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?

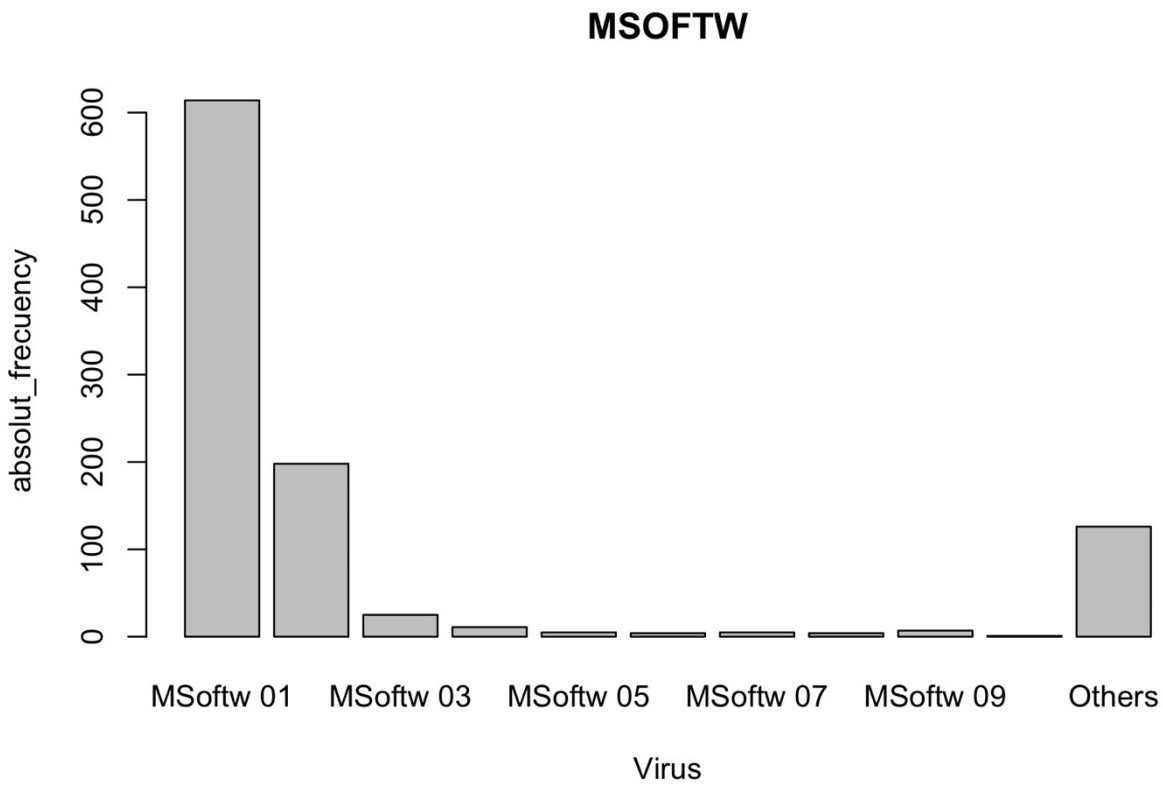
After importing the dataset as exercise1 and executing the command length(exercise1$virus), we can see that the sample size of this dataset is 1000 elements.

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

```
> freq(exercise1$virus,cumul=FALSE)
Frequencies
exercise1$virus
Type: Factor

                Freq   % Valid   % Total
-------------- ------ --------- ---------
     MSoftw 01    614     61.40     61.40
     MSoftw 02    198     19.80     19.80
     MSoftw 03     25      2.50      2.50
     MSoftw 04     11      1.10      1.10
     MSoftw 05      5      0.50      0.50
     MSoftw 06      4      0.40      0.40
     MSoftw 07      5      0.50      0.50
     MSoftw 08      4      0.40      0.40
     MSoftw 09      7      0.70      0.70
     MSoftw 10      1      0.10      0.10
        Others    126     12.60     12.60
         <NA>      0                0.00
         Total  1000    100.00    100.00
```

(c) Display the bar graph of this dataset.



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

In the frequency table (b) we can seen in the column Freq the absolute frequency of each virus of the dataset, if we sum all the values we get the sample size, that is 1000. Then, in the % Valid and % Total columns, the percentage frequency of each virus is shown, if we sum all the values we get 100%, as expected.

In the bar graph (c) the absolute frequency is represented in the Y axis, and each type of virus in the X axis. We can see that there are 11 bars, each one corresponding to one of the viruses of our dataset.

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

If we look in the table provided by the exercise we can see that the I-Worm.Tanatos virus is the MSoftw 03 in our datasheet, so we can conclude that the number of computers infected with this virus were 25 computers, as this is the frequency shown in the table (b).

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

The most frequent one of this sample is the Msoftw 01 or I-Worm.Klez this is the malicious software that is repeated more times so also we can conclude that this is the mode of this sample.

