The FREQ Procedure

PROC FREQ is a procedure that can describe a data set by producing frequency counts and cross tabulation tables. PROC FREQ can also produce tests of hypotheses for cross tabulations or n-way tables. The data to be analyzed in the FREQ procedure are nominal or ordinal data.

The syntax of the FREQ procedure is as follows:

```
PROC FREQ  DATA=tablename  <options>;
TABLES  requests  </options>;
WEIGHT variable;
```

PROC FREQ statement

Some of the options for the PROC FREQ statement are as follows:

PAGE prints only one table per page. Otherwise, PROC FREQ prints multiple tables per page as space permits.

ORDER = *value* specifies the order in which the variable levels are to be reported.

Value of ORDER =	Data is sorted and reported by
DATA	order of appearance in the input data set
FREQ	descending frequency count; levels with the most observations appear first If no ORDER is specified, variable levels are reported in ascending numeric or alphabetic order

NLEVELS displays the number of levels of each variable specified in the TABLES statement.

REMEMBER: If a BY statement is used, use the SORT procedure to sort the data first.

TABLES statement

In a TABLES statement many requests can be made. More than one TABLES statement may be used in a single FREQ procedure, and each TABLES statement can have different options. One-way and n-way tables can be requested using one or more TABLES statements. To illustrate the allowable syntax for making table requests, suppose that you have a data set with variables A, B, C, D, E, and F in it. These variables are read into the data set in this order in the INPUT statement.

For one-way frequency tables

TABLES a b c d e f ;	is equivalent to	TABLES a -- f;

Caution when using double dash --: All variables in a list defined in the INPUT statement beginning with variable a and with last variable f are included.

For two-way frequency tables

TABLES a*b a*c ;	is equivalent to	TABLES a * (b c);
TABLES a*c a*d b*c b*d ;	is equivalent to	TABLES (a b) * (c d) ;
TABLES a*d b*d c*d ;	is equivalent to	TABLES (a -- c) * d ;

The table below groups TABLE statement options into categories. This is not a complete list of options. See the SAS/STAT Help and Documentation for the complete list of options. Each option is described in detail in alphabetic order immediately following the table.

Task	Options
Specify statistical analysis	CHISQ
	EXACT
	MEASURES
Request further information	CELLCHI2
	CUMCOL
	EXPECTED
Control printed output	LIST
	NOCOL
	NOCUM
	NOFREQ
	NOPERCENT
	NOROW
	PLOTS=*requests*

CELLCHI2

prints each cell's contribution to the total χ^2 statistic. This is computed as (frequency - expected)^2 / expected.

CHISQ

For two-way contingency tables, χ^2 tests of independence and computes measures of association based on χ^2 are computed. For one-way tables, a χ^2 test of equal proportions across the classes of the specified variables.

CUMCOL

prints the cumulative column percentages in the cells

EXACT

performs Fisher's exact test for tables that are larger the 2x2.

EXPECTED
prints the expected cell frequencies under the hypothesis of independence. Expected cell frequencies are not printed when the LIST option is specified.

LIST
prints two-way to n-way tables in a list format rather than as cross tabulation tables. The LIST option cannot be used when statistical tests or measures of association are requested.

MEASURES
requests a basic set of measures of association and their asymptotic standard errors.

NOCOL
suppresses printing of the column percentages in cells of a table

NOCUM
suppresses printing of the cumulative frequencies and cumulative percentages of one-way frequencies and for frequencies in the list format.

NOFREQ
suppresses the display of cell frequencies in crosstabulation tables. The NOFREQ option also suppresses row total frequencies. This option has no effect for one-way tables or for crosstabulation tables in list format, which you request with the LIST option.

NOPERCENT
suppresses printing of cell percentages for a cross tabulation. This also suppresses printing of percentages for row totals and column totals in a cross tabulation. For one-way frequencies and frequencies in list format, the NOPERCENT option suppresses printing of percentages and cumulative percentages.

NOROW
suppresses printing of the row percentages in cells of a cross tabulation.

