One-way Tables and Goodness of Fit

Key concepts:

- One-way Frequency Table
- Pearson goodness-of-fit statistic
- Deviance statistic
- Pearson residuals

Objectives:

- Learn how to compute the goodness-of-fit statistics
- Understand how well an observed table of counts corresponds to the multinominal model $\text{Mult}(n, \pi)$ for some vector $\pi$.
- Introduce the FREQ procedure in SAS
Useful links:

- Understand chi-square distribution:
- SAS source on PROC FREQ: 
  http://v8doc.sas.com/sashtml/proc/z0146708.htm
- SAS source on one-way frequency tables: 
  http://v8doc.sas.com/sashtml/proc/zreq-ex2.htm

Readings:

- ch.1 Agresti (Sec. 1.4 and 1.5)
- Advanced topics: Sec. 1.4.4 and 1.4.5
A **frequency table** arises when sample units are classified into mutually exclusive categories; the number of units falling into each category is recorded.

“One way” means that units are classified according to a single categorical variable. The frequencies or counts can be arranged in a single row or column.

*Example 1.* A sample of \( n = 96 \) persons is obtained, and the eye color of each person is recorded.

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>46</td>
</tr>
<tr>
<td>Blue</td>
<td>22</td>
</tr>
<tr>
<td>Green</td>
<td>26</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>96</strong></td>
</tr>
</tbody>
</table>

Brown, blue, green, and other have no intrinsic ordering. The response variable, eye color, is therefore an *unordered categorical* or *nominal* variable.
Example 2. Hypothetical attitudes of $n = 116$ people toward war against Iraq.

<table>
<thead>
<tr>
<th>Attitude</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>35</td>
</tr>
<tr>
<td>Disagree</td>
<td>27</td>
</tr>
<tr>
<td>Agree</td>
<td>23</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>31</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>116</strong></td>
</tr>
</tbody>
</table>

The response categories in this example are clearly ordered, but no objectively defined numerical scores can be attached to the categories. The response variable, attitude, is therefore said to be an ordered categorical or ordinal variable.
Example 3. Same as above, but with additional categories for “not sure” and “refused to answer.”

<table>
<thead>
<tr>
<th>Attitude</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>35</td>
</tr>
<tr>
<td>Disagree</td>
<td>27</td>
</tr>
<tr>
<td>Agree</td>
<td>23</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>31</td>
</tr>
<tr>
<td>Not sure</td>
<td>6</td>
</tr>
<tr>
<td>Refusal</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>130</strong></td>
</tr>
</tbody>
</table>

The first four categories are clearly ordered, but the placement of “not sure” and “refusal” in the ordering is questionable. We would have to say that this response is partially ordered.
Example 4. Number of children in $n = 105$ randomly selected families.

<table>
<thead>
<tr>
<th>Number of children</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4–5</td>
<td>11</td>
</tr>
<tr>
<td>6+</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>105</strong></td>
</tr>
</tbody>
</table>

The original data, the raw number of children, has been coarsened into six categories ($0$, $1$, $2$, $3$, $4–5$, $6+$). These categories are clearly ordered, but—unlike the previous example—the categories have objectively defined numeric values attached to them. We can say that this table represents coarsened numeric data or interval variable.
Example 5. Total gross income of \( n = 100 \) households.

<table>
<thead>
<tr>
<th>Income</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>below $10,000</td>
<td>11</td>
</tr>
<tr>
<td>$10,000–$24,999</td>
<td>23</td>
</tr>
<tr>
<td>$25,000–$39,999</td>
<td>30</td>
</tr>
<tr>
<td>$40,000–$59,999</td>
<td>24</td>
</tr>
<tr>
<td>$60,000 and above</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

The original data (raw incomes) were essentially continuous. Any type of data, continuous or discrete, can be grouped or coarsened into categories.

Grouping data will typically result in some loss of information. How much information is lost depends on (a) the number of categories and (b) the question being addressed. In this example, grouping has somewhat diminished our ability to estimate the mean or median household income. Our ability to estimate the proportion of households with incomes below $10,000 has not been affected, but estimating the proportion of households with incomes above $75,000 is now virtually impossible.
Remember measurement hierarchy?
Assumptions

A one-way frequency table with $k$ cells will be denoted by the vector

$$X = (X_1, X_2, \ldots, X_k),$$

where $X_j$ is the count or frequency in cell $j$. Note that $X$ is a summary of the original raw dataset consisting of $n = \sum_{j=1}^{k} X_j$ observations.

