

09/24/99

# **SAS/INSIGHT WORKSHOPS MANUAL**

**Manitoba Centre for Health Policy and Evaluation**

Sandra Peterson

<b>I. INTRODUCTION .....</b>	<b>2</b>
<b>II. THE SAS/INSIGHT ENVIRONMENT .....</b>	<b>2</b>
A. Getting Started .....	2
B. Description of the SAS/Insight Data Window .....	3
<b>III. EXAMINING DATA .....</b>	<b>4</b>
A. The Scroll Bar .....	4
B. Arranging Variables .....	4
C. Sorting Observations .....	4
D. Finding Observations .....	5
<b>IV. DEFINING VARIABLES .....</b>	<b>6</b>
A. Measurement Level .....	6
B. Default Roles .....	6
C. Transforming Variables .....	7
D. Formats .....	7
E. Saving Modified Data .....	7
<b>V. EXPLORING DATA .....</b>	<b>8</b>
A. Bar Charts .....	8
1. Nominal Data .....	8
2. Interval Data .....	9
B. Box Plots .....	10
<b>VI. EXAMINING DISTRIBUTIONS .....</b>	<b>11</b>
<b>VII. FITTING CURVES .....</b>	<b>12</b>
<b>VIII. OTHER CAPABILITIES OF SAS/INSIGHT .....</b>	<b>13</b>

# I. INTRODUCTION

SAS/Insight software is an interactive tool for data exploration and analysis. It provides a graphical user interface so that the user can point and click on the desired variables and procedures to achieve a number of simple analyses. SAS/Insight is for people with minimal analysis requirements who may only need to look at their data, or for people who have programmers doing the more advanced programming work for them.

SAS/Insight is a component of SAS/Assist, a menu-driven interface to the SAS System. SAS/Insight allows users to explore data through a variety of graphic displays, whereas SAS/Assist is a front-end to more general activities, such as managing and manipulating data, analysing data and writing reports, as well as creating graphic displays.

In this manual we will cover the basics of how to use SAS/Insight to:

- examine data
- define variables
- create histograms (bar charts) and box plots
- identify observations within these charts and plots
- include or exclude certain observations from the analysis
- examine distributions
- perform a regression analysis

A series of examples using the simulated Manitoba Health data set is provided in the manual. Data must be stored in SAS data set form (either permanent or temporary) to use SAS/Insight. Creating a SAS data set from raw data has been covered in the SAS PC Workshops Manual.

## II. THE SAS/INSIGHT ENVIRONMENT

### A. Getting Started

Run the program to create the simulated Manitoba Health data set (called “test”). See the SAS PC Workshops Manual for instructions on how to do this.

To invoke SAS/Insight, type ‘insight’ on the command line, or, using your mouse, click on **Solutions > Analyze > Interactive data analysis**. (Unless otherwise stated, ‘click’ means to click once with the left mouse button). A window will pop up with a list of the available libraries (locations where SAS data sets can be stored), and available data sets. The library called WORK always refers to the temporary data sets that are created in SAS.

Click on **WORK**; in the right-hand column, the data set name **TEST** will appear (plus any other SAS data sets you may have created in this SAS session).



Click on **TEST** (so it is highlighted) and then on **Open**. This will open SAS/Insight with the simulated health data set.

## B. Description of the SAS/Insight Data Window

The SAS/Insight data window is a table with the variables displayed in columns across the top and the observations displayed in rows. To enlarge the window to fill most of the screen, click on the square in the upper right-hand corner. On other systems you may have to double click (that is, click twice with the left mouse button very fast) on the top bar of the window.

A pull-down menu is displayed at the top of the window, and consists of **File**, **Edit**, **Analyze**, **Options**, **Window** and **Help**. Additional menus are Tables, Graphs, Curves and Vars. These last menus are drawn in light grey, which means that they are disabled until they are appropriate to use. These menus will be available for each window opened using SAS/Insight. The File menu contains options to:

- open a **New** data window
- **Open** a data window with a different data set
- **Save** the contents of the window (e.g. the data set)
- **Print** (if the printing is set up)
- **End** the window
- **Exit** the SAS session.

