Chapter 14

Comparing proportions - Chi-square (χ^2) tests

Contents

14.1 Introduction

In previous chapters, the Analysis of Variance (ANOVA) was used as the general methodology for testing hypotheses about population MEANS. In many cases, the response variable can often be dichotomized into classes. Often the proportion of the population that falls into the various classes is of interest. There are two types of hypotheses that are often considered

- Are the observed proportions comparable to a known set of proportions. For example, a sample of habitat usage by animals is sampled, and we wish to know if this is the same as the actual proportion of available habitat as determined using a GIS. Notice that the proportion of habitat available by type is known EXACTLY because the entire area is covered by the GIS system.
- Are the observed proportions across treatment groups the same. For example, both males and female animals could be sample for habitat preference and it is of interest to know if males and females behave similarly. In this case proportions in both groups are determined from samples and NOT known exactly - they include sampling error.

For both cases, the standard methodology to test hypotheses about population proportions is called a **chi-square test**, sometimes shown as χ^2-test.

Warning: The use of the generic term *chi-square test* is unfortunate as there are many statistical tests where the final test statistic is also compared to a chi-squared distribution which are NOT tests of proportions. As well, not all tests of proportions lead to chi-square tests.

Some examples of the type of studies to be considered in this chapter are:

- Resource selection studies where a random sample of habitat usages by a group of animals is compared to a known habitat classification.
- Resource selection studies where random samples are selected for two or more groups and proportions between the groups are to be compared.
- An experiment is conducted to investigate the effects of dissolved gases upon fish survival. Fish are randomly assigned to either a control group or a experimental group. At the end of the experiment each fish is classified as live or dead. Is the proportion of live/dead the same in both treatment groups?
- People are cross-classified by the amount of education (e.g. high school, college, post-graduate) and their socio-economic status (low, middle, high). Is there a relationship?
- Animals are cross-classified by breeding status and age. Is there a relationship between the two variables?
- Students are cross classified by their usage of marijuana and alcohol compared to their parents' usage. Is there evidence of a relationship?

Warning: This chapter limited to single factor CRD. As mentioned many times in the previous chapters, it is important that the analysis match the experimental design. This chapter will review ONLY the analysis of single-factor completely randomized designs – the analysis of blocked designs, multi-factor designs, split-plot designs, or sub-sampling designs when testing proportions is VERY complex. Please consult suitable help before proceeding. For example, if the data are paired, then McNemar's test, rather than the chi-square tests of this chapter, would be used. See, for example, Hoffman, J.I.E.

1976. The incorrect use of Chi-sqare analysis for paired data. Clin. Exp. Journal, 24, 227-229. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538510/?tool=pubmed`.

Warning: Sacrificial pseudo-replication. Hurlbert (1984) found that pseudo-replication is a particular problem with experiments that compare proportions - particularly sacrificial pseudo-replication where studies are (inappropriately) pooled before conducting a chi-square test. A related problem of pooling is Simpson's Paradox. Both are explored in more detail later in this chapter. For example, what is the difference between a study with 10 fish in each tank and 1 fish in each of 10 tanks?

Warning: Power and sample size hard to determine. There is no easy way to determine power and sample size for these tests as the hypothesis is so nebulous. It is possible to use *JMP* to determine power and sample size when testing for changes in a single proportion (e.g. differences in death rates among treatments). Please seek help if you are interested in doing this.

Warning: This is NOT compositional data. This chapter does NOT deal with the analysis of compositional data. For example, you may observe birds for 2 hour periods and compute the proportion of time spent in various activities. Despite the similarities to contingency tables (where row or column percentages add to 100%), compositional data is NOT analyzed using the chi-square test discussed in this chapter.

Warning: Avoid duplicate counts. The techniques presented in this chapter are NOT applicable if a subject can appear in more than one category. For example, you will often see surveys where respondents are asked to check all the sports that they enjoy watching; or animals are watched and ALL of the activities are recorded. A key assumption of the analyses in this chapter is that every experimental unit appears once and only once in the contingency table. Methods have been developed to deal with multi-response data – seek help if you have a research problem in this area.

14.2 Response variables vs. Frequency Variables

The data for a test of proportions can come in two formats:

- individual records
- summarized records

If data are individual records, each observation in the dataset consists of the classification for a single individual. For example, consider a study that examined the condition factor of deer after a particularly nasty winter. As each deer is spotted, its sex and condition factor are noted where the condition factor is classified into two classes (good and poor). Here are some data:

Individual	Sex	Condition
1	m	good
2	f	poor
3	m	poor
4	m	good
5	m	poor
6	f	good
7	f	good
...		

Each deer has two variables recorded: sex (nominal scale) and condition factor (ordinal scale). The study would consider the sex to be the explanatory variable (the X variable) and the condition factor to be the response variable (the Y variable).

In some cases, the data are summarized and information on individual animals is not given. For example, the summary information for the above study could take the form:

Sex	Condition	Count
m	good	10
f	good	10
m	poor	23
f	poor	18

The variables recorded are sex (nominal scale), condition factor (ordinal scale), and a count. The study would now consider sex to be the explanatory variable (the X variable), the condition factor to be the response variable (the Y variable), and the count as a frequency variable. **The count variable is NOT the response variable** – it merely summarizes the information from a number of individuals. This often causes confusion when it comes to an analysis as it is tempting to consider the count variable as the response. The easiest way to avoid this confusion is to ask what variable would be the response variable if information on individuals was recorded – the same variable will be the response variable when summary data are given.

JMP will requires that both the X and Y variables are nominal or ordinal scale.

14.3 Overview

Regardless of the statistical procedure used or the type of response variable examined, it is extremely important that the analysis match the experimental/survey design. As noted earlier, most computer packages assume that you have done the matching of design with the analysis.

Recall that a single factor completely randomized design has the following structures:

- **treatment structure**. A single factor with at least 2 levels describes the population groups for which comparisons of the population parameter are to be made. As there is only one factor, the *treatments* correspond to the levels of the factor.
- **experimental unit structure**. There is a single size of experimental unit; the observational unit is the experimental unit.
- **randomization structure**. Each experimental unit is randomly and independently assigned to each treatment group, or in the case of a analytical studies, the units are a simple random sample from the relevant treatment groups.

There are two variables of interest:

- The explanatory variable (X variable) that defines the treatment groups. This should be nominal or ordinal scale. [1]

[1] If the case where a single group is being compared to known proportions, the explanatory variable is not used

- The response variable (Y variable) which is the category that the unit is classified into. This should be nominal or ordinal scale.

If the data has been summarized, there will also be a *frequency* variable which counts the number of individuals for each combination of the explanatory and response variables.

The response and explanatory variables can also be interval or ratio variables if the interval or ratio variables are first broken into categories and the categories are used as nominal or ordinal variables. For example, after measuring a person's height, you could create a new variable called 'height group' that had the values *short*, *average*, or *tall*; or the actual amount of fiber in a cereal was used to classify a cereal as a *low*, *medium*, or *high* fiber cereal.

The hypotheses of interest can be written in several (equivalent) ways:

- **Equality of proportions**. The null hypothesis is that the proportion in each the response variable categories is the same for all treatment groups. The alternate hypothesis is that the set of proportions differs somewhere among the treatment groups.
- **Independence**. The null hypothesis is that the response category is *independent* of the the treatment group. The alternate hypothesis is that there is some sort of (ill defined) association.

Both of the above hypotheses are exactly equivalent and can be used interchangeably.

It is possible to write the hypotheses in terms of population parameters. Let π_{ij} be the proportion of treatment population i that is found in response category j. There are G treatment groups and K response categories. Note that $\sum_j \pi_{ij} = 1$, i.e., the proportions must add to 100% for each treatment population. The null hypothesis is:

$$\begin{aligned} H :& \pi_{11} = \pi_{21} = \pi_{31} = \ldots = \pi_{G1} \quad and \\ & \pi_{12} = \pi_{22} = \pi_{32} = \ldots = \pi_{G2} \quad and \\ & \ldots \\ & \pi_{1K} = \pi_{2K} = \pi_{3K} = \ldots = \pi_{GK} \end{aligned}$$

The basic summary statistic in a chi-square analysis is the **contingency table** which summarizes the number of observations for each combination of the explanatory and response variable. Sample percentages are often computed (denoted as p_{ij} to distinguish them from the π_{ij} in the population). The basic summary graph is the **mosaic chart** which consists of side-by-side segmented bar charts.

The test statistic can be computed in several ways:

- **Pearson chi-square** is a comparison of the observed and expected counts when the null hypothesis is true.
- **Likelihood ratio chi-square** is a weighted average of the ratio of the observed and expected counts.
- **Fisher's exact test** looks at how many contingency tables are more unusual than the one observed here.

The first two test statistics often are very similar and there is no objective way to choose between them. [It can be shown that in large samples the two are equivalent]. However, both test statistics assume that the counts in the contingency table are reasonably large. An often quoted rule of thumb is that the expected count in each cell of the contingency table should be at least 5. The Fisher test statistic may be more suitable for very sparse tables, i.e., tables with lots of 0's, 1's or 2 counts. For tables that are not sparse, Fisher's test statistic often takes too long to compute even with modern computers.

Regardless of which test procedure is used, the ultimate end-point is the p-value. This is interpreted in exactly the same way as in all previous studies, i.e., it is a measure of how consistent the data is with the null hypothesis. **It does NOT measure the probability that the hypothesis is true!** As before, small p-values are strong evidence that the data are not consistent with the hypothesis – leading to a conclusion against the null hypothesis.

As in the ANOVA procedure, if strong enough evidence is found against the null hypothesis, you still don't know where the hypothesis apparently failed. Some sort of multiple comparison procedure is required. Unfortunately, statistical theory is not yet well developed in this area because of the complication that the percentages within a treatment group must add to 100%. Consequently, it can be confusing (as you will see later) to try and see which category is different among treatment groups. About the best that can be done is to look at the individual cell contributions to the overall test-statistic (called the individual cell chi-squared values) to see where the large contributions have occurred. A rough rule of thumb is that entries larger than 4 or 5 indicate some problems with the fit.

In the ANOVA, a great deal of emphasis was placed on estimation of effect sizes (e.g., confidence intervals for differences). It is possible to construct such intervals for contingency tables, but these are more readily done when *log-linear models* are used – these are beyond the scope of this course.

14.4 Single sample surveys - comparing to a known standard

The simplest survey is when data is collected from a single population via a single sample and the question of interest compares the proportions of each category in the population with a KNOWN set of proportions (the standard). For example, to test if a coin is fair, a series of tosses are performed. The proportion of heads for that coin is tested against the standard of .50 for a fair coin.

The sampling design is a simple random sample from the relevant population. For each unit in the sample, a categorical response (e.g. heads or tails) is recorded. The sample statistics of interest are the observed proportion of these categories (e.g. the observed proportion of heads or tails in the sample). Notice that unlike previous chapters, the summary statistics is a sample proportion (refer back to the creel survey where the the proportion of angling parties with sufficient life jackets was found).

As before, never present naked estimates – some measure of precision needs to be reported, either the *se* or a confidence interval or both.

The hypothesis of interest is H: $\pi_{heads} = .50; \pi_{tails} = .50$. Again notice that we are now testing a population PROPORTION rather than a population mean. A measure of discrepancy of the data relative to the hypothesis is computed (the chi-square test statistic) and this eventually leads to a p-value which is interpreted in the usual fashion. Again, never report a naked p-value – always report an effect size size along with a measure of precision.

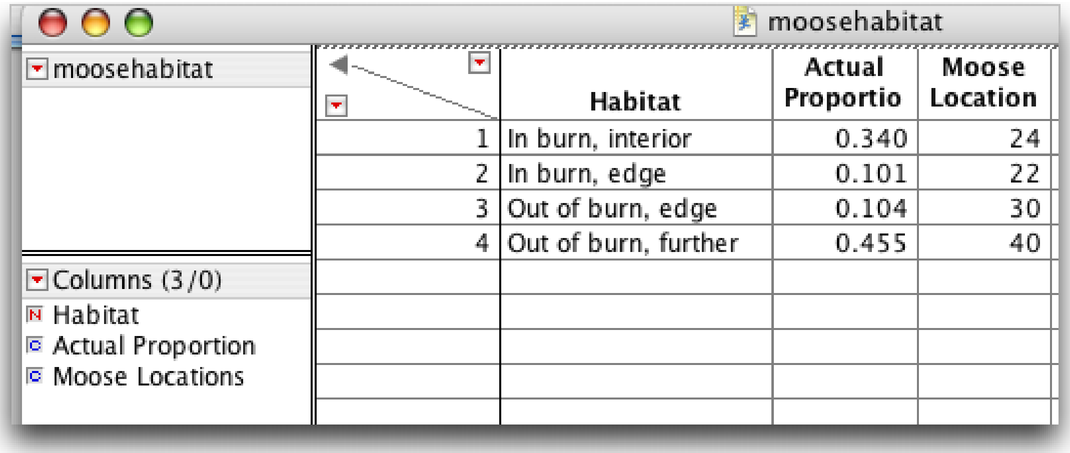
14.4.1 Resource selection - comparison to known habitat proportions

Neu et al. (1974) considered selection of habitat by Moose *Alces alces* in the Little Sioux Burn area of Minnesota in 1971-72. The authors determined the proportion of four habitat categories (see table below) using an aerial photograph, and latter classified moose usage of 116 moose during later aerial surveys.

Habitat	Actual Proportion	Moose observations
In burn, interior	.340	24
In burn, edge	.101	22
Out of burn, edge	.104	30
Out of burn, further	.455	40
Total	1.000	116

The data are entered in the usual fashion in three columns. The data are available in a *JMP* datafile *moosehabitat.jmp* in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

Here is the raw data.



moosehabitat

moosehabitat

Columns (3/0)
Habitat
Actual Proportion
Moose Locations

	Habitat	Actual Proportio	Moose Location
1	In burn, interior	0.340	24
2	In burn, edge	0.101	22
3	Out of burn, edge	0.104	30
4	Out of burn, further	0.455	40

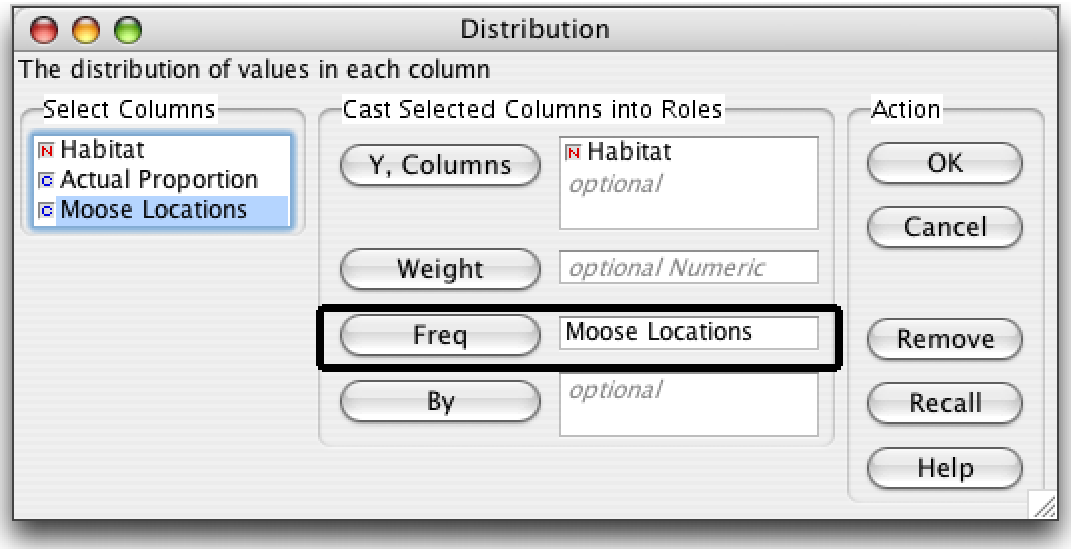
Note that the habitat variable should be be defined as a nominal scale.

The habitat variable is the response variable. The number of moose locations is the frequency of each habitat observed - this is summarized data as noted in the introduction.

Does this survey meet the assumptions required for this procedure:

- The proportions of actual habitat are determined by an aerial photograph and are known exactly (or with negligible error). This seems satisfied.
- The moose sighting are a random sample of moose? The paper is not clear on how moose were sighted but these represent observations collected over a number of days and different animals.
- Is each unit measured only once? Moose are indistinguishable so it unknown if the same moose is measured multiple times. However, the survey was over a number of days, the moose move considerable distances, so it seems reasonable that the observations are independent even though the same moose may have been measured multiple times.

The *Analyze->Distribution* platform is used to analyze the habitat data.

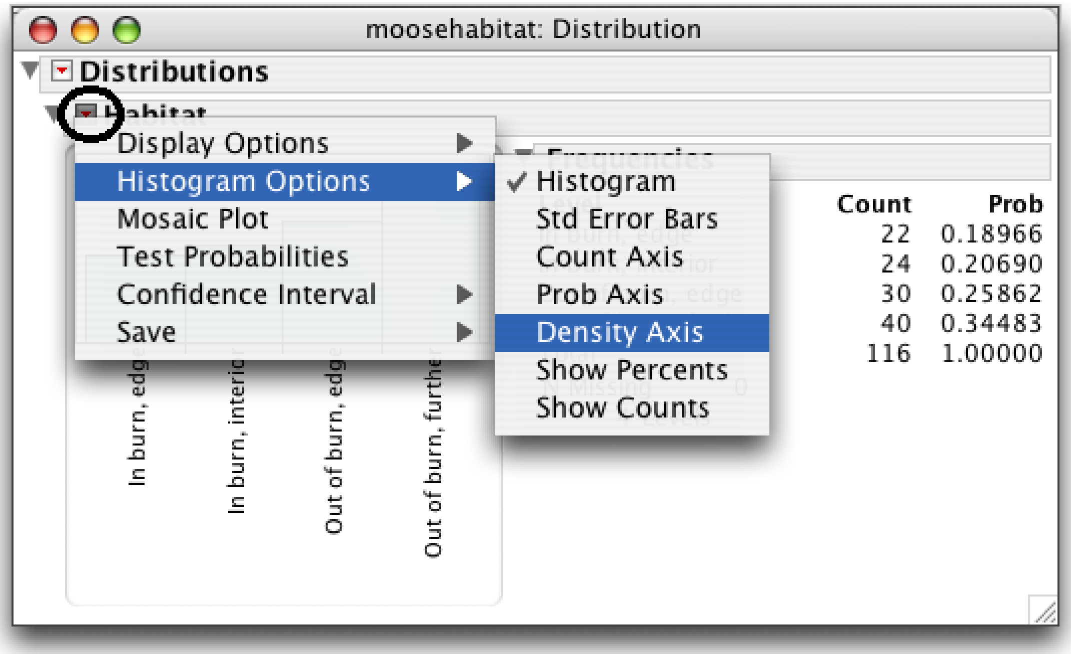


Y, Columns: Habitat

Freq: Moose Locations

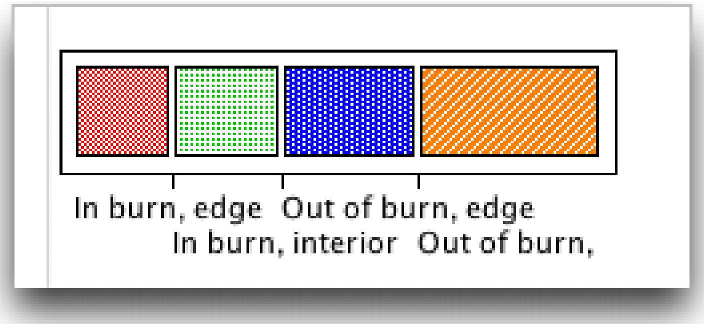
Besure to specify that the moose locations are the frequency variable.

This gives a histogram of the proportion of the various habitat categories. By using the red-triangle pop down menu, the histogram display can be customized.



Count	Prob
22	0.18966
24	0.20690
30	0.25862
40	0.34483
116	1.00000

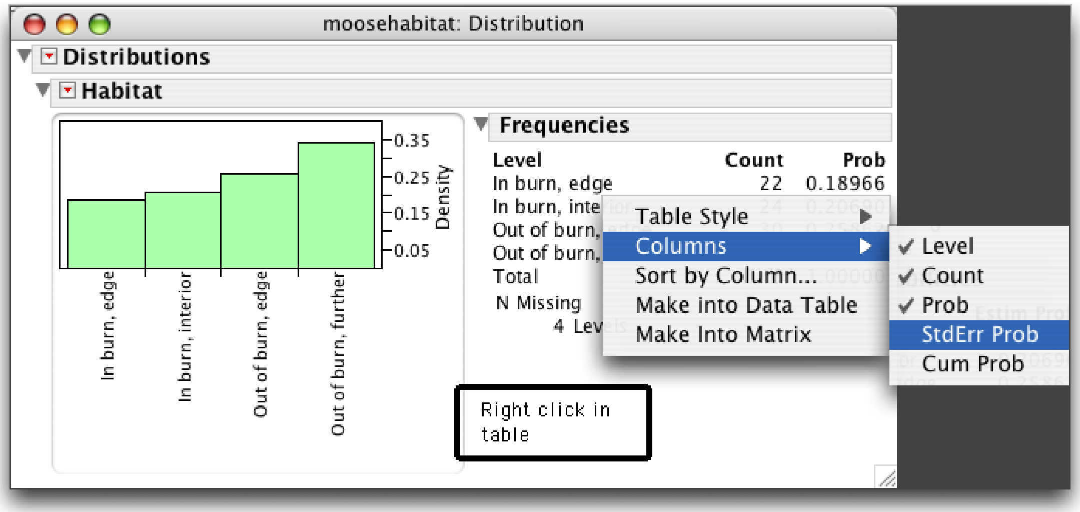
These side-by-side bar charts is not total satisfactory as the bars must add to 100%. A mosaic plot can be created in *JMP* or *R* but is not easily created in *SAS* – contact me for details.



The mosaic plot is a segmented bar chart that divides the bar into habitat usage by the number of moose points observed. It adds to 100% is an alternate display to side-by-side histograms.

We now compute the raw proportions and their standard errors:

By right clicking in the table of observed probabilities, the standard error of each observed proportion of habitat usage can be displayed.

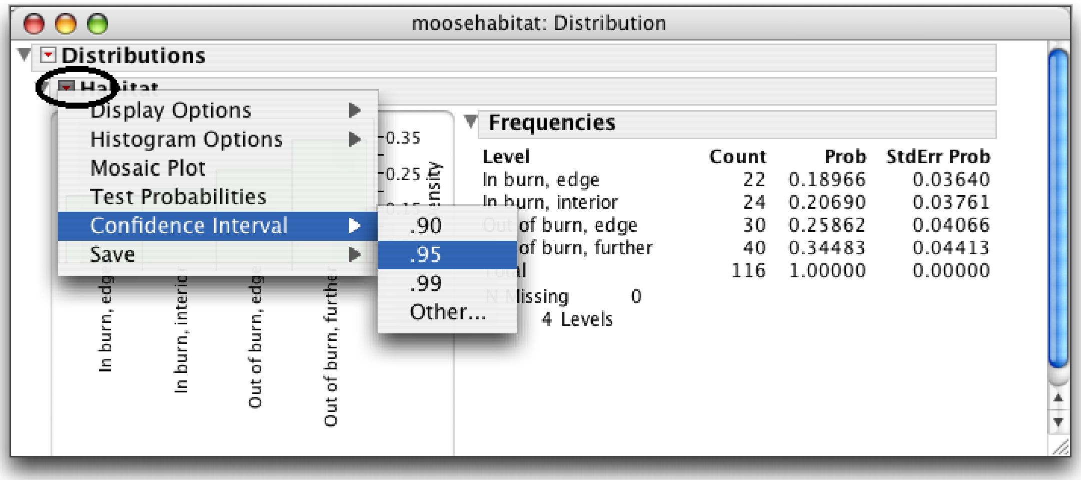


Frequencies

```
Level              Count     Prob
In burn, edge         22  0.18966
In burn, inte
Out of burn,
Out of burn,
Total
 N Missing
        4 Lev
```

Right click in table

Finally, confidence intervals for each proportion are obtained by using the red-triangle pop-down menus

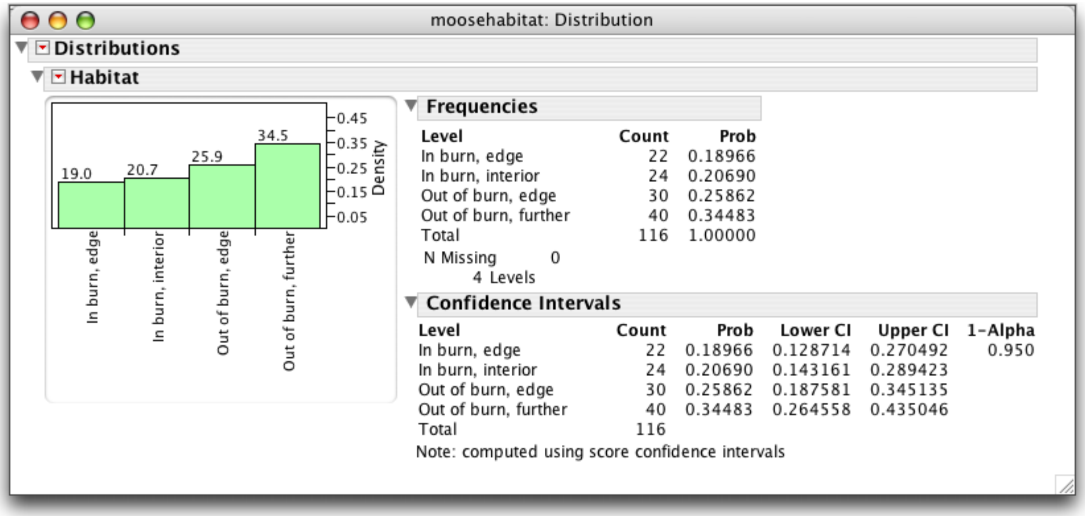


Frequencies

Level	Count	Prob	StdErr Prob
In burn, edge	22	0.18966	0.03640
In burn, interior	24	0.20690	0.03761
of burn, edge	30	0.25862	0.04066
of burn, further	40	0.34483	0.04413
l	116	1.00000	0.00000

lissing 0
4 Levels

This gives the final display:



Frequencies

Level	Count	Prob
In burn, edge	22	0.18966
In burn, interior	24	0.20690
Out of burn, edge	30	0.25862
Out of burn, further	40	0.34483
Total	116	1.00000

N Missing 0
4 Levels

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
In burn, edge	22	0.18966	0.128714	0.270492	0.950
In burn, interior	24	0.20690	0.143161	0.289423	
Out of burn, edge	30	0.25862	0.187581	0.345135	
Out of burn, further	40	0.34483	0.264558	0.435046	
Total	116				

Note: computed using score confidence intervals

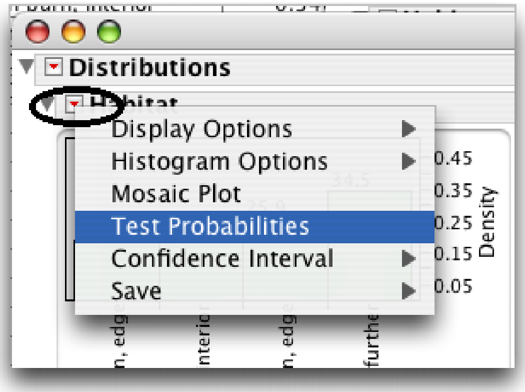
We observe that the sample proportion in *Out of burn, further* appears to be less than the actual proportion as measured by the aerial photographs as its confidence interval does not include the actual proportion of 45%. Similarly, the observed proportion of moose in the *Out of burn, edge* appears to be higher than the actual proportion of 26% as again the confidence interval does not cover the actual value. In fact, none of the confidence intervals cover the actual values.

At this point, one could stop, but often a formal hypothesis test is conducted. The hypothesis of interest is

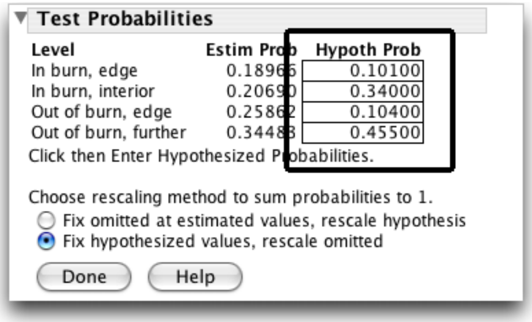
$$H : \pi_{in\ burn\ interior} = .34; \pi_{in\ burn\ edge} = .101; \pi_{out\ of\ burn\ edge} = .104; \pi_{out\ of\ burn\ further} = .455$$

where π represent the proportion of habitat classes used by all the moose (and not just the sample observed). The alternate hypothesis is that the used proportions do not match the known proportions of habitat.

The hypothesis is tested in *JMP* using the red-triangle pop-down menu:



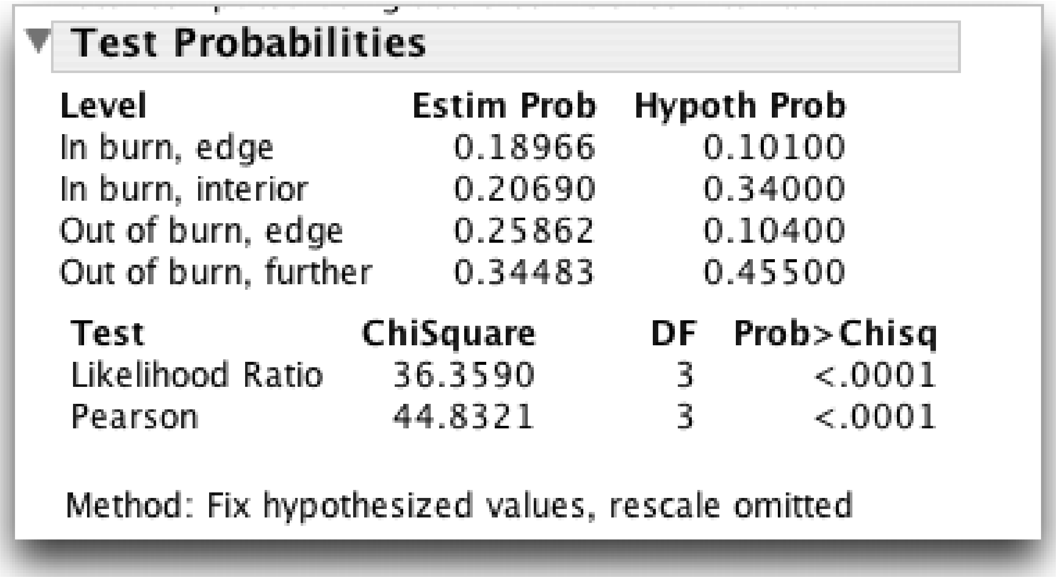
Notice that the hypothesized probabilities are entered into the appropriate boxes.



```
Test Probabilities
Level                  Estim Prob   Hypoth Prob
In burn, edge          0.18966      0.10100
In burn, interior      0.20690      0.34000
Out of burn, edge      0.25862      0.10400
Out of burn, further   0.34483      0.45500
Click then Enter Hypothesized Probabilities.

Choose rescaling method to sum probabilities to 1.
○ Fix omitted at estimated values, rescale hypothesis
● Fix hypothesized values, rescale omitted
```

After the *Done* button is pressed, the results are shown:



```
Test Probabilities
Level                  Estim Prob   Hypoth Prob
In burn, edge             0.18966       0.10100
In burn, interior         0.20690       0.34000
Out of burn, edge         0.25862       0.10400
Out of burn, further      0.34483       0.45500

Test               ChiSquare    DF   Prob>Chisq
Likelihood Ratio     36.3590     3       <.0001
Pearson              44.8321     3       <.0001

Method: Fix hypothesized values, rescale omitted
```

The *p*-values from both the Pearson and Likelihood ratio test are very small. There is strong evidence against the hypothesis that moose use the habitat in the same proportion as availability. The confidence intervals clearly show where the preferences and avoidances lie.

The book:

> Manly, B.F.J, McDonald, L.L., Thomas, D.L., McDonald, T.L. and Erickson, W.P. (2002). Resource Selection by Animals. Kluwer Academic Publisers

continues with this example and shows how to compute resource selection probabilities that can be interpreted as the probability that a randomly selected moose will select each habitat at a randomly chosen point in time.

14.4.2 Example: Homicide and Seasons

Is there a relationship between weather and violent crime? In the paper:

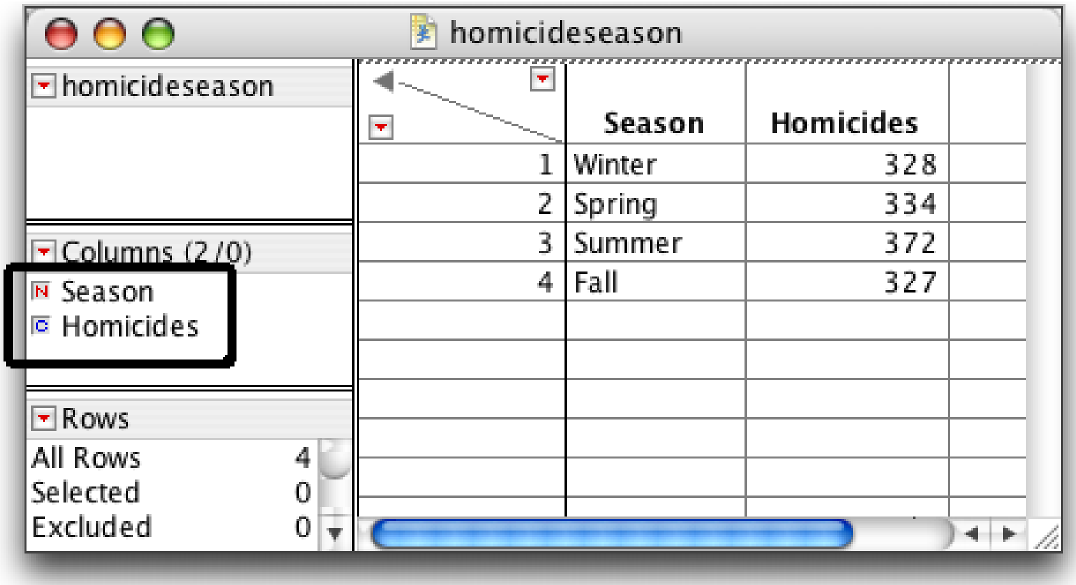
> Cheatwood, D. (1988).
> Is there a season for homicide? Criminology, 26, 287-306.
> `http://dx.doi.org/10.1111/j.1745-9125.1988.tb00842.x`

the author classified 1361 homicides in Baltimore from 1974-1984 by season[2] as follows:

Winter	328
Spring	334
Summer	372
Fall	327

Is there evidence of difference in the number of homicides by season?

The data are entered in *JMP* and are available in a *JMP* data file *homicideseason.jmp* in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.



	Season	Homicides
1	Winter	328
2	Spring	334
3	Summer	372
4	Fall	327

Let

[2] Refer to Table 7 of the paper

- π_{winter}, π_{spring}, π_{summer} and π_{fall} represent the true proportion of homicides.

We first formulate the null hypothesis. This is always in terms of population parameters:

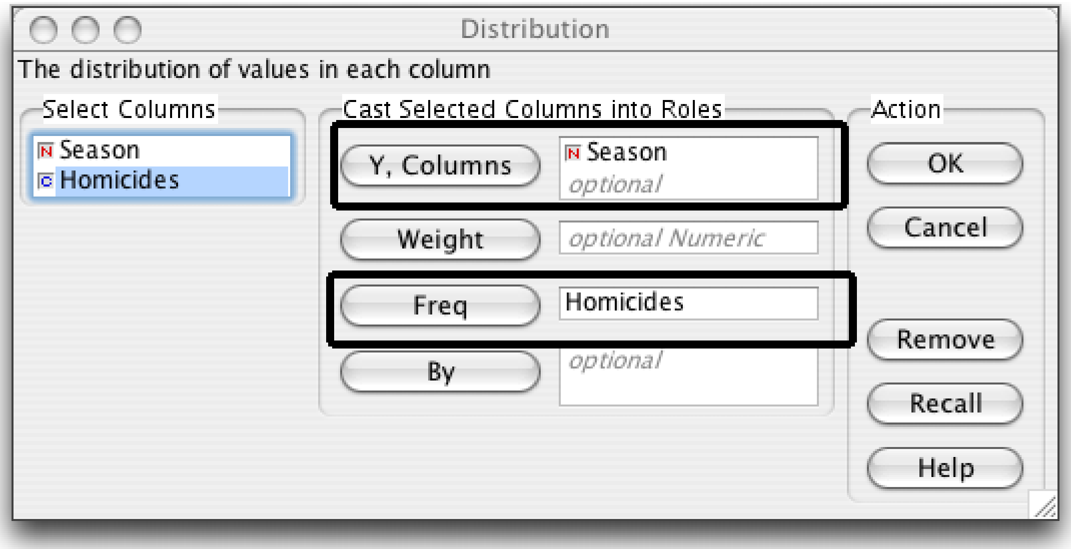
$$H : \pi_{winter} = \pi_{spring} = \pi_{summer} = \pi_{fall} = .25$$

where π_{season} represent the true proportion of all homicides in each of the seasons.

The assumptions seem to be satisfied for this analysis

- The proportion of the year in each season is known exactly
- Each homicide is measured once and only once – presumably multiple murders are counted only once.

We proceed as before



Distribution

The distribution of values in each column

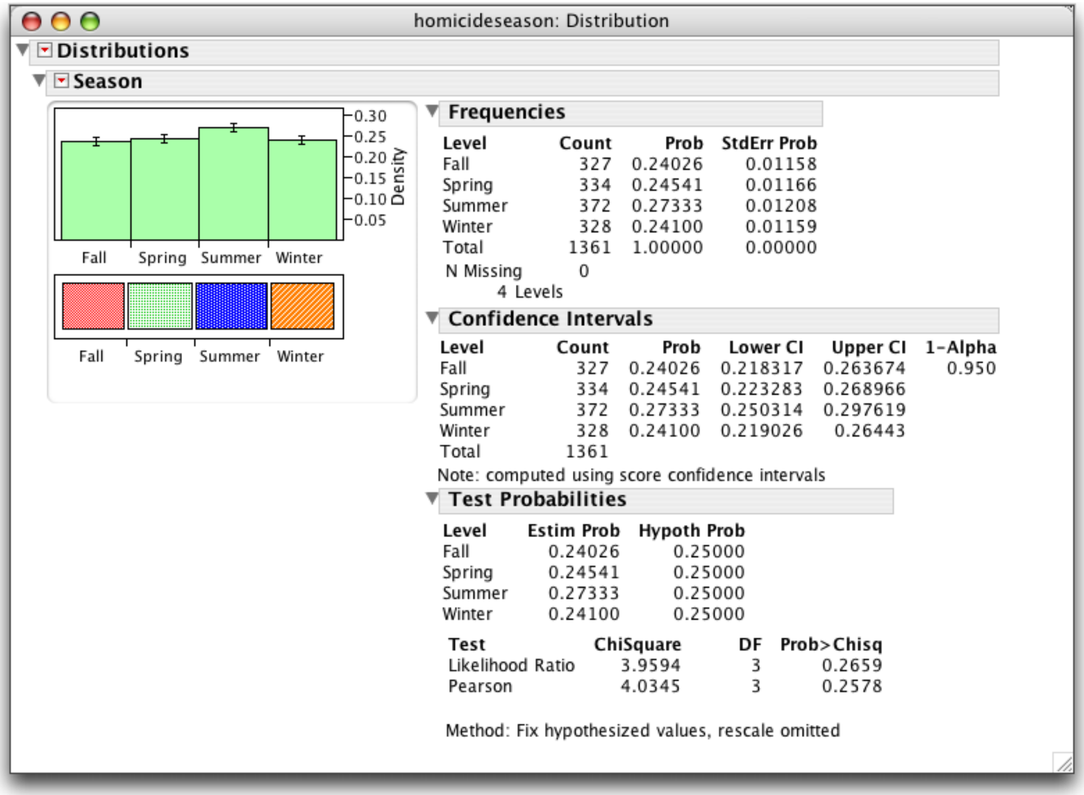
Y, Columns: Season

Weight: optional Numeric

Freq: Homicides

By: optional

and obtain the following output:



```
Frequencies
Level      Count    Prob   StdErr Prob
Fall         327  0.24026     0.01158
Spring       334  0.24541     0.01166
Summer       372  0.27333     0.01208
Winter       328  0.24100     0.01159
Total       1361  1.00000     0.00000
N Missing      0
    4 Levels

Confidence Intervals
Level      Count    Prob   Lower CI  Upper CI  1-Alpha
Fall         327  0.24026  0.218317  0.263674    0.950
Spring       334  0.24541  0.223283  0.268966
Summer       372  0.27333  0.250314  0.297619
Winter       328  0.24100  0.219026   0.26443
Total       1361
Note: computed using score confidence intervals

Test Probabilities
Level    Estim Prob  Hypoth Prob
Fall        0.24026      0.25000
Spring      0.24541      0.25000
Summer      0.27333      0.25000
Winter      0.24100      0.25000

Test              ChiSquare   DF  Prob>Chisq
Likelihood Ratio     3.9594    3      0.2659
Pearson              4.0345    3      0.2578

Method: Fix hypothesized values, rescale omitted
```

As the *p*-value is quite large, there is no evidence that homicides occur unequally over the year. All of the confidence intervals for the individual proportions include the hypothesized value of 0.25. Note that we have made adjustment in computing the confidence intervals for each season for the multiple-testing problem (e.g. similar to the Tukey-adjustment following ANOVA). This can be done, but is beyond the scope of this course.

14.5 Comparing sets of proportions - single factor CRD designs

In many cases, the questions of interest lies if the proportions in the categories is the same across different (treatment) groups. For example, is the proportion of heads the same for pennies produced in the United States and Canada? Now the hypothesis of interest is written as:

$$H : \pi_{heads,US} = \pi_{heads,Canada}; \text{ AND } \pi_{tails,US} = \pi_{tails,Canada}$$

The actual proportion of heads/tails in either population really is NOT of interest (we are not testing if both population are fairs) – we are only interested in testing if BOTH population have the same set of proportions.

In this chapter, only single factor complete randomized designs are considered. This implies that every object is independent of every other object and that the observational unit is the same as the experimental unit. More advanced designs are covered in more advance courses.

A VERY COMMON ERROR is to use the simple χ^2 test presented in this chapter even for more complex designs!

Note that if there are only two categories, an alternative (and equivalent) approach is to use logistic regression and logistic ANOVA. Please refer to the appropriate chapter for more details.

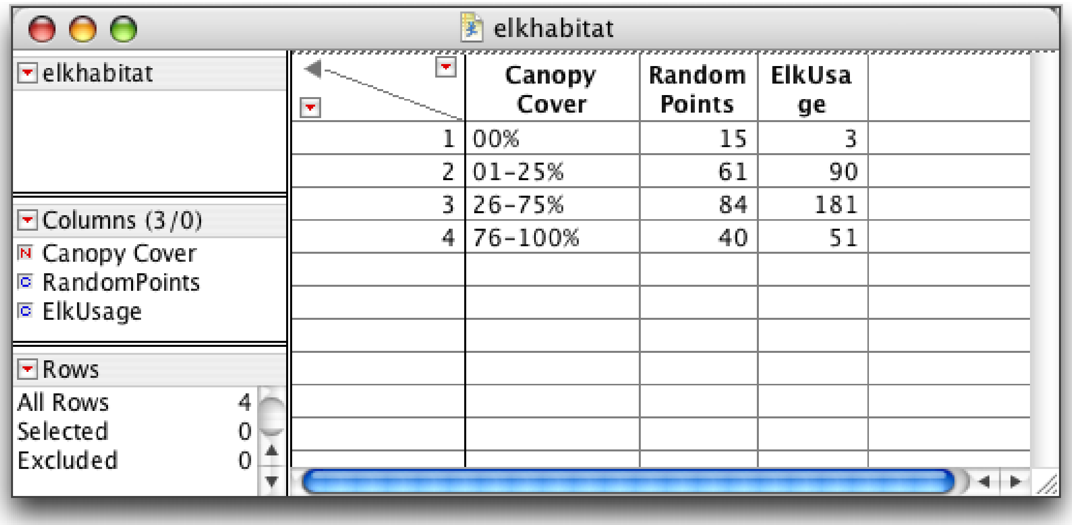
14.5.1 Example: Elk habitat usage - Random selection of points

Marcum and Loftsgaarden (1980) use data from Marcum (1975) where 200 randomly selected points were located on a map of a study area which contained a mixture of forest-canopy cover classes. These were broken into 4 cover classes as shown below.

At the same time, 325 location of elk in the region were located as shown below.

Cover Class	Random Location	Elk Location
00%	15	3
01-25%	61	90
26-75%	84	181
76-100%	40	51
Total	200	325

The data are available in a *JMP* data file called *elkhabitat.jmp* available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`:



	Canopy Cover	Random Points	ElkUsage
1	00%	15	3
2	01-25%	61	90
3	26-75%	84	181
4	76-100%	40	51

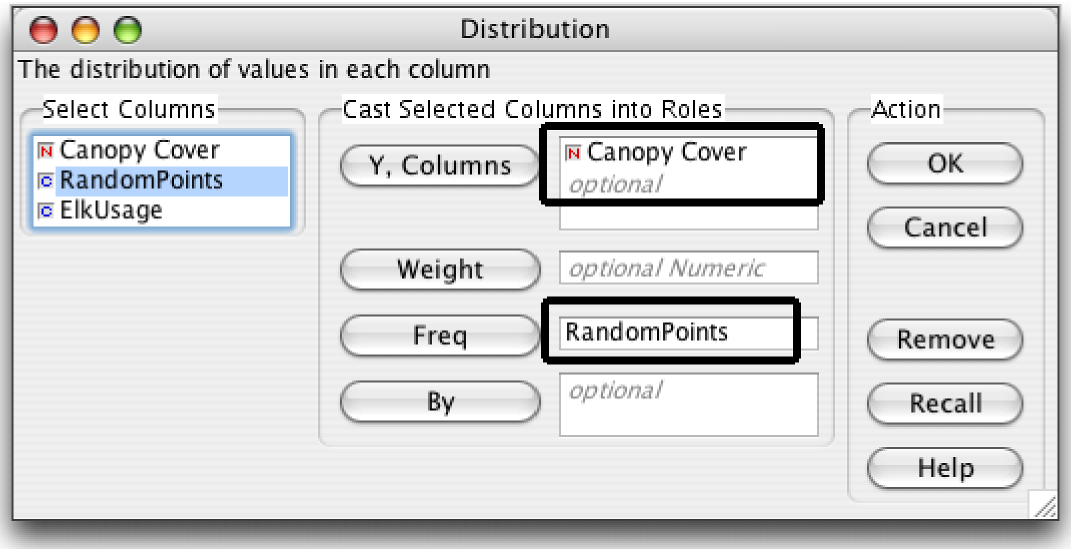
The hypothesis of interest is if the proportion of elk usage of the canopy class matches the actual proportion. Notice how this differs from the moose habitat example earlier in the chapter as the actual proportion of cover class is NOT known exactly and is only estimated using the sample of random points. In modern times with GIS and other aerial methods, there is little to be gained by sampling a random sample of points rather than measuring the entire region.

We again consider how the data were collected to see if it meets the assumptions of the analysis:

- Is this a CRD? The article is not clear, but it appears that if the same elk were measured multiple times, that the timings are far enough apart to treat the locations as independent measurements.
- Observational and experimental unit the same? Yes, the elk or random point.

- Randomization structure is random in location and animals.

We start by looking the individual breakdowns of the proportions obtained by the *Analyze->Distribution* platform applied separately to each set of points. Note that in both cases, the response (Y) variable is the canopy class and that the count is the *Frequency* variable. For example, the analysis of the random points starts as:



The distribution of values in each column

Select Columns: Canopy Cover, RandomPoints, ElkUsage

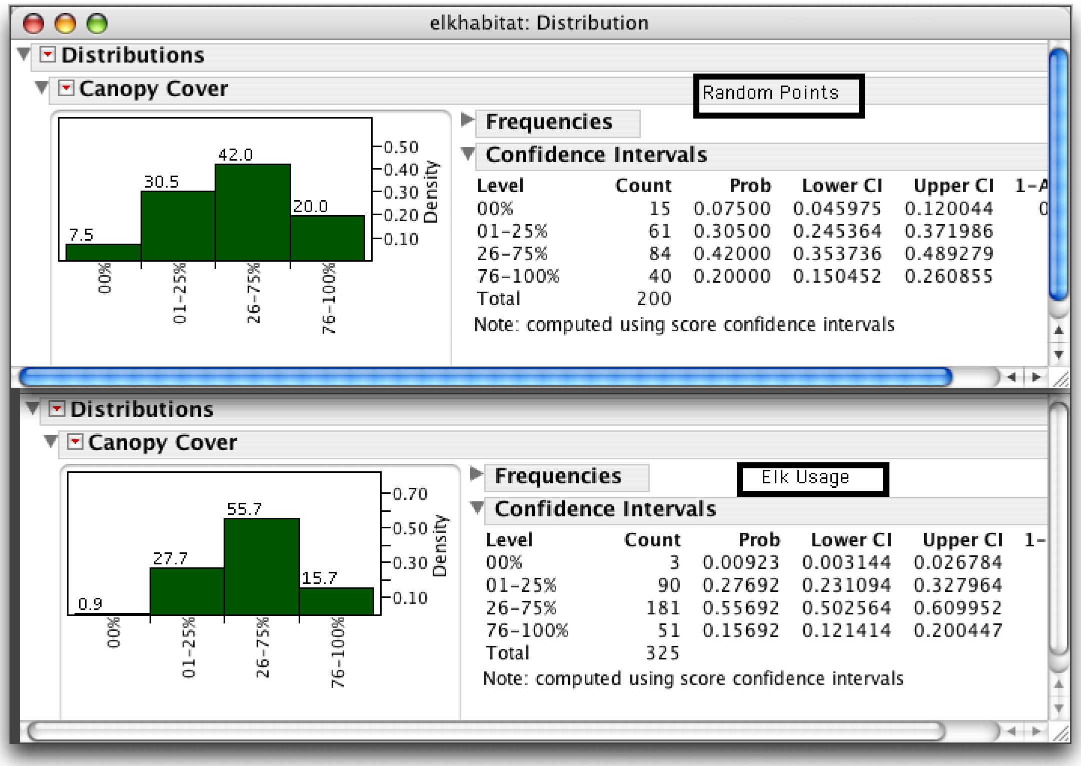
Y, Columns: Canopy Cover

Weight: optional Numeric

Freq: RandomPoints

By: optional

[The ElkUsage is analyzed similarly.] After manipulating the displays [see the example on Elk Habitat Usage earlier in the chapter], we obtain the following two displays:

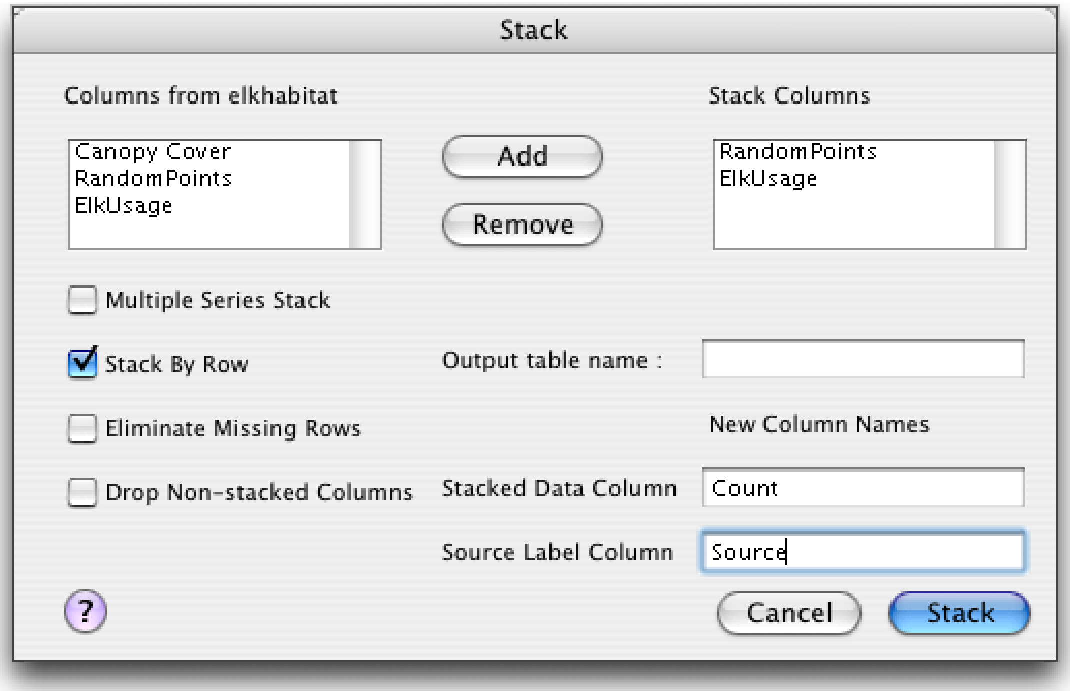


```
Confidence Intervals
Level      Count    Prob      Lower CI   Upper CI  1-A
00%           15  0.07500  0.045975  0.120044     0
01-25%        61  0.30500  0.245364  0.371986
26-75%        84  0.42000  0.353736  0.489279
76-100%       40  0.20000  0.150452  0.260855
Total        200
Note: computed using score confidence intervals
```

```
Confidence Intervals
Level      Count    Prob      Lower CI   Upper CI  1-
00%            3  0.00923  0.003144  0.026784
01-25%        90  0.27692  0.231094  0.327964
26-75%       181  0.55692  0.502564  0.609952
76-100%       51  0.15692  0.121414  0.200447
Total        325
Note: computed using score confidence intervals
```

We notice that the confidence intervals for the 26-75% cover class from the Elk doesn't overlap that from the random points, nor for the 0% cover class.

In order to do a statistical test, we must first stack the data into a format similar to that used for ANOVA:



Stack

Columns from elkhabitat
- Canopy Cover
- RandomPoints
- ElkUsage

Stack Columns
- RandomPoints
- ElkUsage

□ Multiple Series Stack

☑ Stack By Row

□ Eliminate Missing Rows

□ Drop Non-stacked Columns

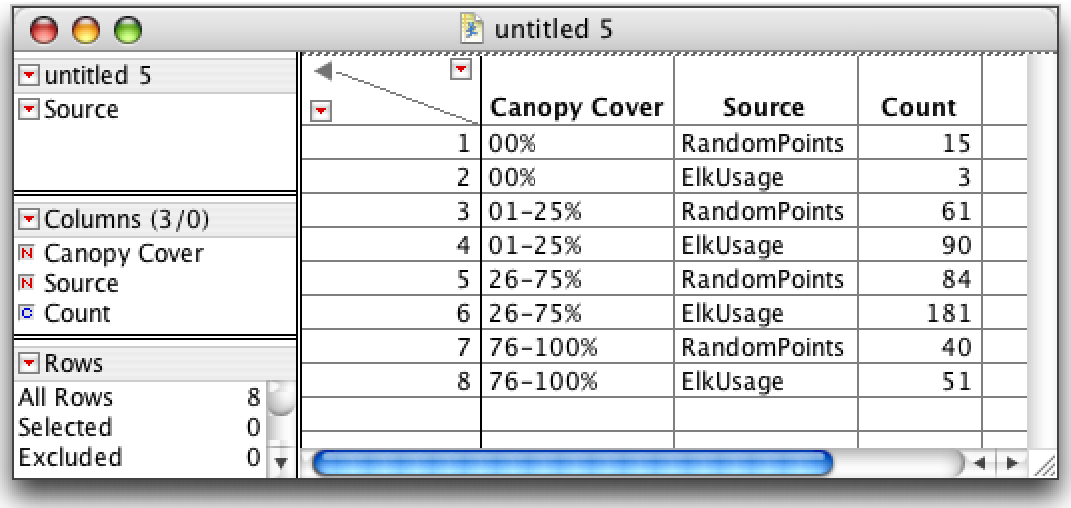
Output table name :

New Column Names

Stacked Data Column: Count

Source Label Column: Source

which gives the following data structure:



untitled 5

- untitled 5
- Source

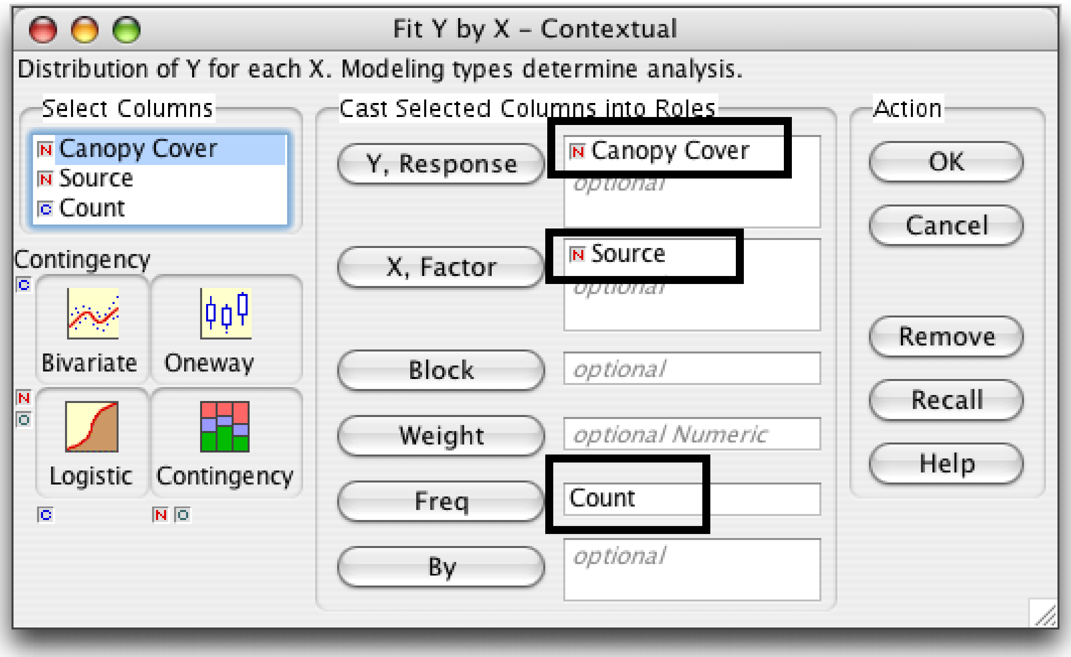
Columns (3/0)
- Canopy Cover
- Source
- Count

Rows

All Rows	8
Selected	0
Excluded	0

	Canopy Cover	Source	Count
1	00%	RandomPoints	15
2	00%	ElkUsage	3
3	01-25%	RandomPoints	61
4	01-25%	ElkUsage	90
5	26-75%	RandomPoints	84
6	26-75%	ElkUsage	181
7	76-100%	RandomPoints	40
8	76-100%	ElkUsage	51

Now use the *Analyze->Fit Y-by-X* platform and notice that the response (Y) variable is the cover class, the explanatory (X) variable is the source of the data (elk vs. random sample of points), and the frequencies of the points under the two groups is the *Frequency* variable.



Fit Y by X – Contextual

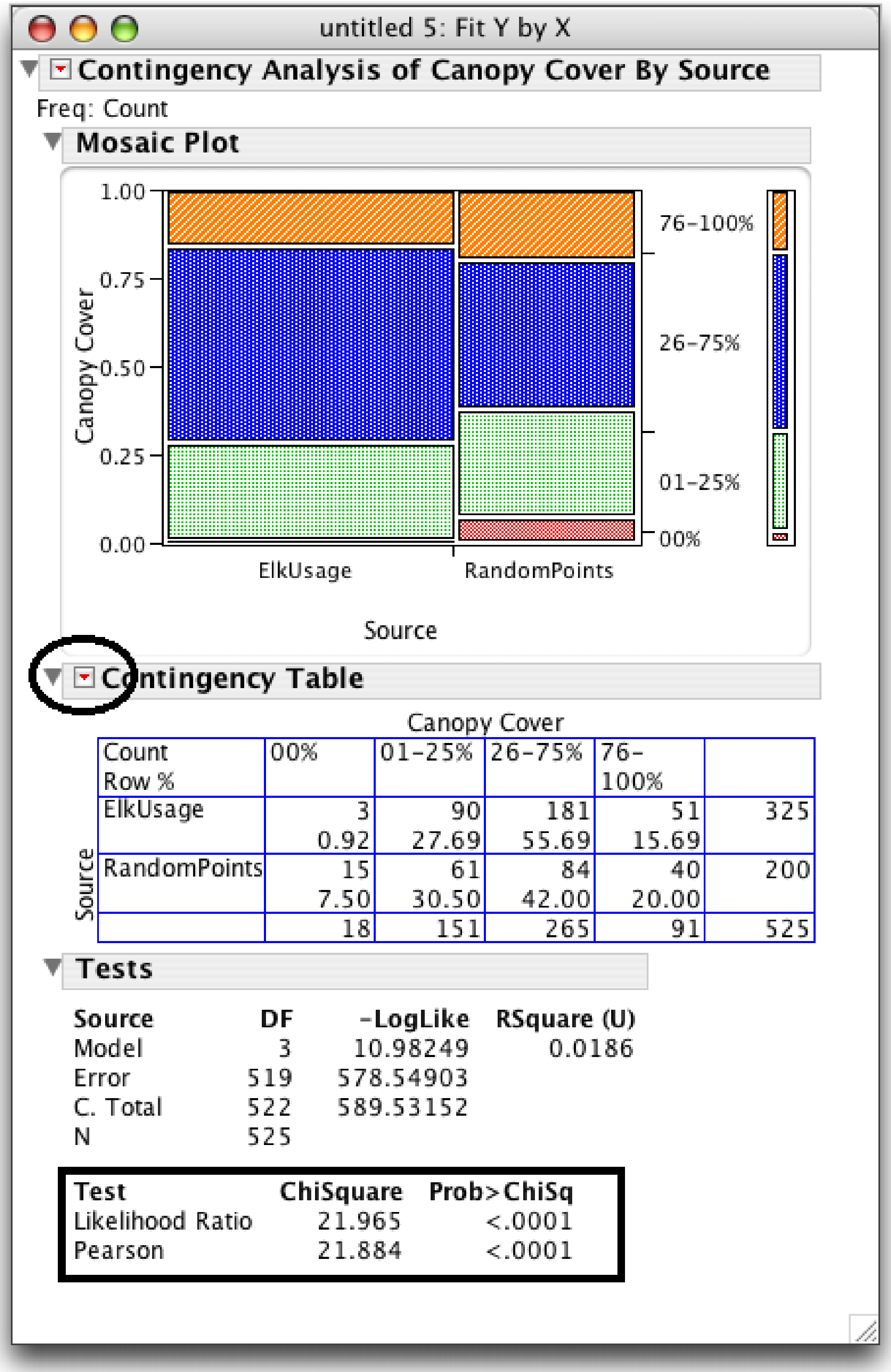
Distribution of Y for each X. Modeling types determine analysis.

Y, Response: Canopy Cover

X, Factor: Source

Freq: Count

This gives the following results (after some manipulations of the displays using the red-triangles)



Canopy Cover

Count Row %	00%	01-25%	26-75%	76-100%	
ElkUsage	3 0.92	90 27.69	181 55.69	51 15.69	325
RandomPoints	15 7.50	61 30.50	84 42.00	40 20.00	200
	18	151	265	91	525

```
Source          DF     -LogLike   RSquare (U)
Model            3     10.98249       0.0186
Error          519    578.54903
C. Total       522    589.53152
N              525

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio     21.965       <.0001
Pearson              21.884       <.0001
```

The mosaic plot shows some evidence of a difference in usage by the elk compared to the random points.

The contingency table and row percentages show where the observed differences lie. Finally, the *p*-value is very small indicating that there is very strong evidence that elk are not found in the canopy classes in proportion to the availability.

To investigate where the differences occurred, look at the mosaic plot, and the table of observed and expected counts.

The book:

> Manly, B.F.J, McDonald, L.L., Thomas, D.L., McDonald, T.L. and Erickson, W.P. (2002)
> Resource Selection by Animals. Kluwer Academic Publisers

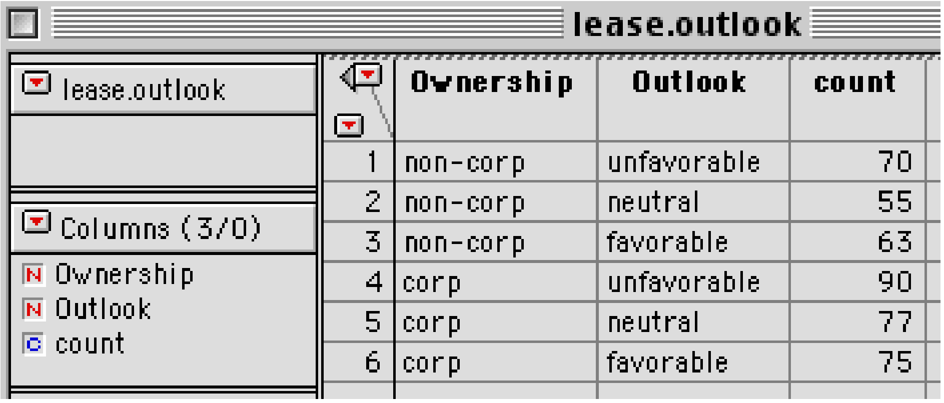
continues with this example and shows how to compute resource selection probabilities that can be interpreted as the probability that a randomly selected moose will select each habitat at a randomly chosen point in time.

14.5.2 Example: Ownership and viability

Does the type of ownership of a natural resource (e.g. oyster leases) affect the long term prospects? A sample of oyster leases was classified by type of ownership and the long-term prospects as shown below:

	Outlook		
	Unfavorable	Neutral	Favorable
Non-corporation	70	55	63
Corporation	90	77	75

The raw data is available in a *JMP* datafile called *oyster-lease.jmp* available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.



	Ownership	Outlook	count
1	non-corp	unfavorable	70
2	non-corp	neutral	55
3	non-corp	favorable	63
4	corp	unfavorable	90
5	corp	neutral	77
6	corp	favorable	75

This is an analytical survey as there is no manipulation of experimental units.

Does this survey meet the conditions of a single factor complete randomized design? There is a single factor (type of ownership) with 2 levels (corporate or non-corporate ownership). There is a single size of survey unit (the lease) and the observational unit is the same as the survey unit. There is insufficient information to assess if the sampled leases are a simple random sample from the respective populations, so we will have to assume that they were selected appropriately.

This data is in the form of a summary table, i.e. there are 430 leases that have been "collapsed" into a simple table.

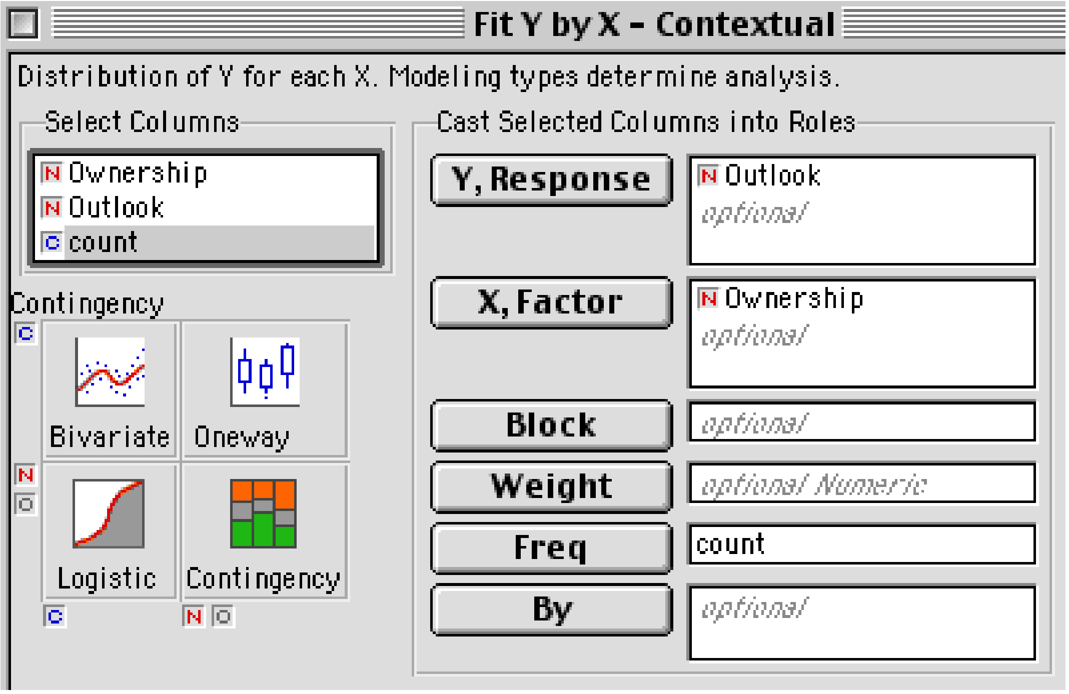
Note that both *Ownership* and *Outlook* are **nominal** scaled variables, and that the variable *Count* records the number of leases with the attributes.

The null hypothesis is that:

H: outlook is independent of ownership or $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ where $\pi_{\cdot j}$ is the overall proportion of leases with outlook j (the columns) and where $\pi_{i\cdot}$ is the overall proportion of leases with ownership i (the rows).

This is yet another way to write out the hypothesis of independence. Because independence implies that if you know the marginal proportions (the $\pi_{i\cdot}$ and $\pi_{\cdot j}$), then you know the individual cell proportions (the π_{ij}).

We begin by computing some summary graphs and statistics using the *Analyze->Fit Y-by-X* platform.



Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Select Columns

- Ownership
- Outlook
- count

Cast Selected Columns into Roles

Y, Response: Outlook

X, Factor: Ownership

Block: optional

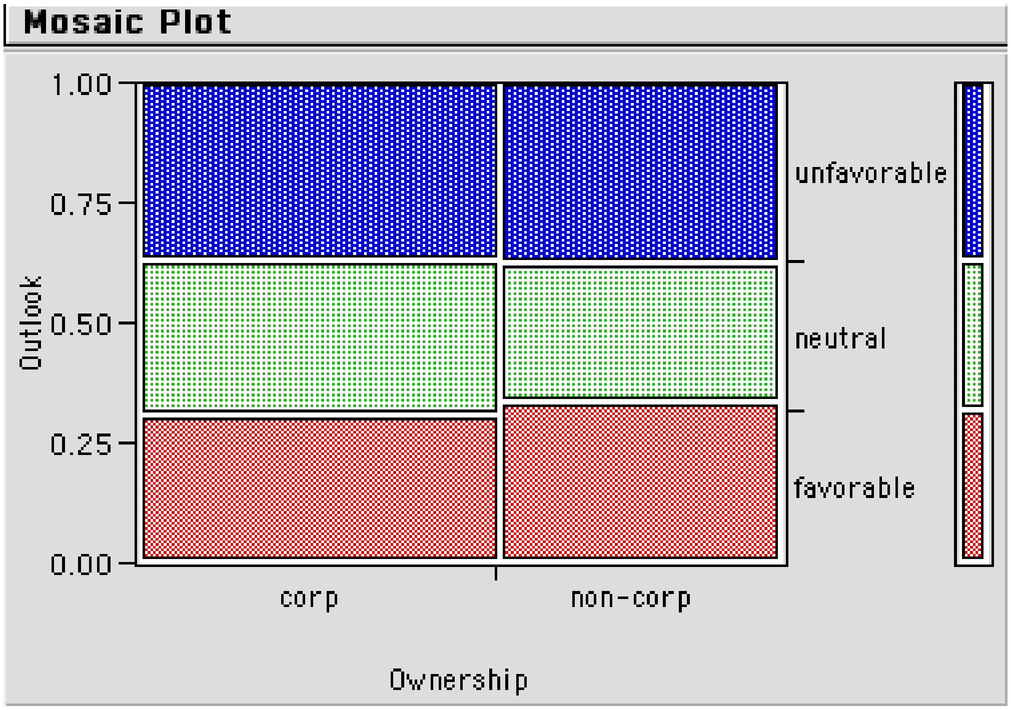
Weight: optional Numeric

Freq: count

By: optional

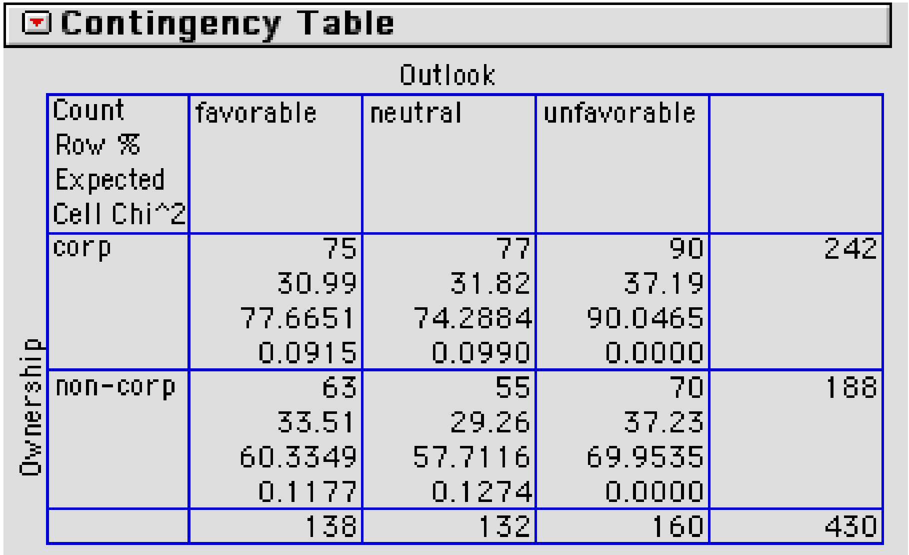
The X variable is the type of ownership; the Y variable is the long-term outlook (favourable, neutral, or unfavourable); the count of the number of leases in each combination of X and Y will be a frequency variable.

The resulting mosaic plot is shown below:



The side-by-side segmented bar charts are very similar; so we don't expect to find much evidence of an association between the long-term outlook and ownership status of the restaurant.

The contingency table can also be computed by *JMP* as shown below.



Contingency Table

Outlook

Ownership

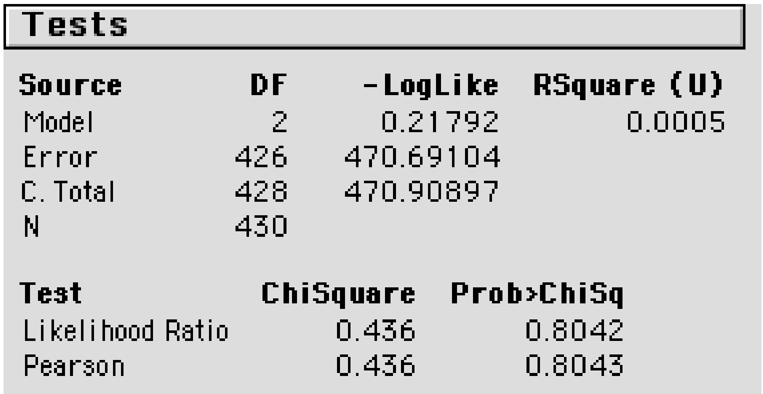
Count Row % Expected Cell Chi^2	favorable	neutral	unfavorable	
corp	75 30.99 77.6651 0.0915	77 31.82 74.2884 0.0990	90 37.19 90.0465 0.0000	242
non-corp	63 33.51 60.3349 0.1177	55 29.26 57.7116 0.1274	70 37.23 69.9535 0.0000	188
	138	132	160	430

The overall percentage (n_{ij}/n) is not very interesting and has been removed. It seems natural that the percentage in each row should be computed (why?), so the column percentages have also been suppressed. I've also added two more numbers as explained in the key in the upper left cell of the table.

The first entry in each cell is the actual count. The 'Row percentage' is the second number in each cell; the third entry is the expected number of leases for that combination of attribute if the hypothesis of independence were true and the last entry is the contribution of each cell to the overall test statistic.

Looking at the row percentages within the table, we see that the percentages for each outlook category are similar for the two types of ownership, confirming our impression from the mosaic chart.

The analysis then proceeds to compute an **Analysis of Deviance table** which is analogous to an ANOVA table. The total variation in the data is partitioned into sources. The actual computations are complex and not discussed in this course.



```
Tests

Source          DF     -LogLike  RSquare (U)
Model            2      0.21792       0.0005
Error          426    470.69104
C. Total       428    470.90897
N              430

Test              ChiSquare  Prob>ChiSq
Likelihood Ratio      0.436      0.8042
Pearson               0.436      0.8043
```

Finally, two test-statistics are computed for the table and *p*-values computed for each test-statistic are shown at the bottom of the test output.

The Pearson chi-square test statistic is computed by comparing the observed (n_{ij}) and expected counts (e_{ij}) when the hypothesis of independence is true. The idea behind the test is that if the data are consistent with the null hypothesis, then the expected and observed counts should be close; if the data are not consistent with the null hypothesis, then the observed and expected counts should be substantially different. The last entry in the cell is a measure of closeness of the two counts.

The Pearson chi-square value is formed as $\frac{(observed-expected)^2}{expected}$ (refer to the technical details in the appendix for derivations). The overall Pearson chi-square test statistic is formed as the sum of these entries over all cells in the table, i.e., $\chi^2 = 0.0915{+}0.1177{+}0.0990{+}0.1274{+}0.0000{+}0.0000 = 0.436$.

The *p*-value is computed by comparing this test statistic to the χ^2 distribution - hence the name of the test (refer to the appendix).

The *p*-value is 0.8043 which is very large. Hence we conclude that there is no evidence of a difference in the proportion of outlooks among the ownership types or that there is no evidence against outlook being independent of ownership type.

An alternate test statistic, the likelihood ratio test is also displayed. It compares the ratio of observed to expected counts and is usually very similar to the Pearson chi-square test-statistic. It is used in more complex modeling situations where the Pearson chi-square cannot be computed.

We also examine that all expected counts are reasonably large (a rough rule of thumb is that they should all be about 5 or greater). This can be relaxed somewhat but, as shown later, the interpretation of *p*-values in cases with many cells with small counts is problematic.

14.5.3 Example: Sex and Automobile Styling

The example we will look at is the relationship between sex and the type of automobile preferred.

A random sample of 303 people were asked for the person's sex, marital status, and type of car preferred (Japanese, European, or American).

Here is part of the raw data:

Sex	**Marital status**	**Age**	**Car pref**
Male	Married	34	American
Male	Single	36	Japanese
Male	Married	23	Japanese
Male	Single	29	American
...			

There are 303 records, one for each person.

The data are available in the *Car Poll.jmp* dataset in one of the *JMP* sample dataset directories. As well it is available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

This is an analytical survey as sex or marital status cannot be (yet) randomly assigned to individuals. The factor of interest is sex that has two levels. The marital status could also be used as a factor. If both variables are to be analyzed together, then this would be a two-factor experiment which is beyond the scope of this course.

The experimental unit is the same as the observational unit in this study and is the person interviewed. Note that only one member of a family, i.e., either the husband, or the wife, or another adult over 18 would be included in the survey, and only one response from a household would be used so that family influences would be minimized. If both husband and wife had been included in the dataset, it would no longer be a completely randomized design as families would now tend to be treated as blocks.

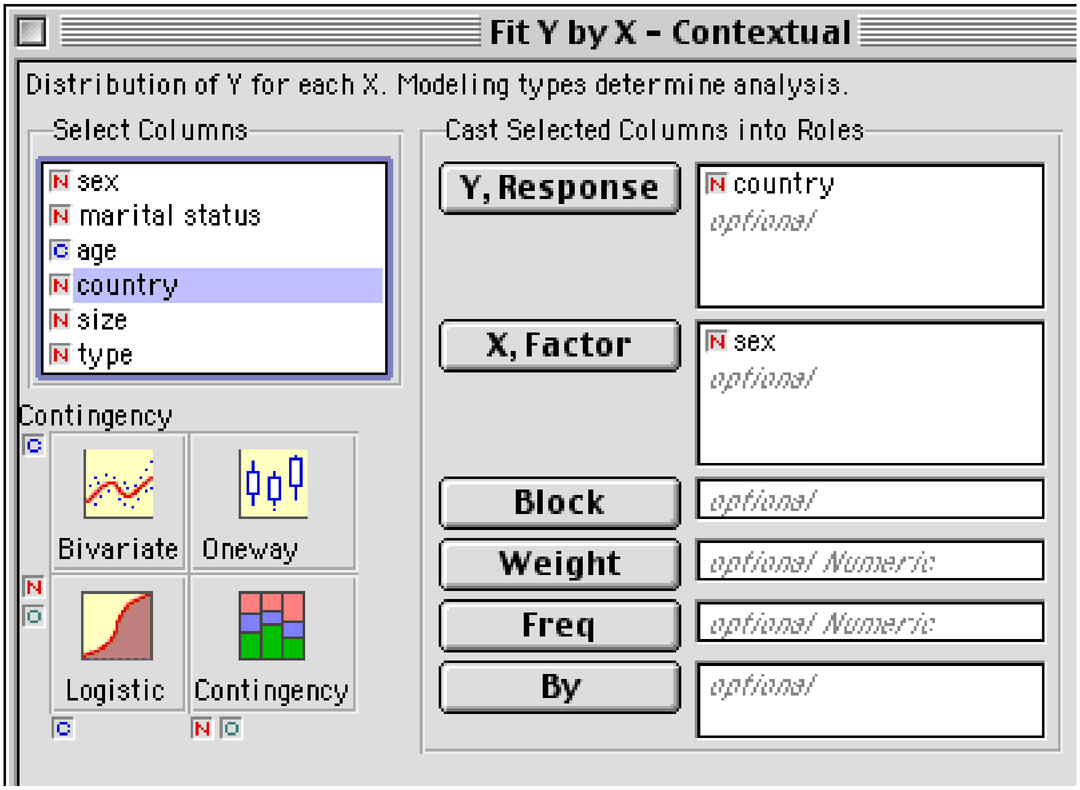
We assume that people have been selected at random from the relevant treatment groups.

The response variable is automobile preference.

One could argue that there is no clear distinction between a response and an explanatory variable. Is sex the explanatory variable to try and explain the car preference response, or is car preference the explanatory variable to try and explain the sex of the subject. This is analogous to correlation between two interval/ratio variables where there is no clear distinction. It turns out that the test statistic and results are identical regardless of which variable is treated as the explanatory or response variable.

The first thing to be done is to create a contingency table, which is a table of counts that cross-classifies the data.

Use the *Analyze->Fit Y-by-X* platform:



Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Select Columns: sex, marital status, age, country, size, type

Cast Selected Columns into Roles

Y, Response: country

X, Factor: sex

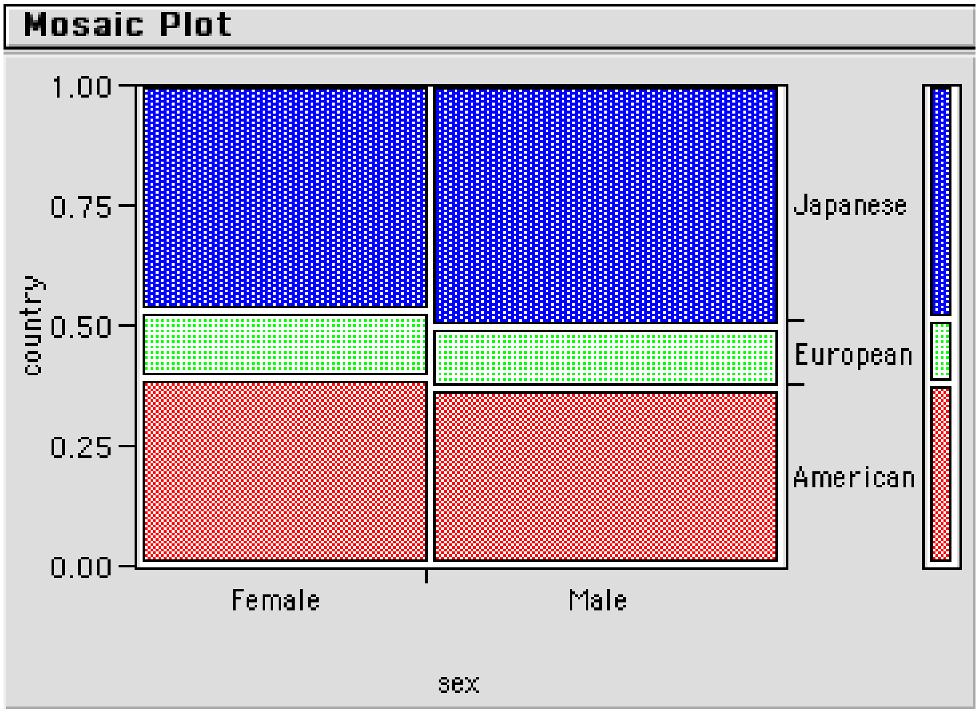
Block: optional

Weight: optional Numeric

Freq: optional Numeric

By: optional

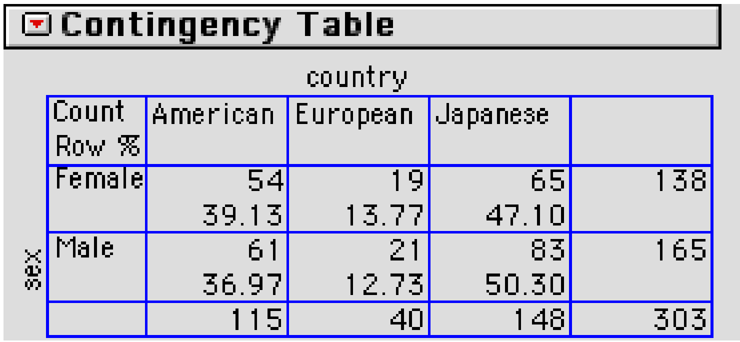
Choose *Country of manufacture* as the Y variable and *Sex* as the X variable. Because the raw data is available, there is no need to specify a frequency variable (by default each row in the table has a frequency of 1). As before, *JMP* provides a hint as to the type of analysis you will get for various combination of X and Y variables by the chart in the lower left of the platform.



The mosaic chart shows that both sexes appear to have similar preferences for country of manufacture: Each level of X has a segmented-bar-chart, and the width of each bar-chart is proportional to the sample size in each level of X. On the right is a segmented-bar-chart for the data pooled over all levels of X. It is easy to compare cumulative percentages using this chart, and it is easy to compare the relative sizes of the top and bottom segments (the American or Japanese brands), but it is more difficult to compare the middle segment (the European brand). For this reason, mosaic charts are often drawn to put the two most important categories at the top and bottom of each bar so that they are easily compared across treatment groups.

If you want a mosaic chart based on row percentages, you will need to reverse the roles of the country of manufacture and sex variables, i.e., make country of manufacture an X variable, and sex a Y variable.

Neither the overall percentage nor the column percentages are very useful here and have been suppressed.



Contingency Table

country

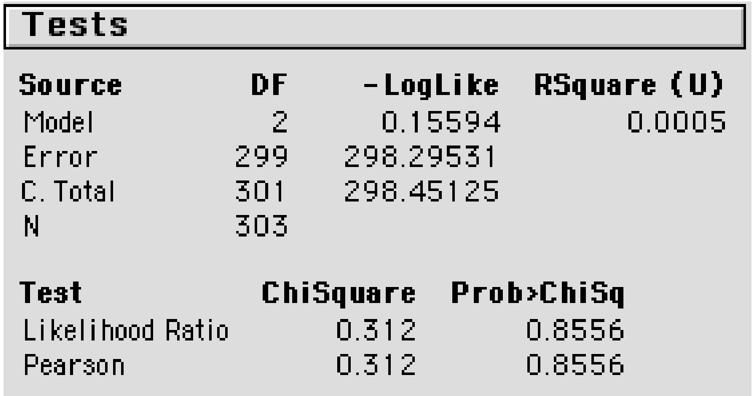
	Count Row %	American	European	Japanese	
sex	Female	54 39.13	19 13.77	65 47.10	138
	Male	61 36.97	21 12.73	83 50.30	165
		115	40	148	303

Each row percentage is computed as the cell divided by the row total. Each row is done separately, and each row's percentages add to 100%. Don't report too many decimal places - I would be inclined only to report the percentages rounded to a whole number rather than two decimal places.

Comparing the row percentages we see that there doesn't seem to be much of a difference between the country of manufacture preferences of males and females as the percentages are fairly close.

Finally, we examine the statistical test. The null hypothesis is that the proportion who choose a car manufactured in the three countries is the same for all levels of sex. Another way to state the null hypothesis is that preference for country of manufacturing is *independent* of sex.

The Analysis of Deviance table and test statistics are shown below:



```
Tests

Source          DF     -LogLike   RSquare (U)
Model            2      0.15594        0.0005
Error          299    298.29531
C. Total       301    298.45125
N              303

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio      0.312       0.8556
Pearson               0.312       0.8556
```

Not unexpectedly, the p-value is quite large (.8556) which indicates that there is little evidence against the null hypothesis, i.e we find no evidence against the hypothesis of independence.

Finally check that the expected counts in each cell are at least 5. **NOTE: the EXPECTED counts need to be checked, not the observed counts**.

14.5.4 Example: Marijuana use in college

The following was taken from the paper "Marijuana Use in College (*Youth and Society* (1979): 323-34). Four hundred and forty five college students were classified according to both frequency of marijuana use and parental use of alcohol and psychoactive drugs.

The individual student responses are not available, only the summary data were given in the paper:

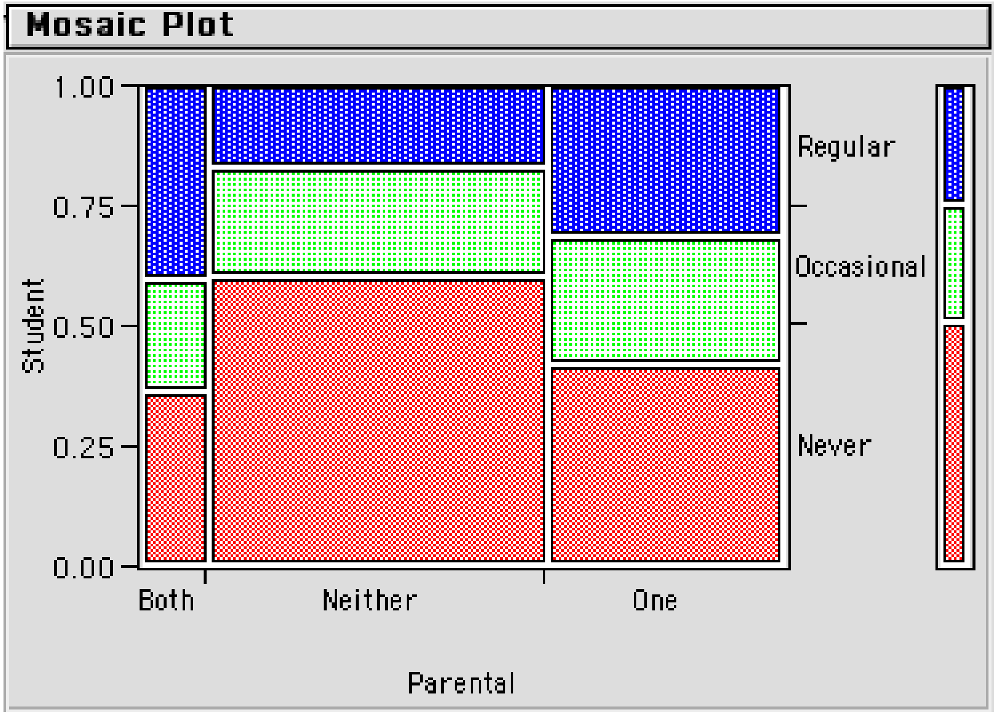
Parental use of Alcohol and Drugs	Student level of marijuana use	Count
Neither	Never	141
Neither	Occasional	54
Neither	Regular	40
One	Never	68
One	Occasional	44
One	Regular	51
Both	Never	17
Both	Occasional	11
Both	Regular	19

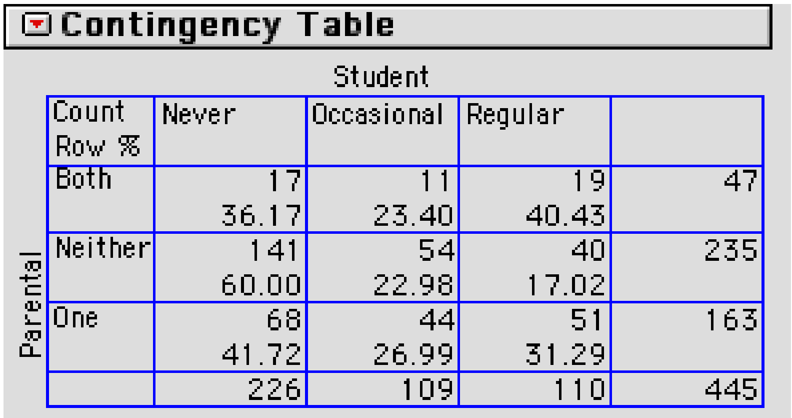
The raw data is available in a *JMP* datafile called *marijuana.jmp* available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

Does this study meet the requirements for a single factor completely randomized design?

The analysis proceeds using the *Analyze->Fit Y-by-X* platform. Assign *Parental use* as the X variable, *Student use* as the Y variable, and *Count* as a frequency variable.

Obtain the mosaic plot and contingency table:



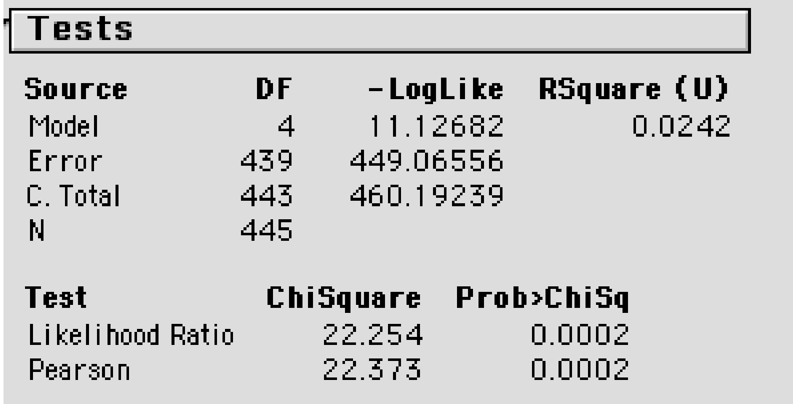


Student

Parental	Count Row %	Never	Occasional	Regular	
	Both	17 36.17	11 23.40	19 40.43	47
	Neither	141 60.00	54 22.98	40 17.02	235
	One	68 41.72	44 26.99	51 31.29	163
		226	109	110	445

The mosaic plot shows substantial differences (how can you tell?). This impression is confirmed in the contingency table (how can you tell?) where it appears that students are more prone to use these substances if the parents also used these substances.

The Analysis of Deviance (no pun intended) table is shown below. The null hypothesis is that student usage is independent of parental usage.



```
Tests
Source          DF     -LogLike   RSquare (U)
Model            4     11.12682        0.0242
Error          439    449.06556
C. Total       443    460.19239
N              445

Test            ChiSquare  Prob>ChiSq
Likelihood Ratio   22.254      0.0002
Pearson            22.373      0.0002
```

The *p*-value is very small (.0002) indicating very strong evidence against the null hypothesis.

Check to see that all expected counts are at least 5 (how is this done?) to ensure that the conditions necessary for the validity of the chi-square test are satisfied.

At this point, it is fairly clear looking at the mosaic plots and contingency table where the non-independence is occurring – there appears to be a “shift” in regular usage among students as the use by parents increases.

A more sophisticated analysis could be performed taking into account that the levels of X and Y are ordinal scale, but this is beyond the scope of this course.

14.5.5 Example: Outcome vs. cause of accident

Accidents along the Trans-Canada highway were cross-classified by the cause of the accident and by the outcome of the accident. Here are some summary statistics:

Cause	**Outcome**	**Number of Accidents**
speeding	death	42
speeding	no death	88
drinking	death	61
drinking	no death	185
reckless	death	20
reckless	no death	74
other	death	12
other	no death	86

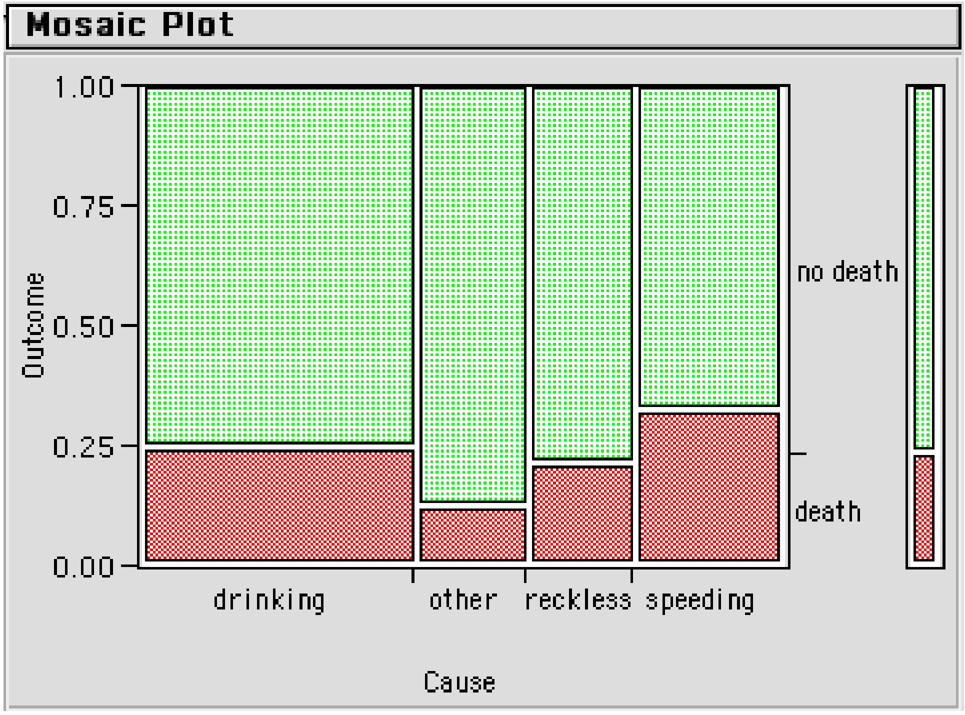
The raw data is available in a *JMP* datafile called *accident.jmp* available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

Does this study meet the assumptions for a single factor completely randomized design? In particular, how are multiple people within a car treated? Is it reasonable to treat these as independent events?

Assuming that the conditions are satisfied, here is a summary table in a more conventional format. Each cell has the observed count and the column percentage.

	Cause				
Outcome	drinking	other	reckless	speeding	
death	61 (25%)	12 (12%)	20 (21%)	42 (32%)	135
no death	185 (75%)	86 (88%)	74 (79%)	88 (68%)	433
	246	98	94	130	568

Here is a mosaic plot of the data:



Our hypotheses are:

H: the outcome (fatal or not fatal) is independent of the cause of the accident.

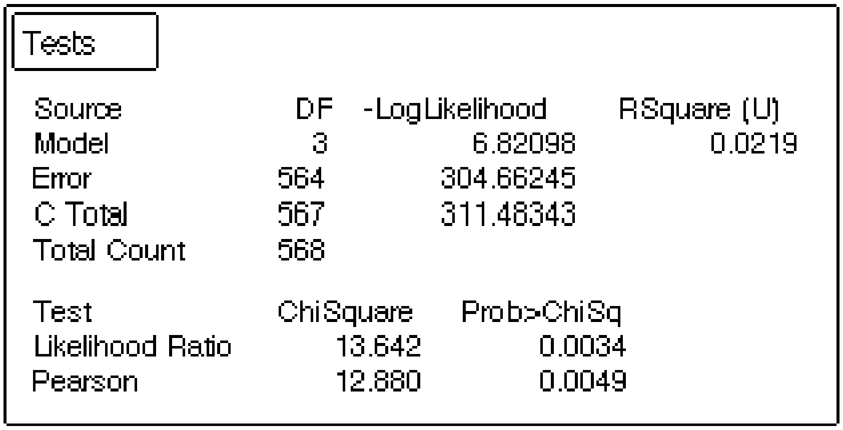
A: the outcome (fatal or not fatal) is not independent of the cause of the accident.

Here is the output from *JMP* (in text format rather than graphical format):

```
                         Cause
         drinking        other    reckless    speeding
-------------------------------------------------------------
death        61          12         20          42           135
expected     58.46       23.29      22.34       30.89
diff          2.53      -11.29      -2.34       11.10
cell chi2     0.10        5.47       0.24        3.98

no death    185          86         74          88           433
expected    187.53       74.70      71.65       99.10
diff         -2.53       11.29       2.34      -11.10
cell chi2     0.03        1.70       0.07        1.24
-------------------------------------------------------------

Total       246          98         94         130           568
```



```
Tests

Source            DF   -LogLikelihood    RSquare (U)
Model              3        6.82098          0.0219
Error            564      304.66245
C Total          567      311.48343
Total Count      568

Test              ChiSquare    Prob>ChiSq
Likelihood Ratio     13.642        0.0034
Pearson              12.880        0.0049
```

The overall test statistic is found by summing the individual cell chi-square values (the square of the residuals if present) and is 12.880. The test statistic will be compared to a chi-square distribution with 3 df. The observed p-value is .0049 which provides strong evidence against the null hypothesis.

We find the evidence sufficient strong against the null hypothesis and conclude that there is evidence against the independence of outcome and cause of accident.

Check to see that the expected counts at least 5 to ensure the validity of the chi-square test.

Looking at the cell chi-square values in the contingency table, we see that the major differences appear to occur with 'other' and 'speeding' as the cause of the accident with a lower and higher fatality rate compared to the other two causes. Their contributions to the overall chi-square test statistics are about 4 or higher.

14.5.6 Example: Activity times of birds

A survey was conducted on feeding behaviour of sand pipers at two sites close to Vancouver, B.C. The two sites are Boundary Bay (BB) and Sidney Island (SI).

An observer went to each site and observed flocks of sand pipers. The observer recorded the time spent by a flock between landing (most often to feed) and departing (either because of a predator nearby or other cues).

The protocol called for the observer to stop recording and to move to a new flock if the time spent exceeded 10 minutes. This is a form of censoring - the actual time the flock spent on the ground is unknown; all that is known in these cases is if the time exceeded 10 minutes.

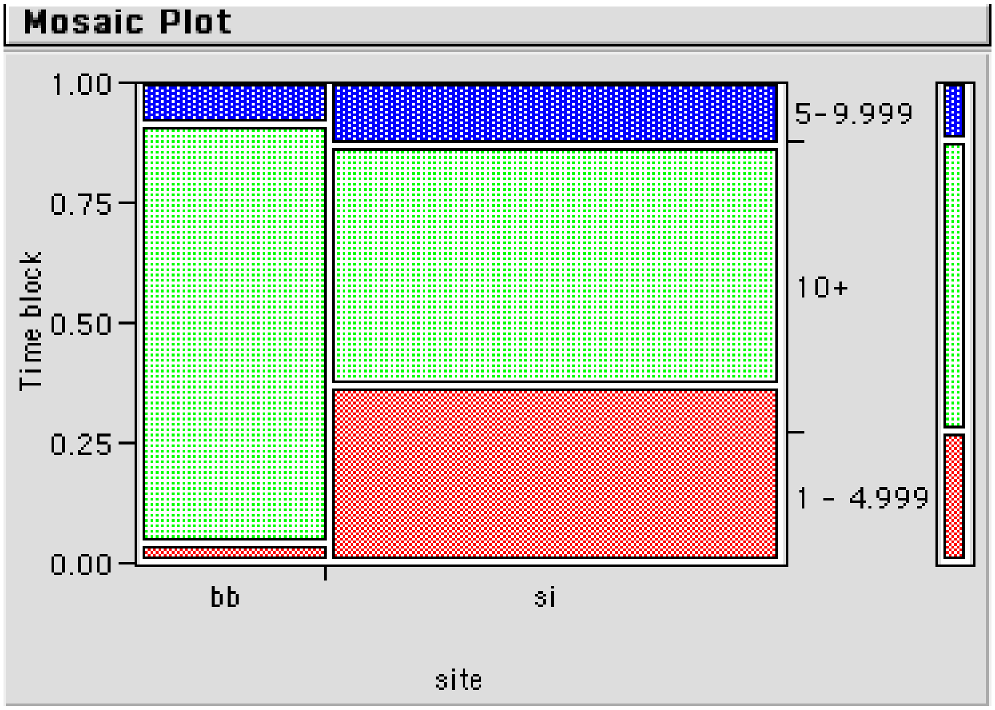
Had the exact times been recorded for every flock, the analysis would be straightforward. It would analyzed as a single factor (site with 2 levels) with a completely randomized design. Of course, the inference is limited to these two sites and we are making assumptions that once flocks leave, their subsequent return, reformation, and activity may be treated as introducing a "new, independent" flock.

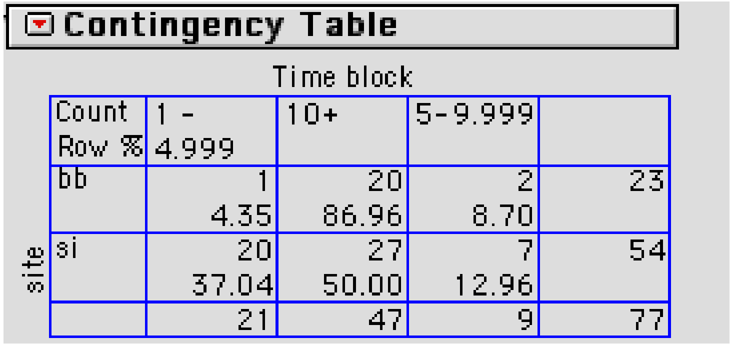
Sophisticated methods are available to deal with the censored data, but a simple analysis can be conducted by classifying the time spent into three categories: 0-5 minutes, 5-10 minutes, and 10+ minutes. [The division into 5 minute time blocks is somewhat arbitrary].

The hypothesis is that the proportion of flocks that appear in each time block is independent of site.

The raw data are available in a *JMP* data table *flocks.jmp* available from Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`

Here are mosaic plots, contingency tables, and analysis of deviance table:

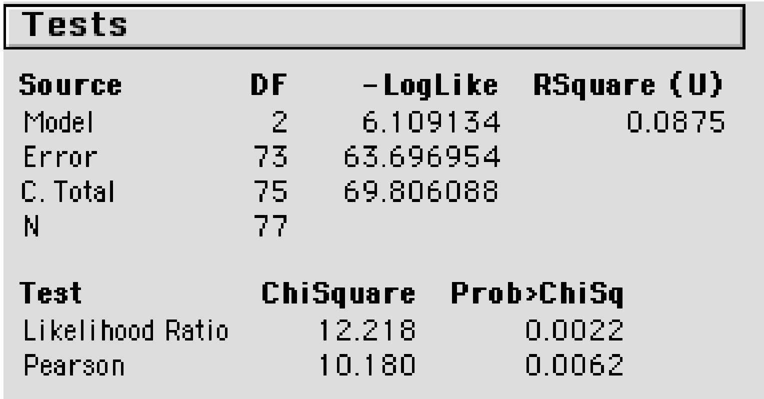




Contingency Table

Time block

site	Count Row %	1 - 4.999	10+	5-9.999	
	bb	1 4.35	20 86.96	2 8.70	23
	si	20 37.04	27 50.00	7 12.96	54
		21	47	9	77



Tests

```
Source          DF    -LogLike   RSquare (U)
Model            2    6.109134        0.0875
Error           73   63.696954
C. Total        75   69.806088
N               77

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio     12.218       0.0022
Pearson              10.180       0.0062
```

The *p*-value is very small - very strong evidence against the null hypothesis and it is quite clear where the difference lies between the two sites. However, before writing up the results, look at the expected counts – one is quite small. Fortunately, the contribution of that cell to the total test statistics is very small – it contributes only .1762 to the overall value of 12.218, so this is not a problem.

Notice that one of the expected values is quite small (less than 5). In cases like this the chi-square test should be treated with caution as very small expected counts tend to inflate the test statistic and make the results look more significant than they really are.

14.6 Pseudo-replication - Combining tables

Hurlbert (1984) states:

> "Chi-square is one of the most misapplied of all statistical procedures."

According to Hurlbert (1984), the major problem is that individual units are treated as independent objects, when in fact, there are not. Experimenters often pool experimental units from disparate sets of observations. He specifically labels this pooling as **sacrificial pseudo-replication**.

Hurlbert (1984) cites the example of an experiment to investigate the effect of fox predation upon the sex ratio of mice. Four plots are established. Two of the plots are randomly chosen and a fox-proof fence is erected around the plots. The other two plots are controls.

Here are the data (Table 6 of Hurlbert (1984)):

	Plot	% Males	Number males	Number females
Foxes	A_1	63%	22	13
	A_2	56%	9	7
No foxes	B_1	60%	15	10
	B_2	43%	97	130

Many researchers would pool over the replicates to give the pooled table:

	Plot	% Males	Number males	Number females
Foxes	$A_1 + A_2$	61%	31	20
No foxes	$B_1 + B_2$	44%	112	140

If a χ^2 test is applied to the pooled data, the *p*-value is less than 5% indicating there is evidence that the sex ratio is not independent of the presence of foxes.

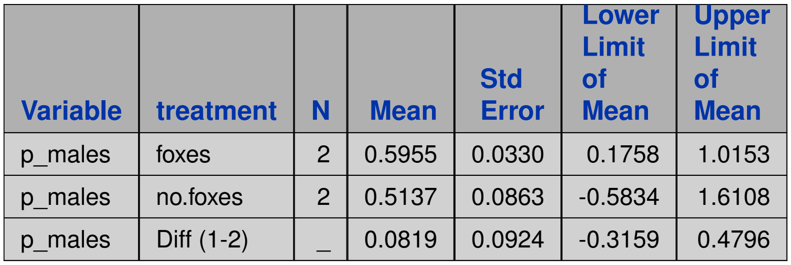
Hurlbert (1984) identifies at least 4 reasons why the pooling is not valid:

- **non-independence of observation**. The 35 mice caught in A_1 can be regarded as 35 observations all subject to a common cause, as can the 16 mice in A_2, as each group were subject to a common influence in the patches. Consequently, the pooled mice are *NOT* independent; they represent two sets of interdependent or correlated observations. The pooled data set violates the fundamental assumption of independent observations.
- **throws away some information**. The pooling throws out the information on the variability among replicate plots. Without such information there is no proper way to assess the significance of the differences between treatments. Note that in previous cases of ordinary pseudo-replication (e.g. multiple fish within a tank), this information is also discarded but is not needed - what is needed is the variation among tanks, not among fish. In the latter case, averaging over the pseudo-replicates causes no problems.
- **confusion of experimental and observational units**. If one carries out a test on the pooled data, one is implicitly redefining the experimental unit to be individual mice and not the field plots. The enclosures (treatments) are applied at the plot level and not the mouse level. This is similar to the problem of multiple fish within a tank that is subject to a treatment.
- **unequal weighting**. Pooling weights the replicate plots differentially. For example, suppose that one enclosure had 1000 mice with 90% being male; and a second enclosure has 10 mice with 10% being male. The pooled data would have $1000 + 10$ mice with $900 + 1$ being male for an overall male ratio of 90%. Had the two enclosures been given equal weight, the average male percentage would be (90%+10%)/2=50%. In the above example, the number of mice captured in the plots varies from 16 to over 200; the plot with over 200 mice essentially drives the results.

Hurlbert (1984) suggests the proper way to analyze the above experiment is to essentially compute a single number for each plot and then do a two-sample t-test on the percentages. [This is equivalent to the ordinary averaging process that takes place in ordinary pseudo-replication.] For example, with the above table, the data for the t-test would be:

Treatment	% males
Foxes	63
Foxes	56
No Foxes	60
No Foxes	43

The results from a simple t-test conducted in *SAS* are:



Variable	treatment	N	Mean	Std Error	Lower Limit of Mean	Upper Limit of Mean
p_males	foxes	2	0.5955	0.0330	0.1758	1.0153
p_males	no.foxes	2	0.5137	0.0863	-0.5834	1.6108
p_males	Diff (1-2)	_	0.0819	0.0924	-0.3159	0.4796

The estimated difference in the sex ratio between colonies that are subject to fox predation and colonies not subject to fox predation is .082 (SE .092) with p-values of .46 (pooled t-test) and .51 (unpooled t-test) respectively. As the p-values are quite large, there is NO evidence of a predation effect.

With only two replicates (the colonies), this experiment is likely to have very poor power to detect anything but gross differences.

The above analysis is not entirely satisfactory. The proportion of males have different variabilities because they are based on different number of total mice. A more "refined" analysis is now available using Generalized Linear Mixed Models (GLIMM).

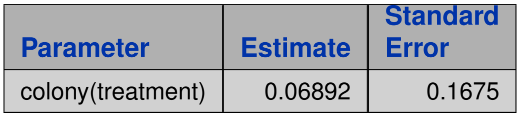
For this experiment, the model would be specified (in the usual short-hand notation) as:

$$logit(p_{males}) = Treatment \;\; Colony(Treatment)(R)$$

where the $Colony(Treatment)$ would be the random effect of the experimental units (the colonies). A logistic type model is used.

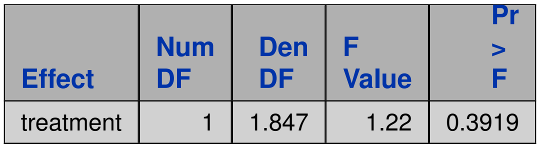
Unfortunately, *JMP* does not have a platform for GLIMs. The output from GLIMMIX (*SAS* 9.3) follows.

First is an estimate of the variability among colonies (on the logit scale):



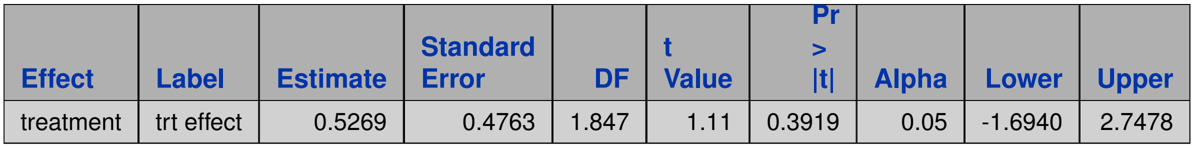
Parameter	Estimate	Standard Error
colony(treatment)	0.06892	0.1675

Next is the test of the overall treatment effect:



Effect	Num DF	Den DF	F Value	Pr > F
treatment	1	1.847	1.22	0.3919

The *p*-value is .39; again no evidence of a predation effect on the proportion of males in the colonies. Finally, an estimate of the treatment effect:



Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treatment	trt effect	0.5269	0.4763	1.847	1.11	0.3919	0.05	-1.6940	2.7478

Some caution is required. The estimate of .53 (SE .47) is for the difference in the logit(proportions) between males and females. If you take $exp(.53) = 1.69$, this is the estimated odds-ratio of males to females comparing colonies with predators to colonies without predators. The 95% confidence interval for the odds-ratio is $exp(-1.6950) = .183$) to $exp(2.7478) = 15.60$ which includes the value of 1 (indicating no effect). Consult the chapter on logistic regression for an explanation of odds and odds-ratios. Consult the chapter on advanced logistic regression for more details.

Hurlbert (1984) then continues:

> The commonness of this type of chi-square misuse probably is traceable to the kinds of examples found in statistics texts, which too often are only from genetics, or from mensurative rather than manipulative experiments, or from manipulative experiments (e.g. medical ones) in which individual organisms *are* the experimental units and not simply components of them, as in the mammal field studies cited. . . . I know of no statistics textbook that provides clear and reliable guidance on this matter.

We (statisticians) are guilty as charged! Hopefully, things will change in the future.

One of the reasons for the urge to pool, is that until recently, good software was not readily available. This is no longer the case - please seek assistance before pooling.

14.7 Simpson's Paradox - Combining tables

Another problem related to pooling of tables is that the pooled table may show different results than the individual tables. This is generically known as **Simpson's Paradox**.

Simpson's paradox is an example of the dangers of lurking variables.

14.7.1 Example: Sex bias in admissions

Is there a sex bias in admissions? Consider the following tables on the number of admissions to an MBA program and a Law program cross-classified by sex. For each table compute the % admitted for each sex (row percentage).

Business School

	Admit	Deny
Male	480=80%	120=20%
Female	180=90%	20=10%

It would appear that females are admitted at a slightly higher rate than males in the Business school.

Similarly, look at a table for Law school.

Law School

	Admit	Deny
Male	10=10%	90=90%
Female	100=33%	200=66%

Again, it appears that females are admitted at a higher rate than males in the Law school.

And what happens when the tables are combined?

Business and Law Schools

	Admit	Deny
Male	490=70%	210=30%
Female	280=56%	220=44%

Now females seem to be admitted at a lower rate than males!

Why has this happened? This is caused by the different percentages in admission in the two tables – they really shouldn't be combined. It is not caused by different sample sizes.

14.7.2 Example: - Twenty-year survival and smoking status

This is based on "Ignoring a covariate: An example of Simpson's Paradox" by Appleton, D.R. French, J.M. and Vanderpump, M.P (1996, American Statistician, 50, 340-341).

In 1972-1994 a one-in-six survey of the electoral roll, largely concerned with thyroid disease and heart disease was carried out in Wichkham, a mixed urban and rural district near Newcastle-upon-Tyne, in the UK. Twenty years later, a follow-up study was conducted.

Here are the results for two age groups of females. Each table shows the twenty-year survival status for smokers and non-smokers.

Age 55-64

	Dead	Alive
Smokers	51=44%	64=56%
Non-smokers	40=33%	81=67%

Age 65-74

	Dead	Alive
Smokers	29=80%	7=20%
Non-smokers	101=78%	28=22%

It appears that smokers die at a higher rate than non-smokers in each table.

And what happens when the tables are combined?

Ages 55-74 combined

	Dead	Alive
Smokers	80=53%	71=47%
Non-smokers	141=56%	109=44%

Now smokers seem to have a lower death rate!

What has happened? Most of the smokers have died off before reaching the older age classes, and so the higher number of deaths (in absolute numbers) for the non-smokers in the older age classes has obscured the result.

The original article had 7 age classes. In each age class the smoker had from 1.2 to 2 times the death rate of non-smokers, yet in the pooled table, the non-smokers had a 50% higher death rate!

14.8 More complex designs

It is possible to extend the above analyses to more complex situations, e.g., multi-factor designs, blocked designs, sub-sampling etc. As with the analysis of means where the general methodology of ANOVA was developed, comparable methods, *log-linear modeling* or *generalized linear modeling*, can handle all of these situations. Routines to do these analyses are available in *SAS* and many other computer packages. As before, be sure that your brain in in gear before engaging the computer!

The hypotheses can remain fairly simple, like "Is there a dependence in the occurrences of the two species on fungus brackets of the same age?" They can also become more complex. One might also be interested in, for example, the possibility of differences in the dependence patterns in young vs. old brackets. Or you may have a four-way breakdown of animals by sex, breeding status, age, and condition factor and are interested in relationships among these attributes.

Unfortunately, these types of analyses are beyond the scope of these notes.

14.9 Final notes

- There are special formula for 2x2 tables. However, the methods presented in this chapter are general to all sizes of tables and will give the same results.
- The concept of one-sided and two-sidedness for the hypothesis of independence does not exist except when the contingency table is a 2x2 table. A one-sided analysis is covered in more advanced classes in statistics.
- **Caution**: if some of the cells have **expected** values < 5 there may be problems since the individual χ^2_{ij} value may be unusually large (e.g. if $e_{ij} = 0.01$, then a difference between the observed and expected values of 1 is expanded to a χ^2_{ij} of 100!) and the total χ^2 statistic may no longer be reliable. This is a particular problem if you conclude that there is evidence **against** the null hypothesis - always check the **expected** counts in each cell and the individual χ^2_{ij} statistics to see if a single cells results are distorting the final statistic.

  If the contingency table has small counts, then the Fisher Exact Test is a better way to compute the test-statistic and p-value as it doesn't make any large sample approximations as in the Pearson chi-square or the likelihood ratio G^2 test.

- The preferred language is to conclude that there is evidence against independence, or there is insufficient evidence against the hypothesis of independence. The following phrases should be avoided:
  - 'conclude that the variables are dependent' - you don't know - all you have done is collected evidence against independence.
  - 'conclude that the variables are independent' - you don't know - all you have done is failed to find evidence against independence.

  It sounds picky, but is the same reason that juries either convict or fail to convict people, rather than declaring if the party is innocent or guilty.
- As always, Type I and II errors are possible. A Type I error (a false positive result) would be to conclude that the two variables are not independent (find evidence against the null hypothesis) when in fact they are independent (null is true). A Type II error (a false negative result) would be to conclude that there is no evidence against independence (fail to find evidence against the null hypothesis) when in fact the variables are not independent.

14.10 Appendix - how the test statistic is computed

This section is not required for Stat 403/650.

Refer to an earlier section, where a study was conducted to determine if the type of ownership of a natural resource (e.g. oyster leases) affects the long term prospects? A sample of oyster leases was classified by type of ownership and the long-term prospects as shown below:

	Outlook			
	Favorable	Neutral	Unfavorable	Total
Corporation	75	77	90	242
Non-corporation	63	55	70	188
Total	138	132	160	430

Define the notation:

- π_{ij} = true proportion of outlook j (favor, neutral, or unfavourable) in group i (corp or non-corp ownership). These are the population parameters.
- n_{ij} = observed count of outlook j (favor, neutral, or unfavourable) in group i (corp or non-corp ownership). These are the sample statistics.
- e_{ij} = expected count in cell (i, j) if the hypothesis of independence is true.

The Pearson chi-square test statistic is computed by comparing the observed (n_{ij}) and expected counts (e_{ij}) when the hypothesis is true. The idea behind the test is that if the data are consistent with the null hypothesis, then the expected and observed counts should be close; if the data are not consistent with the null hypothesis, then the observed and expected counts should be substantially different.

How would an expected count be found when the hypothesis is true? Well it seems reasonable to estimate $\pi_{i\cdot}$ (the proportion of leases of ownership i) by the marginal count (in this case the row sum) divided by the overall sample size. Hence $\widehat{\pi}_{i\cdot} =$ row i marginal total/total sample size and

$\widehat{\pi}_{1\cdot} = r_1/n$ = 242/430

$\widehat{\pi}_{2\cdot} = r_2/n$ = 188/430

where r_1 and r_2 are the marginal total for row 1 and row 2 and n is the total sample size.

Similarly, it seems reasonable (if the hypothesis were true and there was no difference between the proportions in each column) to estimate $\pi_{\cdot j}$ (the proportion of leases with outlook j) by the marginal column total divided by the overall sample size. Hence $\widehat{\pi}_{\cdot j}$ = column j marginal total /total sample size and:

$\widehat{\pi}_{\cdot 1} = c_1/n$ = 138/430

$\widehat{\pi}_{\cdot 2} = c_2/n$ = 132/430

$\widehat{\pi}_{\cdot 3} = c_3/n$ = 160/430.

where c_1, c_2, and c_3 are the marginal total for columns 1-3 and n is the total sample size.

Putting these together, it seems reasonable (if the hypothesis were true and there was no difference between proportions in each column) to estimate $\widehat{\pi}_{ij}$ = $\widehat{\pi}_{i\cdot} \times \widehat{\pi}_{\cdot j}$ and furthermore, and estimate of the expected number in each cell, as total sample size x $\widehat{\pi}_{ij}$ or

$$e_{ij} = n \times \widehat{\pi}_{ij} = n \times \widehat{\pi}_{i\cdot} \times \widehat{\pi}_{\cdot j} = n \times (r_i/n) \times (c_j/n) = r_i \times c_j/n.$$

This is the third entry in each cell labeled ‘Expected’. For example:

- the entry in the (favour, corp) cell is found as 138(242)/430=77.67
- the entry in the (neutral, non-corp) cell is found as 132(188)/430=57.71

The test statistic is computed for that cell as:

$X^2_{ij} = \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$.

which is the last entry in the table. For example:

- $X^2_{11} = (75 - 77.67)^2/77.67$ = 0.0915
- $X^2_{22} = (55 - 57.71)^2/57.71$ = 0.1274.

The overall test statistic is then found by summing all these individual entries

$$\mathrm{X}^2 = \sum\sum \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = \sum\sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

= 0.0915 + 0.1177 + 0.0990 + 0.1274 + 0.0000 + 0.0000 = 0.436.

It is compared to a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom where r=number of rows, c=number of columns. In this case, it would be compared to a chi-square distribution with (2-1)(3-1) = 2 degrees of freedom.

The *p*-value is found as the probability that the appropriate chi-square distribution exceeds the test statistic. [Despite its appearance, this is not a one-sided test in the usual sense of the word. As in ANOVA, the concept of one- or two-sidedness for test really has no meaning in chi-square tests.]

The approximate *p*-value can be obtained from the chi-square distribution and looking at the 2 df line. We see that this table indicates that the *p*-value is larger than 0.30 which is consistent with *JMP*.

```
         0.3000    0.2000    0.1500  0.1000    0.0500    0.0250    0.0100    0.0050    0.0010
 df      ------------------------------------------------------------------------------------------
  1       1.074     1.642     2.072   2.706     3.841     5.024     6.635     7.879    10.828
  2       2.408     3.219     3.794   4.605     5.991     7.378     9.210    10.597    13.816
```

The rough rule of thumb to investigate where differences could occur is derived from the fact that each cells contribution to the overall test statistic has an approximate chi-square distribution with a single degree of freedom and the 95^{th} percentile of such a distribution is about 4.

14.11 Fisher's Exact Test

The traditional χ^2 test for independence relies upon an approximation to the sampling distribution of the test statistic (the X^2 value) in order to obtain the *p*-value. However, this approximation works well only when the EXPECTED count in each cell of the table is reasonably large – the rules of thumb are that each expected count should be at least 5, but the approximation still works well if the expected counts are as low as 3.

In some case with very sparse data, the simple χ^2 test for independence does not work well. For example, consider an experiment similar to that of:

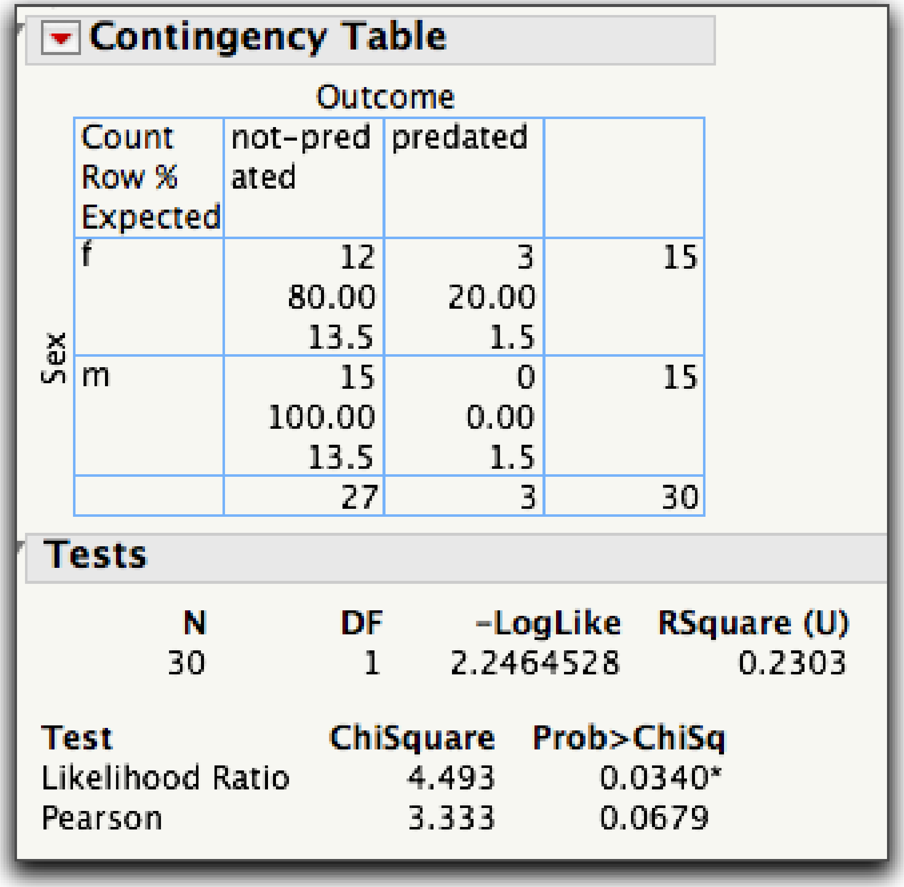
> McCleery, R.A., Lopez, R.R, Silvy, N.J., and Gallant, D.L. (2008).
> Fox squirrel survival in urban and rural environments.
> Journal of Wildlife Management, 72, 133-137.
> `http://dx.doi.org/10.2193/2007-138`

They released fox squirrels with radio collars and tracked their subsequent survival.

Here is a (fictitious) data summary for rural areas:

Site	Sex	Predated	Not-predated
Rural	M	0	15
	F	3	12

The traditional χ^2-test gives:



Contingency Table

Outcome

	Count Row % Expected	not-pred ated	predated	
Sex	f	12 80.00 13.5	3 20.00 1.5	15
	m	15 100.00 13.5	0 0.00 1.5	15
		27	3	30

Tests

N	DF	-LogLike	RSquare (U)
30	1	2.2464528	0.2303

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	4.493	0.0340*
Pearson	3.333	0.0679

i.e. a test statistic of 4.49 with a (likelihood) p-value of .0340 and the Pearson χ^2 test p-value of .068. However, the expected counts[3] are small in the second column and the χ^2 approximation is dubious. The key problem is that when the X^2 is computed, each cell's discrepancy between the observed and expected value is inflated by the expected value (i.e. divided by the expected value). Consequently, small expected counts lead to very large contributions from that particular cell.

Fisher's Exact Test is traditionally used in these circumstances. It is most often used in 2×2 tables, but can be used in larger tables as well, but the computations quickly become tedious. Modern statistical software (e.g. *SAS*) has good algorithms for large problems.

There is also a version of Fisher's Exact Test suitable for goodness-of-fit tests; this is not examined in this section of the notes.

14.11.1 Sampling Protocol

The same sampling protocol as for ordinary χ^2 tests is required, i.e. a completely randomized design with a single observation per experimental unit. A common problem is the unthinking use of Fisher's Exact Test (and other statistics) in cases where they are not appropriate.

[3] The expected count in the cell in row i and column j is obtained as $E_{ij} = \frac{r_i c_j}{n}$ where r_i is the row total, c_j is the column total, and n is the grand total. For example, $E_{11} = \frac{r_1 c_1}{n} = \frac{27 \times 15}{30} = 13.5$

14.11.2 Hypothesis

The null hypothesis is the same in Fisher's Exact Test as in the ordinary χ^2 test, i.e. the null hypothesis is that the row and column variables are independent, or that the proportions in each row (or column) are equal.

In the example of the fox squirrels above, the hypothesis is that the predation rate of males is the same as the predation rate of females in the population of interest. Again, the hypothesis refers to the population of interest and NOT to the observed sample.

The alternate hypothesis in a 2 × 2 table can either be one-sided (predation rate of males is less than the predation rate of females) or two-sided (the predation rate of males is different than the predation rate of females). In larger tables, the concept of one-sided or two-sidedness isn't relevant, and the alternate hypothesis is that there is a difference in proportions somewhere in the experiment.

14.11.3 Computation

This section will outline the conceptual framework for the computation of Fisher's Exact Test using complete enumerations. The algorithms used in modern software do NOT do a complete enumeration and use sophisticated algorithms to speed up the computations.

Fisher's Exact Test starts by enumerating all the possible tables where the row and column total are fixed. In the fox-squirrel example, it enumerates all of the possible tables where there are 27 squirrels not predated; 3 squirrels predated; 15 females; and 15 males.

There are 4 possible tables:

Site	Sex	Predated	Not-predated
Rural	M	0	15
	F	3	12

Site	Sex	Predated	Not-predated
Rural	M	1	14
	F	2	13

Site	Sex	Predated	Not-predated
Rural	M	2	13
	F	1	14

Site	Sex	Predated	Not-predated
Rural	M	3	12
	F	0	15

For each table, the probability of each table is computed (conditional upon the set of possible tables). Note that for the tables above, we can index them by the number of males squirrels that were predated (the number in the upper left corner of the tables). Once n_{11} is specified, all of the other values are found by subtraction since the row and column total are fixed. The probability of each tables is found as:

$$P(N_{11} = n_{11}) = \frac{(\prod n_{i+}!)(\prod n_{+j}!)}{n_{++}! \prod \prod n_{ij}!}$$

where n_{i+} are row totals for row i; n_{+j} are the totals for column j, and n_{++} is the grand total. For example,

$$P(n_{11} = 0) = \frac{15!15!27!3!}{30!0!15!3!12!} = .112$$

The complete table of probabilities is:

n_{11}	Probability
0	.112
1	.388
2	.388
3	.112

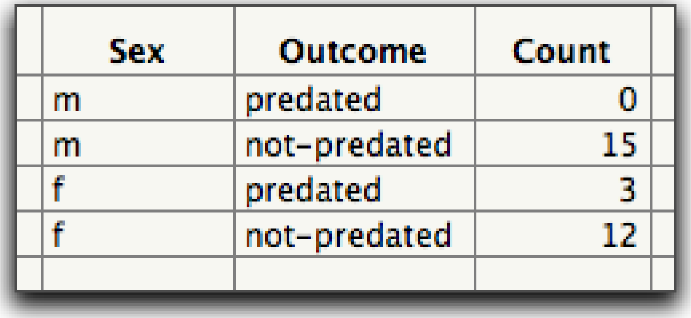
In general, the table is NOT symmetric in the probabilities.

For the one-sided alternatives, we add together all of the probabilities of tables that correspond to the observed outcome or "more extreme". For example, the observed number of predated-male squirrels is 0. If alternate hypothesis is that male squirrels had a lower predation rate than females, then number of predated-males should be 0 or less. There is only 1 table, and hence the one-sided *p*-value for this alternative is .112.

If the alternate hypothesis is that male squirrels had a higher predation rate than females, then the probabilities of tables corresponding to $n_{11} \geq 0$ are added together. This is actually all of the tables, so the one-sided *p*-value for this alternative is 1.00.

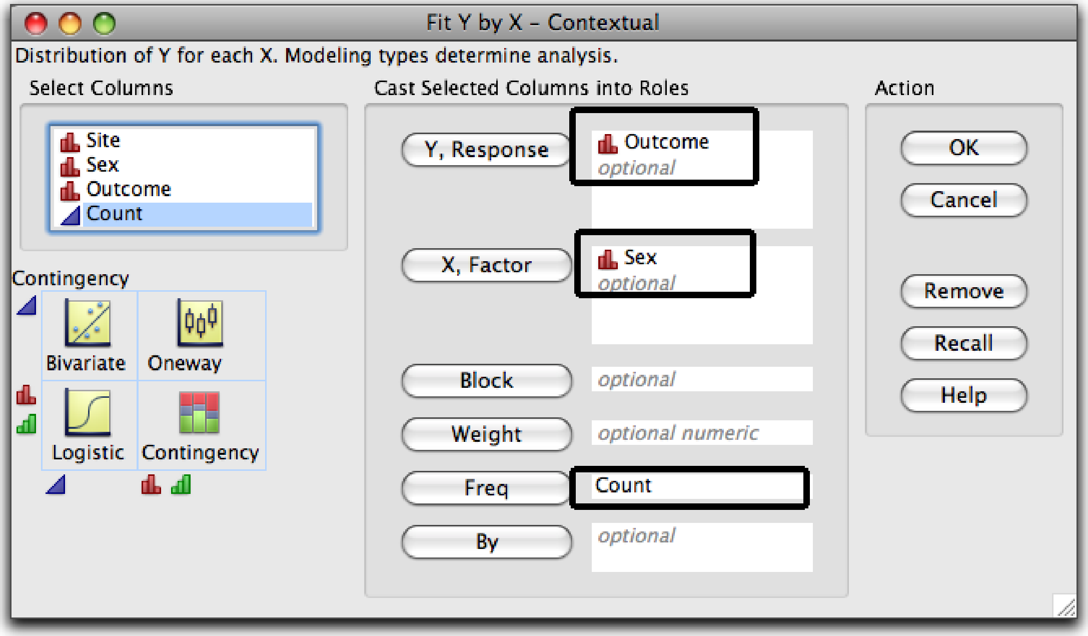
There are several suggested ways to compute the *p*-value for two-sided alternatives - refer to Agresti (2002), p.93. One method is to simply double the smallest one-sided *p*-value using a similar rule in normal theory tests. This would give a two-sided *p*-value of .224. A second method is to add probabilities of all tables with observed probabilities $\leq$ the probability of the observed table. In this case, this corresponds to the tables with n_{11} equal to 0 or 3, again giving a two-sided *p*-value of .224.

The data are entered in *JMP* in the usual fashion:



Sex	Outcome	Count
m	predated	0
m	not-predated	15
f	predated	3
f	not-predated	12

The *Analyze->Fit Y-by-X* platform is used:



Y, Response: Outcome
optional

X, Factor: Sex
optional

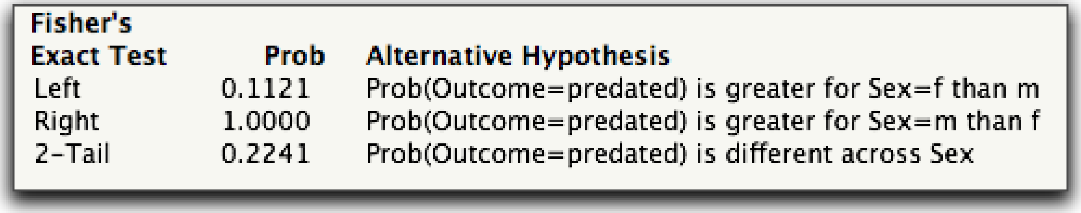
Block: optional

Weight: optional numeric

Freq: Count

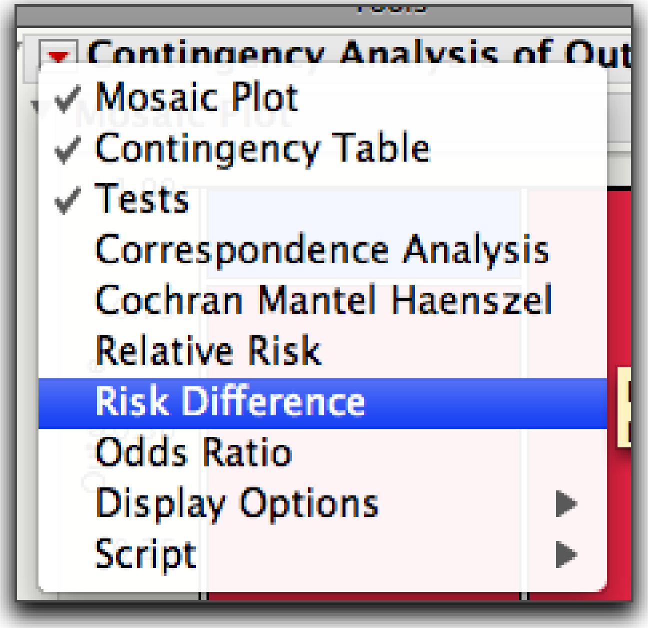
By: optional

JMP automatically computes Fisher's Exact Test for contingency tables with smallish counts. For this example, it reports:



```
Fisher's
Exact Test      Prob    Alternative Hypothesis
Left          0.1121    Prob(Outcome=predated) is greater for Sex=f than m
Right         1.0000    Prob(Outcome=predated) is greater for Sex=m than f
2-Tail        0.2241    Prob(Outcome=predated) is different across Sex
```

It is also possible to estimate the relative risk and odds ratios:



but these are not discussed in this section.

14.11.4 Example: Relationship between Aspirin Use and MI

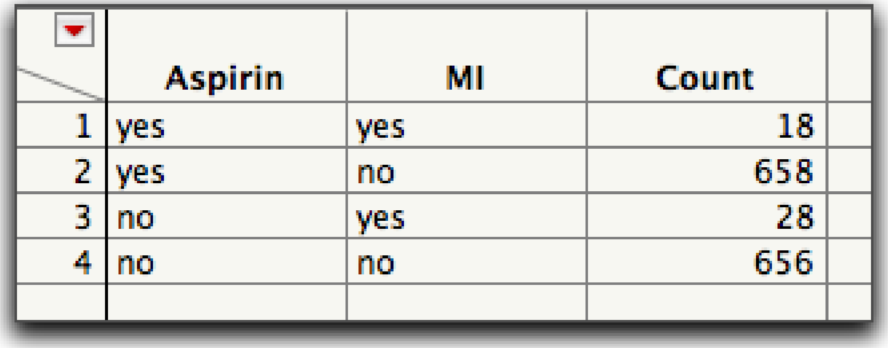
This example is taken from Agresti (2002), p. 72, Table 3.1. It concerns a long term Swedish study to examine the impact of taking aspirin upon subsequent heart attacks (mycardial infarctions).

The data are:

	MI	
Aspirin	Yes	No
Yes	18	658
No	28	656

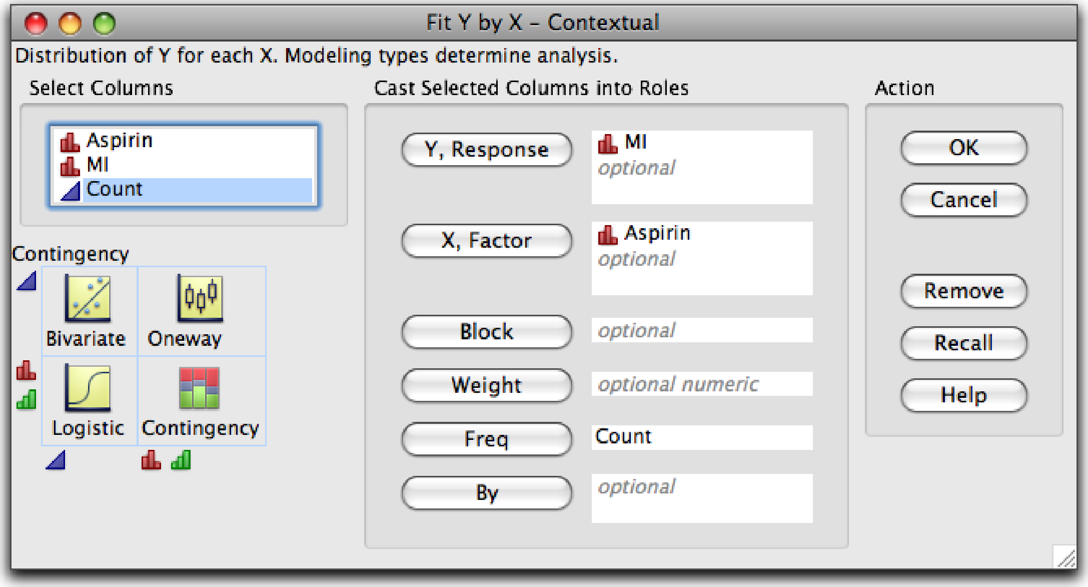
This is a classical randomized design and fulfills the sampling requirements for Fisher's Exact Test.

The data are entered in *JMP* in the usual fashion:



	Aspirin	MI	Count
1	yes	yes	18
2	yes	no	658
3	no	yes	28
4	no	no	656

The *Analyze->Fit Y-by-X* platform is used:



Distribution of Y for each X. Modeling types determine analysis.

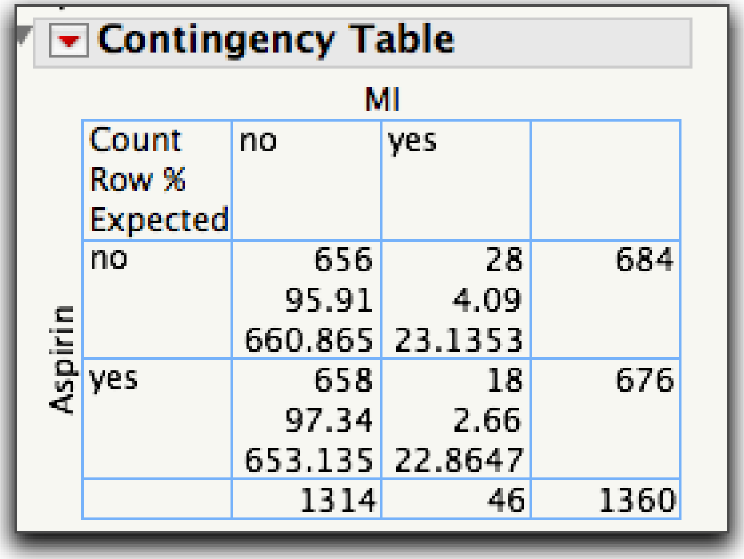
Select Columns: Aspirin, MI, Count

Y, Response: MI

X, Factor: Aspirin

Freq: Count

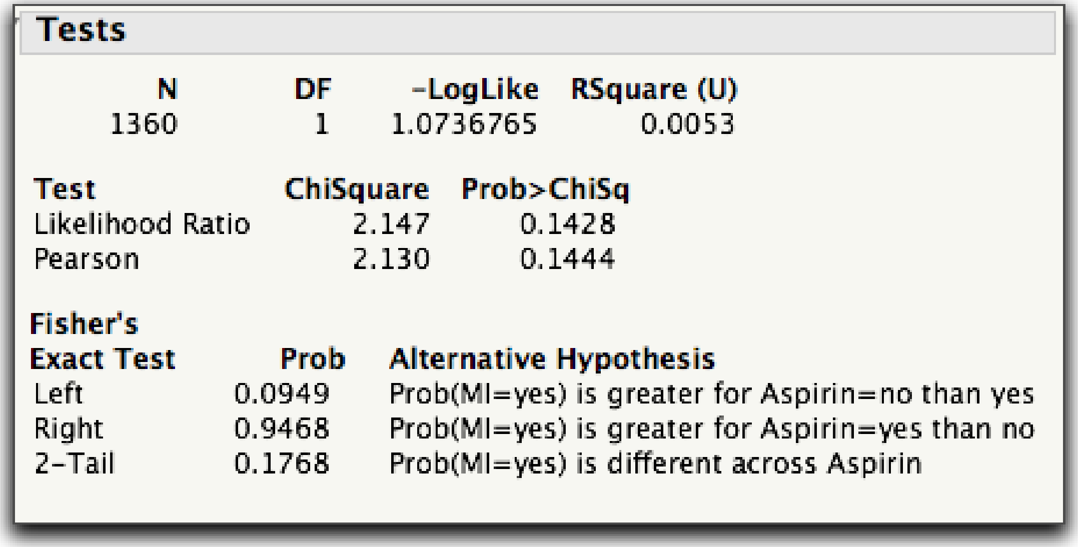
This gives the following contingency table:



Contingency Table

Aspirin	Count Row % Expected	MI: no	MI: yes	
	no	656 95.91 660.865	28 4.09 23.1353	684
	yes	658 97.34 653.135	18 2.66 22.8647	676
		1314	46	1360

The expected number in each cell is sufficiently large that the large-sample approximation for the χ^2 test should be valid. Both the large sample and Fisher Exact Test results are presented:



```
Tests
        N         DF      -LogLike   RSquare (U)
     1360          1    1.0736765        0.0053

Test              ChiSquare  Prob>ChiSq
Likelihood Ratio      2.147      0.1428
Pearson               2.130      0.1444

Fisher's
Exact Test          Prob    Alternative Hypothesis
Left              0.0949    Prob(MI=yes) is greater for Aspirin=no than yes
Right             0.9468    Prob(MI=yes) is greater for Aspirin=yes than no
2-Tail            0.1768    Prob(MI=yes) is different across Aspirin
```

The significance levels are fairly similar. Remember that the Pearson and Likelihood Ratio tests are two-sided alternative tests.

Mechanics of the test

There are now 47 possible tables with the following probabilities where n_{11} is the number of people who took aspirin and had a heart-attack. Our observed value is 18.

n11	Probability
0	8.547087e-15
1	4.159315e-13
2	9.870249e-12
3	1.522164e-10
4	1.715339e-09
5	1.505870e-08
6	1.072153e-07
7	6.364053e-07
8	3.212936e-06
9	1.400604e-05
10	5.334182e-05
11	1.791460e-04
12	5.345671e-04
13	1.426018e-03
14	3.418037e-03
15	7.392312e-03
16	1.447590e-02
17	2.574072e-02
18	4.166081e-02

19	6.148833e-02
20	8.288309e-02
21	1.021500e-01
22	1.152002e-01
23	1.189359e-01
24	1.124306e-01
25	9.729742e-02
26	7.704779e-02
27	5.578509e-02
28	3.688792e-02
29	2.224374e-02
30	1.220853e-02
31	6.084544e-03
32	2.745707e-03
33	1.117974e-03
34	4.090136e-04
35	1.337738e-04
36	3.887400e-05
37	9.961707e-06
38	2.230215e-06
39	4.311260e-07
40	7.088462e-08
41	9.716430e-09
42	1.080170e-09
43	9.354616e-11
44	5.919893e-12
45	2.434601e-13
46	4.882510e-15

The one-sided alternative that the probability of a heart-attack is LOWER when taking aspirin is found by adding up the individual cell probabilities for all tables where $n_{11} \leq 18$ or

$$P = 8.54 \times 10^{-15} + \ldots + .04166 = .09489$$

The two-sided alternative is that the probability of a heart attack is DIFFERENT when taking aspirin compared to not taking aspirin is found by adding up ail the probabilities of tables whose probability is $\leq .04166$. This corresponds to tables with n_{11} in the range of $0, \ldots, 18$ and $29, \ldots, 46$ or a total of .1768.

14.11.5 Avoidance of cane toads by Northern Quolls

This example was created by Nathan Nastili in 2012.

In 1935, the highly toxic cane toad was introduced to Australia to aid with pest control of scarab beetles. The beetles were wreaking havoc on sugarcane crops. Unfortunately, this decision led to an unforeseen and devastating effect on Australia's wildlife due to animal's consuming the toxic toad. Damage has included the possible extinction of the Northern Australian quoll. Although initiatives such as relocating the quoll on the nearby islands have been taken in order to save the species, there is no guarantee the new habitats will remain toadless. Scientists have developed a new plan of attack using conditioned taste aversion (**CTA**) in order to save the quoll population.

The goal of CTA is to have a subject associate a negative experience (illness) with an edible substance. More specifically, CTA is a conditioning technique applied to subjects (predators) in order to deter them from consuming poisonous substances (prey). CTA works as follows: Subjects (quolls) are given a non-lethal dose of the poisonous substance (toad) injected with a nausea-inducing chemical. Scientists hope the subject will remember the experience and avoid the substance in the future. In this case, scientists are hoping to have the quolls avoid the highly toxic toad.

The paper:

> O'Donnell, S., Webb, J.K. and Shine, R. (2010).
> Conditioned taste aversion enhances the survival of an endangered predator imperiled by a toxic invader.
> Journal of Applied Ecology, 47, 558-565.
> `http://dx.doi.org/10.1111/j.1365-2664.2010.01802.x`

discusses an experiment with Northern Quolls begin conditioned to avoid cane tides.

A sample of 62 quolls was taken (32 males and 30 females). The quolls were then split up into two treatment groups; toad smart (quolls that were given the CTA treatment, 15 males and 16 females) and toad naive (Control group, 17 males and 14 females). A samples of 34 quolls (21 males, 13 females) were subjected to a prescreening trial (prior to release into the wild).

The trial proceeded as follows: Both toad naive and toad smart quolls were subjected to a live cane toad in a controlled environment. The quolls were monitored using hidden cameras and their response was recorded. The response variable had three levels; attack (attacked the toad), reject(sniffed but did not pursue) and ignore. The observations for male and female quolls have been tabulated in the contingency tables below.

Male Quolls

	Attack	Avoid	Reject
Toad Naive	4	1	5
Toad Smart	1	0	10

Female Quolls

	Attack	Avoid	Reject
Toad Naive	0	4	1
Toad Smart	0	2	6

Unfortunately, *JMP* does not provide Fisher's Exact test for tables that are larger than 2 × 2. Please consult the results from *SAS* or *R*.

We are interested in determining whether there is a treatment effect on the response of the male and female quolls. In other words, we are testing the row (treatment) and column (response) of each table for independence. Notice in tables bigger than 2×2, the concept of one-side or two-sided tests does not exist and so the alternative is simply "not independent".

H_o : CTA had no effect on the proportion in each reaction class between naive and smart quolls.

H_a : CTA had an effect on the proportion in each reaction class between naive and smart quolls.

We start with the usual χ^2 test for each sex separately and print out the expected counts for each cell:

Unfortunately, *JMP* does not provide Fisher's Exact test for tables that are larger than 2×2. Please consult the results from *SAS* or *R*.

As we can see from above, the estimated expected value in the majority of cells for both contingency tables (male and female) have values less than 5, which suggests that a *chi-squared test* might not be trustworthy.

However, a *Fisher's exact test* may be used. *Fisher's exact test* is a procedure which determines the probability of observing tables with the same row and columns totals of the observed table. Using these probabilities it is possible to determine how much the observed table may deviate from what is expected if the hypothesis of independence is true.

Unfortunately, *JMP* does not provide Fisher's Exact test for tables that are larger than 2×2. Please consult the results from *SAS* or *R*.

For neither sex, is there enough evidence to reject the null hypothesis of independence between the treatment and response variables. This is not surprising given the small sample sizes.

It is possible to combine the results from the two-tables. You should NOT simply pool the data from the two sexes – this would be an example of sacrificial pseudo-replication. Rather you should combine the *p*-values from the two sexes using the discrete version of Fisher's method to combine *p*-values. Yes, this is the same Fisher that created the Exact Test.

For large sample cases, the distribution of the *p*-value will follow a Uniform distribution if the null hypothesis is true. Then you can combined the results using

$$\chi^2 = -2 \sum \log(p_i)$$

where p_i is the *p*-value from the ith test.[4] This is compared to a χ^2 distribution with $2k$ df where k is the number of tests being combined.

This method doesn't work well in the case of sparse contingency tables because now the null distribution of the *p*-value is no longer uniformly distributed. The paper:

> Mielke, P.W., Johnston, J.E, and Berry, K.J. (2004).
> Combining probability values from independent permutations tests: A discrete analog of Fisher's classical method.
> Psychological Reports, 95, 449-458.
> `http://dx.doi.org/10.2466/pr0.95.2.449-458`

[4] Refer to `http://en.wikipedia.org/wiki/Fisher's_method`.

describes a method of combining the two *p*-values. Following this method, the joint *p*-value is found to .0483 and so there is some (but not very strong) evidence of an effect of training the quolls. Consult the *R* code for details on computing the test.

How is the Fisher's Exact test computed? This section will outline the conceptual framework for the computation of Fisher's Exact Test using complete enumerations. The algorithms used in modern software do NOT do a complete enumeration and use sophisticated algorithms to speed up the computations.

Computation for Female data.

The Female data is a 2 × 2 table after dropping the column with all zeros. We start by enumerating all the possible tables where the row and column total are fixed in the same way as previously done. There are six possible tables with n_{11} ranging from 0 to 5 with probabilities

n_{11}	n_{12}	n_{21}	n_{22}	Probability
0	5	6	2	0.016317016
1	4	5	3	0.163170163
2	3	4	4	0.407925408
3	2	3	5	0.326340326
4	1	2	6	0.081585082
5	0	1	7	0.004662005

For each table, the probability of each table is computed (conditional upon the set of possible tables). Note that for the tables above, we can index them by the number of males squirrels that were predated (the number in the upper left corner of the tables). Once n_{11} is specified, all of the other values are found by subtraction since the row and column total are fixed. The probability of each tables is found as:

$$P(N_{11} = n_{11}) = \frac{(\prod n_{i+}!)(\prod n_{+j}!)}{n_{++}! \prod \prod n_{ij}!}$$

where n_{i+} are row totals for row i; n_{+j} are the totals for column j, and n_{++} is the grand total. For example,

$$P(n_{11} = 0) = \frac{5!10!7!6!}{13!0!5!6!2!} = .0163$$

In general, the table is NOT symmetric in the probabilities.

For the one-sided alternatives, we add together all of the probabilities of tables that correspond to the observed outcome or "more extreme". For example, the observed number of Naive-Avoid quolls is 4. If alternate hypothesis is that naive quolls had a lower predation rate than smart quolls, then number of naive-avoid quolls should be 4 or more. There are 2 such tables, and hence the one-sided *p*-value for this alternative is $.08159 + .004662 = .08625$.

There are several suggested ways to compute the *p*-value for two-sided alternatives - refer to Agresti (2002), p.93. One method is to simply double the smallest one-sided *p*-value using a similar rule in normal theory tests. This would give a two-sided *p*-value of .1725. A second method is to add probabilities of all tables with observed probabilities $\leq$ the probability of the observed table. In this case, this corresponds to the tables with n_{11} equal to 0, 4 or 5, giving a two-sided *p*-value of $.08159 + .004662 + .016317 = .1025641$ which is the value reported.

Computation for Male data.

It is possible to do a Fisher Exact test for larger than 2×2 tables, but it can be tedious and good computing algorithms have been developed to enumerate all of the possible tables subject to the row and column constraints. Here, we simply enumerated then by brute force.

n_{11}	n_{12}	n_{13}	n_{21}	n_{22}	n_{23}	Probability
0	1	9	5	0	6	0.014189886
1	1	8	4	0	7	0.091220699
2	1	7	3	0	8	0.182441398
3	1	6	2	0	9	0.141898865
4	1	5	1	0	10	0.042569659
5	1	4	0	0	11	0.003869969
0	0	10	5	1	5	0.008513932
1	0	9	4	1	6	0.070949432
2	0	8	3	1	7	0.182441398
3	0	7	2	1	8	0.182441398
4	0	6	1	1	9	0.070949432
5	0	5	0	1	10	0.008513932

For each table, the probability of each table is computed (conditional upon the set of possible tables). The probability of each tables is found as:

$$P(Table) = \frac{(\prod n_{i+}!)(\prod n_{+j}!)}{n_{++}! \prod \prod n_{ij}!}$$

where n_{i+} are row totals for row i; n_{+j} are the totals for column j, and n_{++} is the grand total. For example,

$$P(\text{Table with } n_{11} = 0) = \frac{10!11!5!1!15!}{21!4!1!5!1!0!10!} = 0.014189886$$

Notice that in generally, you would not do all that multiplication of factorials then followed by division by all those factorials as this is VERY subject to round-off and other numerical difficulties. The above computation is for illustration purposes only.

In larger tables, the concept of one-sided or two-sided doesn't really exist (similar to what happens in ANOVA with 3+ levels). There are several suggested ways to compute the *p*-value for larger tables. One method is to add probabilities of all tables with observed probabilities $\leq$ the probability of the observed table. In this case, this corresponds to the table in the 5th row above, or all tables with probabilities $le.04256966$. This gives a *p*-value of $.04256966 + 0.014189886 + 0.003869969 + 0.008513932 + 0.008513932 = 0.07766$ which is the value reported.

Combining the *p*-values from the two tables.
The joint *p*-value is computed by first finding the product of the probabilities of the observed tables, i.e. $p_{obs,joint} = .08159 \times .04256966 = 0.003473049$. This is compared to the product of all possible pairs of tables (one from M and one from F). In this case there are $12 \times 6 = 72$ possible pairs of tables. We find again the sum of all the joint table probabilities that don't exceed the value 00347 and get a joint *p*-value of 0.0483845.