Chi-square Goodness-of-Fit Test

**Application 1: One-way Frequency Table with k classes and frequency counts**
- Test if the counts conform to sampling from a multinomial population with specified multinomial probabilities $p_0$
- Test $H_0: p_i = p_{0i}, \ i = 1, 2, \ldots, k$
- The Pearson chi-square statistic is
  \[ \chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \]
  - $f_i$ is the observed frequency count in class $i$
  - $e_i$ is the expected frequency calculated under the null hypothesis (i.e., $e_i = p_{0i}N$ where $N$ is the total number of responses).

**NOTE:** Multinomial Distribution
1. $n$ independent and identical trials
2. $k$ possible outcomes in each trial ($i = 1, \ldots k$)
3. $p_i = Pr(\text{outcome } i \text{ in any trial})$

4. Let $x_i$ is the number of trials resulting in the outcome $i$, for $i = 1, \ldots k$
5. Probability distribution of $X_1, X_2, \ldots, X_k$

\[ P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{(x_1!)(x_2!)(\ldots)(x_k!)} p_1^{x_1} \cdots p_k^{x_k} \]

where $x_1, 2, \ldots$ and $x_1 + x_2 + \cdots + x_k = n$

**Example 1:**
This example examines whether the children’s hair color in a certain European geographic region has a specified multinomial distribution. The hypothesized distribution for hair color is 30% fair, 12% red, 30% medium, 25% dark, and 3% black. That is test $H_0: p_1 = .3, p_2 = .12, p_3 = .3, p_4 = .25, p_5 = .03$
Example 2: Goodness-of-Fit of a Probability Model
Suppose that an investigator wants to examine whether a Poisson distribution provides a reasonable model for the number of cell clumps per algae species in a lake. A lake sample was analyzed to determine the number of clumps of cells per microscopic field. The data are summarized below for 150 microscopic fields examined. Here, $x_i$ denotes the number of cell clumps per field and $n_i$ denotes the frequency of occurrence of fields of each cell clump count.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$n_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
</tbody>
</table>

Write a SAS program to perform a chi-square goodness-of-fit at $\alpha = 0.05$ to test the hypothesis that the observed counts were drawn from a Poisson probability distribution. Hint: For this data the Poisson mean $\lambda$ is estimated as

$$\lambda = \frac{\sum n_i x_i}{n} = \frac{486}{150} \approx 3.3.$$ 

To fit the data to a Poisson distribution the probabilities for a Poison distribution with mean=3.3 are needed. These are as follows:

$$P(X = x_i) \quad 0.037 \quad 0.122 \quad 0.201 \quad 0.221 \quad 0.182 \quad 0.120 \quad 0.066 \quad 0.034$$

data algae;
input Clumps $ Count;
label Clumps = "Cell clumps/field";
label Count = "Number of fields";
datalines;
0 6
1 23
2 29
3 31
4 27
5 13
6 8
7 13
;

proc freq data=algae order=data;
weight Count;
tables Clumps/ nocum testp=(3.7 12.2 20.1 22.1 18.2 12.0 6.6 5.1);
title 'Fitting a Poisson to cell clump counts';
run;

Part of the output extracted from executing the above program follows:

### Fitting a Poisson to cell clump counts

<table>
<thead>
<tr>
<th>Clumps</th>
<th>Frequency</th>
<th>Percent</th>
<th>Test Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>19.33</td>
<td>12.20</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>29.33</td>
<td>22.10</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>10.00</td>
<td>18.20</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>8.67</td>
<td>12.00</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5.33</td>
<td>6.60</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>8.67</td>
<td>5.10</td>
</tr>
</tbody>
</table>

Since the p-value is large, we fail to reject the null hypothesis that the cell clump counts fit a Poisson distribution.

Application 2: Two-way Frequency Table with I populations and J categories

- Testing homogeneity of several multinomial populations.
- Random samples are taken from each population and then classified by a categorical variable
- The populations are usually defined by levels of variables such as gender, age group, state, etc.
- Levels of the categorical variable form the $k$ categories of the multinomial populations
- Null hypothesis of homogeneity of populations is $H_0: p_{i1} = p_{i2} = \cdots = p_{ij}, j = 1, 2, \ldots, J$

Example:

- For example, suppose samples are drawn from two populations (say, males and females or persons below and above the age of 40) and they are grouped into three categories (say, according to three levels of support for a certain local bond issue).
- Suppose that the multinomial probabilities for each population are as given in the following table:

<table>
<thead>
<tr>
<th>Groups</th>
<th>$p_{i1}$</th>
<th>$p_{i2}$</th>
<th>$p_{i3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Populations</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$p_3$</td>
</tr>
</tbody>
</table>

- Null hypothesis of homogeneity of populations is $H_0: p_{11} = p_{12} = p_{13}, p_{21} = p_{22} = p_{23}$

Application 3: Contingency tables constructed by multiple classification of a single random sample.

