CS130/230 Lecture 6
Introduction to StatView

Thursday, January 15, 2004

Intro to StatView

StatView is a statistical analysis program that allows:

- Data management in a spreadsheet-like format
- Graphs and Tables
- A broad range of statistical analyses

Goals for this section of the course include:

- Becoming familiar with StatView and what it can do
- Creating new Datasets, importing & exporting Datasets
- Manipulating data in a Dataset
- Basic analysis of data (mainly descriptive statistics)
- Producing professional quality output using Word, Equation Editor, and StatView results.
- An overview of StatView's advanced features

Note: This is not a statistics course such as Math 207. We will only concentrate on basic statistical concepts.

Creating a Simple Dataset

Problem 1:

Let's start a new dataset from the File Menu -> New.

What you see is the main window called (Untitled Dataset #1). The top part of the window is for entering the variable names and properties, and the bottom half of the window is for entering the actual data.

Right now, we only have one column called "Input Column". This is not a variable and does not contain any data. We can use this column to create the variables we want in our dataset.

Part I: Add a variable called Brand by:

- Clicking the "Input Column" to select it
- Type the new name "Brand"
- Press Enter

Press Tab to move to the next column and add variables Name, Serving/pkg, Oz/pkg, Calories, Total Fat g, Saturated Fat g

The attribute pane is the top half of the dataset and consists of several rows of information, the first five rows, which tells you the type, source, class, format, and decimal places for each variable. You need to expand it in order to see all the information The first five rows are:

(1) Type - either integer, real, categories (group membership), string, currency, or date/time. Let's go through each column and make sure that the type is correct.

Part II: Change the Brand to type category by going to the first column, holding the mouse down on the Type pane in the Brand column, and holding the mouse button down. Select category, which brings up the Choose Category window with nothing in it. Select New then add the brand names Hershey, Charms, and M&M/Mars.

If you want to edit a category you've created then you need to go to the Manage menu and select (Edit Category).

Notice that the Format and Decimal Places are automatically set to missing. Since this is not a number.

(2) Source - is where the data came from (e.g. user entered, static formula, dynamic formula, ...). For now, our data will be user entered.

(3) Class - is how the data are to function (continuous measurements such as Serving, Total Fat, ... OR nominal data such as our Brand groups OR as informative data such as label names).

(4) Format - we will discuss a little later.

(5) Decimal Places is self-explanatory.

Part III: Now it's time to enter the data below. The data should be entered in the lower half of your dataset. You should enter the data in the grayed boxes making sure that the data is of the correct data type (i.e. Enter a number if the type is real, text if the type is string and so on).

```
M&M/Mars, Snickers Peanut Butter, 1,   2,    310,  20,  7
Hershey,  Cookies 'n Mint,        1,   1.55, 230,  12,  6
Hershey,  Cadbury Dairy Milk,     3.5, 5,    220,  12,  8
M&M/Mars, Snickers,               3,   3.7,  170,  8,   3
Charms,   Sugar Daddy,            1,   1.7,  200,  2.5, 2.5
```

Note: To expose all summary statistics, drag the attribute pane control (the x with a line above it) in the right hand portion of the window down. What is the mean of the calories?

Importing Data from Other Sources

It is also possible to import data from an Excel worksheet into StatView.

- File Menu -> Open
- In Files of Type -> Excel Worksheet (*.xls)
- Find the file Candy Bars.xls in the Sample Data Folder
- Open file

You should now have the complete candy bars dataset. StatView will convert the data types and formats to the nearest equivalent.

Question: Examine the dataset. Do you see anything that needs to be changed?

Note: Data can also be imported from simple plain text (ASCII) files exported by other applications. StatView can read files delimited by tabs, spaces, commas, returns, ... or any character you specify.

Sorting Data

Many times we want to look at sorted data. How could we sort the candy bar data by brand then by name?

Simply go to Manage Menu -> Sort and then Select the variable brand and Make Key. Do the same for Name. If you want to go decreasing, click on the arrow and it changes.

