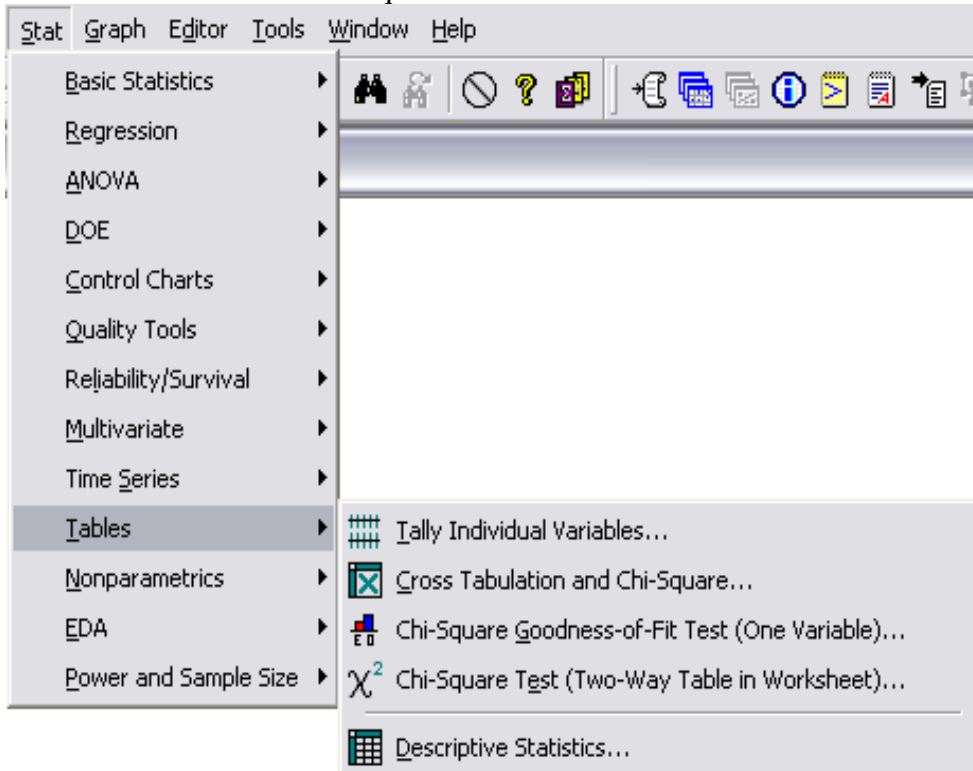


## Stat 538 - Biostatistics I - Fall 2005

**Lab 9***Discrete (Categorical) Data Analysis with Chi-square  
Goodness-of-Fit Tests and (Row-Column) Independence*

Minitab can perform analysis of categorical tables in the Stat/Tables menu. Goodness-of-Fit using “Chi-square Goodness-of-Fit Test (One Variable)” and Independence using “Cross Tabulation and Chi-Square”.



## Goodness-of-Fit Tests

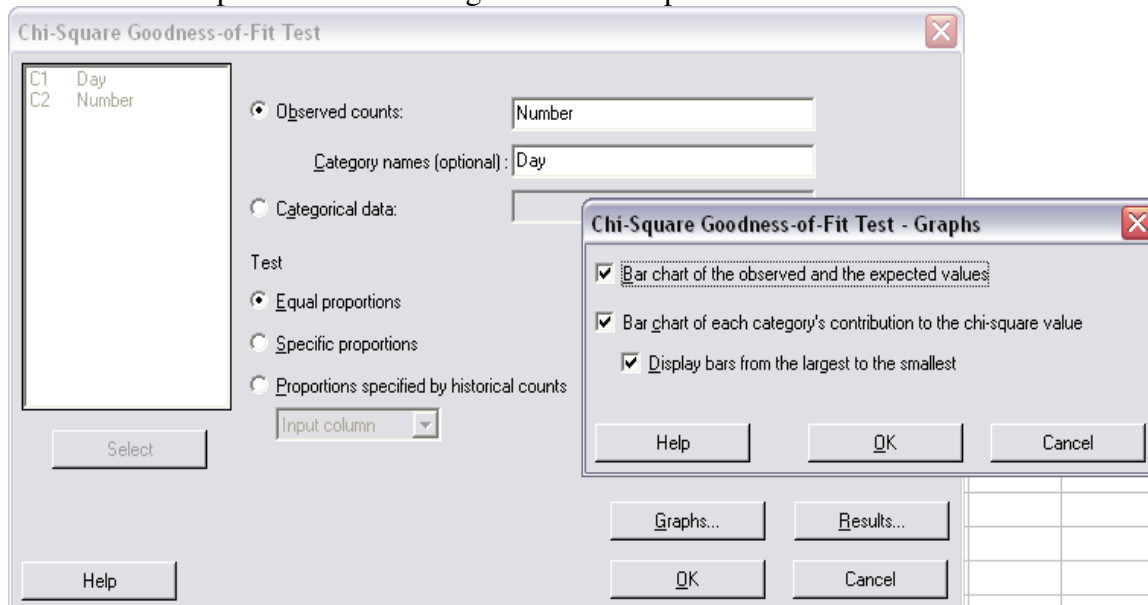
First, we learn how to use Minitab to test whether the proportion of outcomes is equal for a set of groups. Stat/Tables/Chi-square Goodness-of-Fit Test (One Variable)

