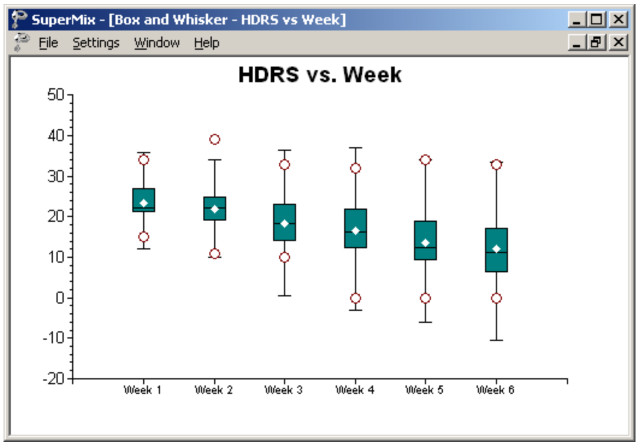
Contents

Data-based graphs: Bivariate graphs

Box-and-whisker plot for two-level data

A box-and-whisker plot is useful for depicting the locality, spread and skewness of a data set. It also offers a useful way of comparing two or more data sets with each other with regard to locality, spread and skewness. To illustrate this feature, we use **reisby.ss3** (see Section 4.3 for a brief description of these data). In the plot shown below, we requested box-and-whisker plots for the HDRS ratings that served as the outcome variable in previous analyses (see Section 3.2) at each of the measurement occasions.

The bottom line of a box represents the first quartile (q_1), the top line the third quartile (q_3), and the in-between line the median (me). The arithmetic mean is represented by a diamond. Here, a decrease in the mean HDRS rating is observed over the course of the study. In addition, the larger distances between the extremes of the boxes at the later measurement occasions indicate more variability in HDRS ratings towards the end of the study.



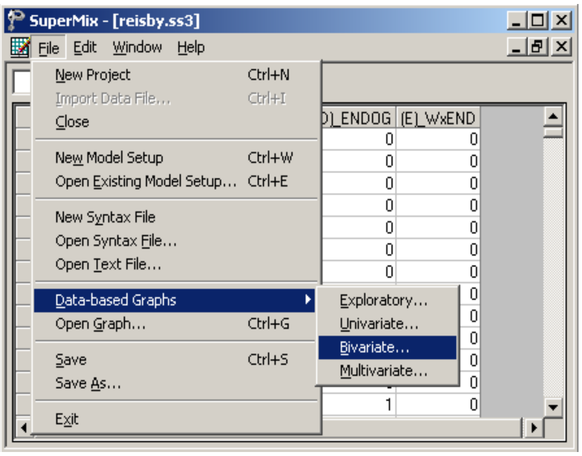
The whiskers of these boxes extend to $1.5(q_3 - q_1)$ on both sides of the median. The length of the whiskers is based on a criterion of Tukey (1977) for determining whether outliers are present. Any data point beyond the end of the whiskers is then considered an outlier. Two red circles are used to indicate the minimum and maximum values.

For symmetric distributions, the mean equals the median and the in-between line divides the box in two equal parts.

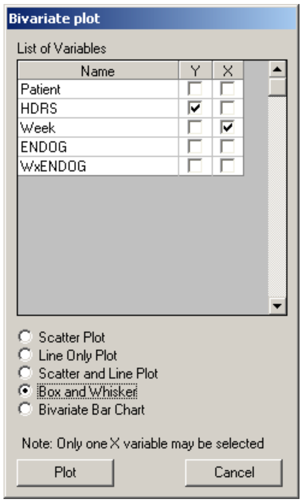
For a large sample from a normal distribution, the plotted minimum and maximum values should be close to the whiskers of the plot. This is not the case for the first box-and-whisker plot corresponding to the first measurement occasion. However, we conclude that for the remainder of the occasions the assumption of a symmetric distribution would be acceptable.

Creating a box-and-whisker chart

To create the box-and-whisker plot shown above, start by opening **Examples\Primer\Graphics\reisby.ss3**. Next, select the **Bivariate** option from the pop-up menu displayed when the **File**, **Data-based Graphs** option is selected from the main menu bar.