Problem 2

To make sure we stay fresh with Excel for the final, here's a little problem. Generate 100 random integer numbers (i.e. the numbers do not contain any decimal places) between 1 and 20. Beside each number output "EVEN" or "ODD". Save this file as random.xls.

Notice, that if we need to generate some random data and transfer it into StatView for testing, using Excel is a great way of doing this.

StatView Tidbits

- Enter data in columns pressing Return
- Enter data in rows pressing Tab
- Use the "Add Multiple Columns" in Manage Menu to add several columns with the same attributes
- Column widths can be modified similar to Excel by placing the cursor on the column divider and dragging
- A new column can be added by positioning the cursor on the vertical line between the columns on the variable row, holding down the command key until the cursor changes, and then click the mouse
- Similar thing with rows

Types of Data Analysis

When doing data analysis, we are interested in two types of summaries:

1) Statistical Summaries (e.g. descriptive, hypothesis testing)

2) Visual Summaries (e.g. tables, graphs)

Statistics is sometimes broken up into two different areas:

1) Descriptive Statistics - a situation is described by the statistics by the collection, summarization, organization and presentation of data.

2) Inferential Statistics - where inferences are made from samples of the population (e.g. smokers smoking a pack of cigarettes per day have a higher cholesterol). In this area we get into Hypothesis testing.

In the Descriptive Statistics world, we are concerned about each of the following. Just give a general description of the meaning of each of the following terms:

- Mean
- Median
- Mode

Here is an interesting problem that Descriptive Statistics can help us get a handle on.

Problem 3

A paint manufacturer tested two experimental brands of paint over a period of months to determine how long they would last without fading. Here are the results:

Brand A	Brand B
10	25
20	35
60	40
40	45
50	35
30	30

What do the descriptive statistics tell us about the paint with regard to fading?

Histogram

Let's see how good the random number generator in Excel really is.

Problem 4

Import the random number file we created at the beginning of class into StatView and let's create a histogram of the random data.

Part I: Import the data.

Part II: Create a histogram of numbers. (1) Analyze Menu -> New View (2) Click on the Frequency Distribution Triangle (3) Select Histogram (4) Select Create Analysis and click ok (you do not to change any of the options at the moment (5) Select the random number from the variables box on the right. If you can't see the random number variable make sure that you have the correct dataset selected in the drop down box.

Question: Based on what you see, how good is the random number generator?

Scatterplots of One Variable

Another type of graph is a Cell Plot. Cell plots are use to show the means for a variable of your choice split by some nominal variable.

Problem 5

There is a sample data file called "Lipid Data". I would like you to take this file and produce a bar chart in the cell plot option showing the mean weight of the people in the file split by Gender. Also make a plot of the mean Cholesterol split by Gender.

These two plots really allow us to examine one variable of interest. What if we want to examine the relationship between two variables?

Using More Than One Variable

In statistics, we can define two types of variables:

(1) independent - "it is what it is" and nothing influences it (e.g. Gender)

(2) dependent - most likely dependent on another variable (e.g. Cholesterol may be dependent on age)

Problem 6

Consider the following table which shows the number of bushels of wheat produced for the given rainfall amounts:

Rainfall	2.5	3	4.5	7.6	9.5	10.3
Bushels	37	43	42	46	48	51

The rainfall amount is given in inches.

We want to plot this data onto a scatterplot (scattergram) and find a trendline that best fits the data. This is similar to the regression exercises that we did in Excel.

Part I: Create a new dataset and add the rainfall and bushel information to this dataset.

Part II: Select New View from the Analyze menu and go to Regression Plot under the Regression option. Select the simple option for the moment. This will draw perform linear regression. Determine which variable is the dependent one and which is independent and plot the data.

Part III: How many bushels of wheat will be produced if the rainfall amount was 6.2 inches?

Part IV: How much rainfall would we need to have to produce 60 bushels of wheat?