## Describing  Your Data Using PROC MEANS

PROC MEANS can be used to compute various univariate descriptive statistics for specified variables including the number of observations, mean, standard deviation, variance, minimum and maximum values, the standard error of the mean, uncorrected sum of squares, corrected sum of squares, the coefficient of variation, and confidence limits for the mean. You can also ask PROC MEANS  to perform a t-test on the hypothesis that the population mean is zero.

The general form of the PROC MEANS statement is

```
        PROC MEANS   options;
```

The simplest form

```
        PROC MEANS;
```

will automatically compute and print the mean, standard deviation, minimum and maximum values for each of the numeric variables in the most recently created data set. If the linesize of the output window is large enough, PROC MEANS  will also print the standard error of the mean, the variance, the sum, and the coefficient of determination. You can specify the data set which PROC MEANS  will process by using the DATA= option in the PROC MEANS statement:

```
    PROC MEANS DATA= datasetname;
```

The VAR statement may be used in conjunction with the PROC MEANS statement to specify the variables for which you want descriptive statistics computed.

```
    DATA example1;
    INPUT  name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22  ;
    SAS programming statements go here
    DATALINES;
    Data records go here
    ;
    RUN;
    PROC MEANS DATA= example1;
```

```
        VAR  age  height;
    RUN;
```

Descriptive statistics will be printed only for the variables specified in the VAR statement.

A BY statement can be used with PROC MEANS to obtain separate analyses on observations in groups defined by the variable(s) in the BY statement. When a BY statement appears as part of the MEANS procedure, the procedure expects the data to be sorted in the order of the BY variables.

```
    DATA example1;
    INPUT  name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22  ;
    SAS programming statements go here
    DATALINES;
    Data records go here
    ;
    RUN;
    PROC SORT DATA =  example1;
        BY sex;
    RUN;
    PROC MEANS DATA= example1;
        VAR   height  weight  age ;
        BY sex;
    RUN;
```

This example will produce a listing of descriptive statistics for females (sex='F') and a separate listing for males (sex='M').

You can ask PROC MEANS to produce only the descriptive statistics you desire. Corresponding to each statistic is a keyword that can be specified in the PROC MEANS statement:

| | |
|---|---|
| N | number of nonmissing observations (in a subgroup) |
| NMISS | number of observations with missing values (in a subgroup) |

| | |
|---|---|
| MEAN | sample mean |
| STD | sample standard deviation |
| MIN | minimum value |
| MAX | maximum value |
| RANGE | range |
| SUM | sum |
| VAR | variance |
| USS | uncorrected sum of squares |
| CSS | corrected sum of squares |
| CV | coefficient of variation |
| STDERR | standard error of the mean |
| CLM | confidence limits on the mean |
| LCLM | lower confidence limit on the mean (one-sided) |
| UCLM | upper confidence limit on the mean (one-sided) |
| T | Student's t statistic for testing $H_0$: $\mu = 0$ |
| PRT | probability of a greater absolute value for Student's t statistic |

The confidence intervals have a 95% confidence level (by default). If you desire a different confidence level, use the ALPHA= option on the PROC MEANS statement. For example, to produce 99% confidence limits,

```
PROC MEANS CLM ALPHA=0.01;
```

Now for a more complicated example, including a t-test, we might use the following program:

```
DATA example1;
INPUT  name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22  ;
SAS programming statements go here
DATALINES;
Data records go here
;
RUN;
PROC SORT DATA =  example1;
        BY sex;
RUN;
PROC MEANS DATA= example1  N MEAN STD STDERR CLM T
        PRT ALPHA=0.10;
        VAR   height  weight  age ;
        BY sex;
RUN;
```

With some clever program statements, you can "trick" PROC MEANS into performing tests involving hypotheses other than $H_0$: $\mu = 0$. Suppose that you wish to test $H_0$: $\mu = 60$ for the variable height. This can be accomplished by

```
DATA example1;
INPUT  name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22  ;
/* set up variable for hypothesis test H0: μ = 60  */
htdiff = height  -  60 ;
DATALINES;
Data records go here
;
RUN;

PROC MEANS DATA= example1  N MEAN STD STDERR T PRT;
        VAR   htdiff ;
RUN;
```

PROC MEANS can also produce an output data set containing the computed descriptive statistics by using an OUTPUT statement, which has the form

```
        OUTPUT  OUT= datasetname    statistics ;
```

For example,

```
        DATA example1;
        INPUT  name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22  ;
        SAS program statements go here
        DATALINES;
        Data records go here
        ;
        RUN;

        PROC MEANS DATA= example1  N MEAN STD STDERR;
                VAR   weight ;
                OUTPUT  OUT= calcstat  N MEAN STD STDERR;
        RUN;
```

The data set calcstat will contain the specified statistics for the variable weight.