Our first example comes from the Dutchess County STOP-DWI Program. *In a study of fatal car crashes, 216 cases are randomly selected from the pool in which the driver was found to have a blood alcohol content over 0.10. These cases are broken down according to the day of the week, which the results listed in the table below (also on website). At the 0.05 significance level, test the claim that such fatal crashes occur on the days of the week with equal frequency.*

Note: I put a number at the beginning of the day category so that they would remain in this order in the output. Otherwise, they are listed alphabetically.

Day	Number
01Sun	40
02Mon	24
03Tue	25
04Wed	28
05Thu	29
06Fri	32
07Sat	38

Below we select Number for the observed counts and Day as the category names (so it labels our plots). Under the graph button, I've selected both plots, and the "category's contribution to the chi-square" chart will be sorted with the largest difference between observed and expected value coming first. The output follows.



The table below gives the category names, the observed values, the proportion we're testing (all equal at  $1/7^{\text{th}}=0.142857$ ), the expected values ( $[\text{total count}] * [\text{test proportion}]$ ), and contribution to chi-square ( $(\text{obs-exp})^2/\text{exp}$ ). Note: the further the observed value is from the expected value, the larger the value of chi-square.

We are testing the hypothesis

H0: all population proportions are equal

HA: at least one population proportion is different.

Based on the p-value below of 0.284, we fail to reject H0 concluding there is insufficient evidence to conclude the population proportions of fatal crashes differ for days of the week.

### Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Number

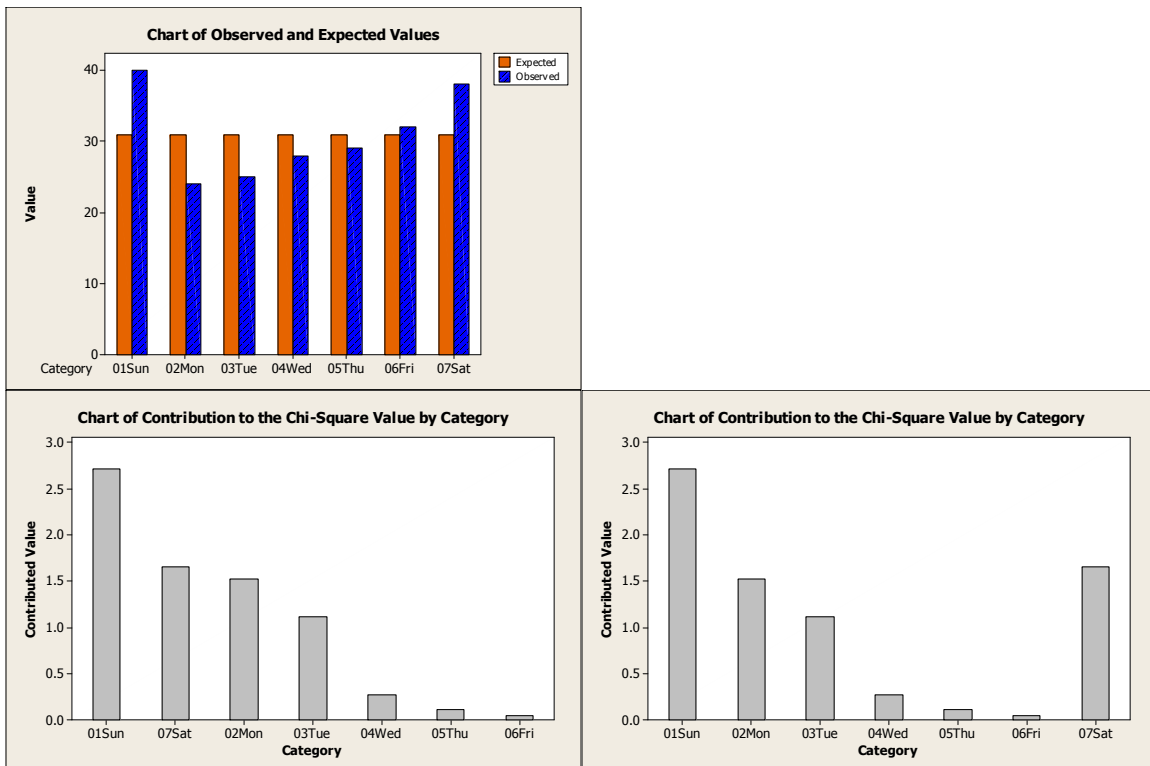
Using category names in Day

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
01Sun	40	0.142857	30.8571	2.70899
02Mon	24	0.142857	30.8571	1.52381
03Tue	25	0.142857	30.8571	1.11177
04Wed	28	0.142857	30.8571	0.26455
05Thu	29	0.142857	30.8571	0.11177
06Fri	32	0.142857	30.8571	0.04233
07Sat	38	0.142857	30.8571	1.65344

N	DF	Chi-Sq	P-Value
216	6	7.41667	<b>0.284</b>

The first chart shows the expected values in orange (constant at 30.8571), and how the observed values compare. We see that Sunday has more fatal crashes than expected, and Monday has fewer than expected. In fact, if we start on Monday, the number of fatal crashes appears to increase through Sunday.

The second and third charts are the same, first it is sorted descending and then kept in group order. They show each observation's contribution to chi-square. See that the values are large when observed is different from expected, and small when they are similar.



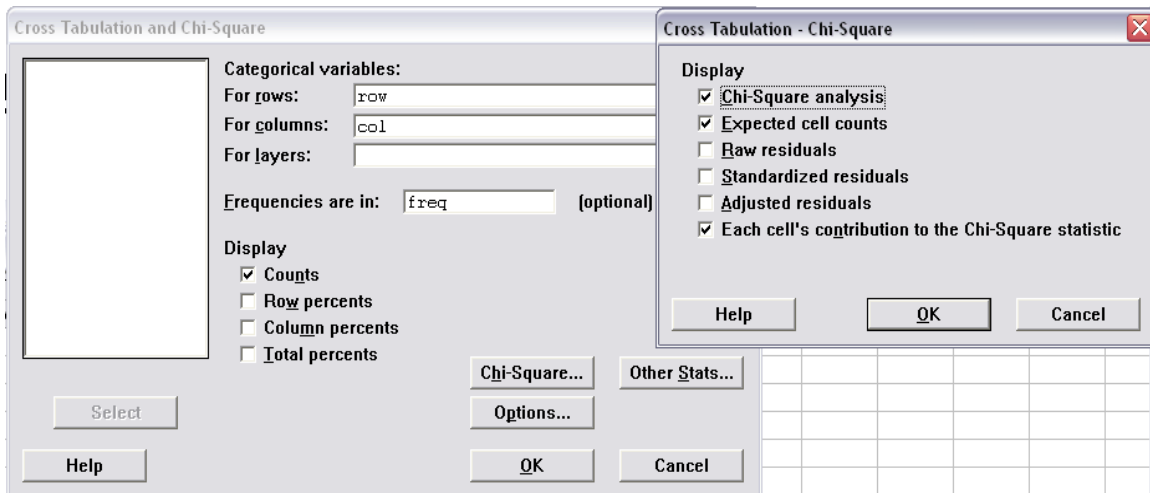
Our next example is a test of a 2x2 table. For a 2x2 table, a test for equal proportions is the same as a test of independence between row and column variables.

*Nicorette is a chewing gum designed to help people stop smoking cigarettes. Tests for adverse reactions yield the results given in the table (based on data from Merrell Dow Pharmaceuticals, Inc.). At the 0.05 significance level, test the claim that the treatment (drug or placebo) is independent of the reaction (whether or not mouth or throat soreness was experienced).*

	Drug	Placebo
Mouth or throat soreness	43	35
No mouth or throat soreness	109	118

```
freq  row  col
43    no_soreness  drug
109   no_soreness  placebo
35    soreness     drug
118   soreness     placebo
```

Use Stat/Tables/Cross Tabulation and Chi-Square. Input the row, column, and frequency columns. Under the Chi-Square button, select Chi-Square analysis, and other output if you wish.