PLOTS = ALL | NONE | FREQPLOT <options>
There are several ODS Graphics available in the FREQ procedure depending upon the calculations requested in a TABLES statement. One can suppress all of the plots (PLOTS=NONE) or request all available plots (PLOTS=ALL) for the type of analysis requested. Two of the basic plots available are presented here.
PLOTS=DEVIATIONPLOT are the default for a one-way classification when CHISQ is also specified on the TABLES statement.
FREQPLOTs may be requested for one-way and two-way classifications.
PLOTS=FREQPLOT produces frequency plots for each level of the variable in a one-way classification, and produces frequency plots for each level of variable2 within each level of variable1 when TABLES variable1*variable2 is written. Switch "variable1" with "variable2" to reverse the roles of the variables in the graph. The FREQPLOTS are bar charts (TYPE=BARCHART) by default. Dot plots can also be requested.
PLOTS=FREQPLOT(TYPE=DOTPLOT).

PLOTS(ONLY) = (*list of plot requests*) ONLY will restrict the default graphics and only those requested will be produced.

WEIGHT statement
Normally, each observation contributes a value of 1 to the frequency counts. When you use a WEIGHT statement, each observation contributes the weighting variable's value for the observation.

Statistical Tests

For a one-way frequency table

H_0: all classes have the same proportion, $p_1 = \ldots = p_t$ ($= p_0$, a specified value)

H_1: at least one of the classes differs

Test Statistic: $\chi^2 = \sum_{i=1}^{t} \frac{(n_i - E_i)^2}{E_i}$ t = # of classes

n_i = observed frequency in class i

E_i = expected frequency for class i = (total sample size) x p

Reject H_0 if $\chi^2 \geq \chi^2_{\alpha,(t-1)}$

For a two-way contingency table

H_0: Two variables are independent.

H_1: Two variables are related.

Test Statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$ r = # of rows, c = # of columns

n_{ij} = observed frequency in row i, column j

E_{ij} = expected frequency in row i, column j
= (total sample size) x (observed proportion in row i, column j)

Reject H_0 if $\chi^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$

Both of these test procedures require:
1. Random Samples
2. Large Samples indicated by the expected frequencies larger than 5.

If the large sample size condition is not met, there are small sample test alternatives. These alternatives methods are taught in categorical data analysis classes.

Objective 4: Conduct a chi-square test for independence for the following problem. Investigate the effects of the WEIGHT statement.

A random sample of 202 residents is taken, and each resident is surveyed regarding his/her opinion on Right to Work issues and the employment status is also recorded. Suppose the results of the survey appear as shown in table below. Is there evidence to indicate that a person's opinion concerning Right to Work depends on his or her employment status? That is, can we conclude that the two variables are dependent?

Opinion on Right to Work

Employment Classification	Favor	Do Not Favor	Undecided	Totals
Industry	20	24	16	60
Business	40	51	9	100
Unemployed	20	15	7	42
Totals	80	90	32	202

```
DM 'LOG; CLEAR; ODSRESULTS; CLEAR; ';

TITLE 'Objective 4';
DATA   A;
INPUT CLASS  $  OPINION $  Y  @@ ;
DATALINES;
I  F 20     I  DNF   24   I   U  16
B  F 40     B  DNF   51   B   U   9
U  F 20     U  DNF   15   U   U   7
;
* The following FREQ procedure will count the number of occurrences *;
* of the levels of the variables CLASS and OPINION, and the         *;
* CLASS/OPINION combinations occurring in the data set.             *;
* Most large survey data sets would require the following steps for *;
* a frequency analysis.  *;
PROC FREQ  DATA=A  ;
TABLES   CLASS  OPINION   CLASS*OPINION / LIST;

* The WEIGHT statement is necessary in order to get a correct analysis ;
* when the counts for each level of a variable are included in the data;
PROC FREQ  DATA=A  ORDER=FREQ;
TABLES  CLASS OPINION CLASS*OPINION/ CHISQ;
WEIGHT  Y;
RUN;

QUIT;
```

Compare the output from the two FREQ procedures. Examine the effect of the LIST option in the first FREQ procedure versus the default output in the second FREQ procedure.