## Frequencies and Crosstabulations Using PROC FREQ

Frequency tables and crosstabulation tables provide a way to summarize data for ordinal and categorical variables. Frequency tables show the distribution of a variable's values. Crosstabulation tables show joint distributions for two or more variables. The SAS procedure FREQ will produce one-way to n-way frequency and crosstabulation tables. PROC FREQ can also compute chi-square statistics, the phi coefficient, the contingency coefficient, and Cramer's V statistic. These statistics measure the degree of association of the values of the variable in a contingency table. Tests for independence can be performed in PROC FREQ. In addition, the procedure can compute kappa statistics, which can be used as a measure of agreement between two classification systems.

The general form of the PROC FREQ statement is

```
PROC FREQ   options;
```

Available options include the DATA= option, which will tell PROC FREQ which data set to process. You can request that PROC FREQ print only one table per page by using the PAGE option (otherwise multiple tables per page will be printed, as space permits).

The simplest form is

```
PROC FREQ;
```

and will produce frequency tables for all the variables in the most recently created data set. There is typically little purpose in obtaining frequency tables for continuous numeric variables, or for character variables whose values are unique, e.g., name. You can ask PROC FREQ to construct and print frequency and crosstab tables for selected variables in the data set by using the TABLES statement. Any number of TABLES statements may be included in one execution of PROC FREQ. If no TABLES statement is included in the procedure, PROC FREQ will construct one-way frequency tables for each of the variables in the data set. The TABLES statement has the form

```
TABLES   requests  /  options  ;
```

To request a one-way frequency table on the variable sex :

```
PROC FREQ;
TABLES     sex  ;
```

If you request a one-way table and specify no options, PROC FREQ produces frequencies, cumulative frequencies, percentages of the total frequency, and cumulative percentages for each level of the variable specified in the TABLES statement.

To request a two-way table on the variables sex and religion :

```
PROC FREQ;
TABLES     sex  * religion;
```

If you request a two-way table and specify no options, PROC FREQ produces a crosstabulation table that includes cell frequencies, cell percentages of the total frequency, cell percentages of the row frequencies, and cell percentages of the column frequencies.

Three - way and general n- way tables requests are specified in a similar fashion:

```
PROC FREQ;
TABLES     a * b * c;
```

will result in a three-way crosstab table.

For each frequency table or crosstabulation table you want, put a table request in the TABLES statement.

Missing values of each variable are typically excluded from the table that PROC FREQ produces, but the total frequency of missing values is printed below each table.

You can include options in the TABLES statement after the slash (/). If you use the option MISSING, PROC FREQ will treat missing values as nonmissing values and include them in the calculation of percentages and other statistics. The OUT= option may be included in a TABLES statement to produce a data set containing the levels of the variables and counts and percentages.

Options can be specified to request additional table information. The option EXPECTED will cause expected cell frequencies (under the hypothesis of independence) to be printed. The option CELLCHI2 tells PROC FREQ to print each cell's contribution to the total chi-square statistic.

Options to request additional statistical analyses can also be specified in the TABLES statement. The option CHISQ requests a chi-square test of homogeneity or independence, along with measures of association based on the chi-square. These include the Pearson chi-square, liklihood ratio chi-square, Mantel-Haenszel chi-square, the phi coefficient, the contingency table coefficient, and Cramer's V. The option EXACT requests Fisher's exact test. The option MEASURES will produce Pearson and Spearman correlation coefficients, Kendall's tau-b, Stuart's tau-c, and Somer's d statistics. The option AGREE can be used to produce kappa statistics.

An example of a program that use PROC FREQ follows:

```
DATA  socio1;
INPUT  name $ 1-10 sex $ 12 race $ 15-25  religion  $ 30-35  ;
SAS programming statements go here
DATALINES;
Data records go here
;
RUN;
PROC FREQ DATA =  socio1;
        TABLES     sex  ;
        TABLES    religion   race  / MISSING ;
        TABLES      sex* race   sex*religion  race*religion /
                EXPECTED  CELLCHI2  CHISQ  ;
        TABLES  sex*race*religion /  EXACT ;
RUN;
```