Below is the resulting table from the options selected above. Just below the table is a key for the values in each cell. In our case, first is listed the observed count, then the expected count, then the contribution to chi-square. The right-most column and bottom row are for totals (which is what you'd use if doing this by hand).

We are testing

H0: rows independent from columns (drug not related to soreness)

HA: rows and columns related (drug related to soreness)

The p-value for the test below is large at 0.279, so we fail to reject H0, concluding the drug does not effect soreness of the mouth or throat.

### Tabulated statistics: row, col

Using frequencies in freq

Rows: row	Columns: col		
	drug	placebo	All
no_soreness	43	109	152
	38.9	113.1	152.0
	0.4383	0.1506	*
soreness	35	118	153
	39.1	113.9	153.0
	0.4355	0.1496	*
All	78	227	305
	78.0	227.0	305.0
	*	*	*

Cell Contents:           Count  
                               Expected count  
                               Contribution to Chi-square

Pearson Chi-Square = 1.174, DF = 1, **P-Value = 0.279**  
 Likelihood Ratio Chi-Square = 1.176, DF = 1, P-Value = 0.278

Because in the 2x2 case this test is the same as test whether the proportions for each group is the same, below are the results from a 2P two-sample test of population proportions. First I do the test down the columns, then across the rows (see the X and N values for what I mean). Notice that both tests have the same p-value (to rounding error). The conclusions from these three tests are the same because they are testing the same thing. Our next example has more rows and columns, so we rely on our cross tabulation chi-square test.

### Down columns:

#### Test and CI for Two Proportions

Sample	X	N	Sample p
1	43	152	0.282895
2	35	153	0.228758

Difference = p (1) - p (2)

Estimate for difference: 0.0541366

95% CI for difference: (-0.0436215, 0.151895)

Test for difference = 0 (vs not = 0): Z = 1.09 **P-Value = 0.278**



Additionally, the output below does not include the row labels, but because they are in the same order as the data, there is no confusion.

We are testing

H0: rows independent from columns (region not related to appeal)

HA: rows and columns related (region related to appeal)

Because the p-value is 0.000, we reject H0 in favor of HA, concluding that region and product appeal are related.

### Chi-Square Test: Like, Dislike, Uncertain

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	Like	Dislike	Uncertain	Total
1	30	15	15	60
	20.43	26.81	12.77	
	4.488	5.201	0.391	
2	10	30	20	60
	20.43	26.81	12.77	
	5.321	0.380	4.099	
3	40	60	15	115
	39.15	51.38	24.47	
	0.019	1.445	3.664	
Total	80	105	50	235

Chi-Sq = 25.008, DF = 4, **P-Value = 0.000**

Having concluded a difference between the proportions of levels of appeal for each region, let's see which regions are different from each other. To do this, we'll transpose the columns (so that our region rows are now columns) using Data/Transpose Columns, select the data columns to transpose, click OK, then fill in the column headers.

Now, we can test whether NE is different from SE by only selecting those columns in the Chi-square analysis.



**Chi-Square Test: NE, W**

Expected counts are printed below observed counts  
 Chi-Square contributions are printed below expected counts

	NE	W	Total
1	30	40	70
	24.00	46.00	
	1.500	0.783	
2	15	60	75
	25.71	49.29	
	4.464	2.329	
3	15	15	30
	10.29	19.71	
	2.161	1.127	
Total	60	115	175

Chi-Sq = 12.364, DF = 2, P-Value = 0.002

**Chi-Square Test: SE, W**

Expected counts are printed below observed counts  
 Chi-Square contributions are printed below expected counts

	SE	W	Total
1	10	40	50
	17.14	32.86	
	2.976	1.553	
2	30	60	90
	30.86	59.14	
	0.024	0.012	
3	20	15	35
	12.00	23.00	
	5.333	2.783	
Total	60	115	175

Chi-Sq = 12.681, DF = 2, P-Value = 0.002

This pairwise comparison is easier with the data left in the table. I originally did this with stacked data for the homework solutions, and in that case you have to copy the original data into sets of three columns, then remove the rows with the observations for the groups you're not interested in. You leave only the rows for two groups of observations and perform the chi-square test. I was happy it was simple when leaving it in table form.

Hint: In the homework, you do not have to transpose the data to do the last question.