In the fourth example (number of children), for instance, the raw data consisted of $n = 105$ integers; 24 of them were 0, 26 of them were 1, etc. For example, $X_1 = 24$ is the count in the first cell, and there are $k = 6$ cells.

We typically assume that $X$ has a multinomial distribution,

$$X \sim \text{Mult}(n, \pi),$$

where $n$ is fixed and known, and

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k)$$

may or may not have to be estimated. If $n$ is random, we can still apply the multinomial model (Lecture 4).
But we do have to worry about other kinds of model failure.

Recall, the critical assumptions of the multinomial are:

(a) the $n$ trials are independent, and

(b) the parameter $\pi$ remains constant from trial to trial.

The most common violation of these assumptions occurs when clustering is present in the data. Clustering means that some of the trials occur in groups or clusters, and that trials within a cluster tend to have outcomes that are more similar than trials from different clusters. Clustering can be thought of as a violation of either (a) or (b).
Example. Recall the first example, in which eye color was recorded for \( n = 96 \) persons.

Suppose that the sample did not consist of unrelated individuals, but that some whole families were present.

Persons within a family are more likely to have similar eye color than unrelated persons, so the assumptions of the multinomial model would be violated.
Now suppose that the sample consisted of “unrelated” persons randomly selected within Pennsylvania. In other words, persons are randomly selected from a list of Pennsylvania residents.

If two members of the same family happen to be selected into the sample purely by chance, that’s okay; the important thing is that each person on the list has an equal chance of being selected, regardless of who else is selected.

Could this be considered a multinomial situation?

For all practical purposes, yes. The sampled individuals are not independent according the common English definition of the word, because they all live in Pennsylvania.

But we can suppose that they are independent from a statistical viewpoint, because the individuals are exchangeable; before the sample is drawn, no two are a priori any more likely to have the same eye color than any other two.
Pearson and deviance statistics

Now we introduce two statistics that measure goodness of fit; that is, they measure how well an observed table $X$ corresponds to the model $\text{Mult}(n, \pi)$ for some vector $\pi$.

Consider our second example (hypothetical attitudes). We want to test the hypothesis that there is an equal probability of four attitudes; that is compare the observed frequencies to the assume model:

$X \sim \text{Multi}(n = 1116, \pi_0 = (1/4, 1/4, 1/4, 1/4))$. 
The **Pearson goodness-of-fit statistic** is

\[ X^2 = \sum_{j=1}^{k} \frac{(X_j - n\pi_j)^2}{n\pi_j}. \]

An easy way to remember it is

\[ X^2 = \sum_{j} \frac{(O_j - E_j)^2}{E_j}, \]

where \( O_j = X_j \) is the observed count and \( E_j = E(X_j) = n\pi_j \) is the expected count in cell \( j \).

The **deviance statistic** is

\[ G^2 = 2 \sum_{j=1}^{k} X_j \log \left( \frac{X_j}{n\pi_j} \right), \]

where “log” means natural logarithm. An easy way to remember it is

\[ G^2 = 2 \sum_{j} O_j \log \left( \frac{O_j}{E_j} \right). \]
In some texts, $G^2$ is also called the *likelihood-ratio test statistic*. A common mistake in calculating $G^2$ is to leave out the factor of 2 at the front; don’t forget the 2.

Note that $X^2$ and $G^2$ are both functions of the observed data $X$ and a vector of probabilities $\pi$. For this reason, we will sometimes write them as $X^2(x, \pi)$ and $G^2(x, \pi)$, respectively; when there is no ambiguity, however, we will simply use $X^2$ and $G^2$. 
Testing goodness of fit

$X^2$ and $G^2$ both measure how closely the model $Mult(n, \pi)$ “fits” the observed data.

- If the sample proportions $p_j = X_j/n$ are exactly equal to the model’s $\pi_j$ for $j = 1, 2, \ldots, k$, then $O_j = E_j$ for all $j$, and both $X^2$ and $G^2$ will be zero.

