**Chapter VIII - Producing Descriptive Statistics**

**PROC MEANS** statement

Provides like mean, minimum, maximum and other data summarization tools, as well as helpful options from controlling your output (Default: mean, minimum, maximum & standard deviation (Quiz 1))

Basic code:

**PROC MEANS** <DATA=*SAS-data-set*> <*statistic-keyword(s)*> <*option(s)*>**;**

**RUN;**

*SAS-data-set* is the name of the data set to be used

*statistic-keyword(s)* specify the statistics to compute

*option(s)* control the content, analysis, and appearance of output.

Eg:

**PROC MEANS** data=perm.survey **MEDIAN RANGE**;

**RUN**;

Specify statistics by adding keywords as options as below:

|  |  |
| --- | --- |
| ***Descriptive Statistics*** | |
| **Keyword** | **Description** |
| CLM | Two-sided confidence limit for the mean |
| CSS | Corrected sum of squares |
| CV | Coefficient of variation |
| KURTOSIS | Kurtosis |
| LCLM | One-sided confidence limit below the mean |
| MAX | Maximum value |
| MEAN | Average |
| MODE | Value that occurs most frequently (new in SAS 9.2) |
| MIN | Minimum value |
| N | Number of observations with nonmissing values |
| NMISS | Number of observations with missing values |
| RANGE | Range |
| SKEWNESS | Skewness |
| STDDEV / STD | Standard deviation |
| STDERR | Standard error of the mean |
| SUM | Sum |
| SUMWGT | Sum of the Weight variable values |
| UCLM | One-sided confidence limit above the mean |
| USS | Uncorrected sum of squares |
| VAR | Variance |
| ***Quantile Statistics*** | |
| **Keyword** | **Description** |
| MEDIAN / P50 | Median or 50th percentile |
| P1 | 1st percentile |
| P5 | 5th percentile |
| P10 | 10th percentile |
| Q1 / P25 | Lower quartile or 25th percentile |
| Q3 / P75 | Upper quartile or 75th percentile |
| P90 | 90th percentile |
| P95 | 95th percentile |
| P99 | 99th percentile |
| QRANGE | Difference between upper and lower quartiles: Q3-Q1 |
| ***Hypothesis Testing*** | |
| **Keyword** | **Description** |
| PROBT | Probability of a greater absolute value for the t value |
| T | Student's t for testing the hypothesis that the population mean is 0 |

当format没有被specified时，PROC MEANS会自动choose BEST*w*. format to display numeric values。BEST*w.* format 会自动选择显示most information about the value according to the available field width

1. **Limiting Decimal Places**

* **MAXDEC**=option

To limit decimal place

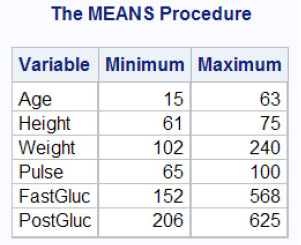
Basic code:

**PROC MEANS** <DATA=*SAS-data-set*> <*statistic-keyword(s)*> **MAXDEC=***n***;**

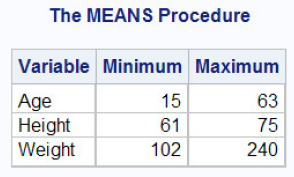
where *n* specifies the maximum number of decimal places.

Eg:

**PROC MEANS** data=clinic.diabetes min max **MAXDEC** =0;

**RUN**;

1. **Specifying Variables in PROC MEANS**

* **VAR** statement

To specify the variables that **PROC** **MEANS** analyzes

Basic code:

Chapter 4, page 1

1. **Group Processing**

* **CLASS** statement (Quiz 2)
* Used to categorize data, can be either character or numeric but contain a limited number of discrete values
* If a CLASS statement is used, then the N Obs statistic is calculated. The N Obs statistic is based on the CLASS variables,

Basic code:

**CLASS** *variable(s)***;**

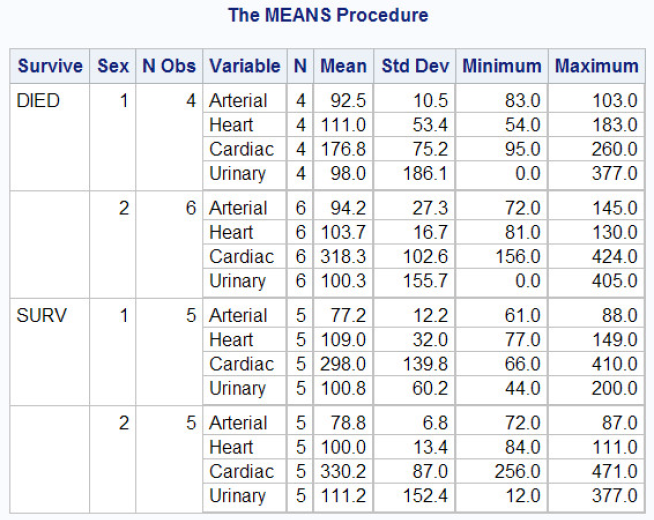
where *variable(s)* specifies category variables for group processing.

