­INDIVIDUAL WRITTEN EXERCISE

STATISTICS FOR BUSINESS MANAGEMENT

MASTER IN STATISTICS AND OPERATIONS RESEARCH

November 3, 2021

Answer briefly to the following 10 questions (answers in Spanish are also accepted). You can use your course notes. Results will be available by November 18 or earlier

(Dr. BOLANCE sessions)

1. Indicate whether the following statements are true or false. If false then explain what statement should be true.

|  |  |  |
| --- | --- | --- |
| **QUESTION** | **TRUE/FALSE** | **IF FALSE EXPLAIN WHY** |
| 1. The ODS and Data Warehouse are associated to the informational environment | True |  |
| 1. The DELTA model classifies the company in five analytical types | True |  |
| 1. The data Marts are associated to the operational environment | False | False, because data marts (and databases) are normally associated with the informational environment, while the operational environment only has current values (non-persistently stored in databases, perhaps only stored day to day and overwritten each day with the current values) |
| 1. A persistent transaction does not have to be stored in the database | False | False, because a persistent transaction is by definition stored in the database. That is (a part of) what it means to be persistent. |
| 1. CRISP-DM is the first step in the roadmap of companies that want to have an informational environment | True |  |

1. The model scoring results from applying a model to a data set. What is the difference between the scoring obtained with a Kohonen and the scoring obtained with a regression?

ANSWER: The difference between the scoring obtained with a Kohonen and a scoring obtained with regression is due to the fact that the Kohonen is used for classification while a regression is used for prediction. This means that the scorings from Kohonen are class predictions and confidence levels regarding the assignment to each class, while regression gives predicted values (continuous) and prediction standard errors.

1. In the six-sigma context, it is known that DPMO can be calculated as:

,

where is the standard normal cumulative distribution function (remember, in standard normal =1).

1. Explain the value of DPMO expressed in the above formula.
2. What is the desirable level?
3. What is the desirable DPMO?

ANSWER:

a) The value of DPMO is the long term defective parts per million when approximating our process with a normal distribution. The formula calculates the probability of the process having “level” number of sigmas that are non-defective, multiplied by 1000000.

b) The desirable level is 6.

c) The desirable DPMO is (less than) 3.4.

1. In a production process, it is known that CTQ is 500, and the desirable six-sigma upper limit is 600. From a data set with 1000 observations, we obtain that the data are distributed as a normal with estimated mean around 500 and estimated standard deviation 20. Is the current process within desirable six-sigma levels? Why?.

ANSWER: No, this is not within desirable six-sigma levels, since the DPMO (using the formula above) is calculated to approximately 233.

(Calculated in R: 1000000\*(1-pnorm(600,500+1.5\*20,20)) )

This means that the six-sigma project (in this case) will consist of analyzing all company data in order to find factors such that the standard deviation can be reduced from 20 to approx. 16.7, which is the k-sigma process associated with the observed data.

( Found in R: s <- (600-500)/6 )

1. The cost of producing a part of an electronic system is random. From a set of observed data, we estimate that the asymmetry of the evaluated costs is positive. To apply the six-sigma method two one-parameter Box-Cox transformations are proposed, the first using =1.5 and the second using =1.5. Given the skewness of the data, what transformation would you use? Why?.

ANSWER: I would use lambda = -1.5, since the data is positively asymmetric, which means it is right skewed. On right-skewed data a negative lambda would work nicely for the Box-Cox transformation, in order to “restore”(/transform) to normality.

(Dr. GUILLEN sessions)

1. Read the following text: “*A manager has analysed the fluctuations of sales in the period 2010-2021 during summer, using a time series of monthly volume of sales*” What type of analysis is this? Descriptive, predictive or prescriptive?

ANSWER: As the text does not say how the analysis will be used, I will say that this still can be regarded as descriptive analysis. He has used a time series of monthly volumes of sales to look at what has happened earlier and to help explain why. However, this can easily be called a predictive analysis also if the goal is to use the time series model to predict future trends based on the historical data.

1. What is the purpose of PMML? Is it possible to estimate one model in R and then implement the results in SAS directly?

ANSWER: The point of PMML is to avoid proprietary issues and incompatibilites, and to gain a clear separation of tasks. It is also used for deployment, since it makes it possible to separate model development from model deployment. This means that one of the advantages of PMML is that the (programming) language (or more generally: software) of estimation can be separated from the language of implementation. Thus, it is possible to estimate a model in R, transfer the estimations to SAS via PMML and implement the results in SAS directly. Lastly, I would like to mention that you also can specify pre- and post-processing of the data in PMML, which means that it contains the entire description of how the data has been used to fit a given data-mining model.

1. I have a group of 5000 clients and I want to estimate the proportion of clients that would consider some offer if I give them a 3% discount. I want to conduct a survey. I assume that the proportion of clients accepting the offer should be around 1/3, with a confidence level equal of 95%, if I want an error margin less than +/- 4%, what should my sample size be?

ANSWER: The sample size should be 483 (according to the formula for calculation of sample size when estimating a proportion).

1. What is the purpose of setting a seed number before a simulation starts?

ANSWER: The seed initializes the pseudo random number generator, which means that it determines what numbers are generated. This means that you will get the same sequence of numbers each time you initialize the generator with the same seed. The purpose is reproduction – in order to test code that contains generation of (pseudo-)random numbers one needs to be able to reproduce the same results over and over again.

1. Write an R program to simulate and calculate the probability of throwing 2 dices and the sum of the result is equal or larger than 10.

ANSWER: Added in the zip file with this document.