

Graphical Displays and Examining Relationships

Proc Univariate, Plot, Corr, and Freq

1 Graphical Displays

1.1 Proc Univariate

- In proc univariate, the plots option will produce low resolution plots (histogram or stem and leaf, normal probability plot, boxplot). The histogram (histogram), probplot (normal probability plot) and qqplot (quantile-quantile plot) statements in proc univariate will produce high resolution graphs.
- Proc boxplot will produce a high resolution boxplot.
- The 'Class' option can be used for plotting by a grouping variable.
- Many types of high quality plots can be obtained interactively through the analyst. Some of these are option under summary statistics and distributions, other are under graph.

Example: Basic graphical displays

```
libname temp "P:\";

title 'univariate with plots';
proc univariate data=temp.hanes plots;
var sysbp;
run; /* creates low resolution plots.*/

proc univariate data=temp.hanes;
var diabp;
histogram diabp;
qqplot diabp;
run; /* creates high resolution plots.
There are 23 pages of documentation in the on-line
    help describing options for this */

proc univariate data=temp.hanes;
class sex;
var diabp sysbp;
histogram diabp sysbp;
run;
```

1.2 Scatterplot for two variables.

A scatterplot simply plots one variable versus another. A low resolution plot can be obtained via proc plot. The vpercent and hpercent will control the percent of the page used.

```

proc plot data=tmp1.nhanes hpercent=50 vpercent=50;
plot har2*sysbp;
run;

data new;
set tmp1.nhanes;
run;

proc sort;
by sex;
run;

proc plot hpercent=50 vpercent=50;
plot diabp*sysbp;
by sex;
run;
quit;

```

Notice that a new data set has to be created in order to sort the variables of tmp1.nhanes. This has to be done since nhanes is a permanent data set.

2 Examining relationships among variables.

2.1 Correlation

For examining the relationship between two quantitative variables, we can graphically use scatterplots while numerical correlations (either Pearson or Spearman rank correlation) are popular. Two-dimensional scatterplots can be obtained via proc plot or via SAS/GRAPH. SAS/GRAPH will also give three-dimensional plots. Two and three dimensional plots as well as other graphics are available under graph in the analyst.

Proc corr provides correlation and other descriptive statistics involving multiple variables. These are useful for quantitative variables.

Correlation.

$\rho = corr(X, Y)$ where X and Y are random variables (quantitative). This is the population correlation and measures the strength of the linear relationship between X and Y . If X and Y are independent then $\rho = 0$, but the converse is not always true.

The sample (Pearson) correlation is r . Spearman correlation is obtained by ranking each of the X and Y variables and computing Pearson correlation on the ranks. This is a nonparametric measure of the association between X and Y that is not effected by outliers.

And remember to **Always plot the data.**

```

proc plot data=tmp1.nhanes hpercent=100 vpercent=100;

```

```

plot weight*height;
plot weight*sysbp;
run;

proc corr pearson spearman;
var weight height sysbp diabp;
run;

```

2.2 Two-way contingency tables.

Suppose two random variables X and Y are each discrete (they can be categorical), say with X having possible values x_1, x_2 , and x_3 (these could be labels for a categorical variable) and Y with possible values y_1, y_2 , and y_3 . A contingency table can be created to view the frequency counts of the row variables across column variables.

Example: Using the hanes data set we consider the relationship of race with har1. Race is a categorical variable with 1=White, 2=Black, and 3=Other. Har1 is an indicator variable of, whether the respondent has smoked more than 100 cigarettes in their life time (1=yes, 2=no).

```

proc freq data=tmp1.nhanes;
tables race*har1/nocol norow nopercnt;
run;

```

- The 'nocol' option will tell SAS not to print the column percentages for each subgroup.
- The 'norow' option will tell SAS not to print the row percentages for each subgroup.
- The 'nopercnt' option will tell SAS not to print the overall percentage of observations for each subgroup.