Lecture 32: Chapter 12, Sections 1-2
Two Categorical Variables
Chi-Square

- Formulating Hypotheses to Test Relationship
- Test based on Proportions or on Counts
- Chi-square Test
- Confidence Intervals

Looking Back: Review

- 4 Stages of Statistics
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing (Lectures 5-12)
  - Probability (discussed in Lectures 13-20)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 21-23)
    - 1 quantitative (discussed in Lectures 24-27)
    - cat and quan: paired, 2-sample, several-sample (Lectures 28-31)
    - 2 categorical
    - 2 quantitative

Inference for Relationship (Review)

- \( H_0 \) and \( H_a \) about variables: not related or related
  - Applies to all three C\( \rightarrow \)Q, C\( \rightarrow \)C, Q\( \rightarrow \)Q
- \( H_0 \) and \( H_a \) about parameters: equality or not
  - C\( \rightarrow \)Q: pop means equal?
  - C\( \rightarrow \)C: pop proportions equal?
  - Q\( \rightarrow \)Q: pop slope equals zero?

Example: 2 Categorical Variables: Hypotheses

- Background: We are interested in whether or not smoking plays a role in alcoholism.
- Question: How would \( H_0 \) and \( H_a \) be written
  - in terms of variables?
  - in terms of parameters?
- Response:
  - in terms of variables
    - \( H_0 \): smoking and alcoholism _____ related
    - \( H_a \): smoking and alcoholism _____ related
  - in terms of parameters
    - \( H_0 \): Pop proportions alcoholic _____ for smokers, non-smokers
    - \( H_a \): Pop proportions alcoholic _____ for smokers, non-smokers
**Example: Summarizing with Proportions**

- **Background:** Research Question: Does smoking play a role in alcoholism?
- **Question:** What statistics from this table should we examine to answer the research question?
- **Response:** Compare proportions (response) for (explanatory).

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>Not Alcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>30</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1,000</td>
</tr>
</tbody>
</table>

**Example: Test Statistic for Proportions**

- **Background:** One approach to the question of whether smoking and alcoholism are related is to compare proportions.

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>Not Alcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>30</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1,000</td>
</tr>
</tbody>
</table>

$\hat{p}_1 = \frac{30}{230} = 0.130$

$\hat{p}_2 = \frac{10}{770} = 0.013$

- **Question:** What would be the next step, if we’ve summarized the situation with the difference between sample proportions 0.130-0.013?
- **Response:** ______ the difference between sample proportions 0.130-0.013.

 Stan. diff. is normal for large $n$: __________

**z Inference for 2 Proportions: Pros & Cons**

- **Advantage:** Can test against one-sided alternative.
- **Disadvantage:**
  - 2-by-2 table: comparing proportions straightforward
  - Larger table: comparing proportions complicated, can’t just standardize one difference $\hat{p}_1 - \hat{p}_2$

**Another Comparison in Considering Categorical Relationships (Review)**

- Instead of considering how different are the proportions in a two-way table, we may consider how different the counts are from what we’d expect if the “explanatory” and “response” variables were in fact unrelated.
- Compared observed, expected counts in wasp study:

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>A</th>
<th>NA</th>
<th>T</th>
<th>Exp</th>
<th>A</th>
<th>NA</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>16</td>
<td>15</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>24</td>
<td>7</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>40</td>
<td>22</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inference Based on Counts

To test hypotheses about relationship in \( r \)-by-\( c \) table, compare counts observed to counts expected if \( H_0 \) (equal proportions in response of interest) were true.

Example: Table of Expected Counts

- **Background:** Data on smoking and alcoholism:
  - |   | Alcohol | Not Alcohol | Total |
  - |-----------------|-----------|------------|
  - | Smoker          | 30        | 200        | 230   |
  - | Non-smoker      | 10        | 760        | 770   |
  - | **Total**       | **40**    | **960**    | **1,000** |

- **Question:** What counts are expected if \( H_0 \) is true?
- **Response:** Overall proportion alcoholic is ______

If proportions alcoholic were same for S and NS, expect

- \((40/1,000)(230)=____\) smokers to be alcoholic
- \((40/1,000)(770)=____\) non-smokers to be alcoholic; also
- \((960/1,000)(230)=____\) smokers not alcoholic
- \((960/1,000)(770)=____\) non-smokers not alcoholic