- If the sample proportions deviate from the $\pi_j$’s, then $X^2$ and $G^2$ are both positive. Large values of $X^2$ and $G^2$ mean that the data do not agree well with the model.
How can we judge the sizes of $X^2$ and $G^2$? The answer is provided by this result:

If $x$ is a realization of $X \sim Mult(n, \pi)$, then as $n$ becomes large, the distributions of both $X^2(x, \pi)$ and $G^2(x, \pi)$ approach $\chi^2_{k-1}$.

This means that we can easily test a null hypothesis $H_0: \pi = \pi_0$ against the alternative $H_1: \pi \neq \pi_0$ for some prespecified vector $\pi_0$. An approximate $\alpha$-level test of $H_0$ versus $H_1$ is:

Reject $H_0$ if $X^2(x, \pi_0)$ or $G^2(x, \pi_0)$ exceeds $\chi^2_{k-1}(1 - \alpha)$.

Here, $\chi^2_{k-1}(1 - \alpha)$ denotes the $(1 - \alpha)$th quantile of the $\chi^2_{k-1}$ distribution, the value for which the probability that a $\chi^2_{k-1}$ random variable is less than or equal to it is $1 - \alpha$. The $p$-value for this test is the area to the right of $X^2$ or $G^2$ under the $\chi^2_{k-1}$ density curve.

http://www.statsoft.com/textbook/stathome.html

http://www.ruf.rice.edu/~lane/stat_sim/chisq_theor/index.html
Here are some comments on this test.

- When $n$ is large and the model is true, $X^2$ and $G^2$ tend to be approximately equal. For large samples, the results of the $X^2$ and $G^2$ tests will be essentially the same.

- An old-fashioned rule of thumb is that the $\chi^2$ approximation for $X^2$ and $G^2$ works well provided that $n$ is large enough to have $E_j = n\pi_j \geq 5$ for every $j$. Nowadays, most agree that we can have $E_j < 5$ for some of the cells (say, 20% of them). Some of the $E_j$s can be as small as 2, but none of them should fall below 1. If this happens, then the $\chi^2$ approximation isn’t appropriate.

- In practice, it’s a good idea to compute both $X^2$ and $G^2$ to see if they lead to similar results. If the resulting $p$-values are similar, then we can be fairly confident that the large-sample approximation is working well.

- If it is apparent that one or more of the $E_j$s are too small, we can sometimes get around the problem by collapsing or combining cells until all the $E_j$s are large enough.
In most applications, we will reject the null hypothesis $X \sim Mult(n, \pi)$ for large values of $X^2$ or $G^2$. On rare occasions, however, we may want to reject the null hypothesis for unusually small values of $X^2$ or $G^2$. That is, we may want to define the $p$-value as $P(\chi^2_{k-1} \leq X^2)$ or $P(\chi^2_{k-1} \leq G^2)$. Very small values of $X^2$ or $G^2$ suggest that the model fits the data too well, i.e. that the data may have been fabricated or altered in some way to fit the model closely. This is how R.A. Fisher figured out that some of Mendel’s experimental data must have been fraudulent (e.g. see page 23 in Agresti).
Example 6 Suppose that we roll a die 30 times and observe the following table.

<table>
<thead>
<tr>
<th>Face</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
</tr>
</tbody>
</table>
We want to test the null hypothesis that the die is fair. Under this hypothesis, $X \sim Mult(n = 30, \pi_0)$ where

$$\pi_0 = \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right).$$

Under $H_0$, the expected counts are $E_j = n\pi_j = 5$ for all $j$. The goodness-of-fit measures are:

$$X^2 = \frac{(3 - 5)^2}{5} + \frac{(7 - 5)^2}{5} + \frac{(5 - 5)^2}{5}$$

$$+ \frac{(10 - 5)^2}{5} + \frac{(2 - 5)^2}{5} + \frac{(3 - 5)^2}{5}$$

$$= 9.2$$

$$G^2 = 2 \left( 3 \log \frac{3}{5} + 7 \log \frac{7}{5} + 5 \log \frac{5}{5} \right.$$

$$+ \left. 10 \log \frac{10}{5} + 2 \log \frac{2}{5} + 3 \log \frac{3}{5} \right)$$

$$= 8.8.$$ 

The $p$-values are $P(\chi^2_5 \geq 9.2) = .10$ and $P(\chi^2_5 \geq 8.8) = .12$. The fair-die model doesn’t fit the data very well, but the fit isn’t bad enough to conclude beyond a reasonable doubt that the die is unfair.

ref: SAS code lec5ex6.sas
Example. Tall cut-leaf tomatoes were crossed with dwarf potato-leaf tomatoes, and $n = 1611$ offspring were classified by phenotype.