1. Observations in a sample may be cross-classified by variables categorical variables.
   - ordinal values: age, income, temperature
   - nominal values: gender, marital status, region
2. These variables form natural subsets or strata of the data.
3. These may also be generated from other quantitative variables such as population or income.
   - $4$ income groups (say “low,” “middle,” or “high”)
   - according as income is $< $30,000, between $30,000 and $70,000, or $> $70,000

Example: Two-way $r \times c$ contingency table to test whether the two categorical variables are independent

<table>
<thead>
<tr>
<th></th>
<th>Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rows</strong></td>
<td>$p_{11}$</td>
</tr>
<tr>
<td></td>
<td>$p_{21}$</td>
</tr>
<tr>
<td></td>
<td>\vdots</td>
</tr>
<tr>
<td>$p_{r1}$</td>
<td>$p_{r2}$</td>
</tr>
</tbody>
</table>

1. The null hypothesis tested is of the form $H_0: p_{ij} = p_i p_j, i = 1, 2, \ldots, r, j = 1, 2, \ldots, c$.
   - $p_{ij}$ are the probabilities considering that the entire sample is from a multinomial population,
   - $p_i$ and $p_j$ are probabilities for multinomial populations defined by each categorical variable. (called the marginal probabilities)
2. The chi-square statistic for the test of this hypothesis is given by
\[ \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \]
where
- \( f_{ij} \) is the observed frequency in the \( ij \)th cell
- \( e_{ij} = f_i f_j / N \)
- \( f_i \) observed row marginal frequency
- \( f_j \) observed column marginal frequency
3. Row and Column variables are independent: \( \chi^2 \) has an asymptotic chi-square distribution with \((r-1)(c-1)\) degrees of freedom.
4. Instead of \( \chi^2 \), the likelihood ratio chi-square statistic, usually denoted by \( G^2 \), that has the same asymptotic null distribution may be computed.
5. Row and Columns are ordinal variables, the Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between them.
6. For small sample sizes and for sparse tables: Fisher’s exact test; does not depend on asymptotic theory.
7. Measures of Association between variables: May or may not depend on the chi-square test of independence.

- **Kendall’s tau-b**: the ratio of the difference between the number of concordant and discordant pairs to the total number of pairs. It is scaled to be between \(-1\) and \(+1\) when there are no ties, but not otherwise.
- **Somers’ D**: is an ordinal measure; adjusts for ties so that the value of the statistic lies between \(-1\) and \(+1\). Two values of this statistic are computed:
  1. when the row variable is considered the independent variable (Somers’ D \( C(R) \))
  2. when the column is considered the independent variable (Somers’ D \( R(C) \))
- **Asymmetric lambda** \( \lambda(R|C) \): a nominal measure; interpreted as the proportional reduction in error (PRE) in predicting the dependent (row) variable given the independent (column) variable. The range \( 0 \leq \lambda(R|C) \leq 1 \), values around 0.3 are considered high.
- **Asymmetric lambda** \( \lambda(C|R) \): similarly interpreted
- **Stuart’s tau c**: both variables in ordinal scale; adjusts for table size and corrects for ties. In the range \(-1 \leq \tau_c \leq 1 \).

**Measures of Association:**

Statistics suitable for measuring the strength of the dependency between nominal variables but are also applicable for ordinal variables. The above three measures of association are all derived from the Pearson chi-square statistic.

- **phi coefficient**: The range is \( 0 < \phi \leq \min(\sqrt{r-1}, \sqrt{c-1}) \).
- **contingency coefficient** \( C \):
  1. value of \( C \) is zero if there is no association
  2. value that is less than 1 even with perfect dependence
  3. value is dependent on the size of the table
  4. a maximum value of \( \sqrt{(r-1)(c-1)} / r \) for an \( r \times c \) table
- **Cramer’s V** is a normed measure, so its value is between 0 and 1

**Other Measures of Association:** Many of these statistical measures also require the assignment of a dependent variable and an independent variable, as the goal is to predict a rank (category) of an individual on the dependent variable given that the individual belongs to a certain category in the independent variable.

For calculating the following measures need to define pair of observations as concordant or discordant:
- the pair \((12, 2.7)\) and \((15, 3.1)\) are discordant
- the pair \((12, 2.7)\) and \((10, 3.1)\) are discordant
- **Gamma**: is a normed measure; based on the numbers of concordant and discordant pairs.
  1. no discordant pairs, Gamma is \(+1\): perfect positive association
  2. if there are no concordant pairs, Gamma is \(-1\): perfect negative association
  3. values in between \(-1\) and \(+1\) measure the strength of negative or positive association.
  4. the numbers of discordant and concordant pairs are equal, Gamma is zero; rank of the independent variable cannot be used to predict the rank of the dependent variable.

**Example:**

In the following example, factor \( B \) could be a factor such as Marital Status and factor \( A \) could be a response to a question with three possibilities. Consider the table of frequencies is as follows:

<table>
<thead>
<tr>
<th></th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( b_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>8</td>
<td>16</td>
<td>34</td>
<td>58</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>9</td>
<td>18</td>
<td>74</td>
<td>101</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>34</td>
<td>23</td>
<td>17</td>
<td>74</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>57</td>
<td>122</td>
<td>220</td>
</tr>
</tbody>
</table>

In this case, the column variable is called the independent variable with categories being \( classes \), \( groups \), or \( strata \).

data ex8;
input A $ B $ count @@;
datalines:
a1 b1 8 a1 b2 16 a1 b3 31
a2 b1 9 a2 b2 18 a2 b3 74
a3 b1 34 a3 b2 23 a3 b3 17
run;
proc freq;
  weight count;
  tables A*B/chisq expected cellchi2 nocel nopercent norow measures;
title 'Example B8: Illustration of Tables Options';
run;

1. again, the cell frequencies are directly input to proc freq instead of raw data.
2. the use of the the weight statement allows SAS to construct the two-way cross-tabulation using the cell counts.
3. the percentage of the total frequency, percentage of row frequency, and percentage of column frequency are suppressed using tables statement.
4. so only observed count \( f_i \), the expected frequency \( e_{ij} \), and its contribution to the total chi-squared statistic will be displayed in each cell.
5. the measures option requests various measures of association between the two variables be calculated.