DEPARTAMENTO DE MATEMÁTICAS
UNIVERSIDAD AUTÓNOMA DE MADRID

Statistics and Probability - 2019-2020

Bachelor degree in Computer Sciences and Engineering

Assignment 1

Exercise 1. An article in *Information Security Technical Report* ["Malicious Software-Past, Present and Future" (2004, Vol. 9, pp. 6-18)] provided some data on the top malicious software instances for 2002.

To study the spread of the malicious software, a random sample from all the registered instances for 2002 was drawn and the malicious software reported in each incidence was recorded. The dataset is available in the file `msoftw_data.txt`. To ease the representation of the dataset, the names of the malicious software were encoded as follows:

Malicious software code	Malicious software names
MSoftw 01	I-Worm.Klez
MSoftw 02	I-Worm.Lentin
MSoftw 03	I-Worm.Tanatos
MSoftw 04	I-Worm.BadtransII
MSoftw 05	Macro.Word97.Thus
MSoftw 06	I-Worm.Hybris
MSoftw 07	I-Worm.Bridex
MSoftw 08	I-Worm.Magistr
MSoftw 09	Win95.CIH
MSoftw 10	I-Worm.Sircam
Others	Others

(a) What is the sample size of the dataset?

(b) Construct the frequency table of the dataset considered.

(c) Display the bar graph of this dataset.

(d) Describe the results that you obtained in *(b)* and *(c)*.

(e) How many computers of the sample considered were infected by *I-Worm.Tanatos*?

(f) What malicious software was the most frequent one in this sample?

(g) Find the proportion of computers of this sample infected by *Macro.Word97.Thus*.

(h) What is the percentage of computers of the sample infected by *I-Worm.Bridex*?

Exercise 2. An article in the *IEEE Transactions on Instrumentation and Measurement* ["Direct, Fast, and Accurate Measurement of V_T and K of MOS Transistor Using V_T-Sift Circuit" (1991, Vol. 40, pp. 951-955)] described the use of regression models to express drain current y (in milliamperes) as a function of ground-to-source voltage x (in volts). The datasets corresponding to certain sample are gathered in the file `transistor_data.txt`.

(a) Draw a scatter diagram of these datasets and describe the relationship that you observe between the two variables in this sample.

(b) Find the value of the sample (linear) correlation coefficient and interpret it.

(c) Determine the regression line for this dataset and plot it.

(d) Find the value of the sample linear coefficient of determination.

(e) What can we say about the accuracy of the predictions made by using the regression line?

Assignment rules:

- Solve the previous exercises using `R` (strongly recommended), `Calc` or `Excel`.
- You can solve the exercises in groups of no more than 3 people.
- All answers must be explained.
- Only a compressed folder must be sent. This folder must contain:
  - A `pdf` file with all the solutions. Please, indicate the full names clearly in this file.
  - When using `R`, the `.RData` file where you ran the commands to obtain the solutions that you provide in the `pdf` file.

  - When using `R`, the `.Rhistory` file with the commands you have run to obtain the solutions that you provide in the pdf file.
  - If you use `Calc` or `Excel`, the `Calc sheet` or `Excel sheet`.

- The pdf file must be submitted in Times New Roman font in 12 point type, single-space with 2.5cm margins.
- The pdf file length is limited to 4 pages maximum.
- The assignments must be sent via the on-line campus Moodle.
- The name of the compressed file folder must contain authors' names. For example, if the author is a single student:

  `Assignment1_FamilyName1_FamilyName2_Name`

- Deadline: **20 March 2020**.