The other options will be examined throughout the manual.

The number of observations and the number of variables in the data set appear in the upper left corner of the data window. The “test” data set has 500 observations and 33 variables.

	33	Nom	Nom	Int
500		NCASE	GENDER	AGE
■	1	00001	2	38
■	2	00002	2	18
■	3	00003	1	0

The observations are numbered down the left-hand side of the window; the variable names are displayed across the top. The variable type is given above the variables' names; the variable type determines how a variable is treated in graphs and analyses. Two types of variables are possible: interval (Int) or nominal (Nom). Interval variables vary across a continuous range, for example, age. Nominal variables contain a discrete set of values or categories, for example, gender. More information will be given about interval and nominal classification in section IV (Defining Variables).


### III. EXAMINING DATA

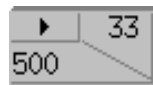
#### A. The Scroll Bar

Many data sets are too large to fit in the data window, so the window contains scroll bars to scroll through the data. Clicking on the arrow button on the bottom of the vertical scroll bar will move you down one observation at a time. Clicking within the scroll bar will scroll the height of the window. Dragging the slider on the vertical scroll bar all the way up or down is a quick way to get to the last or first observation. The horizontal scroll bar works similarly, allowing you to see all of the variables.


#### B. Arranging Variables

Using the scroll bar allows us to view all the data, but the variables or observations may not be arranged in the order we want. For example, suppose we want age to be in the first column. To move age to the first column, do the following:

- Click on the **AGE** variable name (it should become highlighted).
- Click on the little arrow below File (just above the number of observations and to the left of the number of variables). This arrow button will be denoted by  in the following text.




- Select **Move to First**.

This will move the age column to the first column. Note that multiple variables can be selected at the same time by pressing the control button (Ctrl) on the keyboard and, while holding it down, clicking with the left mouse button on the variables you want to select to move. The  menu includes other options that will be looked at later.

#### C. Sorting Observations

It is often useful to examine data by the values of a variable. To sort the health data set by the patients' age, do the following:

- Click on the **AGE** variable.
- Click , then choose **Sort**.

The whole data set is then sorted by age, with the records for the younger patients being first. Note that multiple variables can be selected and sorted by following the method outlined above in B (Arranging Variables). The data set will be sorted by the first variable selected and at each level of the first variable, the data will be sorted by the second variable selected, etc.

## D. Finding Observations

It is possible to find all the observations that share a characteristic. For example, to find all the patients who spend 1 or more days in ICU:

- Click on **Edit > Observations > Find**.
- In the box on the left of the pop-up-menu, click on **NDAYICU**, which is the variable that contains the number of days spend in ICU for the “test” data set.
- In the middle box of the pop-up-menu, click on “>=”.
- All values of NDAYICU are shown in the third box. Click on **1**, then press **Apply**. The FIND window will remain open (although it may be hidden behind the data window).



All of the observation numbers that correspond to records with  $NDAYICU \geq 1$  are highlighted. To look at the records with 1 or more days in ICU, we can either scroll through using the vertical scroll bar, or we can select **Find Next** (by clicking on the arrow in the WORK.TEST data set window). The ‘next’ observation that has 1 or more days in the ICU will be moved to the top of the screen (without rearranging the order of the records). If the sort order of the records does not matter, then choose **Move to First**. This moves all highlighted records to the top of the window where they can all be examined together. Click **OK** in the FIND window to close it. Note that the FIND window should not be closed until you are finished examining the records with 1 or more days in ICU, because if you accidentally click anywhere in the data window, the highlights on these selected observations would be lost. Having the FIND window still open makes it easier to repeat the process of finding the observations.