- **Note:** Each expected count is \( \frac{Column \; total \times \; Row \; total}{Table \; total} \)

**Expect:**

- \((40)(230)/1,000 = 9.2\) smokers to be alcoholic
- \((40)(770)/1,000 = 30.8\) non-smokers to be alcoholic; also
- \((960)(230)/1,000 = 220.8\) smokers not alcoholic
- \((960)(770)/1,000 = 739.2\) non-smokers not alcoholic
Chi-Square Statistic

- Components to compare observed and expected counts, one table cell at a time: \[
\text{component} = \frac{(\text{observed} - \text{expected})^2}{\text{expected}}
\]
Components are individual standardized squared differences.

- Chi-square test statistic \(\chi^2\) combines all components by summing them up:
\[
\text{chi-square} = \sum \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right)
\]
Chi-square is sum of standardized squared differences.

Example: Chi-Square Statistic

- **Background**: Observed and Expected Tables:

<table>
<thead>
<tr>
<th>Obs</th>
<th>A</th>
<th>NA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>30</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>NS</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exp</th>
<th>A</th>
<th>NA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>9.2</td>
<td>220.8</td>
<td>230</td>
</tr>
<tr>
<td>NS</td>
<td>30.8</td>
<td>739.2</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1000</td>
</tr>
</tbody>
</table>

- **Question**: What is the chi-square statistic?
- **Response**: Find \(\text{chi-square} = \sum \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right)\)

Example: Assessing Chi-Square Statistic

- **Background**: We found \(\text{chi-square} = 64\).
- **Question**: Is the chi-square statistic (64) large?
- **Response**: 

Chi-Square Distribution

- \(\text{chi-square} = \sum \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right)\) follows a predictable pattern (assuming \(H_0\) is true) known as \(\text{chi-square distribution}\) with \(\text{df} = (r-1) \times (c-1)\)
  
  - \(r = \text{number of rows (possible explanatory values)}\)
  - \(c = \text{number of columns (possible response values)}\)

Properties of chi-square:

- Non-negative (based on squares)
- Mean = df \([=1 \text{ for smallest } (2 \times 2) \text{ table}]\)
- Spread depends on df
- Skewed right
Chi-Square Density Curve

For chi-square with 1 df, \( P(\chi^2 \geq 3.84) = 0.05 \)
\( \Rightarrow \) If \( \chi^2 > 3.84 \), \( P\)-value < 0.05

Properties of chi-square:
- Non-negative
- Mean = df
  - df=1 for smallest [2×2] table
- Spread depends on df
- Skewed right

Example: Assessing Chi-Square (Continued)

- **Background:** In testing for relationship between smoking and alcoholism in 2×2 table, found \( \chi^2 = 64 \)
- **Question:** Is there evidence of a relationship in general between smoking and alcoholism (not just in the sample)?
- **Response:** For \( df=(2-1)\times(2-1)=1 \), chi-square considered “large” if greater than 3.84
  \( \Rightarrow \) chi-square=64 large? **____** \( P\)-value small? **____**
  Evidence of a relationship between smoking and alcoholism? **____**

Example: Relating Chi-Square & z

- **Background:** We found chi-square = 64 for the 2-by-2 table relating smoking and alcoholism.
- **Question:** What would be the \( z \) statistic for a test comparing proportions alcoholic for smokers vs. non-smokers?
- **Response:**

\[ z^2 = \chi^2 \]
- \( z \) statistic (comparing proportions) \( \Rightarrow \) combined tail probability=0.05 for \( z = 1.96 \)
- chi-square statistic (comparing counts) \( \Rightarrow \) right-tail prob=0.05 for \( \chi^2 = 1.96^2 = 3.84 \)
Assessing Size of Test Statistics (Summary)

When test statistic is “large”:
- \( z \): greater than 1.96 (about 2)
- \( t \): depends on \( df \); greater than about 2 or 3
- \( F \): depends on DFG, DFE
- \( \chi^2 \): depends on \( df=(r-1)(c-1) \); greater than 3.84 (about 4) if \( df=1 \)

Explanatory/Response: 2 Categorical Variables

- Roles impact what summaries to report
- Roles do not impact \( \chi^2 \) statistic or \( P \)-value

Example: Summaries Impacted by Roles

- **Background**: Compared proportions alcoholic (resp) for smokers and non-smokers (expl).