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>tall cut-leaf</td>
<td>926</td>
</tr>
<tr>
<td>tall potato-leaf</td>
<td>288</td>
</tr>
<tr>
<td>dwarf cut-leaf</td>
<td>293</td>
</tr>
<tr>
<td>dwarf potato-leaf</td>
<td>104</td>
</tr>
</tbody>
</table>

Genetic theory says that the four phenotypes should occur with relative frequencies 9 : 3 : 3 : 1. Do the observed data support this theory?

Under the null hypothesis, the probabilities are

$$\pi_1 = \frac{9}{16}, \quad \pi_2 = \pi_3 = \frac{3}{16}, \quad \pi_4 = \frac{1}{16},$$

and the expected frequencies are

$$E_1 = 1611(9/16) = 906.2,$$

$$E_2 = E_3 = 1611(3/16) = 302.1,$$

and

$$E_4 = 1611(1/16) = 100.7.$$ The fit statistics are

$$X^2 = 1.47$$ and $$G^2 = 1.48.$$ Note that $X^2$ and $G^2$ are very close. The $p$-values from the $\chi^2_3$ distribution are approximately .69, so the data are consistent with the genetic theory.
Here is an example of how to do the computations in Minitab:

```
MTB > # put observed counts into C1
MTB > set c1
DATA> 926 288 293 104
DATA> end
MTB > # check that they add up
MTB > sum c1
   SUM      =   1611.0
MTB > # calculate expected counts into c2
MTB > set c2
DATA> 9 3 3 1
DATA> end
MTB > let c2 = 1611*c2/16
MTB > # look at results
MTB > print c1-c2

   ROW  C1     C2
      1  926    906.188
      2  288    302.062
      3  293    302.062
      4  104    100.688

MTB > # now calculate Pearson's X^2
MTB > let k1 = sum((c1-c2)**2/c2)
MTB > print k1

   K1      1.46872

MTB > # and likelihood-ratio G^2
MTB > let k2 = 2*sum(c1*log(c1/c2))
```
MTB > print k2

K2  1.47760
MTB > # get p-value for X^2
MTB > cdf k1;
SUBC> chisq 3.
   1.4687    0.3105
MTB > # p-value is 1-.3105 = .6895
MTB > # now get p-value for G^2
MTB > cdf k2;
SUBC> chisq 3.
   1.4776    0.3126
MTB > # p-value is 1-.3126 = .6874

Here's the same thing in S-PLUS/R:

> ob<-c(926,288,293,104)
> ex<-1611*c(9,3,3,1)/16
> X2<-sum((ob-ex)^2/ex)
> X2
[1] 1.468722
> G2<-2*sum(ob*log(ob/ex))
> G2
[1] 1.477587
> 1-pchisq(X2,3)
[1] 0.6895079
> 1-pchisq(G2,3)
[1] 0.6874529
Residuals.

The Pearson goodness-of-fit statistic can be written as

\[ X^2 = \sum_{j=1}^{k} r_j^2, \]

where

\[ r_j^2 = \frac{(X_j - n\pi_j)^2}{n\pi_j} = \frac{(O_j - E_j)^2}{E_j} \]

represents the contribution to \( X^2 \) by cell \( j \). The quantity

\[ r_j = \frac{X_j - n\pi_j}{\sqrt{n\pi_j}} = \frac{O_j - E_j}{\sqrt{E_j}} \]

is called the Pearson residual for cell \( j \).

The sign (positive or negative) indicates whether the observed frequency in cell \( j \) is higher or lower than expected under the model, and the magnitude indicates the degree of departure. When data do not fit a model, examination of the Pearson residuals often helps to diagnose where the model has failed.