## IV. DEFINING VARIABLES

### A. Measurement Level

As previously mentioned, each variable has a “measurement level”, which can be either interval (Int) or nominal (Nom). By default, all character variables are assigned as “nominal” and numeric variables are assigned as “interval”. The measurement level can be either interval or nominal for numeric variables, depending on the intended use of the variable. For example, to change the measurement level from interval to nominal for the variable DISCRE, which is a 1/0 variable:

- Use the horizontal scroll bar until the variable DISCRE is in the data window.
- Click on the variable **DISCRE**.
- Click on the **Int** measurement level indicator above the variable DISCR.
- Click on **Nominal**.

One of the optional questions that is provided with this manual examines the effects of defining DISCRE as interval or nominal.

### B. Default Roles

It is possible to define default roles for variables. Default roles are applied to every analysis done during the SAS session.

Four choices are available: Group, **Label**, **Freq** or **Weight**:

- i) **Group** - Separates analyses by each value of the group variable.
- ii) **Label** - Labels identify observations in plots. The default is the observation number.
- iii) **Freq** - This option is used if a variable in your data represents the frequency of occurrence for other values in each observation.
- iv) **Weight** - Supplies weights for each observation.

To see these choices, click on **Age** (or any other variable), then click on the empty box beside **Int** (or **Nom**) above the selected variable.

Weight, Freq and Label roles can only be assigned to one variable at a time. The Group role can be assigned to more than one variable (the order that you assign the group role determines the order in which the variables are used to define the groups). Only interval variables can be assigned the Freq or Weight roles.

Each individual analysis includes the option of assigning Group, Label, Freq or Weight variables, where appropriate.

## C. Transforming Variables

New variables can be created from existing variables. The most common transformations are available in the **Edit > Variables** menu. For example, to create a variable that is the log transformation of LOS:

- Click on **LOS**.
- Choose **Edit > Variables > log (Y)**. A new variable, L\_LOS, will appear as the last column in the data window. Selecting **Edit > Variables > Other** allows more complex transformations to be done.

## D. Formats

If labelling formats were attached to values in your SAS data set, SAS/Insight displays these by default. If, however, the data set contains numeric variables with no formats, or if new numeric variables are created, SAS/Insight chooses an appropriate format based on that variable's values. For example, for L\_LOS that we created above, SAS/Insight has chosen to display the values with 4 decimal places. To change this so that only 2 decimal places are displayed:

- Click on L\_LOS.
- Select **Edit > Formats**.
- Click on **10.2**.

Labelling formats must be attached before the SAS/Insight window is opened (that is, they must be attached within the program that you submitted in the program window before starting SAS/Insight). It is then possible to use the labelling formats that were discussed in the SAS PC Workshops Manual, such that MALE and FEMALE are displayed for GENDER instead of the values 1 and 2.

## E. Saving Modified Data

If you have created any new variables, edited any of the data values, or changed the state of any observations using the show/hide or include/exclude features (these features are discussed in the next section), you may want to save the revised data set to a permanent SAS data set. To save the data:

- Choose **File > Save > Data**.
- A dialogue box will appear with a list of available libraries and a spot to enter a data set name. If you do not already have a library created for either your a: drive or c: drive, you will have to create one as outlined in the SAS PC Workshops Manual. This must be done in the program editor window, and can be done while running SAS/Insight.
- Select **OK** to save the data, once you have selected the appropriate library and data set name.



Note that the WORK library is a temporary library, active only for your current SAS session. Therefore you should not save your data to this library, unless you need to use it only for this session.