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>Not Alcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>30</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1,000</td>
</tr>
</tbody>
</table>

  \( \hat{p}_1 = \frac{30}{230} = 0.130 \)
  \( \hat{p}_2 = \frac{10}{770} = 0.013 \)

  \( \frac{30}{40} = 0.75 \quad \frac{200}{960} = 0.21 \)

- **Question**: What summaries would be appropriate if alcoholism is explanatory variable?

- **Response**: Compare proportions _______ (resp) for ____________________________ (expl).

Example: Comparative Summaries

- **Background**: Calculated proportions for table:

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>Not Alcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>30</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>960</td>
<td>1,000</td>
</tr>
</tbody>
</table>

  \( \hat{p}_1 = \frac{30}{230} = 0.130 \)
  \( \hat{p}_2 = \frac{10}{770} = 0.013 \)

  \( \frac{30}{40} = 0.75 \quad \frac{200}{960} = 0.21 \)

- **Question**: How can we express the higher risk of alcoholism for smokers and the higher risk of smoking for alcoholics?

- **Response**: Smokers are ___ times as likely to be alcoholics compared to non-smokers. Alcoholics are ______ times as likely to be smokers compared to non-alcoholics.
Guidelines for Use of Chi-Square Procedure

- Need random samples taken independently from several populations.
- Confounding variables should be separated out.
- Sample sizes must be large enough to offset non-normality of distributions.
- Need populations at least 10 times sample sizes.

Rule of Thumb for Sample Size in Chi-Square

- Sample sizes must be large enough to offset non-normality of distributions.
  Require expected counts all at least 5 in 2×2 table
  (Requirement adjusted for larger tables.)

Looking Back: Chi-square statistic follows chi-square distribution only if individual counts vary normally. Our requirement is extension of requirement for single categorical variables \(np \geq 10\), \(n(1-p) \geq 10\) with 10 replaced by 5 because of summing several components.

Example: Role of Sample Size

- **Background:** Suppose counts in smoking and alcohol two-way table were 1/10\(^{th}\) the originals:

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>Not Alcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>3</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>1</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4</strong></td>
<td><strong>96</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

- **Question:** Find chi-square; what do we conclude?
- **Response:** Observed counts 1/10\(^{th}\) \(\rightarrow\) expected counts 1/10\(^{th}\) \(\rightarrow\) chi-square ________ instead of 64.

But the statistic does not follow \(\chi^2\) distribution because expected counts (0.92, 22.08, 3.08, 73.92) are __________; individual distributions are not normal.

Confidence Intervals for 2 Categorical Variables

Evidence of relationship \(\rightarrow\) to what extent does explanatory variable affect response?

Focus on **proportions:** 2 approaches

- **Compare confidence intervals** for population proportion in response of interest (one interval for each explanatory group)
- **Set up confidence interval for difference** between population proportions in response of interest, 1\(^{st}\) group minus 2\(^{nd}\) group
Example: Confidence Intervals for 2 Proportions

- **Background:** Individual CI’s are constructed:
  - **Non-smokers** 95% CI for pop prop \( p \) alcoholic \((0.005, 0.021)\)
  - **Smokers** 95% CI for pop prop \( p \) alcoholic \((0.09, 0.17)\)
- **Question:** What do the intervals suggest about relationship between smoking and alcoholism?
- **Response:** Overlap?____
  Relationship between smoking and alcoholism?____ (____ likely to be alcoholic if a smoker).

Example: Difference between 2 Proportions (CI)

- **Background:** 95% CI for difference between population proportions alcoholic, smokers minus non-smokers is \((0.088, 0.146)\)
- **Question:** What does the interval suggest about relationship between smoking and alcoholism?
- **Response:** Entire interval ______suggests smokers ______ significantly more likely to be alcoholic \(\rightarrow\) there ______ a relationship.

Lecture Summary

*(Inference for Cat \(\rightarrow\) Cat; Chi-Square)*

- Hypotheses in terms of variables or parameters
- Inference based on proportions or counts
- Chi-square test
  - Table of expected counts
  - Chi-square statistic, chi-square distribution
  - Relating \( z \) and chi-square for \( 2 \times 2 \) table
  - Relative size of chi-square statistic
  - Explanatory/response roles in chi-square test
- Guidelines for use of chi-square
- Role of sample size
- Confidence intervals for 2 categorical variables