How large should a “typical” value of \( r_j \) be? Recall that the expectation of a \( \chi^2 \) random variable is its degrees of freedom. Thus if a model is true,

\[ E(X^2) \approx E(G^2) \approx k - 1, \]
and the typical size of a single $r_j^2$ is $(k - 1)/k$. Thus, if $|r_j|$ is much larger than $\sqrt{(k - 1)/k}$—say, 2.0 or more—then the model does not appear to fit for cell $j$. 
Like $X^2$, the deviance can be regarded as the sum of squared residuals,

$$G^2 = \sum_{i=1}^{k} \epsilon_i^2,$$

where

$$\epsilon_i = \sqrt{\left| 2X_i \log \frac{X_i}{n\pi_i} \right|} \times \text{sign}(X_i - n\pi_i).$$

When the expected counts $n\pi_i$ are all fairly large (much greater than 5) the deviance and Pearson residuals resemble each other quite closely.
Effects of zero cell counts. If an \( X_i \) is zero, \( X^2 \) can be calculated without any problems, provided that the \( \pi_i \)'s are all positive. But \( G^2 \) has problems. If \( X_i = 0 \) then the deviance residual is undefined, and if we try to use the formula

\[
G^2 = 2 \sum_{i=1}^{k} X_i \log \frac{X_i}{n\pi_i},
\]

an error will result. But if we write the deviance as

\[
G^2 = 2 \log \frac{L(p; X)}{L(\pi; X)} = 2 \log \prod_{i=1}^{k} \left( \frac{X_i/n}{\pi_i} \right)^{X_i},
\]
a cell with $X_i = 0$ contributes 1 to the product and may be ignored. Thus we may calculate $G^2$ as

$$G^2 = 2 \sum_{i: X_i > 0} X_i \log \frac{X_i}{n\pi_i}.$$ 

Alternatively, we can set the deviance residuals to zero for cells with $X_i = 0$ and take $G^2 = \sum_i \epsilon_i^2$ as before. But if we do, $\epsilon_i = 0$ should not be interpreted as “the model fits well in cell $i$.” The fit could be quite poor, especially if $n\pi_i$ is large.

If any element of $\pi$ is zero, then $X^2$ and $G^2$ both break down.
Example. Below is a table of observed counts, expected counts, and residuals for the fair-die example.

<table>
<thead>
<tr>
<th>cell $j$</th>
<th>$O_j$</th>
<th>$E_j$</th>
<th>$r_j$</th>
<th>$\epsilon_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>-0.89</td>
<td>-1.75</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>5</td>
<td>+0.89</td>
<td>2.17</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>+0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
<td>+2.24</td>
<td>3.72</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
<td>-1.34</td>
<td>-1.91</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>+0.89</td>
<td>-1.75</td>
</tr>
</tbody>
</table>

The only cell that seems to deviate substantially from the fair-die model is $j = 4$. If the die is not fair, then it may be “loaded” in favor of the outcome 4. But recall that the $p$-value was about .10, so the evidence against fairness is not overwhelming.
Goodness-of-fit tests with unknown parameters

For many statistical models, we do not know the vector of probabilities $\pi$, but can only specify it up to some unknown parameters.

*Example: Hardy-Weinberg.* Suppose that a gene is either dominant ($A$) or recessive ($a$), and the overall proportion of dominant genes in the population is $p$. If we assume mating is random (i.e. members of the population choose their mates in a manner that is completely unrelated to this gene), then the three possible genotypes—$AA$, $Aa$, and $aa$—should occur in the so-called Hardy-Weinberg proportions:

<table>
<thead>
<tr>
<th>genotype</th>
<th>proportion</th>
<th>no. of dominant genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AA$</td>
<td>$\pi_1 = p^2$</td>
<td>2</td>
</tr>
<tr>
<td>$Aa$</td>
<td>$\pi_2 = 2p(1 - p)$</td>
<td>1</td>
</tr>
<tr>
<td>$aa$</td>
<td>$\pi_3 = (1 - p)^2$</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that this is equivalent to saying that the number of dominant genes that an individual has ($0$, $1$, or $2$) is distributed as $\text{Bin}(2, p)$, where $p$ is not specified.
Example: The Poisson model. Suppose that we observe the following numbers of children in \( n = 100 \) families:

<table>
<thead>
<tr>
<th>no. of children:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>19</td>
<td>26</td>
<td>29</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

Could this be a Poisson distribution? Recall that if a random variable \( Y \) has a Poisson distribution with mean \( \lambda \), then

\[
P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}
\]

for \( y = 0, 1, 2, \ldots \). Therefore, under the Poisson model, the proportions should be

<table>
<thead>
<tr>
<th>no. of children</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \pi_1 = e^{-\lambda} )</td>
</tr>
<tr>
<td>1</td>
<td>( \pi_2 = \lambda e^{-\lambda} )</td>
</tr>
<tr>
<td>2</td>
<td>( \pi_3 = \lambda^2 e^{-\lambda}/2 )</td>
</tr>
<tr>
<td>3</td>
<td>( \pi_4 = \lambda^3 e^{-\lambda}/6 )</td>
</tr>
<tr>
<td>4+</td>
<td>( \pi_5 = 1 - \sum_{j=1}^{4} \pi_j )</td>
</tr>
</tbody>
</table>

for some unknown \( \lambda \).
In each of these examples, the null hypothesis is that the multinomial probabilities depend on one or more unknown parameters in a known way. That is, the model specifies that

\[
\begin{align*}
\pi_1 &= g_1(\theta), \\
\pi_2 &= g_2(\theta), \\
\vdots \\
\pi_k &= g_k(\theta),
\end{align*}
\]

where \( g_1, g_2, \ldots, g_k \) are known functions but \( \theta \) is unknown. Let \( S_0 \) denote the set of all \( \pi \) that satisfy these constraints for some \( \theta \). We want to test

\[ H_0 : \pi \in S_0 \]

versus the alternative

\[ H_1 : \pi \in S \]

Where \( S \) denotes the simplex of all possible values of \( \pi \). (Notice that \( S \) is a \((k - 1)\)-dimensional space, but the dimension of \( S_0 \) is the number of free parameters in \( \theta \).)
The method for conducting this test is as follows.

1. Estimate $\theta$ by an efficient method (e.g. maximum likelihood). Call the estimate $\hat{\theta}$.

2. Calculate estimated cell probabilities
   \[ \hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k), \]
   where
   \begin{align*}
   \hat{\pi}_1 &= g_1(\hat{\theta}), \\
   \hat{\pi}_2 &= g_2(\hat{\theta}), \\
   &\vdots \\
   \hat{\pi}_k &= g_k(\hat{\theta}).
   \end{align*}

3. Calculate the goodness-of-fit statistics $X^2(x, \hat{\pi})$ and $G^2(x, \hat{\pi})$. That is, calculate the estimated expected cell counts $E_1 = n\hat{\pi}_1$, $E_2 = n\hat{\pi}_2$, $\ldots$, $E_k = n\hat{\pi}_k$, and find
   \[ X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} \quad \text{and} \quad G^2 = 2 \sum_j O_j \log \frac{O_j}{E_j} \]
   as usual.
An approximate test is provided by the following result.

If $X^2(x, \hat{\pi})$ and $G^2(x, \hat{\pi})$ are calculated as described above, then the distribution of both $X^2$ and $G^2$ under the null hypothesis approaches $\chi^2_\nu$ as $n \to \infty$, where $\nu$ equals the number of unknown parameters under the alternative hypothesis minus the number of unknown parameters under the null hypothesis,

$$\nu = (k - 1) - d,$$

where $d = \text{dim}(\theta)$.

The difference between this result and the previous one is that the expected cell counts $E_1, E_2, \ldots, E_k$ used to calculate $X^2$ and $G^2$ now contain unknown parameters. Because we need to estimate $d$ parameters to find $E_1, E_2, \ldots, E_k$, the large-sample distribution of $X^2$ and $G^2$ has changed; it’s still a chi-square, but the degrees of freedom have dropped by $d$. 
Example. Are the data below consistent with a Poisson model?