Also, examine the default ODS Graphics from this procedure for one-way and two-way classifications with and without the LIST option.

Objective 4

TABLES statement with a LIST option. Note no ODS Graphics are produced.

As you read these tables, note that it appears that only 9 observations are in the data which is definitely not true. There were 202 observations, not 9! This is due to the format of the data we entered. The "preliminary counts" of y are not considered in this FREQ procedure. We need to include the "Y" information in the analysis.

Though we do not have the correct frequency counts, we can observe the effects of the LIST option.

The next two tables are one-way classifications with a LIST option.

The FREQ Procedure

CLASS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	3	33.33	3	33.33
I	3	33.33	6	66.67
U	3	33.33	9	100.00

OPINION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DNF	3	33.33	3	33.33
F	3	33.33	6	66.67
U	3	33.33	9	100.00

The following table is a two-way classification with a LIST option.

CLASS	OPINION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	DNF	1	11.11	1	11.11
B	F	1	11.11	2	22.22
B	U	1	11.11	3	33.33
I	DNF	1	11.11	4	44.44
I	F	1	11.11	5	55.56
I	U	1	11.11	6	66.67
U	DNF	1	11.11	7	77.78
U	F	1	11.11	8	88.89
U	U	1	11.11	9	100.00

Objective 4

This is the output from the second FREQ procedure.

One-way classification tables are formatted the same as with the LIST option. Note the effects of the WEIGHT y ; statement. We now have 202 total observations included in the analysis.

The FREQ Procedure

CLASS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	100	49.50	100	49.50
I	60	29.70	160	79.21
U	42	20.79	202	100.00

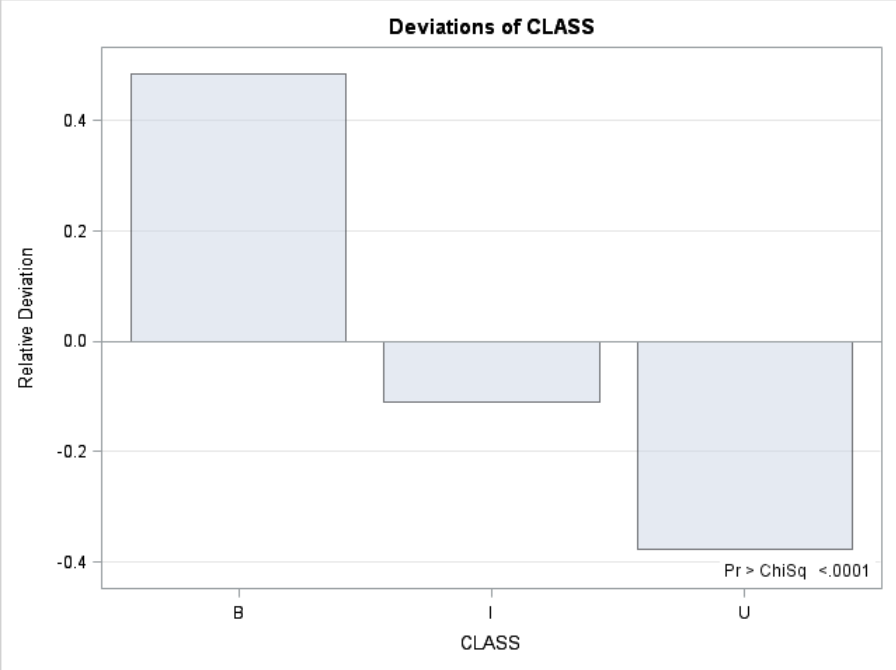
The table below results from the CHISQ option and a one-way classification request. The hypotheses being tested are