The **Bivariate Plot** dialog box is displayed. Select HDRS as the **Y** variable of interest, and Week as the **X** variable. Click the radio button next to the **Box and Whisker** option and then click **Plot** to display the box-and-whisker plot of HDRS at the measurement occasions.



List of Variables

Name	Y	X
Patient	□	□
HDRS	☑	□
Week	□	☑
ENDOG	□	□
WxENDOG	□	□

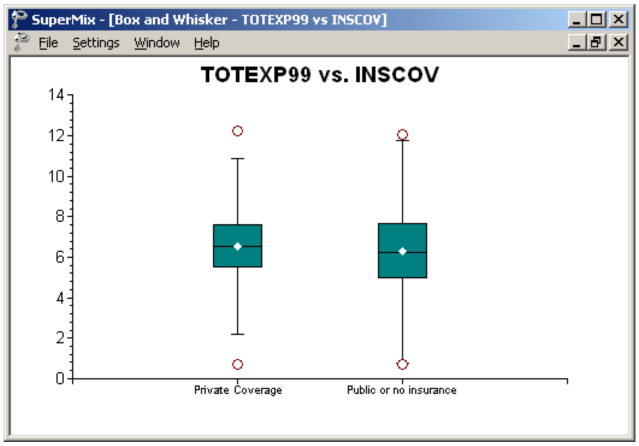
○ Scatter Plot
○ Line Only Plot
○ Scatter and Line Plot
● Box and Whisker
○ Bivariate Bar Chart

Note: Only one X variable may be selected

Box-and-whisker plot for three-level data

The data set used next (see Section 3.3 for a detailed description) forms part of the data library of the Medical Expenditure Panel Survey (MEPS). Collected in 1999, these data from a longitudinal national survey were used to obtain regional and national estimates of health care use and expenditure based on the health expenditures of a sample of US civilian non-institutionalized participants. The data is in the file **Examples\Primer\Graphics\meps.ss3**. A description of the variables TOTEXP99 and INSCOV is repeated below.

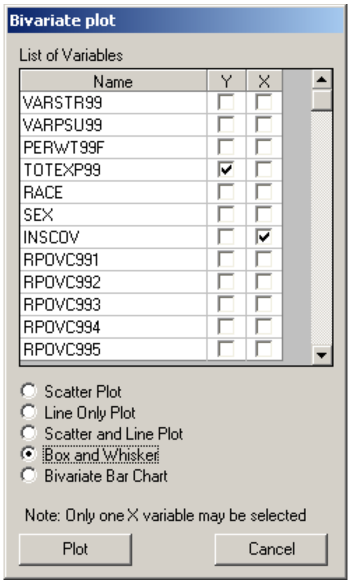
- TOTEXP99 is the natural logarithm of the total health expenditure of a respondent in 1999, ranging between 0 and 12.24 and representing actual expenditure of between $1 and $206,721.
- INSCOV is an indicator of the level of insurance coverage, where 0 indicates private coverage any time during 1999, and 1 indicates public coverage or no insurance at all during 1999.



The plot shows that for both types of insurance coverage the distributions of TOTEXP99 are close to symmetrical. This follows from the fact that, for each category of INSCOV, the median and mean values are close to each other and the median line divides the box into equal parts. It is also evident that the mean (or median) of TOTEXP99 is larger for the private coverage category. On the other hand, the variation in TOTEXP99 values for the public or no insurance categories is larger.

Creating a box-and-whisker chart

To create the box-and-whisker plot shown above, start by opening **meps.ss3**. From the **File**, **Data-based Graphs** menu, select the **Bivariate** option to activate the **Bivariate plot** dialog box. Select TOTEXP99 as the Y-variable and INSCOV as the X-variable. Click **Plot** to obtain the box-and-whisker plot shown above.