## V. EXPLORING DATA

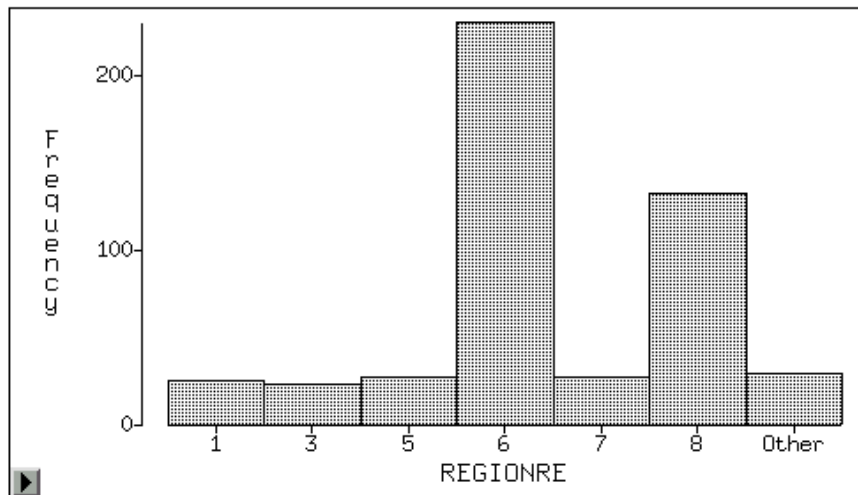
Using SAS/Insight, graphical distributions of variables can be displayed as bar charts or box plots. Bar charts display distributions of interval or nominal data. Box plots are concise displays of interval variable distributions and emphasise any extreme values.


### A. Bar Charts

#### 1. Nominal Data

To create a bar chart of REGIONRE, the patient's region of residence, do the following steps:

- Click on **REGIONRE** (you may have to use the horizontal scroll bar to find REGIONRE).
- Select **Analyze > Histogram/Bar Chart (Y)**.



A pop-up window will appear with a bar chart of REGIONRE. By clicking on any of the bars, the bar will be labelled with its frequency, and all the observations in the data window with that value of REGIONRE will also be highlighted. Since windows in SAS/Insight software are just different views of the same data, observations selected in one window will be selected in all the other windows. If both windows are not visible at once, click in the upper right-hand corner on the symbol that looks like: . Note that the active window is the one that is highlighted. To activate a window, simply click on it.

To attach the frequencies to all the bars at once, click on the **▸** button at the bottom left of the bar chart window, and click on **Values**. To remove the frequencies, click on **▸ Values** again. You can toggle **▸ Axes** and **▸ Observations** as well, to turn off and on the display of the axes and observations (although the resulting graph is not very useful). The **▸ Ticks** option only works for interval data.

To end this window, click on the **✕** in the upper right-hand corner of the bar chart window.

## 2. Interval Data

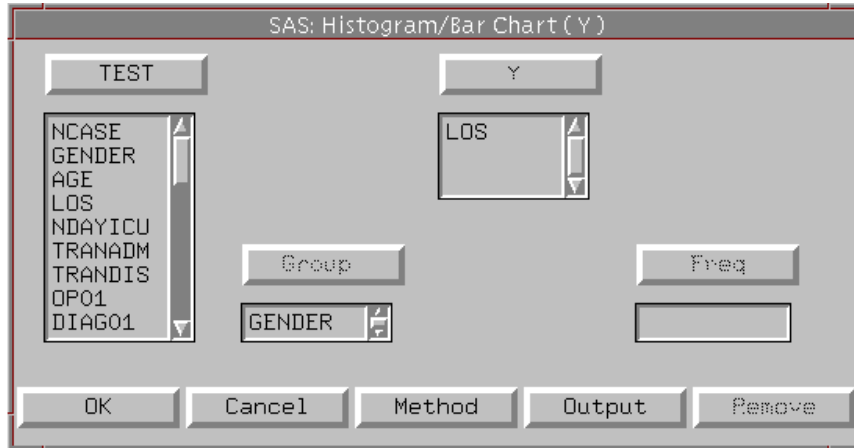
A bar chart can also be created for an interval variable, for example LOS (length of stay):

