>$\frac{R_1 \times C_c}{n}$</td>
<td>$R_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{R_2 \times C_1}{n}$</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$R_2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i^{th}$ row</td>
<td>$\frac{R_i \times C_1}{n}$</td>
<td>$\frac{R_i \times C_j}{n}$</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>$R_i$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>$\frac{R_r \times C_1}{n}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$C_1$</td>
<td>$C_2$</td>
<td>...</td>
<td>$C_j$</td>
<td>...</td>
<td>...</td>
<td>$n$</td>
</tr>
</tbody>
</table>

$\chi^2 = \sum_{all \ cells} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} < \chi^2((r - 1)(c - 1), \alpha)$

Figure from Johnson & Kuby, 2008.

Rowe, D.B.
11: Applications of Chi-Square
11.4 Inferences Concerning Contingency Tables

Test of Homogeneity
Is the distribution within all rows the same for all rows?

<table>
<thead>
<tr>
<th>Residence</th>
<th>Governor's Proposal</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Favor</td>
<td>Oppose</td>
<td>Total</td>
</tr>
<tr>
<td>Urban</td>
<td>143</td>
<td>57</td>
<td>200</td>
</tr>
<tr>
<td>Suburban</td>
<td>98</td>
<td>102</td>
<td>200</td>
</tr>
<tr>
<td>Rural</td>
<td>13</td>
<td>87</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>254</td>
<td>246</td>
<td>500</td>
</tr>
</tbody>
</table>

If so, then

\[ P(F \text{ and } \text{Urban}) = P(F)P(U) \]

\[ E(F \text{ and } \text{Urban}) = nP(F)P(U) \]

\[ E(F \text{ and } \text{Urban}) = 500\left(\frac{254}{500}\right)\left(\frac{200}{500}\right) \]
11: Applications of Chi-Square

11.4 Inferences Concerning Contingency Tables

Test of Homogeneity

Is the distribution within all rows the same for all rows?

\[
E_{ij} = \frac{R_i C_j}{n}
\]

\[
\chi^2 = \sum_{all \ cells} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} < \chi^2((r - 1)(c - 1), \alpha)
\]

\[\alpha = 0.05\]

\[df = (r - 1)(c - 1) = (2 - 1)(3 - 1)\]

<table>
<thead>
<tr>
<th>Residence</th>
<th>Favor</th>
<th>Oppose</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>143</td>
<td>57</td>
<td>200</td>
</tr>
<tr>
<td>Suburban</td>
<td>98</td>
<td>102</td>
<td>200</td>
</tr>
<tr>
<td>Rural</td>
<td>13</td>
<td>87</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>254</td>
<td>246</td>
<td>500</td>
</tr>
</tbody>
</table>
Chapter 11: Applications of Chi-Square

Questions?

Homework: Chapter 11 None.