Computer Lab Project No. 9

Goodness-of-Fit, Contingency Tables, Linear Regression

Today's topics are Goodness-of-Fit, Contingency Tables and Linear Regression. First, here are the general procedures for a **Goodness-of-Fit** analysis:

1. Start StatCrunch.
2. Put the observed frequencies in one column and the hypothesized frequencies in another, so that they line up.
3. Click on Stat→Goodness-of-Fit→Chi-Square test.
4. In the popup-window, put the observed frequencies column in the text field labeled "Observed", and put the hypothesized/expected frequencies column in the field labeled "Expected".
5. Click on the "Calculate" button on the bottom right.

Contingency Tables:

1. Enter the row labels in a column, then enter the frequencies in separate columns. So the row labels go in the spreadsheet body, unlike column headings.
2. Click on "Stat" in the menu bar.
3. Click on "Tables", choose "Contingency", and click the option "with summary".
4. In the popup-window, under "Select column(s)", select all columns containing observed frequencies (by ctrl-clicking them in turn), and then in the next box titled "Row labels:", select the column with the row labels.
5. Make sure under "Hypothesis tests:", the Chi-Square test for independence is chosen.
6. Click on the "Compute!" button on the bottom right.

Here are the projects concerning goodness of fit and contingency tables you can work on today:

1. Two dice are rolled repeatedly. The frequency distribution of the results are as follows.

   Die 1:

Number	1	2	3	4	5	6
Frequency	20	12	11	8	9	0

   Die 2:

Number	1	2	3	4	5	6
Frequency	13	12	8	12	8	7

   For each die, test the hypothesis that it is fair.

2. Refer to your textbook for a discussion of the following data describing the fate of passengers and crew on the Titanic:

	Men	Women	Boys	Girls
Survived	332	318	29	27
Died	1360	104	35	18

   Perform a test for independence between the categories of surviving/dying and passenger type.

Now let's turn to Linear Regression. The general procedure, using StatCrunch, is as follows:

1. Load the (two-dimensional) data you want to analyze into the StatCrunch table.

2. Click on Stat→Regression, and select the option "Simple Linear".

3. In the popup window, select the columns you want to analyze as X- and Y-variable.

4. Scroll down until you reach the "Graphs" field. If you want to see the data as a scatter plot, fitted with the regression line, then choose "Fitted line plot".

5. Click the "Compute!" button on the bottom right. If you chose the graphics option, then you can go back and forth between the numerical linear regression results and the plot using the "Next>" and "<Back" buttons.

Here is what you can work on:

1. Load the Data Set titled *Bear Measurements* into StatCrunch. This data set contains measurements of wild bears such as age, sex, headlength, weight, etc.

2. Use linear regression to analyze the relationship between the age of a bear and at least two of the measurements (other than month or sex). Make plots.

3. Compare with the results of your peers. Which measurement seems to be the most reliable/accurate indicator of the age of a wild bear?

4. If you have time, you can try to analyze the data for male and female wild bears separately. Does this improve the linear approximation? Are there outliers? Can you guess why? If you take away potential outliers, what is the best correlation coefficient you get?