Bivariate plot

List of Variables

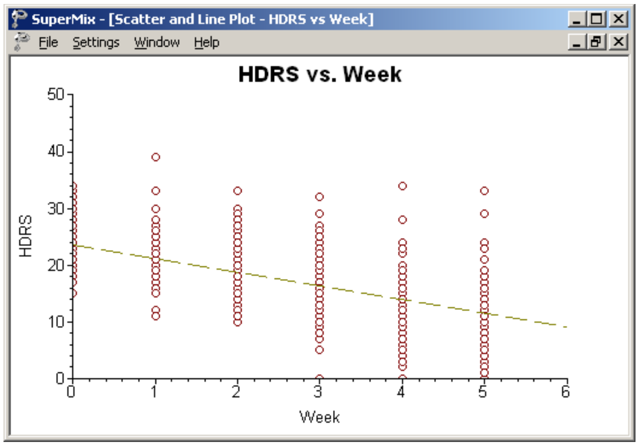
Name	Y	X
VARSTR99	□	□
VARPSU99	□	□
PERWT99F	□	□
TOTEXP99	☑	□
RACE	□	□
SEX	□	□
INSCOV	□	☑
RPOVC991	□	□
RPOVC992	□	□
RPOVC993	□	□
RPOVC994	□	□
RPOVC995	□	□

○ Scatter Plot
○ Line Only Plot
○ Scatter and Line Plot
● Box and Whisker
○ Bivariate Bar Chart

Note: Only one X variable may be selected

Scatter/line plot

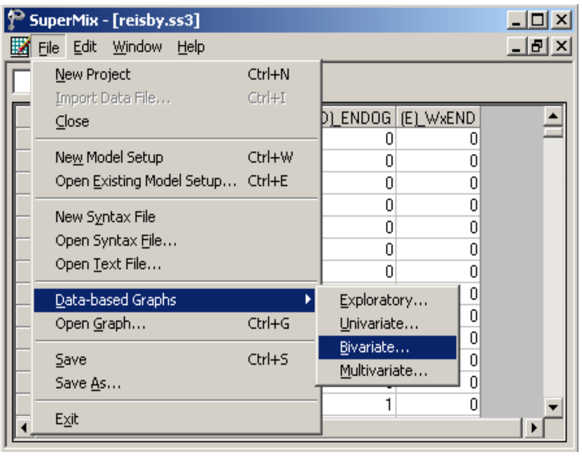
A graph frequently used as part of initial exploratory analysis of data is the scatter and/or line plot. This type of plot is used to examine potential relationships between a continuous outcome variable and possible predictor variables. Scatter plots are particularly useful for the study of repeated measurements data.



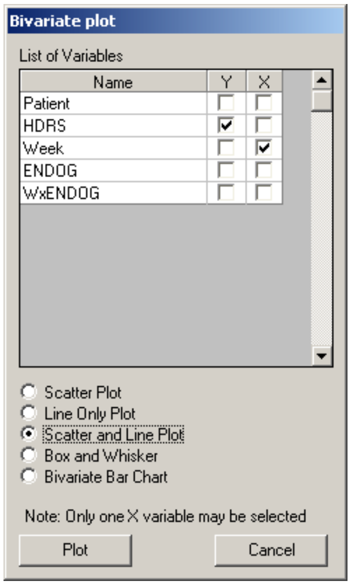
For the Reisby data, we looked at the patients' HDRS ratings at six time points. Previously, we plotted box-and-whisker plots of the HDRS ratings at the six measurement occasions. For these plots, all measurements at each of the six time points were summarized in the box-and-whisker plot for each of the occasions. A scatter/line plot allows us to also look at the trajectories of individual patients' HDRS ratings over the course of the study. The plot below shows these trajectories, with circles representing the actual measurements of the 66 patients (the scatter plot component) and a dashed line representing the fitted linear regression curve. We note that, generally speaking, the ratings seem to decrease over the study period.

Creating a scatter/line plot

To create the scatter and line plot of HDRS ratings at the measurement occasions, open the **Examples\Primer\Graphics\reisby.ss3** spreadsheet, and select the **File**, **Data based Graphs**, **Bivariate** option from the main menu bar.



The **Bivariate Plot** dialog box is displayed. Select HDRS as the **Y** variable of interest, and Week as the **X** variable. Click the radio button next to the **Scatter and Line Plot** option and then click **Plot** to display the combined scatter and line plot of HDRS at the measurement occasions. Note that similar plots, displaying either the observed data or the average line, can be obtained by using the **Scatter Plot** or **Line Only Plot** options.