<table>
<thead>
<tr>
<th>no. of children:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>count:</td>
<td>19</td>
<td>26</td>
<td>29</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

Let’s test the null hypothesis that these data are Poisson. First, we need to estimate $\lambda$, the mean of the Poisson distribution. Recall that if we have an iid sample $y_1, y_2, \ldots, y_n$ from a Poisson distribution, then the ML estimate of $\lambda$ is just the sample mean, 

$$
\hat{\lambda} = n^{-1} \sum_{i=1}^{n} y_i.
$$

Based on the table above, we know that the original data $y_1, y_2, \ldots, y_n$ contained 19 values of 0, 26 values of 1, and so on; however, we don’t know the exact values of the original data that fell into the category 4+. Suppose for now that of the 13 values that were classified as 4+, ten were equal to 4 and three were equal to 5. Then the ML estimate of $\lambda$ is

$$
\hat{\lambda} = \frac{19(0) + 26(1) + 29(2) + 13(3) + 10(4) + 3(5)}{100} = 1.78.
$$

Under this estimate of $\lambda$, the expected counts for the first four cells (0, 1, 2, and 3 children, respectively) are
\[ E_1 = 100e^{-1.78} = 16.86, \]
\[ E_2 = 100(1.78)e^{-1.78} = 30.02, \]
\[ E_3 = 100(1.78)^2 e^{-1.78}/2 = 26.72, \]
\[ E_4 = 100(1.78)^3 e^{-1.78}/6 = 15.85. \]

The expected count for the 4+ cell is most easily found by noting that \( \sum_j E_j = n \), and thus
\[ E_5 = 100 - (16.86 + 30.02 + 26.72 + 15.85) = 10.55. \]

This leads to
\[ X^2 = 2.08 \quad \text{and} \quad G^2 = 2.09. \]

Since the general multinomial model has \( k - 1 = 4 \) parameters and the Poisson model has just one parameter, the degrees of freedom for this test are \( \nu = 4 - 1 = 3 \), and the \( p \)-values are
\[ P(\chi_3^2 \geq 2.08) = 0.56, \]
\[ P(\chi_3^2 \geq 2.09) = 0.55. \]

The Poisson model seems to fit well; there is no evidence that these data are not Poisson.
Below is an example of how to do these computations in S-PLUS. The function `dpois()` calculates Poisson probabilities.

```r
> ob_c(19,26,29,13,13)
> # find estimated expected probabilities
> lambdahat_(19*0+26*1+29*2+13*3+10*4+3*5)/100
> lambdahat
[1] 1.78
> kids_c(0,1,2,3)
> pihat_dpois(kids,lambdahat)
> pihat
[1] 0.1686381 0.3001759 0.2671566 0.1585129
> # attach probability for the 4+ cell
> pihat_c(pihat,1-sum(pihat))
> ex_100*pihat
> X2_sum((ob-ex)^2/ex)
> X2
[1] 2.084625
> G2_2*sum(ob*log(ob/ex))
> G2
[1] 2.088668
> 1-pchisq(X2,3)  # p-value for X^2
[1] 0.5550296
> 1-pchisq(G2,3)  # p-value for G^2
[1] 0.5542087
```
Here is the same thing in Minitab.

MTB > set c1 # enter observed counts
DATA> 19 26 29 13 13
DATA> end
MTB > # find lambdahat
MTB > let k1 = (19*0+26*1+29*2+13*3+10*4+3*5)/100
MTB > print k1

K1 1.78000
MTB > # find estimated expected probabilities
MTB > set c2
DATA> 0 1 2 3
DATA> end
MTB > pdf c2 c3;
SUBC> poisson k1.
MTB > print c2-c3

<table>
<thead>
<tr>
<th>ROW</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.168638</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.300176</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.267157</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.158513</td>
</tr>
</tbody>
</table>

MTB > # now add on the 4+ cell
MTB > let c3(5) = 1-sum(c3)
MTB > # calculate X^2
MTB > let c3 = 100*c3
MTB > let k2 = sum((c1-c3)**2/c3)
MTB > print k2

K2 2.08463
MTB > # and G^2
MTB > let k3 = 2*sum(c1*log(c1/c3))
MTB > print k3

    K3          2.08867
MTB > # get p-values
MTB > cdf k2;
SUBC> chisq 3.
         2.0846       0.4450
MTB > cdf k3;
SUBC> chisq 3.
         2.0887       0.4458
MTB > # p-values are 1-.4450 = .5550 and 1-.4458 = .5542