INTRODUCTION TO PROCS

YI-TING HWANG
DEPARTMENT OF STATISTICS
NATIONAL TAIPEI UNIVERSITY
PROCEDURES

PROCs perform specialized tasks and thus have unique options and statements. One of SAS’s strengths is its consistency of syntax among disparate tasks and the procedures’ ability to use many similar statements.
EXAMPLE 1

data birthday;
input @4 lastname $10. @14 firstnme $10. @25 birthday date7.;
format birthday mmddyy10.;
cards;
  JONES     FRANK     22MAY53
  MCVADE    CURTIS    25DEC54
  SMITH     VIRGINIA  14NOV49
  TURNER    BECKY     26APR50
run;
PROC FORMAT

The FORMAT procedure enables you to define your own informats and formats for variables. For example

```sas
proc format;
value trtf
0='Placebo'
1='Aspirin';
value mif
1='Fatal attack'
2='Nonfatal attack'
3='No attack';
run;
```
EXAMPLE 2

A data were collected to investigate the relationship between aspirin use and heart attacks by Physician’s Health Study Research Group at Harvard Medical School.

<table>
<thead>
<tr>
<th></th>
<th>Myocardial Infarction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fatal Attack</td>
<td>Nonfatal Attack</td>
<td>No Attack</td>
</tr>
<tr>
<td>Placebo</td>
<td>18</td>
<td>171</td>
<td>10845</td>
</tr>
<tr>
<td>Aspirin</td>
<td>5</td>
<td>99</td>
<td>10933</td>
</tr>
</tbody>
</table>
SAS PROGRAM

proc format;
value trtf
  0='Placebo'
  1='Aspirin';
value mif
  1='Fatal attack'
  2='Nonfatal attack'
  3='No attack';
run;
data death;
  input trt mi count;
  format trt trtf. mi mif.;
cards;
  0 1 18
  0 2 171
  0 3 10845
  1 1 5
  1 2 99
  1 3 10933
run;
PROC SORT

Many situations call for rearranging the order of the observations in a dataset. For example, a list of name must be in alphabetical order, scores must be listed in descending order. Combining two or more SAS datasets may require sorting by one or more variables common to the datasets. The syntax is

PROC SORT [DATA=dataset] [OUT=dataout] [NODUP];
BY [descending] byvarlist;

9/28/2006 ch2 SAS PROC 7
PROC PRINT

One of most useful applications for a dataset is a simple listing. The PRINT procedure can list all or just a few variables. The syntax is

PROC PRINT [DATA=dataset] [NOOBS];
[BY byvarlist;]
[VAR varlist;]
EXAMPLE 3

Sort the students’ grades by the math score in ascending order.
# DATA TYPES

<table>
<thead>
<tr>
<th>Type</th>
<th>Sub-description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical data</td>
<td>Nominal data</td>
<td>Gender (Female/Male)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Residence (Rural/City)</td>
</tr>
<tr>
<td></td>
<td>Ordinal data</td>
<td>Preference (Very high, high, moderate, low, very low)</td>
</tr>
<tr>
<td>Continuous data</td>
<td>Interval / Ratio data</td>
<td>Blood pressure, Age, Income, Height, Weight, Tax</td>
</tr>
</tbody>
</table>
DESCRIPTIVE STATISTICS

• Discrete variables
  – Relative frequency
  – E.g. sex (F/M), location (South/North), Treatment(Placebo/new drug)

• Continuous variables
  – Mean, Variance, Standard deviation (STD), Median, Mode, Range
  – E.g. Age, height, weight, …
PROC FREQ

PROC FREQ produces frequency (tables). The syntax is

PROC FREQ [DATA=dataset] ;
[BY byvarlist;]
TABLES varlist;
EXAMPLE 4

43 bone marrow transplant patients at the Ohio State University Bone Marrow Transplant Unit were recruited.

- All patients
  - had either Hodgkin's disease (HOD=1) or non-Hodgkin's lymphoma (NHL=0)
  - were given either an allogeneic (同種異體的, allo=0) transplant from an HLA-matched sibling donor or an autogeneic (auto=1) transplant where their own marrow was cleansed (清洗) and returned to them after a high dose of chemotherapy.

- Two other factors are also available, patients' Karnosky score and waiting time in months.
UNIVARIATE STATISTICS

Both the MEANS and UNIVARIATE procedures can be used to compute means, standard deviations, minimum and maximum values and other measures.

Differences. Univariate’s output usually uses one page per variable. Most of its statistics are computed automatically. On the other hand, output from the MEAN procedure is much more compact.
PROC MEANS

The syntax is

PROC MEANS [DATA=dataset] [STATISTICS];
[CLASS groupvar;]
[BY byvar;]
VAR analysisvar;

Note that both CLASS and BY do the same execution. However, a CLASS statement is usually run faster than a BY statement.
PROC UNIVARIATE

The syntax is

PROC UNIVARIATE [DATA=dataset] [NORMAL] [PLOT];
[BY byvarlist;]
VAR varlist;
## COMPARISONS

<table>
<thead>
<tr>
<th>Statistics</th>
<th>PROC MEANS</th>
<th></th>
<th>PROC UNIVARIATE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Available</td>
<td>Default</td>
<td>Available</td>
<td>Default</td>
</tr>
<tr>
<td># of nonmissing obs</td>
<td>N</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td># of missing obs</td>
<td>NMISS</td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Total # of obs</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Mean</td>
<td>MEAN</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Mode</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>STD</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Variance</td>
<td>VAR</td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Minimum/Maximum</td>
<td>MIN/MAX</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Range</td>
<td>RANGE</td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Percentiles</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Box plot/ Stem &amp; Leaf plot/ Normal plot</td>
<td></td>
<td></td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>Test normality</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>
EXAMPLE 5

Consider a class with 50 students. Compute the relative frequencies for discrete variables and mean for continuous variables. In addition, examine whether there is any difference between female and male.