Recall we have studied binomial data, in which each trial falls into one of 2 categories (success/failure).

Many studies allow for more than 2 categories.

Example 1: Voters are asked which of 6 candidates they prefer.

Example 2: Residents are surveyed about which part of Columbia they live in. (Downtown, NW, NE, SW, SE)

Multinomial Experiment
(Extension of a binomial experiment → from 2 to \( k \) possible outcomes)

(1) Consists of \( n \) identical trials
(2) There are \( k \) possible outcomes (categories) for each trial
(3) The probabilities for the \( k \) outcomes, denoted \( p_1, p_2, \ldots, p_k \), are the same for each trial
(and \( p_1 + p_2 + \ldots + p_k = 1 \))
(4) The trials are independent

The cell counts, \( n_1, n_2, \ldots, n_k \), which are the number of observations falling in each category, are the random variables which follow a multinomial distribution.
Analyzing a One-Way Table

Suppose we have a single categorical variable with $k$ categories. The cell counts from a multinomial experiment can be arranged in a one-way table.

Example 1: Adults were surveyed about their favorite sport. There were 6 categories.

$p_1 = \text{proportion of U.S. adults favoring football}$
$p_2 = \text{proportion of U.S. adults favoring baseball}$
$p_3 = \text{proportion of U.S. adults favoring basketball}$
$p_4 = \text{proportion of U.S. adults favoring auto racing}$
$p_5 = \text{proportion of U.S. adults favoring golf}$
$p_6 = \text{proportion of U.S. adults favoring “other”}$

It was hypothesized that the true proportions are $(p_1, p_2, p_3, p_4, p_5, p_6) = (.4, .1, .2, .06, .06, .18)$.

95 adults were randomly sampled; their preferences are summarized in the one-way table:

<table>
<thead>
<tr>
<th>Favorite Sport</th>
<th>Football</th>
<th>Baseball</th>
<th>Basketball</th>
<th>Auto</th>
<th>Golf</th>
<th>Other</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37</td>
<td>12</td>
<td>17</td>
<td>8</td>
<td>5</td>
<td>16</td>
<td>95</td>
</tr>
</tbody>
</table>

We test our null hypothesis (at $\alpha = .05$) with the following test:
Test for Multinomial Probabilities

\( H_0: \ p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0} \)

\( H_a: \) at least one of the hypothesized probabilities is wrong

The test statistic is:

\[ \chi^2 = \sum \frac{(n_i - E(n_i))^2}{E(n_i)} \]

where \( n_i \) is the observed “cell count” for category \( i \) and \( E(n_i) \) is the expected cell count for category \( i \) if \( H_0 \) is true.

Rejection region: \( \chi^2 > \chi^2_{\alpha} \) where \( \chi^2_{\alpha} \) based on \( k - 1 \) d.f.

(large values of \( \chi^2 \) => observed \( n_i \) very different from expected \( E(n_i) \) under \( H_0 \))

Assumptions: (1) The data are from a multinomial experiment. (2) Every expected cell count \( E(n_i) \) is at least 5. (large-sample test)

Finding expected cell counts: Note that \( E(n_i) = np_{i,0}. \)
For our data,

\begin{tabular}{ccc}
 i & $n_i$ & E($n_i$) \\
\end{tabular}

Test statistic value:

From Table VII:
Analyzing a Two-Way Table

Now we consider observations that are classified according to two categorical variables.

Such data can be presented in a two-way table (contingency table).

Example: Suppose the people in the “favorite-sport” survey had been further classified by gender:

Two categorical variables: Gender and Favorite Sport.

Question: Are the two classifications independent or dependent?

For instance, does people’s favorite sport depend on their gender? Or does gender have no association with favorite sport?
Observed Counts for a $r \times c$ Contingency Table
($r = \# \text{ of rows}, \ c = \# \text{ of columns}$)

<table>
<thead>
<tr>
<th>Column Variable</th>
<th>1</th>
<th>2</th>
<th>…</th>
<th>c</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>…</td>
<td>$n_{1c}$</td>
<td>$R_{1}$</td>
</tr>
<tr>
<td>Row 2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>…</td>
<td>$n_{2c}$</td>
<td>$R_{2}$</td>
</tr>
<tr>
<td>Variable</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$r$</td>
<td>$n_{r1}$</td>
<td>$n_{r2}$</td>
<td>…</td>
<td>$n_{rc}$</td>
<td>$R_{r}$</td>
</tr>
<tr>
<td>Col. Totals</td>
<td>$C_{1}$</td>
<td>$C_{2}$</td>
<td>$C_{c}$</td>
<td>:</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Probabilities for a $r \times c$ Contingency Table:

<table>
<thead>
<tr>
<th>Column Variable</th>
<th>1</th>
<th>2</th>
<th>…</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>…</td>
<td>$p_{1c}$</td>
</tr>
<tr>
<td>Row 2</td>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
<td>…</td>
<td>$p_{2c}$</td>
</tr>
<tr>
<td>Variable</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$r$</td>
<td>$p_{r1}$</td>
<td>$p_{r2}$</td>
<td>…</td>
<td>$p_{rc}$</td>
</tr>
<tr>
<td>Col. Totals</td>
<td>$p_{\text{col 1}}$</td>
<td>$p_{\text{col 2}}$</td>
<td>$p_{\text{col c}}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: If the two classifications are independent, then:

$p_{11} = (p_{\text{row 1}})(p_{\text{col 1}})$ and $p_{12} = (p_{\text{row 1}})(p_{\text{col 2}})$, etc.

So under the hypothesis of independence, we expect the cell probabilities to be the product of the corresponding marginal probabilities.
Hence the (estimated) expected count in cell \((i, j)\) is simply:

\[
\chi^2 \text{ test for independence}
\]

\(H_0\): The classifications are independent  
\(H_a\): The classifications are dependent

Test statistic:

where the expected count in cell \((i, j)\) is 

\[
\hat{E}(n_{ij}) = \frac{R_i C_j}{n}
\]

Rejection region: \(\chi^2 > \chi^2_\alpha\),  
where \(\chi^2_\alpha\) is based on \((r - 1)(c - 1)\) d.f.  
and \(r = \#\) of rows, \(c = \#\) of columns.
Note: We need the sample size to be large enough that every expected cell count is at least 5.

Example: Does the incidence of heart disease depend on snoring pattern? (Test using $\alpha = .05$.) Random sample of 2484 adults taken; results given in a contingency table:

<table>
<thead>
<tr>
<th>Snoring Pattern</th>
<th>Never</th>
<th>Occasionally</th>
<th>≈Every Night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>24</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td>No</td>
<td>1355</td>
<td>603</td>
<td>416</td>
</tr>
</tbody>
</table>

Expected Cell Counts:

| 1379 | 638 | 467 | 2484 |

Test statistic: