EXST SAS Lab
Lab #5: Observations with frequencies

Objectives

1. Infile an external dataset (CSV file)
2. Use PROC FREQ to obtain the frequencies and compare the results with and without a WEIGHT statement
3. Use PROC CHART (VBAR) to get a histogram using the SUMVAR = option
4. Use PROC FREQ again to do a Chi square analysis for both a two-way analysis and a one-way analysis.

Analyzing data with frequencies

Assignment 5 will require running two procedures (FREQ and CHART) on data that has a variable for the number of occurrences of each observation. Most datasets have one line per observation. However, for some data sets it is not necessary to write out 200 lines for 200 observations. If the data is in 2 categories (e.g. In-state Resident and Gender) then there may only be 4 combinations of values and the 200 observations may be expressed on 4 lines. For example;

Resident	Gender	Number
Yes	Male	60
No	Male	30
Yes	Female	70
No	Female	40

This coding of the data is obviously much more efficient than 200 separate lines with only the two variables RESIDENT and GENDER. Since most SAS procedures expect to encounter one observation per line, an additional mechanism is needed to indicate when multiple observations are present. Many procedures use a WEIGHT statement (like PROC FREQ) while others have an option in the procedure to specify category frequency (like PROC CHART). We will look at both types.

Obs	CareType	Failures	Students
1	Relative	Yes	54
2	Non-relative	Yes	5
3	Center	Yes	53
4	Self	Yes	124
5	Parent	Yes	176
6	Multiple	Yes	68
7	Relative	No	366
8	Non-relative	No	76
9	Center	No	393
10	Self	No	897
11	Parent	No	1824
12	Multiple	No	330

The example data file is from a 2010 Ph. D. dissertation in Education by Barbara Ann Baisley at George Mason University titled *After School Care Arrangements and Student Academic Performance and Misbehavior in Middle School*. This dataset contains only 12 different categories, a two way table of the variable FAILURE (having failed a class prior to the grades covered by the study, grades 6 – 8) and the CareType (type of after-school care available to the student). Levels of care are: care by a relative, a nonrelative, a care center, a parent, cared for self or some combination of these, listed as “multiple sources”. This dataset consists of observations on 4366 students by can be summarized as the frequency of occurrence in 12 combinations of these categories. The SAS output of the data is given below.

The data was input with the usual SAS statements at the head of the program. Data input was done by an INFILE statement from a comma separated value file (.CSV). A PROC PRINT was used to produce the listing above, but including an asterisk (*) in the front of the statement insures that it will be only a comment and not produce output every time the program is run.

Note that a LENGTH statement was included in the program. By default, SAS reads character variables with 8 characters. The value "Non-relative" would be read as "Non-Rela". In order to get the full variable value, the length statement specifies that the variable caretype is a character value (i.e. $) and is to be read and stored with 12 characters.

PROC FREQ: The FREQUENCY procedure obviously does a count of the frequency of whatever variables are specified in the TABLE statement and will list those frequencies with row, column and overall percentages. If PROC FREQ is run on this dataset we get something of a surprise, every combination of the variables CareType and Failures occurs just once and the "1" in every cell constitutes 8.33% of the table total, 50% of the row total and 16.67% of the column total.

```
proc freq data=Failures order=data;
   Title3 'Two-way frequencies of CareType and Failures';
   table CareType * Failures;
run;
```

This occurs, of course, because by default the procedure **just counts the frequency of each observation.** In our dataset each combination of the variables CareType and Failures occurs just once. In order to correctly register the frequency of each category we need a mechanism to convey to the program the fact that there is an additional variable, "**Students**", that contains frequencies for the categories. This is accomplished with an additional statement called the "WEIGHT" statement that applies a weight to the calculation of each cell equal to the value specified in the weight variable. For this dataset the weight variable was called "STUDENTS". The following program will produce a properly calculated table for the 4366 students.

```
proc freq data=Failures order=data; weight Students;
   Title3 'One-way frequencies of CareType and Failures';
   table CareType * Failures;
run;
```

Other procedures have other mechanisms to indicate that each observation represents a multiple of individuals. In addition to the WEIGHT statement, there is a FREQ statement that is used by most other procedures. Some procedures have both a WEIGHT statement and a FREQ statement with slightly different ways of impacting the calculations. While a FREQ statement, when present, always counts as multiple observations, the WEIGHT statement may only give extra emphasis to a value in the calculations without counting as multiple observations.