Eg:

**PROC** **MEANS** data=clinic.heart maxdec=1;

**VAR** arterial heart cardiac urinary;

**CLASS** survive sex; **Result of CLASS statement**

**RUN**;

* **BY** statement

Like the **CLASS** statement, the **BY** statement specifies variables to use for categorizing observations

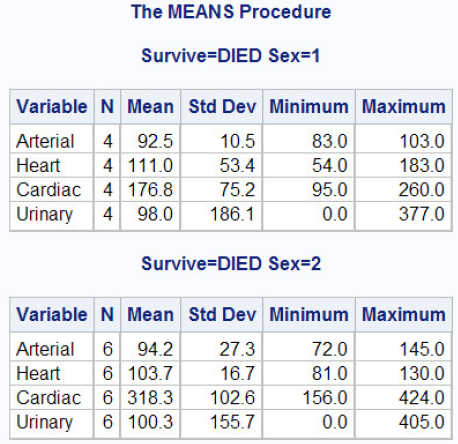
Basic code:

Chapter 4, all pages

Difference between **BY** statement and **CLASS** statement

* 当使用**BY** statement的时候，数据必须是已经sort完毕或者已经标上了index了。所以在**PROC MEANS** with any **BY** group之前需要一个**PROC SORT** step
* 两者的result layout是不同的

**BY** statement in the program below creates four small tables; a **CLASS** statement would produce a single large table **Result of BY statement**

Eg:

**PROC SORT** data=clinic.heart out=work.heartsort;

**BY** survive sex;

**RUN**;

**PROC** **MEANS** data=work.heartsort maxdec=1;

**VAR** arterial heart cardiac urinary;

**BY** survive sex;

**RUN**;

注意：因为CLASS statement不需要SORT step，CLASS statement比BY statement更加高效。但是但数据量比较大的时候BY statement比CLASS statement会相对更加高效

1. **Creating a Summarized Data**

* **PROC MEANS**

Basic code:

**OUTPUT OUT=***SAS-data-set statistic=variable(s)***;**

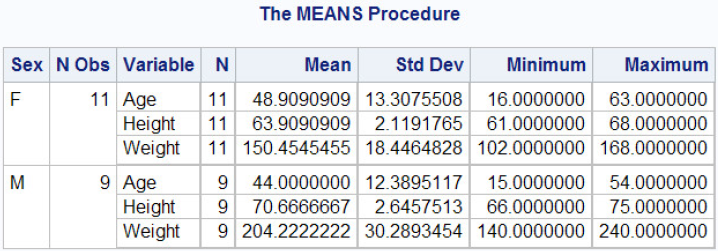
OUT= specifies the name of the output data set

*statistic*= specifies the summary statistic written out

*variable(s)* specifies the names of the variables to create. These variables represent the statistics for the analysis variables that are listed in the VAR statement.

N, MEAN, STD, MIN, and MAX are produced for all of the numeric variables or for *all* of the variables that are listed in a VAR statement with the same order

Eg: ***Report Created by PROC MEANS***

* Typical PROC MEANS report and also creates a summarized output data set

**PROC** **MEANS** data=clinic.diabetes;

**VAR** age height weight;

**CLASS** sex;

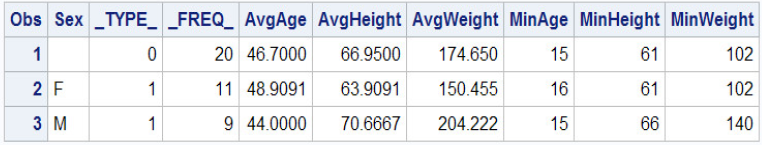
**OUTPUT OUT** =work.sum\_gender

mean=AvgAge AvgHeight AvgWeight

min=MinAge MinHeight MinWeight;