Bivariate plot

List of Variables

Name	Y	X
Patient	□	□
HDRS	☑	□
Week	□	☑
ENDOG	□	□
WxENDOG	□	□

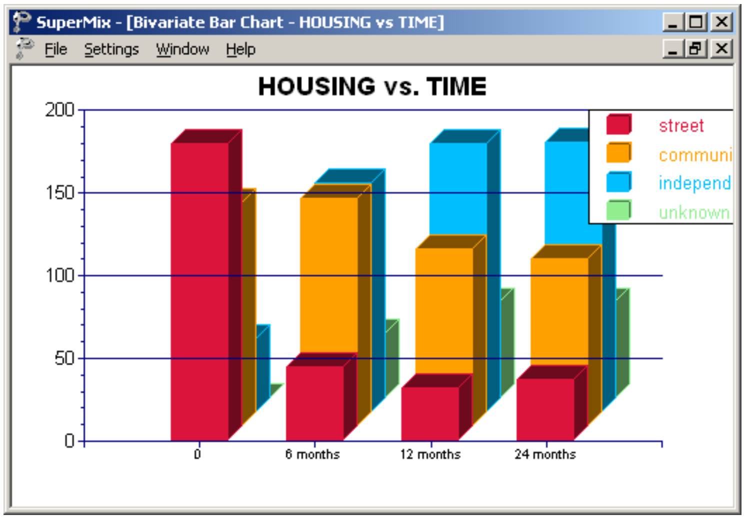
○ Scatter Plot
○ Line Only Plot
● Scatter and Line Plot
○ Box and Whisker
○ Bivariate Bar Chart

Note: Only one X variable may be selected

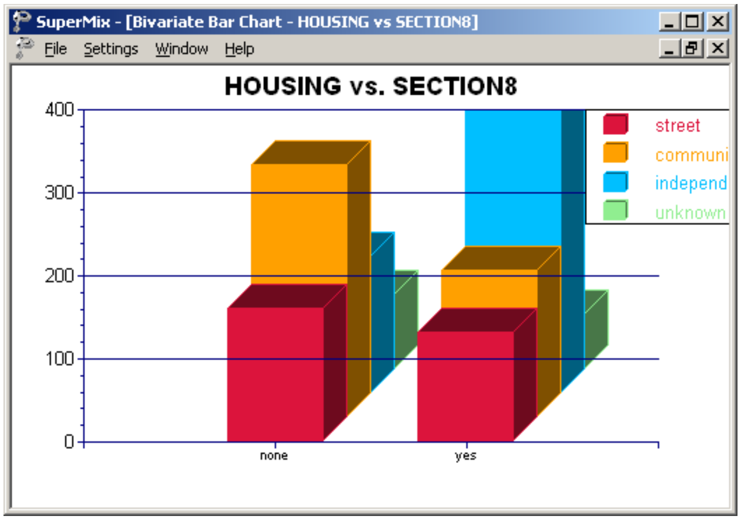
3D bar chart

The data from the McKinney Homeless Research Project study (see Section 3.7 and 4.3) were used for this example. Recall that the data file **sdhouse.ss3** contains subjects' housing status as recorded at 4 time points. From the pie chart, we noted, for example, that over the course of the study, subjects reported living in independent housing in 39.6% of the observations. In 22.8% of the measurements, subjects were living on the street.

A 3D bar chart allows us to graphically display a two-way frequency table. The 3D bar chart shown below is a visual display of the bivariate distribution of the variables HOUSING and TIME, where TIME represents the four measurement occasions. By including the TIME variable, we are essentially acknowledging the longitudinal nature of these data. The row of bars in the front of the graph, associated with the code "0" for HOUSING, represent the number of subjects who were living on the street at each of the four occasions. The second row of bars represent subjects living in community housing, and the third the subjects living in independent housing at the time of measurement. Over the study period, there was a marked decrease in the number of subjects living on the streets. The number of subjects living in community housing showed a decrease between the first two and the last two time points, while the subjects living in independent housing increased rapidly over time.

