Homework 1

Here is the list of problems constituting the first homework assignment. Please, first try to find your own solutions and after this effort consult these solutions with the ones presented during tutorials and, finaly, check the solutions that will be posted on the course webpage. Remember that problems can have several different but correct ways of solving them.

Multiple choice questions

1. In the weekly sales in a sports club, we have seen a seasonal pattern (a), exceptional high values during Christmas week and long weekends(b), and changes of deviation from seasonal pattern on the week-to-week basis (c). Out of these three [(a), (b), (c)] sources of variation in the data

   A (a) represents an assignable cause of variation, while (b) and (c) represent variation caused by chance;

   B (a) represents variation caused by chance, while (b) and (c) represent assignable cause of variation;

   C all three (a), (b), and (c) represent assignable causes of variation;

   D (a) and (b) represent assignable causes of variation, while (b) represents variation caused by chance;

   E all three (a), (b), and (c) represent variation caused by chance.

2. The process view of data

   A displays data in a time ordered line plot;

   B displays data in histogram format;

   C processes data by summarising the main features in numerical form and displays the resulting summaries;

   D uses statistical data processing software to provide a professional level summary of the data;

   E summarises the data in the shape of a Normal curve.

3. There are available data on weekly precipitation in a town and weekly beer sales in a local bar over a period of ten years. To see if considering weather prediction in running the bar business is worth of an effort one could

   A draw histograms of these two data sets and examine if their shapes are similar;

   B draw a scatterplot view of the data examine its shape and detect any form of dependence between the variables;

   C check during which week the most beer was sold and check how much rain there was during this week;

   D write data into the table and examine if there are any relation between values;

   E check if the histograms for both data sets resemble a Normal curve.

4. Assignable causes of variation

   A always follow the Normal model;

   B follow the Normal model when there are no chance causes of variation;

   C correspond to the most frequent classes in a histogram;

   D are the many unpredictable but collectively influential factors that affect a process or system;

   E are the few factors, individually influential and with predictable effect, that affect a process or system.

5. The Normal model for chance variation

   A is a flexible family of frequency distributions with mean and standard deviation as parameters;

   B is the usual model for the linear part of a simple linear regression equation;

   C has a frequency distribution which is evenly spread across the scale;

   D is the standard model for the assignable causes in a chance system;

   E is the result of a series of orderly experiments.

6. Which one of the following is correct?

   A response relationships reflect the flow in a process;

   B a response variable is seen as responding to changes in an explanatory variable;

   C an explanatory variable is one in which the explanation changes from time to time;

   D an explanatory variable is one whose values are explained by the order of a process;

   E explanatory variables may be correlated with the response variable but not with each other.

7. A simple linear regression model

   A is one where the vertical axis moves closer to the horizontal axis;

   B is one where all the points fall on a straight line;

   C attempts to explain the variation in a response variable in terms of distance from a straight line;

   D attempts to explain the variation in a response variable in terms of a linear relation with an explanatory variable and chance variation;

   E is another name for a scatter plot.

8. The prediction error associated with a simple linear prediction formula

   A is conventionally set at plus or minus twice the standard error of the mean;

   B is always less than the prediction itself;

   C is less than the prediction itself for large values and greater for small values;

D is conventionally set at plus or minus twice the standard deviation of chance variation;

E improves as the degree of extrapolation increases.

9. The regression coefficients in a multiple linear regression

   A are not readily interpretable when the explanatory variables are related;

   B are determined in advance through the process of problem formulation;

   C are only meaningful in an observational study;

   D are also referred to as indicator variables;

   E indicate by their values which version of the prediction formula is appropriate.

10. Explanatory variables in a statistical prediction model

    A explain why the prediction did not work;

    B explain the prediction error involved in forecasting;

    C are related to the variable being predicted;

    D are the coefficients in a multiple regression equation;

    E are variables whose values explain why prediction is necessary.

Problems

1. In the process of manufacturing of tennis balls three new presses have been acquired. After running several batches of tennis balls through the production process, it has been established that the mean values of the tennis ball core diameters for the three prersses are 95, 77, 90, respectively, and their standard deviations are 1.5, 4.1, 5.5. If the quality bands set by the Tennis Federation are: the lower bound is 70 while the upper one is 100, and the manufacturer sets the two standard deviations rule for the press, find out which of the three presses will pass the quality inspection. What is the answer to this question in the case when the manufacturer puts a more stringent three sigma rule as the quality requirement?

2. On a graph of y against x, plot the points $(3, 4)$, $(0, 3)$, $(9, 6)$ and the line with equation $y = 3 + \frac{1}{3}x$. If the simple linear model is represented by

   $$Y = 3 + \frac{1}{3}X + \epsilon$$

   and for values of $X = 0.5, 2, 9, 6$ the error term ϵ took values $0.8, 0.5, -0.3, -0.9$, respectively, plot on the graph the observed values of Y.

3. The simple linear regression model that was proposed in the textbook for the relation between manpower and volume of processed mail has been given by

   $$\hat{Y} = 50 + 3.3X + \hat{\epsilon},$$

   where the prediction error $\hat{\epsilon}$ is in ±20 range (this according to the two sigma rule so the standard deviation of chance variation is 10). Use the above prediction formula to estimate the loss incurred through equipment breakdown in period 6, fiscal 1963, and to predict the *extra* manpower requirement during the Christmas period, based on the experience of period 7, fiscal 1962 and 1963. We refer here to Table 1.3, p.24.

4. Consider the data in Table 1.4, pp. 33. Using a pencil and paper construct a scatter plot of the stamp sales vs. meter sales for the data from 1964 until 1977. Do you observe any relation between these two variables? If yes, draw (by hand) a straight line that in your opinion the best represents this relation. After that write an approximate equation for this line in the form

$$Y = a + bX,$$

i.e. find approximate values of coefficients a and b from your graph. Finally, using this equation evaluate prediction errors for the data from the years 1964-1977.

5. The prediction formula for stamp sales that was found in the textbook, see page 41, is given by

$$Y = 340 - 0.0316X_1 - 70.2X_2 \pm 8,$$

(where the bounds for error are given by the two sigma rule so the standard deviation σ is equal to 4). Here X_1 denotes Gross National Product (GNP) and X_2 represents Real Letter Price (RLP), the latter being defined as the standard stamp value divided by the Consumer Price Index (CPI).

Suppose that the progonosis for 1984 of GNP is 1484.5 and of RLP is 1.835. For 1985, the Central Bank predicts 1.5% increase of GNP and the inflation rate of 5.5% (inflation is increase of CPI expressed as the precentage of the previous year CPI). Use these values to obtain the predicted stamp sales for 1985.

6. Predict the effects on stamp sales in 1984 and 1985 of 5% and 10% increases in stamp prices and 5% decreases in stamp prices.

7. In the analysis of housing completions in the "good years" 1993-2000 of Irish economy, the following model has been proposed for the quarterly data

$$Y = 3937 + 259t,$$

where t represents time period and Y represents predicted completions (see pages 47-50 of the textbook). To account for difference in completions in each of the quarter due to seasonal effect, the following more general model has been proposed

$$Y = 3248Q1 + 3901Q2 + 4174Q3 + 5031Q4 + 250t \pm 500,$$

where $Q1, Q2, Q3, Q4$ are dummies variables equal either zero or one depending from which quarter data are considered (see also Table 1.7, p. 49). Write down separate prediction formulas for each of the four quarters. Evaluate predictions for each quarter of 2001 and of 2002. Do you think that you can use the same approach to predict completion for the year 2009? Support your answer with proper arguments.