- Click on **LOS**, then **Analyze > Histogram/Bar Chart (Y)**.
- A window with a bar chart of LOS will appear, but due to some large values of length of stay the graph is very skewed.
- To find out how many large values there are, click on **▸ Values** – the frequency in each interval will be printed on the top of each bar. There are just a few observations past 45 days.
- To find these observations in the data window, click on **Edit > Observations > Find**.
- Click on **LOS** in the first box.
- Click on **>=** in the second box.
- Scroll down in the third box until you can click on **46**.
- Click **Apply**. This will find all the observations with LOS > 45 days in both the data window and the histogram.
- The observations with LOS > 45 days are spread throughout the data window. These can be moved to the top of the data window by clicking on **▸ Move to First** in the data window (leaving the other windows open).
- To remove these observations from the graph, click on **Edit > Observations > Hide in Graphs**. Notice that the little black boxes beside the records with LOS > 45 disappear, which means that the observations will no longer be included in the graph. Note that these observations will remain hidden in other graphs that you create, until you reverse this process by re-selecting these observations and selecting on **Edit > Observations > Show in Graphs**.
- Close the FIND pop-up window by selecting **Cancel**.
- Notice that the histogram is now somewhat more informative; in some versions of SAS it can be enlarged to fill the screen by dragging on a corner.
- Close the histogram by clicking on the **✕** at the upper right-hand of the histogram window.

To obtain a male/female breakdown in LOS, a histogram can be created of length of stay for each sex. To do this:

- Make sure no variable names are highlighted in the data window.

- Click on **Analyze > Histogram/Bar Chart**. A pop-up window will appear with a list of the variables on the left-hand side.
- Choose **LOS**, then click on the box with the **Y** in it (Y-axis). LOS will appear in the box under the Y. Y is conventionally used as the symbol to represent the dependent variable.
- Now click on **GENDER**.
- Click on the box with **GROUP** written on it. A histogram of LOS will be displayed for every value of GENDER.
- Click **OK**.



A window will open with a histogram of LOS, for GENDER=1 (males). Use the horizontal scroll bar to view the histogram for the females (or, alternatively, open the window to fill the screen). Close the histogram by clicking on the **X**.

If your data were aggregated, you would select a variable for **Freq** in the above dialogue box. Each observation would be assumed to represent  $n$  observations, where  $n$  is the value of the **Freq** variable.

## B. Box Plots

A box plot is a graphical representation of the data, showing such things as the mean, median and the 25<sup>th</sup> and 75<sup>th</sup> percentiles. To create a box plot of LOS, for each gender:

- Select **Analyze > Box Plot/Mosaic Plot (Y)**.
- Click on **LOS**, then the box with the **Y** in it.
- Click on **GENDER**, then the box with the **X** in it – a box plot of LOS will be created for each value of gender, on the same plot.
- Other fields are available in the box plot window:
  - i) **Group** - This will create a separate box plot for each value of the grouping variable. Recall that we did this with the histogram above (grouped by gender).

- ii) **Label** - This can be used to label the observations in the graph. The default is the observation number, unless you have selected a different default label as discussed in section IV (Defining Variables).
- iii) **Freq** - This field can be used for aggregated data (not applicable in this case).
  - Select **OK**

A window will appear with a box plot of LOS for each gender. The solid green horizontal line in each box marks the median, or 50<sup>th</sup> percentile. The bottom and top edges of the box mark the quartiles, or the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The ‘whiskers’ (narrow boxes that extend above and below the box) extend from the quartiles to the farthest observation that is not farther than 1.5 times the distance between the quartiles. Individual observations beyond the whiskers indicate extreme values. Clicking on an extreme value causes the observation number of that observation to be displayed. Note that if we had chosen a variable for the Label field mentioned in ii) above, that the value of that variable would appear.

