Topic - 13
Analysis of Categorical Data
1. Goodness-of-Fit Test

- Many experiments result in measurements that are **qualitative** or **categorical** rather than quantitative.
  - People classified by ethnic origin
  - Cars classified by color
  - M&M®'s classified by type (plain or peanut)
- These data sets have the characteristics of a **multinomial experiment**.
The Multinomial Experiment

1. The experiment consists of \( n \) identical trials.
2. Each trial results in one of \( k \) categories.
3. The probability that the outcome falls into a particular category \( i \) on a single trial is \( p_i \) and remains constant from trial to trial. The sum of all \( k \) probabilities, \( p_1 + p_2 + \ldots + p_k = 1 \).
4. The trials are independent.
5. We are interested in the number of outcomes in each category, \( O_1, O_2, \ldots O_k \) with \( O_1 + O_2 + \ldots + O_k = n \).
The Binomial Experiment

- A special case of the multinomial experiment with $k = 2$.
- Categories 1 and 2: success and failure
- $p_1$ and $p_2$: $p$ and $q$
- $O_1$ and $O_2$: $x$ and $n-x$
- We made inferences about $p$ (and $q = 1 - p$)

In the multinomial experiment, we make inferences about all the probabilities, $p_1, p_2, p_3 \ldots p_k$. 


We have some preconceived idea about the values of the $p_i$ and want to use sample information to see if we are correct.

The **expected number** of times that outcome $i$ will occur is $E_i = np_i$.

If the **observed cell counts**, $O_i$, are too far from what we hypothesize under $H_0$, the more likely it is that $H_0$ should be rejected.
We use the Pearson chi-square statistic:

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]

- When \( H_0 \) is true, the differences \( O-E \) will be small, but large when \( H_0 \) is false.
- Look for **large values of** \( \chi^2 \) based on the chi-square distribution with a particular number of degrees of freedom.
Degrees of Freedom

- These will be different depending on the application.
  1. Start with the number of categories or cells in the experiment.
  2. Subtract 1 df for each linear restriction on the cell probabilities. (You always lose 1 df since $p_1 + p_2 + \ldots + p_k = 1$.)
  3. Subtract 1 df for every population parameter you have to estimate to calculate or estimate $E_i$. 
Assumptions

Assumptions for Pearson’s Chi-Square:

1. The cell counts $O_1, O_2, \ldots, O_k$ must satisfy the conditions of a multinomial experiment, or a set of multinomial experiments created by fixing either the row or the column totals.

2. The expected cell counts $E_1, E_2, \ldots, E_k \geq 5$.

If not (one or more is < 5)

1. Choose a larger sample size $n$.
   The larger the sample size, the closer the chi-square distribution will approximate the distribution of your test statistic $\chi^2$.

2. It may be possible to combine one or more of the cells with small expected cell counts, thereby satisfying the assumption.
The Goodness of Fit Test

- The simplest of the applications.
- A single categorical variable is measured, and exact numerical values are specified for each of the $p_i$.
- Expected cell counts are $E_i = np_i$
- Degrees of freedom: $df = k-1$

Test statistic: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
Example 1

- Toss a die 300 times with the following results. Is the die fair or biased?

<table>
<thead>
<tr>
<th>Upper Face</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of times</td>
<td>50</td>
<td>39</td>
<td>45</td>
<td>62</td>
<td>61</td>
<td>43</td>
</tr>
</tbody>
</table>

A multinomial experiment with $k = 6$ and $O_1$ to $O_6$ given in the table.

We test:

$H_0: \hat{p}_1 = 1/6; \hat{p}_2 = 1/6; \cdots \hat{p}_6 = 1/6$ (die is fair)

$H_1$: at least one $\hat{p}_i$ is different from $1/6$ (die is biased)
Example 1 – Solution

• Calculate the expected cell counts:

\[ E_i = np_i = 300(1/6) = 50 \]

<table>
<thead>
<tr>
<th>Upper Face</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_i )</td>
<td>50</td>
<td>39</td>
<td>45</td>
<td>62</td>
<td>61</td>
<td>43</td>
</tr>
<tr>
<td>( E_i )</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Test statistic and rejection region:

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(50 - 50)^2}{50} + \frac{(39 - 50)^2}{50} + \ldots + \frac{(43 - 50)^2}{50} = 9.2 \]

Reject \( H_0 \) if \( \chi^2 > \chi^2_{0.05} = 11.07 \) with \( k - 1 = 6 - 1 = 5 \) df.

Do not reject \( H_0 \). There is insufficient evidence to indicate that the die is biased.
Some Notes

- The test statistic, $\chi^2$ has only an approximate chi-square distribution.
- For the approximation to be accurate, statisticians recommend $E_i \geq 5$ for all cells.
- Goodness of fit tests are different from previous tests since the experimenter uses $H_0$ for the model he thinks is true.