**RUN**;

* Contents of the output data set ***Data Set Created by PROC PRINT***

**PROC** **PRINT** data=work.sum\_gender;

**RUN**;

* **PROC SUMMARY**
* Also can use **PROC SUMMARY** to create a summarized output data set (same code as **PROC MEANS**), must include a **PRINT** option in the **PROC SUMMARY** statement. Without **PRINT** option in **PROC SUMMARY** step, program creates an output data set but does not create a report
* Different between **PROC SUMMARY** and **PROC MEANS** is **PROC MEANS** produces a report by default (use the **NOPRINT** option to suppress the default report)

Basic code:

**PROC SUMMARY** data=clinic.diabetes print;

**VAR** age height weight;

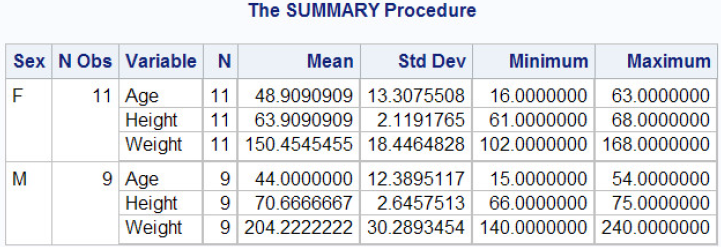
**CLASS** sex;

**OUTPUT OUT** =work.sum\_gender

mean=AvgAge AvgHeight AvgWeight;

**RUN**;

***Output Created by the SUMMARY Procedure with the PRINT Option Specified***



1. **Producing Frequency Tables**

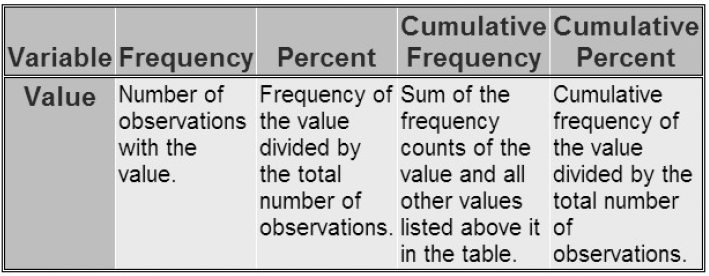
* **PROC FREQ** (Both numerical and character variables)
* **One-way table:**
* Create crosstabulation tables that summarize data for two or more categorical variables by showing the number of observations for each combination of variable values
* Display frequency, percent, cumulative frequency, and cumulative percent. This can produce excessive or inappropriate output. It is recommended that you always use a TABLES statement with PROC FREQ.

Basic code:

**PROC FREQ** <DATA=*SAS-data-set* >**;**

**RUN;**

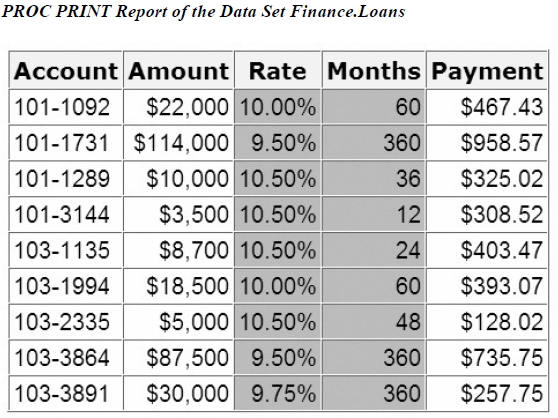
where *SAS-data-set* is the name of the data set to be used.



* Frequency distributions work best with variables whose values can be described as categorical, and whose values are best summarized by counts rather than by averages.
* **TABLES** statement

To specify the variables to be processed by the FREQ procedure

Basic code: **Data set list**

 **TABLES** *variable(s)***;**

where *variable(s)* lists the variables to include

(Chapter 6, Page 3)

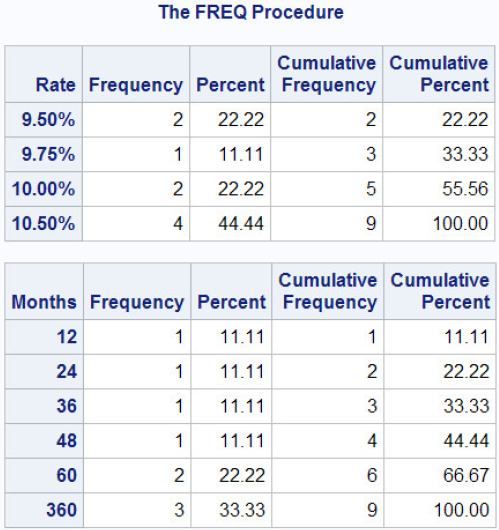
Eg:

**PROC FREQ** data=finance.loans;

**TABLES** rate months;

**RUN**;

**After the program**



* **NOCUM** option

Suppresses the display of cumulative frequencies and cumulative percentages in one-way frequency tables and in list output.