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

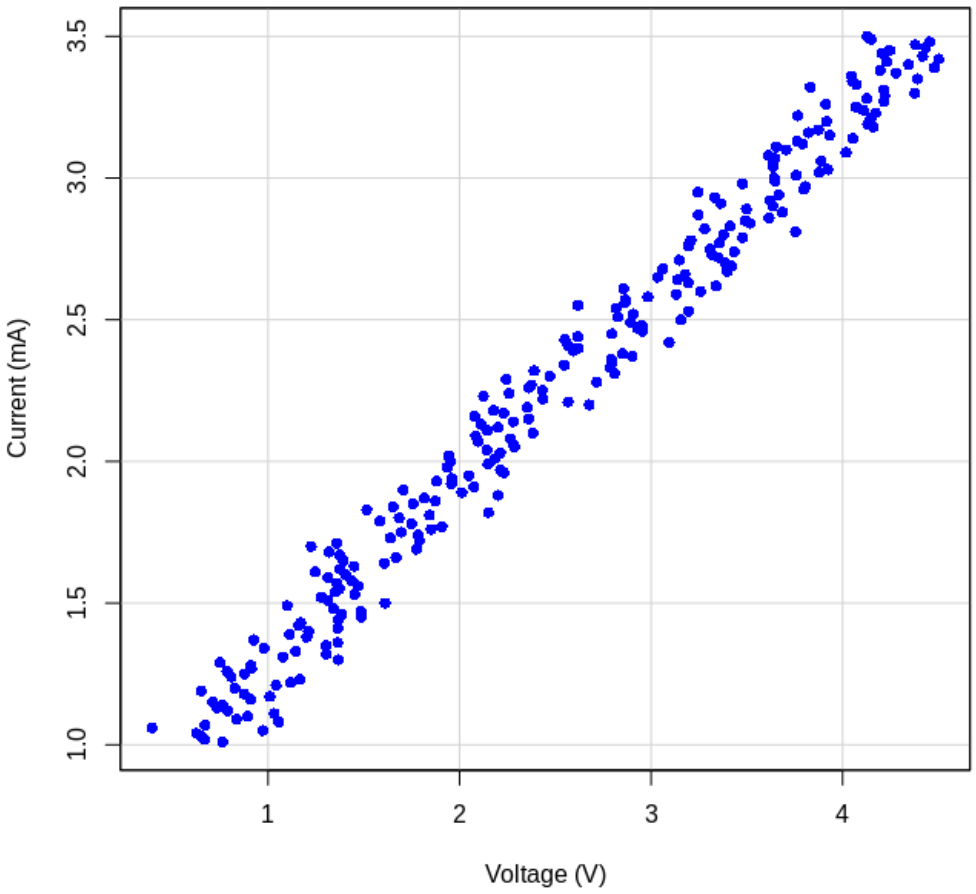
If we look to our frequency table we can see that it corresponds to the 0,5% of the total.

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

If we look to our frequency table we can see that it corresponds to the 0,5% of the total.

Exercise 2. An article in the *IEEE Transactions on Instrumentation and Measure- ment* ["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 volt- age 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.



We can observe a linear relationship between voltage and current, as the changes in voltage are associated with changes in current in a linear way.

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

```
> cor(exercise2$x,exercise2$y)
[1] 0.9895202
```

We can see that the correlation coefficient is nearly equal to 1, that means that it exists a strong positive linear relationship between x and y.

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

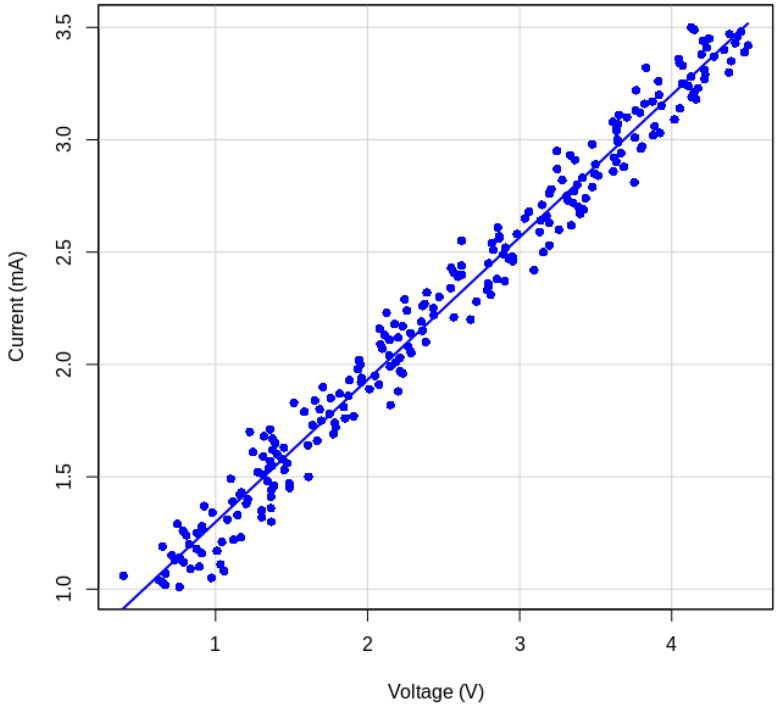
```
Call:
lm(formula = x ~ y, data = exercise2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.253858 -0.078789 -0.000671  0.080117  0.258286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.665114   0.016150   41.18   <2e-16 ***
y           0.633700   0.005872  107.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1046 on 248 degrees of freedom
Multiple R-squared:  0.9792,    Adjusted R-squared:  0.9791
F-statistic: 1.165e+04 on 1 and 248 DF,  p-value: < 2.2e-16
```

The regression line is given by the following equation y=0,6337*x+0,665114



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

The coefficient of determination is given by the square of the correlation coefficient, thus $0{,}9895202^2 = 0.9791502$.

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

We can be very confident saying that the accuracy of the predictions made by using the regression line is pretty accurate. We can conclude that because the coefficient of determination is very near to one, thus the proportion of variability of the Voltage that can be “explained” by a linear relationship between Current and Voltage is almost 1.