PROC CHART: Specifying a frequency value is handled differently in some other procedures. In the PROC CHART procedure the option that specifies the name of the variable with frequencies is called the SUMVAR= statement. If this statement is not used the program assumes that each observation occurs only once.

```
proc chart data=Failures;
Title4 'Histogram (horizontal) of Failures and CareType';
hbar CareType / subgroup=Failures sumvar=Students
midpoints = "Relative" "Non-Relative" "Center" "Self"
          "Parent" "Multiple";
run;
```

More PROC FREQ

The Chi Square Test of Independence: We will run two additional examples of PROC FREQ. It may be a bit premature, but while we are using this procedure we should look at a Chi Square analysis. As you will have noticed, the procedure put 4 values in each *cell* (i.e. each combination of the row and column variables). When we do Chi square we are going to add several additional values to each cell. In order to avoid the confusion of too many variables in each cell we will suppress the inclusion of the row, column and overall percentages. This is accomplished by adding the options "norow", "nocol" and "nopercent" to the TABLE statement after the slash (i.e. /) that typically starts the list of options in SAS. We will also add three new options dealing with the Chi square. The option "chisq" asks for a chi square test and the additional options "cellchi2" and "expected" request the individual chi square contribution to the total chi square and the expected frequency based on the marginal totals. These calculations will be explained in more detail in class when the Chi square test is covered.

```
proc freq data=Failures order=data; weight Students;
   Title3 'Two-way frequencies of CareType and Failures';
   table CareType * Failures / norow nocol nopercent
                          chisq cellchi2 expected;
run;
```

This yields two-way table and produces a Chi square "test of independence" of the type of after school care and the occurrence of an early grade failure. Chi square tests compare the observed frequencies against some "expected value" for the frequency. For the test of independence the "expected value" is derived from the marginal. For example, in the two way table there are 4366 total observations. The first cell row has 9.62% of the total, so we expect the first row to have 9.62% of 4366 = 420, which is the first sum in the margin (e.g. marginal total). This number in the first row is split between two columns. Based on the marginal totals at the bottom of the table, the first column has 10.99% of the total and the second column has 89.01% of the total. Therefore, since the first row has a frequency of 420, the expected values for the first column is 10.99% of 420 (=46.175) and the expected value for the second column is 89.01% of 420 (=373.83). Expected values are calculated his way for each cell.

The chi square calculation is the $\sum \frac{(Observed - Expected)^2}{Expected}$. If every expected were to exactly match the observed then the sum would be zero and we could conclude that the each cell occurred in proportion to its frequency. This is consistent with the null hypothesis. If, however, the chi square sum was quite large it would suggest that some observed values were not in line with what was expected from the marginal totals. We would then reject the null hypothesis and try to determine which cells were inconsistent with the overall pattern. In this case we would examine the table that has the observed and expected frequencies and the individual cell chi square values. Since the chi square sum was unexpectedly large if we rejected we want to determine which cells most contributed to making that sum large.

In our example the sum of the chi square values was 30.0871 with 5 degrees of freedom. The P-value indicates that a value that large with 5 d.f. would occur less that once in 10,000 observations (i.e. P<0.0001). Degrees of freedom for this test are calculated as (number of rows – 1)(number of columns – 1) = (6 – 1)(2 – 1) = 5.

So we have observed a rare event and readily reject the null hypothesis. So which cells are out of line with the overall pattern? The two biggest contributors to the chi square are the PARENT and MULTIPLE categories, both in the Failure = YES column. For the PARENTS we see that we observed 176 failures and expected 219.88. So, students appear to fail less often than expected if after school care is provided by the parents. For the MULTIPLE-care group, the observed value was 68 and the

expected was only 43.756. For this group there are more failures than would be expected based on the overall patterns.

Those are the statistical results. At this point the researcher would draw conclusions. We aren't qualified to draw conclusions about this study, but we will anyway. It is easy to imagine why students receiving after school care from parents would have fewer failures than other groups. The parents have both more authority over the students and are more intensely interested in the student's success. It is not as clear why the MULTIPLE-care category had a significantly higher number of failures. One might speculate that students with multiple sources of care do not have as constant and as reliable sources of after-school care as other groups, and that may lead to more failures. However, without a better understanding of the study this is pure speculation.

The Chi Square test of Goodness of Fit: One other point on Chi square tests with PROC FREQ. If you request a one way table then the procedure will do a chi square test of "goodness of fit". You can provide expected probabilities base on some preconceived expectations on the frequencies. Many genetics experiments expect to see back crosses exhibiting a ratio of traits of 9:6:1 or 9:3:3:1. The expectations of 9/16 = 0.5625, 3/16=0.1875 and 1/16=0.0625 can be specified and tested in SAS. So the goodness of fit test does not use the marginal to derive expected values. The goodness of fit test requires that the investigator have some idea of what the distribution is and can expectations based on those values hypothesized by the investigator.

