SECTION 7.1

- Testing the distribution of a single categorical variable: $\chi^2$ goodness of fit (7.1)

Multiple Categories

- So far, we've learned how to do inference for categorical variables with only two categories
- Today, we'll learn how to do hypothesis tests for categorical variables with multiple categories

Chi-Square Goodness-of-Fit Test

Question of the Day

Are children with ADHD younger than their peers?

ADHD or Just Young?

- In British Columbia, Canada, the cutoff date for entering school in any year is December 31st, so those born late in the year are almost a year younger than those born early in the year.
- Is it possible that younger students are being over-diagnosed with ADHD?
- Are children diagnosed with ADHD more likely to be born late in the year, and so be younger than their peers?
- Data: Sample of kids aged 6-12 in British Columbia who have been diagnosed with ADHD

Hypothesis Testing

1) State Hypotheses
2) Calculate a statistic, based on your sample data
3) Create a distribution of this statistic, as it would be observed if the null hypothesis were true
4) Measure how extreme your test statistic from (2) is, as compared to the distribution generated in (3)

Categories

- For simplicity, we'll break the year into four categories:
  - January – March
  - April – June
  - July – September
  - October – December
- Are birthdays for kids with ADHD evenly distributed among these four time periods?
Hypotheses

- Define parameters as
  - $p_1 =$ Proportion of ADHD births in January – March
  - $p_2 =$ Proportion of ADHD births in April – June
  - $p_3 =$ Proportion of ADHD births in July – Sep
  - $p_4 =$ Proportion of ADHD births in Oct – Dec

$H_0: \ p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$

$H_a: \ $ At least one $p_i \neq \frac{1}{4}$

Observed Counts

- The observed counts are the actual counts observed in the study

<table>
<thead>
<tr>
<th></th>
<th>Jan-Mar</th>
<th>Apr-Jun</th>
<th>July-Sep</th>
<th>Oct-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>6880</td>
<td>7982</td>
<td>9161</td>
<td>8945</td>
</tr>
</tbody>
</table>

Test Statistic

Why can't we use the familiar formula

$$\frac{\text{sample statistic} - \text{null value}}{\text{SE}}$$

to get the test statistic?

We need something a bit more complicated...

Expected Counts

- $n = 32,968$ boys with ADHD
- $H_0: p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$
- Expected count for each category, if the null hypothesis were true:

$$32,968 \cdot (1/4) = 8242$$

<table>
<thead>
<tr>
<th></th>
<th>Jan-Mar</th>
<th>Apr-Jun</th>
<th>July-Sep</th>
<th>Oct-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>6880</td>
<td>7982</td>
<td>9161</td>
<td>8945</td>
</tr>
<tr>
<td>Expected</td>
<td>8242</td>
<td>8242</td>
<td>8242</td>
<td>8242</td>
</tr>
</tbody>
</table>

Chi-Square Statistic

- Need a way to measure how far the observed counts are from the expected counts...
- Use the chi-square statistic:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

<table>
<thead>
<tr>
<th></th>
<th>Jan-Mar</th>
<th>Apr-Jun</th>
<th>July-Sep</th>
<th>Oct-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>6880</td>
<td>7982</td>
<td>9161</td>
<td>8945</td>
</tr>
<tr>
<td>Expected</td>
<td>8242</td>
<td>8242</td>
<td>8242</td>
<td>8242</td>
</tr>
</tbody>
</table>
Chi-Square Statistic

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
<th>Contribution to $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>6880</td>
<td>8242</td>
<td></td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>7982</td>
<td>8242</td>
<td></td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>9161</td>
<td>8242</td>
<td></td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>8945</td>
<td>8242</td>
<td></td>
</tr>
</tbody>
</table>

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

StatKey

- More Advanced Randomization Tests: $\chi^2$ Goodness-of-Fit

Minitab

- Stat > Tables > Chi-Square Goodness-of-Fit Test

Chi-Square Goodness-of-Fit Test for Observed Counts

<table>
<thead>
<tr>
<th>Category</th>
<th>Observed</th>
<th>Proportion</th>
<th>Expected</th>
<th>Contri. to Chi-Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>6880</td>
<td>0.25</td>
<td>8242</td>
<td>255.012</td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>7982</td>
<td>0.25</td>
<td>8242</td>
<td>3.822</td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>9161</td>
<td>0.25</td>
<td>8242</td>
<td>102.470</td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>8945</td>
<td>0.25</td>
<td>8242</td>
<td>59.962</td>
</tr>
</tbody>
</table>

N DF Chi-Sq p-Value

59868 3 598.707 0.000

What Next?

We have a test statistic. What else do we need to perform the hypothesis test?

Upper-Tail p-value

- To calculate the p-value for a chi-square test, we always look in the upper tail.
- Why?
  - Values of the $\chi^2$ statistic are always positive.
  - The higher the $\chi^2$ statistic is, the farther the observed counts are from the expected counts, and the stronger the evidence against the null.