H_0: all classes have the same proportion, $p_1 = \ldots = p_t$

H_1: at least one of the classes differs

Chi-Square Test for Equal Proportions	
Chi-Square	26.1782
DF	2
Pr > ChiSq	<.0001

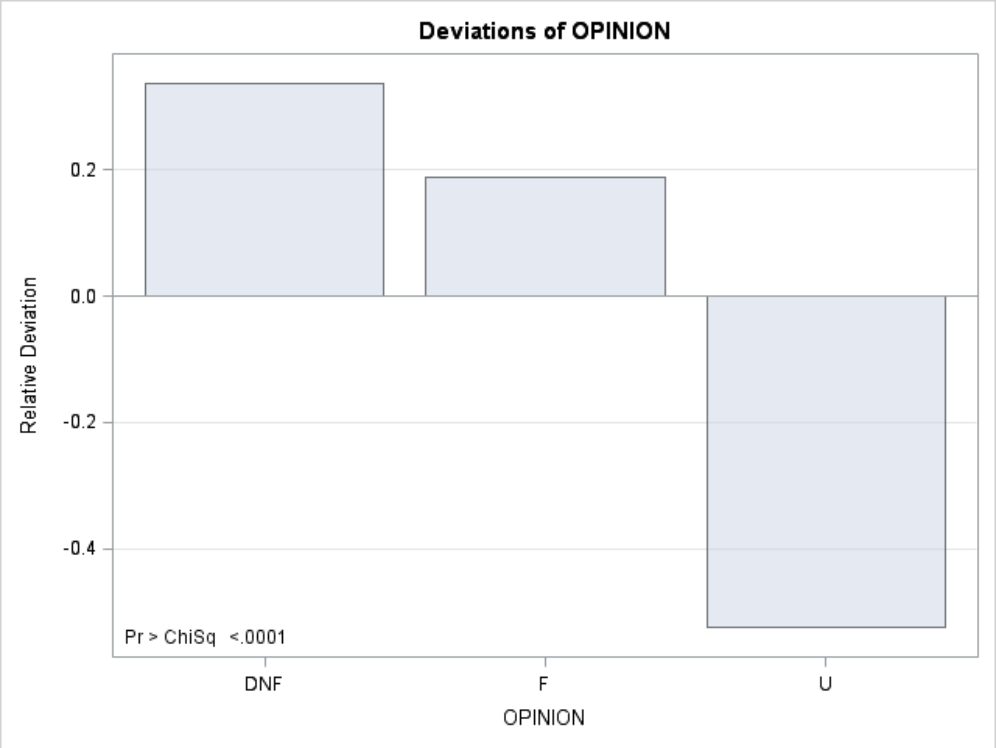
Below: This is called a deviation plot. This is the default ODS Graphic for a one-way classification. Each bar height is the value of $(O_i - E_i)/E_i$ where $E_i = np = 202(0.333333) = 67.3333$



Sample Size = 202

OPINION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DNF	90	44.55	90	44.55
F	80	39.60	170	84.16
U	32	15.84	202	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	28.5545
DF	2
Pr > ChiSq	<.0001



Sample Size = 202

These one-way classification results for Opinion are similarly interpreted.

This the default output for a two-way classification. Note the first variable specified in determines the rows of the table. That is, TABLES class*opinion . Note the levels of each variable are put in ascending order by default.

The upper left cell of the table is the legend for the table. Some of the TABLES statement options can change this information.

Frequency Percent Row Pct Col Pct	Table of CLASS by OPINION			
CLASS	OPINION			
	DNF	F	U	Total
B	51	40	9	100
	25.25	19.80	4.46	49.50
	51.00	40.00	9.00	
	56.67	50.00	28.13	
I	24	20	16	60
	11.88	9.90	7.92	29.70
	40.00	33.33	26.67	
	26.67	25.00	50.00	
U	15	20	7	42
	7.43	9.90	3.47	20.79
	35.71	47.62	16.67	
	16.67	25.00	21.88	
Total	90	80	32	202
	44.55	39.60	15.84	100.00

Statistics for Table of CLASS by OPINION

Statistic	DF	Value	Prob
Chi-Square	4	10.6405	0.0309
Likelihood Ratio Chi-Square	4	10.4423	0.0336
Mantel-Haenszel Chi-Square	1	4.5802	0.0323
Phi Coefficient		0.2295	
Contingency Coefficient		0.2237	
Cramer's V		0.1623	

Sample Size = 202

The first statistic "Chi-Square" is the test statistic for the test of independence. For the remaining test statistics, one should review categorical data procedures. Those tests are not overviewed here.

