CS130/230 Lecture 7
Graphing and Regression

Tuesday, February 24, 2004

Last time we covered comparison operators, logical functions and the IF function. Colleagues

Problem 1

Consider the following data:

Name	ID#	Quiz1	Quiz2	Midterm	Final
Adams	0001	14	23	82	76
James	0002	12	21	76	68
Jones	0003	15	24	91	93
Mann	0004	14	19	88	73
Smith	0005	11	16	79	71
Tolls	0006	10	13	62	65
Wells	0007	5	10	43	55
Points		**15**	**25**	**100**	**100**

Part I: Add two additional columns as follows: (a) Average is a person's total points are divided by the total points possible (b) Grade is 90-100 A, 80-90 B, 70-80 C, 60-70 D, 0-60 F.

Part II: Create a Pie Chart that shows the percentage of A's, B's, etc. Hint: you will need to use the COUNTIF function. You can look up how it works in Excel help.

Graphing with Large Amounts of Data

Problem 2

The file ‘education.txt’ in the ‘CS130 Pub’ folder contains information about education in the United States divided by state. The numbers listed next to each state represent the following:

- Cost per pupil
- Pupil/teacher ratio
- Mean annual teacher salary
- Percentage of students that take the SAT test
- Mean verbal score on the SAT test
- Mean math score on the SAT test
- Mean total score on the SAT test

Copy the data in this file into an Excel worksheet. You will need to format the data using the ‘Text to Columns’ option in the ‘Data’ menu.

A useful feature in Excel is the 'Freeze Panes' feature in the 'Window' menu. Select the top left number in the table and click on 'Freeze Panes'. This will make any rows above this cell and any columns to the left of the cells freeze, while you can still scroll the rest of the spreadsheet.

Add a column chart to your spreadsheet that will show the differences in SAT scores (verbal, math and total) between two states of your choice. Label your chart correctly.

Regression Analysis

Regression analysis is a form of statistical analysis used for forecasting. Regression analysis estimates the relationship between variables, so that a particular variable can be predicted from one or more other variables. During regression analysis, we need to fit functions to data.

Trendlines are used to graphically display trends in data and to analyze problems of prediction. In other words we try to draw a line that best fits the data. By using regression analysis, you can extend a trendline in a chart beyond the actual data to predict future values.

This subject usually falls under statistics and mathematical modeling and can be applied to many different scientific and business applications. Understanding the various formulas for regression is beyond the scope of this class. However, you should understand that the line should be placed such that the distance or variation from each data point to the line is minimized.

Linear Regression

In linear regression we try to find a straight line that best fits our data. We first need to plot our data using Excel's XY or scatter chart. We then add the trendline to the chart and use the function to predict future values for our data.

The detailed steps are:

- Enter the data in an Excel worksheet and select the data you want to plot.
- Click on the chart wizard.
- Choose XY (scatter) plot.
- Check that the data range is correct.
- Enter the titles and labels.
- Click on the chart then select CHART from menu bar and ADD TRENDLINE from this menu.
- From the menu that appears, select the type of function that you would like to use for your model. In this example we will use the default, which is LINEAR REGRESSION.
- In order to have Excel display the equation of our regression line and the correlation coefficient, you need to click on the OPTIONS tab within the Add Trendline screen above, click on these two options, and then press OK.

You should be rewarded with a graph, equation and regression coefficient.

Problem 3

In the CS130 Pub folder is a file called Candy Bars.xls. Copy this file to your folder, open it and do the following.

Part I: Create a ScatterPlot of the data Carbohydrates and Sugars.

Part II: Add a trendline to your chart and display the function or equation.

Part III: What is the amount of sugars (in grams) that we can expect from a candy bar with 60 grams of carbohydrates?

Part IV: Add an empty column after name. In that column, place an asterisk for foods that have a carbohydrate count of 40grams or higher and a sugar count of 35 grams or higher.

Part V: Turn on the AutoFilter and find out the number of M&M/Mars candy that fits these criteria.

Regression Coefficient

The regression coefficient, also known as the R-squared value, is an indicator that ranges in value from 0 to 1 and reveals how closely the estimated values for the trendline correspond to your actual data. A trendline is most reliable when its R-squared value is at or near 1.