**H$_0$:** model is correct (as specified)
**H$_1$:** model is not correct

- Be careful not to accept $H_0$ (say the model is correct) without reporting $\beta$. 
Example 2: Finger Lakes Homes

Finger Lakes Homes manufactures four models of prefabricated homes, a **two-story colonial**, a **ranch**, a **split-level**, and an **A-frame**. To help in production planning, management would like to determine if previous customer purchases indicate that there is a preference in the style selected. The number of homes sold of each model for 100 sales over the past two years is shown below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Colonial</th>
<th>Ranch</th>
<th>Split-Level</th>
<th>A-Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td># Sold</td>
<td>30</td>
<td>20</td>
<td>35</td>
<td>15</td>
</tr>
</tbody>
</table>
Example 2: Finger Lakes Homes

- **Notation**
  - $p_C = \text{popul. proportion that purchase a colonial}$
  - $p_R = \text{popul. proportion that purchase a ranch}$
  - $p_S = \text{popul. proportion that purchase a split-level}$
  - $p_A = \text{popul. proportion that purchase an A-frame}$

- **Hypotheses**
  - $H_0$: $p_C = p_R = p_S = p_A = .25$
  - $H_1$: The population proportions are not $p_C = .25, p_R = .25, p_S = .25, \text{ and } p_A = .25$
Example 2: Finger Lakes Homes

- **Expected Frequencies**
  
  \[ E_1 = .25(100) = 25 \quad E_2 = .25(100) = 25 \]
  
  \[ E_3 = .25(100) = 25 \quad E_4 = .25(100) = 25 \]

- **Test Statistic**

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

  \[
  = \frac{(30 - 25)^2}{25} + \frac{(20 - 25)^2}{25} + \frac{(35 - 25)^2}{25} + \frac{(15 - 25)^2}{25}
  \]

  \[= 1 + 1 + 4 + 4 = 10 \]

- **Rejection Rule**

  \[ \chi^2_{\alpha;df} = \chi^2_{0.05; k-1} = \chi^2_{0.05; 3} = 7.815 \]

  \[ \chi^2 = 10 > \chi^2_{0.05; 3} = 7.815 \]

  We reject the assumption that there is no home style preference, at the 0.05 level of significance.
Example 2: Finger Lakes Homes

Conclusion Using the \( p \)-Value Approach

<table>
<thead>
<tr>
<th>Area in Upper Tail</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 ) Value (df = 3)</td>
<td>6.251</td>
<td>7.815</td>
<td>9.348</td>
<td>11.345</td>
<td>12.838</td>
</tr>
</tbody>
</table>

- Because \( \chi^2 = 10 \) is between 9.348 and 11.345, the area in the upper tail of the distribution is between .025 and .01.
- The \( p \)-value \( \leq \alpha \). We can reject the null hypothesis.
2. Contingency Tables: A Two-Way Classification

The test of independence of variables is used to determine whether two variables are independent when a single sample is selected.

- The experimenter measures two qualitative variables to generate bivariate data.
  - Gender and colorblindness
  - Age and opinion
  - Professorial rank and type of university
- Summarize the data by counting the observed number of outcomes in each of the intersections of category levels in a contingency table.
The contingency table has $r$ rows and $c$ columns = $rc$ total cells.

- We study the relationship between the two variables. Is one method of classification contingent or dependent on the other?

Does the distribution of measurements in the various categories for variable 1 depend on which category of variable 2 is being observed? **If not, the variables are independent.**
**Chi-Square Test of Independence**

**H₀:** classifications are independent  
**H₁:** classifications are dependent

- Observed cell counts are \( O_{ij} \) for row \( i \) and column \( j \).
- Expected cell counts are \( E_{ij} = np_{ij} \)
  
  ✓ If \( H₀ \) is true and the classifications are independent,  
  ✓ \( p_{ij} = p_i p_j = \text{P(falling in row } i)\text{P(falling in row } j) \)
The test statistic has an approximate chi-square distribution with \( df = (r-1)(c-1) \).
EXAMPLE

Furniture defects are classified according to type of defect and shift on which it was made.

<table>
<thead>
<tr>
<th>Type</th>
<th>Shift</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>15</td>
<td>26</td>
<td>33</td>
<td>74</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>21</td>
<td>31</td>
<td>17</td>
<td>69</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>45</td>
<td>34</td>
<td>49</td>
<td>128</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>13</td>
<td>5</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>94</td>
<td>96</td>
<td>119</td>
<td>309</td>
</tr>
</tbody>
</table>

Do the data present sufficient evidence to indicate that the type of furniture defect varies with the shift during which the piece of furniture is produced? Test at the 1% level of significance.