Basic code:

TABLES variable(s) / NOCUM;

* **Two-way table & N-way table:**
* The simplest crosstabulation is a two-way table. To create a two-way table, join two variables with an asterisk (\*) in the TABLES statement of a PROC FREQ step.
* Basic code:

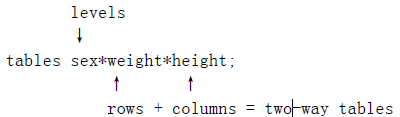
**TABLES** *variable-1 \*variable-2* <*\* ... variable-n*>**;**

where (for two-way tables)

*variable-1* specifies table rows

*variable-2* specifies table columns.

* When crosstabulations are specified, PROC FREQ produces tables with cells that contain
* cell frequency
* cell percentage of total frequency
* cell percentage of row frequency
* cell percentage of column frequency

Eg (Use N-way table as example):

**PROC FORMAT**;

**VALUE** wtfmt low-139='< 140'

140-180='140-180'

181-high='> 180';

**VALUE** htfmt low-64='< 5''5"'

65-70='5''5-10"'

71-high='> 5''10"';

**RUN**;

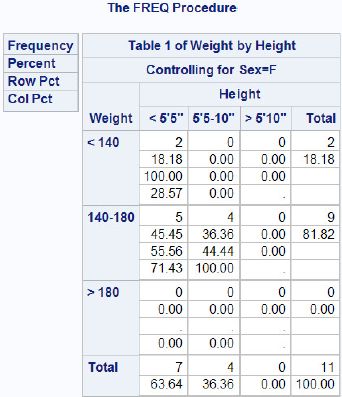
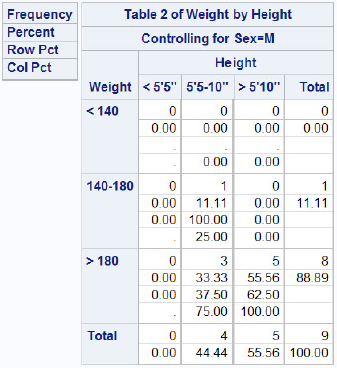
**PROC** **FREG** data=clinic.diabetes;

**TABLE** sex\*weight\*height;

**FORMAT** weight wtfmt. height htfmt.;

**RUN**;

**Result**

1. **Creating Tables in Different Format**

* **LIST** Format

To generate list output for crosstabulations, add a slash (/) and the LIST option to the TABLES statement in your PROC FREQ step.

Basic code:

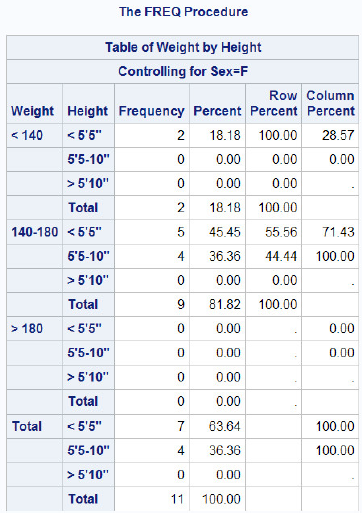
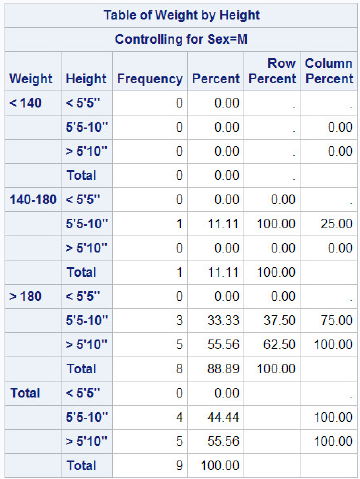
**TABLES** variable-1 \*variable-2 <\* ... variable-n> / **LIST**;

***Table Created by Using the LIST Option***



* **CROSSLIST** Format

***Table Created by Using CROSSLIST Option: Sex=F Table Created by Using CROSSLIST Option: Sex=M***



1. **Suppressing Table Information**

* When crosstabulations are specified, PROC FREQ produces tables with cells that contain (**Page 5**)
* Options to the TABLES statement

|  |  |
| --- | --- |
| **Options** | **Description** |
| NOFREQ | Suppresses cell frequencies |
| NOPERCENT | Suppresses cell percentages |
| NOROW | Suppresses row percentages |
| NOCOL | Suppresses column percentages |

Eg:

**PROC** **FORMAT**;

**VALUE** wtfmt low-139='< 140'

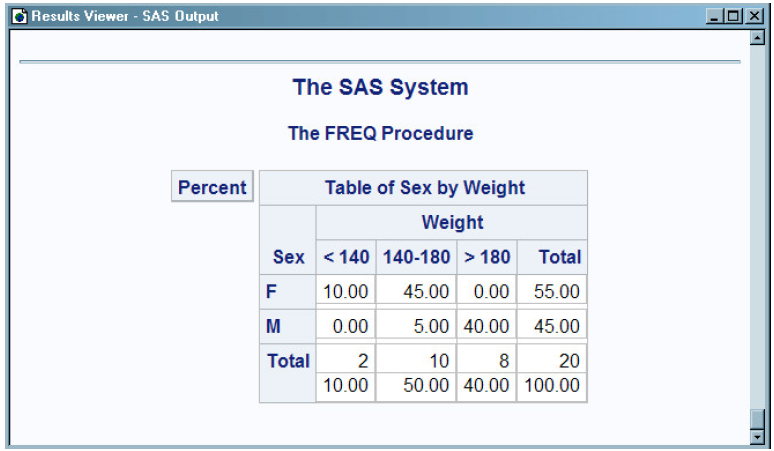
140-180='140-180'

181-high='> 180';

**RUN**;

**PROC** **FREQ** data=clinic.diabetes;

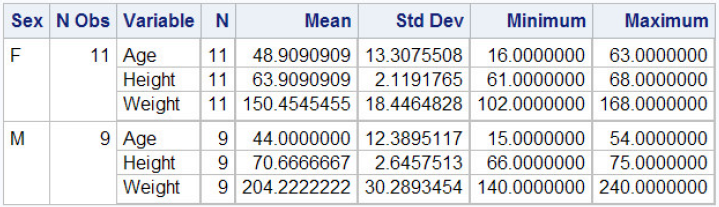
**TABLES** sex\*weight / nofreq norow nocol;

**FORMAT** weight wtfmt.;

**RUN**;

练习

1. The data set Survey.Health includes the following variables. Which is a poor candidate for PROC MEANS analysis?
2. IDnum
3. Age
4. Height
5. Weight
6. Which of the following statements is true regarding **BY** group processing?
7. **BY** variables must be either indexed or sorted.
8. Summary statistics are computed for **BY** variables.
9. **BY** group processing is preferred when you are categorizing data that contains few variables.
10. **BY** group processing overwrites your data set with the newly grouped observations.
11. Which program can be used to create the following output?



1. **PROC** **MEANS** data=clinic.diabetes;

**VAR** age height weight;

**CLASS** sex;

**OUTPUT** **OUT**=work.sum\_gender

mean=AvgAge AvgHeight AvgWeight;

**RUN**;

1. **PROC** **SUMMARY** data=clinic.diabetes **PRINT**;

**VAR** age height weight; class sex;

**OUTPUT** **OUT** =work.sum\_gender

mean=AvgAge AvgHeight AvgWeight;

**RUN**;

1. **PROC** **MEANS** data=clinic.diabetes **NONPRINT**;

**VAR** age height weight;

**CLASS** sex;

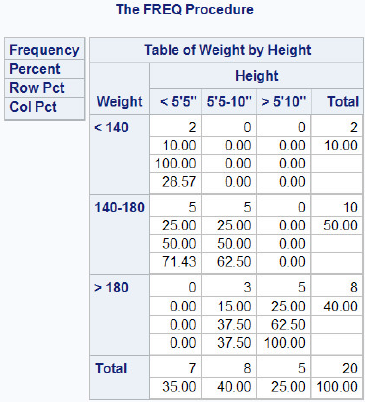
**OUTPUT** **OUT** =work.sum\_gender

mean=AvgAge AvgHeight AvgWeight;

**RUN**;

1. Both a and b.
2. Frequency distributions work best with variables that contain
3. continuous values.
4. numeric values.
5. categorical values.
6. unique values
7. Which PROC FREQ step produced this two-way table?
8. PROC FREQ data=clinic.diabetes;

TABLES height weight;

FORMAT height htfmt. weight wtfmt.;

RUN;.

1. PROC FREQ data=clinic.diabetes;

TABLES weight height;

FORMAT weight wtfmt. height htfmt.;

RUN;

1. PROC FREQ data=clinic.diabetes;

TABLES height\*weight;

FORMAT height htfmt. weight wtfmt.;

RUN;

1. PROC FREQ data=clinic.diabetes;

TABLES weight\*height;

FORMAT weight wtfmt. height htfmt.;

RUN;