Laboratory 2: Data management and Proc Steps in SAS

This session we will practice Modifying SAS Data and Proc Steps. You will work on the dataset ‘case-control978.dat’ and use the SAS program ‘case-control978.sas’.

Modifying SAS Data

Next you will use a variety of SAS statements to modify a SAS dataset.

1. Subsetting Data Sets

It is convenient to create a new SAS dataset containing only the relevant observations. The if statement and where statement have the same form and can be used to achieve the same effect. where statement can also be used with proc steps.

```sas
data cases;
set case_control978;
if status=1;
run;
```

```sas
data cases;
set case_control978;
where status=1;
run;
```

2. Deleting Observations

```sas
if age>100 then delete;
if status>1 then delete;
```

3. Deleting Variables

Variables can be removed from the data set being created by using the drop or keep statement. So drop x y z; results in a data set that does not contain the variables x, y, and z, whereas keep x y z; results in a data set that contains only those three variables.
4. Creating and Modifying Variables

agec=age – 52;
alcohol2=alcohol/1000;

tobacco2=tobacco;
if (tobacco>7) then tobacco2=.;

Please use the case-control978.sas program you ran last time to create 3 dichotomous (binary) variables from the age, alcohol, and tobacco variables by using ‘if … then…’ statements. Name the new variables as ‘agegrp’, ‘alcgrp’, and ‘tobgrp’ respectively and define the codes for various categories as follows:

agegrp 0 = <60 years
       1 = 60+ years

alcgrp 0 = 0-39 gm/day
        1 = 40+ gm/day

tobgrp 0 = 0-9 gm/day
        1 = 10+ gm/day
        . = missing

data case_control978;
set case_control978;
agegrp=0;
if (age>=60) then agegrp=1;
alcgrp=0;
if (alcohol>39) then alcgrp=1;
tobgrp=0;
if (tobacco>7) then tobgrp=.;
if (tobacco>=3 and tobacco<8) then tobgrp=1;

Please save this updated program with codes for creating dichotomous variables in your floppy for future use.

Proc Steps

Once data have been read into a SAS dataset, SAS procedures can be used to analyze the data. Each SAS procedure performs a specific type of analysis. The step begins with a proc statement and ends with a run statement. The most important option is the data= option, which names the data set to be analyzed. If the option is omitted, the procedure uses the most recently created data set.
var Statement

The var statement specifies the variables that are to be processed by the proc step. For example:

```proc print data=case_control978;
var age status;
run;
```

This restricts the printout to the two variables mentioned, whereas the default would be to print all variables.

where Statement

The where statement selects the observations to be processed.

```proc print data=case_control978;
where tobacco>7;
run;
```

This identifies subjects with missing values in tobacco variable (did not answer or unknown).

by Statement

The by statement is used to process the data in groups. The observations are grouped according to the values of the variable named in the by statement and a separate analysis is conducted for each group. To do this, the data set must first be sorted in the by variable.

```proc sort data=case_control978;
by status;
run;
proc freq;
tables age tobacco alcohol sex;
by status;
run;
```

This will create tables with age, tobacco, alcohol, and sex reported separately by the case-control status.
**class Statement**

The *class* statement is used with many procedures to name variables that are to be used as classification variables, or factors.

```
proc ttest data = case_control978;
class status;
var alcohol;
run;
```

**Proc Corr -- correlation**

```
proc corr data = case_control978 pearson spearman;
var age alcohol;
run;
```

**Proc Reg -- regression**

```
proc reg data = case_control978;
model alcohol = age status;
run;
```

**Proc Means -- simple univariate descriptive statistics for numeric variables**

```
proc means data = case_control978;
var age alcohol;
by status;
run;
```

**Proc Univariate -- examine the distributions of numeric variables**

```
proc univariate data = case_control978 normal;
var alcohol age;
histogram alcohol age / normal;
run;
```

The *normal* option results in a test for the normality of the variables. The *histogram* statement produces histograms for both variables and the */normal* option requests a normal distribution curve.
**Proc Ttest** – perform independent t-test

```sas
proc ttest data = case_control978;
class status;
var age alcohol;
run;
```

**Proc Freq** – produce contingency tables and analyze them

```sas
proc freq data = case_control978;
tables age tobacco alcohol status sex;
run;

proc freq data = case_control978;
tables status*agegrp / chisq;
run;
```
Homework 2 (using case-control978.dat)

1. Produce frequency distributions for all the variables of the dataset and check for any inconsistent (or improbable) and out of range values (or outliers). Examine frequency distributions for every variable for the cases and controls separately. If there are any inconsistencies (or improbable values) print the distribution of the variable(s) and mark the inconsistent/improbable value(s) on the printout.

2. If any of the inconsistencies are correctable based on the information on the study you were given last week, correct them in the following manner:

When you have edits that need to be made to the original data set because of data entry errors or other problems you have two options:
1) Overwrite the original raw data with the corrections
2) Make all data edits within your SAS program so you have a record of your changes.

We recommend the second option – unless you are the data manager and can make global changes to the raw data file and get these changes to all users.

The second option will provide a permanent record of all changes you make to the raw file and will make it easy for anyone to take your program and replicate your results.

For organization, I suggest that you put all the changes close to the top of your SAS program so that every time you run the program the changes are made to the original raw data file.

Check whether the inconsistencies are corrected by producing frequency distribution of the corrected variable(s).

3. Are there any inconsistencies in the data that are not correctable without examining the original questionnaire or any outlier(s) (or improbable values) that you would like to verify with the original questionnaire? If yes, indicate which ones.

4. Are there any missing values in the data? If yes, on how many subjects and for which variable(s)?
5. Examine the effects of continuous variables on esophageal cancer. Perform independent t-test procedure, and fill in Table 1. Do cases differ from controls with respect to mean age and alcohol consumption? Write a brief interpretation of your results.

Table 1. Mean age and alcohol consumption on esophageal cancer among 978 men

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>t value</th>
<th>DF</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. alcohol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Examine the effects of categorical variables (‘agegrp,’ ‘alcgrp,’ and ‘tobgrp’) on esophageal cancer. Generate 2 x 2 tables, perform $\chi^2$ test, and fill in Table 2. Do cases differ from controls with respect to age, alcohol and tobacco consumptions? Write a brief interpretation of your results. Discuss some drawbacks, particularly in light of what you have read in Rothman and Greenland, of classifying the variables as you have using agegrp, tobgrp, and alcgrp.

Table 2. Effects of age, alcohol and tobacco consumptions on esophageal cancer among 978 men

<table>
<thead>
<tr>
<th></th>
<th>Percentages in Controls</th>
<th>Percentages in Cases</th>
<th>$\chi^2$</th>
<th>DF</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. agegrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60+ years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. alcgrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40 gm/day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;39 gm/day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. tobgrp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10 gm/day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;9 gm/day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>