Randomization Distribution

Can we reject the null?

a) Yes

b) No
**Chi-Square ($\chi^2$) Distribution**

- If each of the expected counts are at least 5, AND if the null hypothesis is true, then the $\chi^2$ statistic follows a $\chi^2$-distribution, with degrees of freedom equal to
  
  \[ df = \text{number of categories} - 1 \]

- ADHD Birth Month:
  
  \[ df = 4 - 1 = 3 \]

**$\chi^2$ Null Distribution**

**p-value**

**Conclusion**

- This is a TINY p-value
- We have incredibly strong evidence that birthdays of boys with ADHD are not evenly distributed throughout the year.

**Chi-Square Test for Goodness of Fit**

1. State null hypothesized proportions for each category, $p_i$
2. Alternative is that at least one of the proportions is different than specified in the null.
3. Calculate the $\chi^2$ statistic:
   \[ \chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \]
4. Compute the p-value as the proportion above the $\chi^2$ statistic for either a randomization distribution or a $\chi^2$ distribution with $df = (# \text{of categories} - 1)$ if expected counts all $> 5$
5. Interpret the p-value in context.
Births Equally Likely?

- Maybe all birthdays (not just for boys with ADHD) are not evenly distributed throughout the year
- We have the overall proportion of births for all boys during each time period in British Columbia:
  - January – March: 0.244
  - April – June: 0.258
  - July – September: 0.257
  - October – December: 0.241

Hypotheses

- Define parameters as
  - \( p_1 \) = Proportion of ADHD births in January – March
  - \( p_2 \) = Proportion of ADHD births in April – June
  - \( p_3 \) = Proportion of ADHD births in July – Sep
  - \( p_4 \) = Proportion of ADHD births in Oct – Dec
- \( H_0: p_1 = 0.244, p_2 = 0.258, p_3 = 0.257, p_4 = 0.241 \)
- \( H_a: \) At least one \( p \) is not as specified in \( H_0 \)

ADHD or Just Young? (BOYS)

<table>
<thead>
<tr>
<th>Birthday</th>
<th>Proportion of Births</th>
<th>ADHD</th>
<th>Expected Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>0.244</td>
<td>6880</td>
<td></td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>0.258</td>
<td>7982</td>
<td></td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>0.257</td>
<td>9161</td>
<td></td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>0.241</td>
<td>8945</td>
<td></td>
</tr>
</tbody>
</table>

\( n = 32,968 \)

The expected count for the Jan-Mar cell is

- a) 8044.2
- b) 8505.7
- c) 8472.8
- d) 7945.3

ADHD or Just Young? (BOYS)

<table>
<thead>
<tr>
<th>Birthday</th>
<th>Proportion of Births</th>
<th>ADHD Diagnoses</th>
<th>Expected Counts</th>
<th>Contribution to ( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>0.244</td>
<td>6880</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>0.258</td>
<td>7982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>0.257</td>
<td>9161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>0.241</td>
<td>8945</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( n = 32,968 \)

The contribution to \( \chi^2 \) for the Oct-Dec cell is

- a) 32.2
- b) 55.9
- c) 125.8
- d) 168.5

ADHD or Just Young (Boys)

We have VERY (!!!) strong evidence that boys with ADHD do not have the same distribution of birthdays as all boys in British Columbia.

\( p\)-value \( \approx 0 \)
A Deeper Understanding

- The p-value only tells you that the counts provide evidence against the null distribution.
- If you want to know more, look at:
  - Which cells contribute most to the $\chi^2$ statistic?
  - Are the observed or expected counts higher?

ADHD or Just Young? (BOYS)

<table>
<thead>
<tr>
<th>Birth Date</th>
<th>Proportion of Births</th>
<th>ADHD Diagnoses</th>
<th>Expected Counts</th>
<th>Contribution to $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>0.244</td>
<td>6880</td>
<td>8044.2</td>
<td>168.5</td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>0.258</td>
<td>7982</td>
<td>8505.7</td>
<td>32.2</td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>0.257</td>
<td>9161</td>
<td>8472.8</td>
<td>55.9</td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>0.241</td>
<td>8945</td>
<td>7945.3</td>
<td>125.8</td>
</tr>
</tbody>
</table>

- Big contributions to $\chi^2$: beginning and end of the year.
- Jan-Feb: Many fewer ADHD than would be expected.
- Oct-Dec: Many more ADHD cases than would be expected.

ADHD and Birth Year

- Stop and think about the implications of this!
- In British Columbia, 6.9% of all boys 6-12 are diagnosed with ADHD.
- 5.5% of boys 6-12 receive medication for ADHD.
- How many of these diagnoses are simply due to the fact that these kids are younger than their peers???
- ARE YOU CONCERNED?

ADHD or Just Young? (Girls)

<table>
<thead>
<tr>
<th>Birth Date</th>
<th>Proportion of Births</th>
<th>ADHD Diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Mar</td>
<td>0.243</td>
<td>1960</td>
</tr>
<tr>
<td>Apr-Jun</td>
<td>0.258</td>
<td>2358</td>
</tr>
<tr>
<td>Jul-Sep</td>
<td>0.257</td>
<td>2859</td>
</tr>
<tr>
<td>Oct-Dec</td>
<td>0.242</td>
<td>2904</td>
</tr>
</tbody>
</table>

- Want more practice?
- Here is the data for girls. ($\chi^2 = 236.8$)

To Do

- Read Section 7.1
- Do HW 7.1 (due Friday, 11/20)