Two-Sample and c-Sample Tests with Categorical Data

Categorical variables are organized into categories and the data are the frequency counts of the various categories. When frequency count data are classified according to two or more variables or populations, they are called cross-tabulated data and are displayed in a contingency table.

<table>
<thead>
<tr>
<th>Categorical variable</th>
<th>category 1</th>
<th>category 2</th>
<th>. . . . .</th>
<th>category c</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>category or pop. 1</td>
<td>freq. 1</td>
<td>freq. 2</td>
<td>. . . .</td>
<td>freq c</td>
<td></td>
</tr>
<tr>
<td>Populations or</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>category or pop. 2</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Categorical variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>category r</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Column Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hypothesis Tests on such data are performed using a $\chi^2$ test whose test statistic is approximated by a chi-square distribution with degrees of freedom, $D = (r - 1) (c - 1)$ where $r = \# \text{ of rows}$ and $c = \# \text{ of columns}$

$$\chi^2 = \sum \left( \frac{\text{observed frequency} - \text{expected frequency}}{\text{expected frequency}} \right)^2 = \sum \frac{(O - E)^2}{E}$$

where expected frequency $= \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}} = \frac{R \times C}{n}$

Chi-Square Tests of Homogeneity:

- subjects from two populations are classified according to a single categorical variable
- used to determine whether two populations are similar or homogeneous with respect to the categorical variable (the proportion of the population in each category of the variable are the same for each population).
- contingency table: populations versus the categories of the variable

Chi-Square Tests of Independence:

- subjects in a single population are classified according to two different variables
- used to determine if there is any dependency between the two classifying variables
- contingency table: categories of first variable versus categories of second variable

TI-83 Procedures:

- Enter dimensions of table and frequencies into matrix A: 2nd | MATRIX | EDIT | 1:[A]
- Perform test: STAT | TESTS | C: $\chi^2$ -Test then go to Calculate and press ENTER
- Output will contain the test statistic, the p-value, and the degrees of freedom (df)
- Critical values: MATH Solver - replace tcdf by $\chi^2$cdf and use method for upper tail test.
- p-values: 2nd DISTR | 7: $\chi^2$cdf ( test statistic, 1e99, D )
Steps for $\chi^2$ Hypothesis Testing:

1. **Given**: Show Contingency Table and list $\alpha$ level of significance

2. **Hypothesis**: Need to choose between one of the following

   - **2 x 2** contingency table – looking for difference between two proportions - Test of Homogeneity
     
     $\text{Ho: } p_1 = p_2$ (homogeneous or the same)
     $\text{H}_1: p_1 \neq p_2$ (not homogeneous or not the same)

   - **2 x c** contingency table – looking for difference between c proportions - Test of Homogeneity
     
     $\text{Ho: } p_1 = p_2 \ldots = p_c$ (homogeneous or the same) must list all of the proportions in Ho
     $\text{H}_1: \text{not all proportions are equal}$ (not homogeneous or not the same)

   - **r x c** contingency table – looking for a relationship or association - Test of Independence
     
     $\text{Ho: (name of first variable) and (name of second variable) are independent}$ (no relationship)
     $\text{H}_1: \text{(name of first variable) and (name of second variable) are not independent}$ (relationship exists)

3. **Test Statistic and P-value**

   - Enter dimensions of table and frequencies into matrix A: 2\textsuperscript{nd} | MATRIX | EDIT | 1:[A]
   - Perform test: STAT | TESTS | C: $\chi^2$ -Test then go to Calculate and press ENTER
   - Output contains test statistic, the p-value, and degrees of freedom (df)

4. **Conclusion**

   - Compare the P-value to $\alpha$
     $\text{Is p-value} \leq (\alpha) \text{?}$

     State your decision: **Yes, Reject Ho** or **No, Do Not Reject Ho** based on the answer to the question.

   - Write a summary sentence that restates the question asked and includes the answer to the question.
     Begin with "**Yes, there is sufficient statistical evidence**" if you said "**Yes, Reject Ho**"
     or
     Begin with "**No, there is not sufficient statistical evidence**" if you said "**No, Do Not Reject Ho**"