Comparing Counts

The following tutorial illustrates how to analyze and perform inference on two-way contingency tables using STATA. This can be done using the `tabulate` command.

A. Analyzing Contingency Tables

A survey on the severity of rodent problems in commercial poultry houses studied a random sample of poultry operations. Each operation was classified by type (EGG=1 and TURKEY=2) and by the extent of the rodent problems (1=MILD, 2=MODERATE, 3=SEVERE).

The data consists of 124 observations on the two variables `type` and `rodent`. You can access the data set by typing the command:

```
use http://www.stat.columbia.edu/~martin/W1111/Data/Poultry
```

in the STATA command window.

The type of poultry operation is the explanatory variable in this particular problem. One can calculate the conditional distribution of rodent problem for each of the two types of operation, by using the command:

```
tabulate type rodent, row
```

This results in the following output:

```
. tabulate type rodent, row

Key
  frequency row percentage

<table>
<thead>
<tr>
<th>type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>32</td>
<td>8</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>45.75</td>
<td>43.24</td>
<td>10.81</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>21</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>44.00</td>
<td>42.00</td>
<td>14.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>53</td>
<td>15</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>45.16</td>
<td>42.74</td>
<td>12.10</td>
<td>100.00</td>
</tr>
</tbody>
</table>
```

This command gives the conditional distribution of rodent problem for each of the two types of poultry operations. It also gives the marginal distribution for the variable `rodent` (bottom line of the table). If we had used the option `column`
instead of row we would have obtained the conditional distribution of poultry operation for each type of rodent problem and the marginal distribution for the variable type.

Suppose we want to test the null hypothesis that the severity of the rodent problem is independent of the type of operation. We can conduct a chi-square test for this hypothesis by typing:

```
tabulate type rodent, chi2
```

This results in the following output:

```
. tabulate type rodent, chi2

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>32</td>
<td>8</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>21</td>
<td>?</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>53</td>
<td>15</td>
<td>124</td>
</tr>
</tbody>
</table>

Pearson chi2(2) = 0.2867   Pr = 0.866
```

This gives us the two-way table for the problem together with the results of the test of significance. The output shows that $X^2=0.2867$ and the P-value=0.866. Hence, we would not reject the null hypothesis.

**B. Immediate Command**

The equivalent immediate command for displaying and analyzing two-way tables is `tabi`.

Ex. Voter registration data contains information on gender and party affiliation.

```
<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>400</td>
<td>450</td>
</tr>
<tr>
<td>Republican</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Other</td>
<td>200</td>
<td>150</td>
</tr>
</tbody>
</table>
```

$H_0$: The party affiliation for males and females follow the same distribution.

$H_a$: The distributions are not the same.

To test this hypothesis directly using STATA type the command:

```
tabi 400 450 \ 400 400 \ 200 150, chi
```

Note that we enter the data row-wise, with the symbol \ representing a new row.
This command gives rise to the following output:

```
. tabi 400 450 \ 400 400 \ 200 150, chi

<table>
<thead>
<tr>
<th>row</th>
<th>col</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>450</td>
<td>850</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>150</td>
<td>350</td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>1,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Pearson chi2(2) = 10.0840  Pr = 0.006
```

The output shows that $X^2=10.0840$ and the P-value=0.006. Hence, we would reject the null hypothesis.

**HOMEWORK:**

**Q1.** The Carstab data set gives the price class and the weight in pounds for all 1991 model four-door sedans listed in a particular auto guide. American made cars are coded with a "0" and foreign brands are coded with a "1". Cars that cost less than $15,000 are coded with a "0", cars that cost in the range $15,000-$30,000 are coded with a "1" and cars that cost more than $30,000 are coded with a "2"

Read the car data using the command:

```
use http://www.stat.columbia.edu/~martin/W1111/Data/Carstab
```

(a) Calculate the conditional distribution of price class for each of the two types of cars.

(b) How many foreign cars cost less than $15,000?

(c) Test the hypothesis that there the price class is independent of the type of car. What is the p-value of the test?

**Hand in your log file and the answers to the questions above.**

**Q2.** Do problem 26.30 from the textbook.

Solve the problem using STATA and the `tabi` command. Make sure to hand in your log file and answers to any questions in the text.