Clicking on **Means** displays a diamond shape, with the horizontal bar indicating the mean value; the height of the diamond is 2 standard deviations high (one standard deviation to either side of the mean). Notice that the standard deviations seem quite large. Recall all observations with a length of stay of over 45 days are hidden in the graphs. These observations, are, however, still being used in the calculations. To exclude them from the calculations, they need to be identified:

- Click on **Edit > Observations > Find**.
- Click on **LOS** in the first box.
- Click on **>=** in the second box.
- Scroll down in the third box until you can click on **46**.
- Click **Apply**. This will find all the observations with LOS > 45 days in the data window and the box plots.
- To remove these observations from the calculations, click on **Edit > Observations > Exclude in Calculations**. Notice that the observation numbers of the records with LOS > 45 are now written in light grey. This indicates that the observations will no longer be included in the calculations unless we repeat this process choosing **Include in Calculations**.
- Close the FIND pop-up window by selecting **Cancel**.
- Close the box plot window.

## **VI. EXAMINING DISTRIBUTIONS**

Distributions of interval variables can be examined using both graphs (such as the bar charts and box plots) and statistical tables. SAS/Insight allows the user to look at all of these in one simple step. To examine the distribution of LOS:

- Select the variable **LOS** by clicking on its name in the data window.

- Choose **Analyze > Distribution (Y)**.

A pop-up window will appear with a box plot, a histogram, a Moments table and a Quantiles table, by default. See the SAS/Insight manual, or SAS **Help > Reference > Distribution** for detailed descriptions of each of the statistics present in the Moments and Quantiles tables.

Notice that additional pull-down menus are available at the top of the pop-up window. They are **Tables**, **Graphs** and **Curves**. They can be used to do additional analyses, such as adding density curves or testing for specific distributions.

Close the Distribution Window.

## VII. FITTING CURVES

Several methods for examining the relationship between a response (dependent) variable and a set of explanatory (independent) variables are available. You can use least squares methods for simple and multiple linear regression when the response is normally distributed. Generalised linear models and nonparametric regression methods are also available. The following simple example uses the least squares method and involves one response variable and one explanatory variable.

The variable RDRGWT is a variable for which higher values are associated with longer hospital stays. LOS can be used as the response variable – or preferably L\_LOS (the log of LOS) since it will be more normally distributed – and RDRGWT as the independent variable.

- Click on **Analyze > Fit (Y X)** (make sure no variable names are selected before you do this).
- Select **L\_LOS** and then click on the box with the **Y** (dependent variable) in it.
- Select **RDRGWT** and then click on the box with the **X** (independent variable) in it.
- Next select the box labelled **Output**. A dialog box appears with a list of tables and residual plots that can be produced. Click **OK** to select the default tables.
- Click on **Apply** to run the analysis. A dialog box will appear with the selected tables and graphs. Note that the **Tables**, **Graphs**, **Curves** and **Vars** menus are now available to be used. Extra tables and graphs can be added to the Fit output window and variables such as the residuals can be added to your data set.
- Close the Fit results window and the Fit (Y X) window.

## **VIII. OTHER CAPABILITIES OF SAS/INSIGHT**

This manual just touches the surface of what SAS/Insight is capable of doing. Using SAS/Insight it is possible to produce 2-D and 3-D scatter plots, to do correlations and ANOVAs, and to do complicated analysis such as multiple regression.

SAS/Insight is a valuable tool for those with limited programming knowledge whom only need to look at their data (data that has previously been cleaned and otherwise prepared for use). It is not recommended for large projects, since no record (log output) is kept of data manipulations, and some data manipulations are not possible, such as creating LOS groups.

## **Appendix 1: References and Acknowledgements**

### Reference:

SAS Institute Inc.: SAS/INSIGHT User's Guide – Version 6 Third Edition.  
Cary, NC: SAS Institute Inc.; 1995.

### Acknowledgements:

Assistance in developing this manual was provided by Charles Burchill and Ruth Bond. Charles Burchill, Ruth Bond, Akash Bedi and Robert Friesen provided assistance in the testing of the manual.

The “test” data set for the health data examples uses the “fivet93m.raw” data set prepared by Carmen Steinbach of the Manitoba Centre for Health Policy and Evaluation.