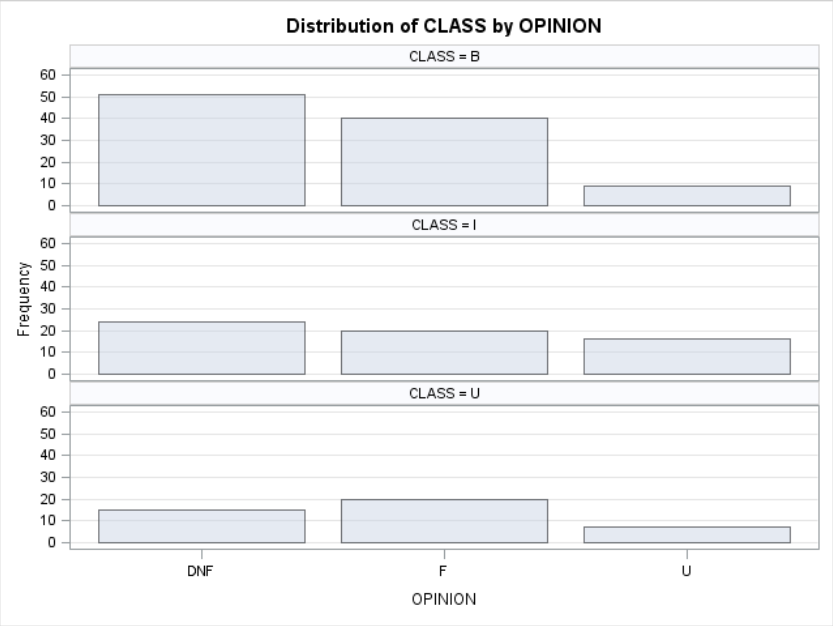
Notice there is no default ODS Graphics for a two-way classification.

Objective 5: In the second FREQ procedure, experiment with the options on the TABLES statement: NOROW NOCOL EXPECTED and other printed output control options.

Objective 6: In the second FREQ procedure, use the ORDER=DATA option. Note the rows and columns occur in the same order as given in the text of Objective 3 when the DATA step mimics that order.

Objective 7: Add PLOTS=FREQPLOT to the TABLES statement of the second FREQ procedure. That is,

```
ODS HTML ;
ODS GRAPHICS ON;
PROC FREQ  DATA=A  ORDER=FREQ;
TABLES  CLASS OPINION CLASS*OPINION/ CHISQ PLOTS=FREQPLOT;
WEIGHT  Y;
RUN;
```



Note that the deviation plots for the one-way classifications are still produced in addition to the above FREQPLOT for the two-way classification. The frequency histograms for the one-way classifications are also produced, but not shown here.
Rerun the program modifying the PLOTS option: PLOTS(ONLY) = FREQPLOT

Try the option: PLOTS(ONLY)=FREQPLOT(TYPE=DOTPLOT)

The chi-square goodness of fit test for the one-way tables and chi-square for the test of independence of variables require large samples. When this requirement is not met, SAS still computes the test statistic but also prints a warning message. Other categorical data analysis methods must be computed in cases when this occurs.

Below - SAS warning for insufficient sample size. ## will be specified for a particular data analysis.

WARNING: ##% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

For the correct analysis and description of a data set that does not require a WEIGHT statement, a large data set is necessary. In these larger data sets, each observation is separately entered. In the illustration above, there would be 202 observations in a data table. The data table would contain 30 observations (lines) with the S F combination, 15 with S DNF, and so on.

In general, the data do NOT have to be sorted for the FREQ procedure to count the number of occurrences of each value of the variable. If the BY statement is used on the FREQ procedure, then the data would, of course, need to be sorted first.