H₀: type of defect is independent of shift  
H₁: type of defect depends on the shift
EXAMPLE

- Calculate the expected cell counts. For example:

\[
\hat{E}_{12} = \frac{r_1 c_2}{n} = \frac{74(96)}{309} = 22.99
\]

Test Statistic: \( \chi^2 = \sum \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \)

\[
= \frac{(15 - 22.51)^2}{22.51} + \frac{(26 - 22.99)^2}{22.99} + \ldots + \frac{(20 - 14.63)^2}{14.63} = 19.18
\]

Reject \( H_0 \) if \( \chi^2 > \chi^2_{0.01;6} = 16.812 \)

with \((r - 1)(c - 1) = 6 \) df

Reject \( H_0 \). There is sufficient evidence to indicate that the proportion of defect types vary from shift to shift.
**Example**

- Calculate the expected cell counts. For example:

<table>
<thead>
<tr>
<th>Chi-Square Test: 1, 2, 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected counts are printed below observed counts</td>
</tr>
<tr>
<td>Chi-Square contributions are printed below expected counts</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

**Chi-Sq = 19.178, DF = 6, P-Value = 0.004**

Reject $H_0$. There is sufficient evidence to indicate that the proportion of defect types vary from shift to shift.
3. Comparing Multinomial Populations

- Sometimes researchers design an experiment so that the number of experimental units falling in one set of categories is **fixed in advance**.

**Example:** An experimenter selects 900 patients who have been treated for flu prevention. She selects 300 from each of three types—no vaccine, one shot, and two shots.

<table>
<thead>
<tr>
<th></th>
<th>No Vaccine</th>
<th>One Shot</th>
<th>Two Shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td></td>
<td></td>
<td></td>
<td>( r_1 )</td>
</tr>
<tr>
<td>No Flu</td>
<td></td>
<td></td>
<td></td>
<td>( r_2 )</td>
</tr>
<tr>
<td><em>Total</em></td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>( n = 900 )</td>
</tr>
</tbody>
</table>

The column totals have been fixed in advance!
Each of the \( c \) columns (or \( r \) rows) whose totals have been fixed in advance is actually a single multinomial experiment.

The chi-square test of independence with \((r-1)(c-1)\) df is equivalent to a test of the equality of \( c \) (or \( r \)) multinomial populations.

Three binomial populations—no vaccine, one shot and two shots.

Is the probability of getting the flu independent of the type of flu prevention used?
Example

Random samples of 200 voters in each of four wards were surveyed and asked if they favor candidate A in a local election.

<table>
<thead>
<tr>
<th></th>
<th>Ward</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Favor A</td>
<td>76</td>
<td>53</td>
<td>59</td>
<td>48</td>
<td>236</td>
</tr>
<tr>
<td>Do not favor A</td>
<td>124</td>
<td>147</td>
<td>141</td>
<td>152</td>
<td>564</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
</tbody>
</table>

Do the data present sufficient evidence to indicate that the fraction of voters favoring candidate A differs in the four wards?

H₀: fraction favoring A is independent of ward
H₁: fraction favoring A depends on the ward

H₀: \( p_1 = p_2 = p_3 = p_4 \)

where \( p_i \) = fraction favoring A in each of the four wards
Calculate the expected cell counts. For example:

\[ \hat{E}_{12} = \frac{r_1 c_2}{n} = \frac{236(200)}{800} = 59 \]

Test statistic: \( X^2 = \sum \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \)

\[ \frac{(76 - 59)^2}{59} + \frac{(53 - 59)^2}{59} + \ldots + \frac{(152 - 141)^2}{141} = 10.722 \]

Reject H_0 if \( X^2 > \chi^2_{0.05} = 7.81 \) with \( (r-1)(c-1) = 3 \) df.

Reject H_0. There is sufficient evidence to indicate that the fraction of voters favoring A varies from ward to ward.
Example - Solution

Since we know that there are differences among the four wards, what are the nature of the differences?

Look at the proportions in favor of candidate A in the four wards.

<table>
<thead>
<tr>
<th>Ward</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor A</td>
<td>76/200 = 0.38</td>
<td>53/200 = 0.27</td>
<td>59/200 = 0.30</td>
<td>48/200 = 0.24</td>
</tr>
</tbody>
</table>

Candidate A is doing best in the first ward, and worst in the fourth ward. More importantly, he does not have a majority of the vote in any of the wards!
Equivalent Statistical Tests

- A **multinomial experiment** with two categories is a **binomial experiment**.
- The data from **two binomial experiments** can be displayed as a two-way classification.
- There are statistical tests for these two situations based on the statistic of Chapter 9:

\[
 k = 2
\]

| One sample: \( z = \frac{p\hat{y} - p_0}{\sqrt{\frac{p_0q_0}{n}}} \)
| Two samples: \( z = \frac{p\hat{y}_1 - p\hat{y}_2}{\sqrt{p\hat{y}q\hat{y} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \) | \( r = c = 2 \)

<table>
<thead>
<tr>
<th>Successes</th>
<th>Failures</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successes</td>
<td>Successes</td>
<td>Failures</td>
<td>Failures</td>
</tr>
</tbody>
</table>
Assumptions

When you calculate the expected cell counts, if you find that one or more is less than five, these options are available to you:

1. **Choose a larger sample size \( n \).** The larger the sample size, the closer the chi-square distribution will approximate the distribution of your test statistic \( X^2 \).

2. It may be possible to **combine one or more of the cells** with small expected cell counts, thereby satisfying the assumption.