

SAS/INSIGHT.....A point and click approach to Exploratory Data Analysis
Ottawa Area SAS Users Society (OASUS) Meeting. November 14, 2007

Synopsis

Point and click approach to data analysis is good for the programmers as well as the data analysts. Programmers don't have to code existing programs and several iterations of analytical probing of the data. Analysts don't have to spend time specifying their requirements and explore the data themselves before finalising their analytical plans.

Type INSIGHT in the window and click check next to window. There are other ways but ...

clk WORK in the libraries.....a spread sheet opens

clk help and then create samples---Help manual has a description of all sample data sets

clk file-open-SASUSER-Baseball

Open Baseball dataset by double clicking in the explorer window and notice the differences in the two data tables. View table, in main SAS is static whereas Insight table is dynamic.

SAS offers two choices to analyse a data set; codes or point and click.

Let's try univariate **analysis** of salary using a p & c approach.

TL menu-move to first-salary & cr-home

Analyse-Distribution Y-salary.....A complete univariate analysis appears with two graphs Familiar with ratio statistics look no further but look carefully. N=263 whereas dataset has 322 observations. The difference is the missing values requiring possible imputation! The two graphs, histogram and boxplot provide additional information about the variable.

Box plot identifies the outliers. Draw a box around these values and d.clk in the box.

Selected values appear for examination. These values are highlighted in the datatable. TL menu and clk extract. The selected values appear in a new table.

In the boxplot, BL menu select means and a red mean diamond appears in the box plot.

Mean is greater than median, suggesting outliers or skewness to the right.

Histogram is plotted with default intervals. Many histograms can be viewed of this variable by using hand tools.

Edit-windows-tools-hand tool and move it over the histogram; the intervals and density change revealing the underlying structure of the data. BL choose the tick option to select that interval.

Clk in a quartile of the boxplot and notice a portion of the histogram highlighted. This interactivity **between histogram and boxplot** provides further information about the subsets of data.

Let's do a **Bivariate Analysis** of salary and cr-home using interactivity.

Analyse-distribution Y-cr-home will produce an analysis window. Resizing the windows so that they appear side by side notice that N=322 for cr-home. Create a box around outliers in the box plot and notice that these do not necessarily correspond to salary and vice versa.

Maximize the salary analysis window. Draw a box next to the boxplot.

Draw a scatter plot. Analyse-scatter plot-salary/cr-home. Examine the linear relationship.

Draw a histogram of positions next to the histogram of salary. Notice the absence of Pitcher in the data.

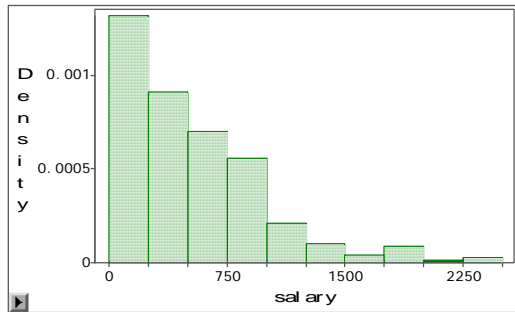
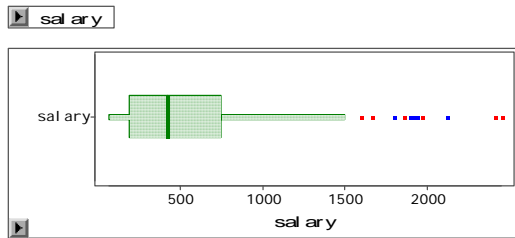
Click on a position and you can see its distribution on the other three graphs.

Edit- windows-tools click on red box and select American and blue for National. The scatter plot now has those colors adding a fourth variable to analysis.



Follow-up

- **The following takes you to the STC training catalogue:**
 - STC-intranet/Human Resources/Training/Course Catalogue/13.2 subject-matter/
 - Code 0430E: Introduction to Exploratory Data Analysis
 - Code 0430F: Introduction à l'Analyse Exploratoire des Données
 - Code 0432E: Intermediate Exploratory Data Analysis
 - Code 0432F: Analyse exploratoire intermédiaire
- **To register contact :**
 - Joelle L'Heureux, Human Resource Development, 4-M RH Coats Bldg, 951-8113, email: joelle.lheureux@statcan.ca
- **For further information contact:**
 - John McVey, mcvey@statcan.ca, 951-3610
 - Anis Ashraf, anis.ashraf@statcan.ca, 951-0020
 - Jocelyn Tourigny, jocelyn.tourigny@statcan.ca, 951-6667



Moments			
N	263.0000	Sum Vgts	263.0000
Mean	535.9259	Sum	140948.507
Std Dev	451.1187	Variance	203508.064
Skewness	1.5890	Kurtosis	3.0590
USS	128857066	CSS	53319112.8
CV	84.1756	Std Mean	27.8172