If there is no list of expected probabilities and a one way table is given with a Chi square test, then the analysis tests to see if the frequency of each category is equal. In this example there were 6 categories so the test would test to see if the observations in each category were 1/6 of the total, 16.67% or the proportion 0.1667 of the total. Since there was a total of 4366 the expected value for each cell would be 01667*4366 = 727.667.

```
proc freq data=Failures order=data; weight Students;
   Title3 'One-way frequencies of CareType and Failures';
   table CareType / norow nocol nopercent chisq cellchi2 expected;
run;
```

Clearly, we don't expect that after school day care for these 6 types occurs in equal proportion. However, this is tested in the example as a demonstration. Again, P<0.0001, so if the null hypothesis were true we would see a value of 3306.0870 with 5 d.f. less that once in 10,000 trials. Degrees of freedom for goodness of fit tests are calculated as (number of categories – 1) = (6 – 1) = 5.

The PROC FREQ can also be run on to test a 50:50 ratio, such as a sex ratio on many animal populations. If we do a frequency table with only two levels, such as our failed and not failed, and if we do not specify proportions, SAS will test for equal proportions (i.e. a 50:50 ratio). To make things more interesting and realistic, since failures were many fewer than non-failures the test in the example is to see if failures equal 10% of the total. The results show that the observed proportion is pretty close to 10%, actually 10.99%. However, with only 1 d.f. the observed, the Chi square 4.7935 was large enough to cause rejection of the null hypothesis at the P=0.0286 level of significance. This means that if the null hypothesis was true (i.e. there is actually a 10% to 90% split) then we would observe a Chi square value as large as 4.7935 about 2.86 percent of the time. By convention we usually reject the null hypothesis for P-values of less that 5% ($\alpha = 0.05$) in favor of the alternate hypothesis (that the split is not a 10:90 ratio).

Chi square tests are a future class topic and are included here to take advantage of our presentation of PROC FREQ.

Assignment 5

We have data from a questionnaire on Laptop purchases. The objective is to describe several factors affecting this purchase. Complete the following assignment and turn in a print of your LOG and your results from either the OUTPUT window or RESULTS viewer window.

1) You will want to include in your program the "usual statements" with option, comments and titles similar to those in previous assignments. (1 point)

2) Also, include appropriate title statements as usual. (1 point)

> The dataset is from a Ph. D. Dissertation, Capella University, December 2006 by Rachel V. McClary titled "An Evaluation of Consumer Buying Criteria and Its Impact on the Purchase of Commoditized Laptops". The data set was downloaded on 30Sept2008 http://www.drjimmirabella.com/dissertations/Dissertation-RachelMcClary.pdf

The study covered a number of variables that might influence a decision to purchase a particular BRAND of Laptop computer such as education, computer expertise gender and the IMPORTANCE of the particular brand to the purchaser. We will analyze the latter variable, importance.

3) The dataset is stored as a CSV dataset. Input the dataset as an external dataset using an INFILE statement. (1 point)

The BRANDs considered were: Apple, Compaq, Dell, HP and Toshiba. The IMPORTANCE of choosing a particular brand was categorized by the purchasers as: Not at all Important, Minimally Important, Somewhat Important, Important, Most important. The third variable in the dataset is the number of respondents in each category (i.e. the frequency).

4) Produce frequency table of the two variables, BRAND and IMPORTANCE without a weight statement (1 point)

5) Add a weight statement of the variable RESPONDENTS to the previous frequency table (1 point)

6) Produce a horizontal bar chart with PROC CHART including the option SUMVAR= for respondents. In order to keep the rank of this ordinal variable, specify the midpoints as: "NotAtAll" "Minimally" "Somewhat" "Important" "Most". (1 point)

7) Produce frequency table of the two variables, BRAND and IMPORTANCE with a weight statement and with a Chi square test and the EXPECTED and CELLCHI2 options. Suppress the listing of percentages with the options "/ norow nocol nopercent" and get the chi square and ancillary values with the additional options "chisq cellchi2 expected". (1 point)

8) Observe that each cell contains an observed frequency and an expected frequency. When a Chi square test is statistically significant there should be a notable discrepancy between these values. The third value, the cell Chi square value, provides a relative measure the size of the discrepancy in each cell. Notice that the frequency of people buying Toshiba is generally more than expected for a rating of "low importance" and are lower than expected for a rating as "important" or "most important". So, Toshiba purchasers are not much influenced by brand.
Question: People who feel that BRAND is the most important consideration are most likely to buy which one of the 5 brands? (1 point)

9) Produce a one-way frequency table of the variable BRAND, also with a weight statement and a Chi square test (chisq) of equal frequency of the various brands. (1 point)
Question: Do the 5 brands occur with equal frequency in this group of consumers? (1 point)