2. EXPLORING AND DESCRIBING DATA

The first thing to do with any set of data is to plot and inspect it visually. Inspection affords an opportunity to determine the shape of a distribution. This information is of interest on its own, but will also help to determine the type of analysis to carry out on the data. There are a number of useful tools for this, including descriptive statistics, histograms and boxplots. JMP IN offers all of these plus a number of additional exploratory tools. For now we will limit ourselves to these three methods, all widely used.

Histograms

A **histogram** plots the frequency of observations falling into different intervals of a continuous variable Y. The number and width of Y intervals should be chosen carefully, although there are no set rules for determining how many classes to use. If the range of Y values is divided into too many intervals, many of the intervals will contain no observations, and the histogram will resemble the skyline of a city dominated by skyscrapers. Someone viewing such a histogram will have difficulty determining the shape of the true distribution. Using fewer, larger classes can alleviate this problem. Holes in the distribution are smoothed over, giving a better picture of the shape of the distribution. It is possible to go overboard with smoothing. Histograms consisting of a few very wide classes may hide significant features of a distribution.

Boxplots

Histograms indicate the whole frequency distribution of a variable, whereas the boxplot summarises its most prominent features. These features include median and spread as well as the extent and nature of departures from symmetry, and the possible presence of observations having extreme values (**outliers)**. The ends of the box represent the lower and upper **quartiles** of the data, and the line across the middle of the box is the **median**. The median is the middle observation of a set of data. The lower quartile marks the median of the lower half of the observations, and the upper quartile is the median of the upper half. If the distribution of the data is symmetric (i.e. from a uniform or normally distributed population), then the box will appear to be divided equally into two halves by the overall median. The lines protruding from the box are the "whiskers". The length of each whisker is up to 1.5 times that of the length of the box (the whisker extends only to the last data point within this 1.5 limit). Beyond the whiskers live the outliers.

Outliers are extreme observations, those that lie unusually far from the main body of the data. These unusual observations may simply reflect the tail end of a highly skewed distribution, but sometimes they are errors of measurement or transcription, or represent individuals from a population other than the one under study. You might be tempted to delete outliers but this is usually justified only when there are errors. If an outlier is deleted that is not an error, valuable information is lost and a bias is introduced into later analyses. Yet including an erroneous measurement also has harmful consequences, data entry methods should be reviewed and specimens re-measured if possible. One strategy is to repeat every analysis with **and** without a suspicious observation and compare the results. If the conclusions from the two analyses are different then this should be reported.

Descriptive Statistics

Several descriptive statistics measured on the variable of interest will also appear next to the histogram. These include the mean and standard deviation. (Also included are the standard error of the mean and the 95% confidence interval).

Accessing Files from the Server

The procedure to access files from the shared drive is to choose the **Open** from the **File** menu at the top of the JMP IN window or click **Open Data Table** on the JMP Starter. Choose the **shared** drive S to grab files from the server. From the shared directory choose the file that you want.

Using the Program

After opening or creating a data table, select the **Distribution** option from the **Analyze** pull-down menu. Or, click "**Distribution"** on the **Basic Stats** tab of the JMP Starter. Select a column in the window that pops up, and then click "Y, columns". Finally, click OK to generate the histogram. The graph that appears provides an estimate of the distribution of the Y variable you chose.

You can modify the style of the graph and the information provided in various ways. By clicking the red "▼" symbol beside the variable name above the histogram. For example, **Display Options -> Horizontal Layout** to change the orientation. Select **Histogram Options -> Count Axis** to have the actual counts plotted along the side of the graph. Experiment with other options.

The same actions generate a **boxplot** next to the histogram. The type of boxplot produced is called an **outlier boxplot** (we will concern ourselves only with this type). It includes a number of features besides those mentioned above. To read about these extra features, select **Index** from the **Help** pull-down menu. Type "outlier box plot" as your keyword and press Enter on your keyboard. The description will appear in the right side of the help window.

Problems

1. Open the data file **poverty** from the shared directory (most files we will use will be located there). For 97 countries in the world, data are given for birth rates, death rates, infant death rates, life expectancies for males and females, and Gross National Product. These data were collected from *The Annual Register* 1992 (data for 1990) and the U.N.E.S.C.O. 1990 *Demographic Year Book* by M. Rouncefield. The variables are:
    - Live birth rate per 1,000 of population
    - Death rate per 1,000 of population
    - Infant deaths per 1,000 of population under 1 year old
    - Life expectancy at birth for males
    - Life expectancy at birth for females
    - Gross National Product per capita in U.S. dollars
    - Country Group
        1 = Eastern Europe
        2 = South America and Mexico

3 = Western Europe, North America, Japan, Australia, New Zealand
4 = Middle East
5 = Asia
6 = Africa

- Country

a) Create a histogram for the variable "death rate per 1000". Describe the general shape of the data distribution: normal (bell-shaped), uniform, skewed with a long tail to the left or right, middle-heavy (platykurtic) or tail-heavy (leptokurtic), or bimodal.

b) Choose the hand tool (click the open hand just below the top pull-down menus) and move it over the histogram. Hold the first mouse button down and describe what happens as you slide the mouse left to right. Describe what happens when you slide the mouse up and down instead?

c) How strongly is the histogram affected by changes in interval start points?

d) What are the consequences of too few intervals in a histogram? Too many?

e) Does the box plot change when you manipulate the histogram? Why?

f) Revert back to the pointer tool (click the arrow button on the tool bar just below the pull-down menus). Try highlighting one bar of the histogram by clicking on it. Now examine the original data table. What effect do you notice? Use this method to identify the countries having the highest death rates (in 1990).

g) Try the reverse: select several rows in the data table using the pointer or selection tool and inspect the effect on the histogram. Use this method to determine how Canada's death rate compares with that of the rest of the world.

h) Display a normal (bell-shaped) curve over your histogram (click the red "▼" symbol beside the variable name, and select **Fit Distribution -> normal**). The normal curve is one of the most useful distributions for statistical analyses. This is the shape that we hope to find in a plot of our data. How well does your histogram approximate a normal curve? (In future sessions we will learn about more powerful tools for testing normality.)

i) Examine the outlier boxplot included with the histogram. Do the data include any outliers? Use the selection tool (arrow) to click the outlying observation(s) and highlight the corresponding value(s) in the data set.

j) Are the interquartiles (the ends of the box) symmetric about the overall median? Does the range of the data set extend equally on either side of the box? Can you tell from the box plot whether the data are well described by a normal distribution?

k) How similar are the values for the mean and median of the death rate data? Which is larger, and why? Under what distributions would you expect the mean and median to be more similar?

2. Keep the histogram for death rate, but go back to the JMP IN menus to generate a second histogram, this time for "live birth rate per 1000". Put the two histogram windows side by side (you may need to look behind some open windows to find the first histogram again; use the **Window** pull-down menu to bring hidden windows forward).

   a) Describe the general shape of the birth rate distribution: normal (bell-shaped), uniform, skewed with a long tail to the left or right, middle-heavy (platykurtic) or tail-heavy (leptokurtic), or bimodal. Compare its shape with that for death rate. Are there any outliers in the birth rate distribution? Can you think of an explanation for why birth and death rates have such different distributions?

   b) With the pointer tool, click some of the bars in the death rates histogram corresponding to high death rate and observe the effect on the birth rates histogram. Do countries with high death rates tend also to have high birth rates?

   c) Repeat the above but click some of the bars corresponding to low death rates, and observe the effect on the birth rates histogram. Do countries with low death rates tend to have low birth rates? Contrast the result here with your finding in (b). Can you think of a reason for the difference?

3. Return to the data table. The variable "country group" is erroneously listed as a **continuous** variable in the columns list on the left of the data table window. It should be a **nominal** variable instead. Click the first mouse button over the **c** at the left of the variable name and change the variable type. DO NOT SAVE THE DATA.

   a) Using the same steps you used above to produce a histogram, produce a distribution of the nominal variable, "country group." Instead of a histogram or a picture of Hank Williams Jr. you will see a bar graph (unfortunately, without a gap between the classes). The counts in this case refer to the number of countries falling in each group. Instead of a boxplot you will see a "mosaic plot", which vertically stacks the same bars to provide easy comparison of relative frequency. We will work more with this type of plot in a later lab.

4. If time permits, experiment with the program to produce other kinds of graphs. For example, try producing a bivariate "scatter plot" (a plot of one variable against another) for pairs of variables such as GNP and infant death rate, birth and death rates, etc. Can you produce a separate box plot of GNP for each country group?