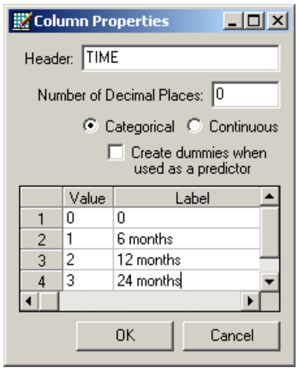


A key concern in the McKinney study was to evaluate the effectiveness of using Section 8 certificates to provide independent housing to the severely mentally ill homeless. The variable SECTION8, contained in the **sdhouse3.ss3** spreadsheet in the **Examples\Primer\Graphics** folder was used to distinguish between subjects with (SECTION8 = 1) and subjects without (SECTION8 = 0) Section 8 certificates. By portraying the variable **SECTION8** against the type of housing (HOUSING) in the form of a 3D bar chart, we can get some idea of the relationship between the type of housing reported and the use or not of a Section 8 certificate. From the 3D bar chart below, it seems as if subjects with Section 8 certificates were approximately twice as likely to report living in independent housing over the course of the study than those without, whom were more likely to report living in community housing. A slight decrease in the number living on the street is also observed in the case of reports by subjects with Section 8 certificates. Ideally, one would like to combine the information in these two 3D bar charts into one. One way to do so would be to prepare model-based graphs (see Section 4.6), based on an analysis where both the longitudinal nature of the data and the availability of Section 8 certificates are taken into account.



Creating a 3D bar chart

To create the 3D bar charts shown above, start by opening the data file **Examples\Primer\Graphics\sdhouse3.ss3**. Right-click on the column header of the variable TIME and select the **Column Properties** option from the pop-up menu to open the **Column Properties** dialog box. This dialog box is used to define the type of variable (nominal, ordinal or continuous) and to provide labels for the categories of nominal and ordinal variables. Indicate HOUSING as a nominal variable by clicking the appropriate radio button, and enter the labels (street, community, independent and unknown) for each category in the **Label** column. Do the same for the ordinal variable TIME as shown below. Click **OK** to return to the spreadsheet window and save the changes to the spreadsheet using the **File**, **Save** option.



Header: TIME

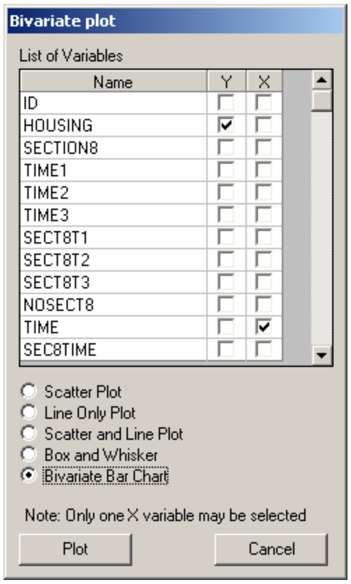
Number of Decimal Places: 0

● Categorical ○ Continuous

□ Create dummies when used as a predictor

	Value	Label
1	0	0
2	1	6 months
3	2	12 months
4	3	24 months

Select the **File**, **Data based Graphs**, **Bivariate** option from the main menu bar. The **Bivariate Plot** dialog box is now displayed. Select HOUSING as the **Y** variable of interest, and TIME as the **X** variable. To obtain a bivariate 3D bar chart of the housing status at the measurement occasions, click the radio button next to the **Bivariate Bar Chart** option and then click **Plot** to obtain the bivariate bar chart.



Bivariate plot

List of Variables

Name	Y	X
ID	☐	☐
HOUSING	☑	☐
SECTION8	☐	☐
TIME1	☐	☐
TIME2	☐	☐
TIME3	☐	☐
SECT8T1	☐	☐
SECT8T2	☐	☐
SECT8T3	☐	☐
NOSECT8	☐	☐
TIME	☐	☑
SEC8TIME	☐	☐

○ Scatter Plot
○ Line Only Plot
○ Scatter and Line Plot
○ Box and Whisker
● Bivariate Bar Chart

Note: Only one X variable may be selected

To obtain a bivariate bar chart of housing status and Section 8 certification, simply close the graphing window to return to the **Bivariate Plot** dialog box. Retain HOUSING as the **Y** variable, but uncheck TIME as the **X** variable and replace it with the variable SECTION8. Click the **Plot** button. The bivariate bar chart of HOUSING versus Section 8 certification, as shown earlier, is now displayed.