DISI – University of Trento

Master in Computer Science AA 2016/2017

Simulation and Performance Evaluation

Assignment 1

Characterization of a population parameters

Renato Lo Cigno, Michele Segata

October 17, 2016

An electronic probe measures the output of a circuitry, which is known to generate i.i.d. samples y_i taken from a population with a logistic distribution:

$$f_Y(y) = \frac{e^{-\frac{y-\mu}{s}}}{s(1+e^{-\frac{y-\mu}{s}})^2}$$

Probing is done with a frequency of 20 kHz, and the measure lasts 3 s, so that the sample contains 60,000 points.

Unfortunately the probe is affected by noise and it also introduces a sinusoidal bias, so that the actual measured samples can be described as

$$x_i = A\sin(2\pi ft) + Y + Z \quad ; \quad i = 1, \ldots 60,000$$

where Y is the random variable describing the population while Z is Gaussian noise with zero mean and standard deviation σ_n. To be more formal:

$$Y \sim Logis(\mu, s), \quad Z \sim N(0, \sigma_n). \tag{1}$$

Remember that, for the logistic distribution, the relation between variance and s is

$$\sigma_Y^2 = \frac{s^2\pi^2}{3}. \tag{2}$$

All we know of the sinusoidal disturbance is that its frequency is low (in the order of tens of Hz) and that its value is a natural number (it might be 10 Hz, 13 Hz, but not 17.429 Hz).

The goal is to find estimates of the five parameters A, f, μ, s, σ_n and to estimate the confidence interval of $\hat{\mu}$, which is the most important parameter for the characterization of the circuitry. For the sake of interpretation, you can imagine that all circuits that have μ below a certain threshold are discarded, while the others are kept and sold for their purposes, so knowing the confidence interval is fundamental to build the commercial agreement to sell them.

Fig. 1 reports a snapshots with 6,000 points from one measure. It is clear that "visual inspection" helps very little in this case. Each of you will work on a different dataset and each solution is different. Download the zip file from Classroom (roughly 14 MB) and pick your dataset[1]. The dataset (named like `yourlastname.csv`) includes a single row with all the 60,000 points.

Some hints on how to solve the assignment:

- As far as we know, there exists no `findAmplitudeandFrequencyandAllTheRest` function in statistical tools, neither in R nor in other tools, so don't waste time searching for it;

[1] We had to make a single file to avoid Classroom limitations.

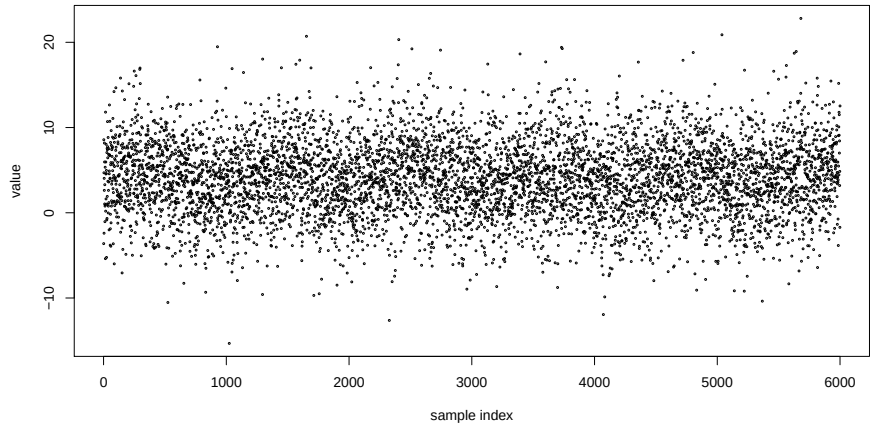

Figure 1: Plot of the initial data (6,000 points x_i) of a single measure.

- Reason on what functions of the data you can compute, and do this in light of the properties and the structure of the measured points x_i;
- Try to understand if there is a parameter you can compute in a straightforward way, meaning that it is not "coupled" to the other in manipulations you can do on data (this is very, very easy!), and start computing it;
- Think what you can do manipulating the data once you know the parameter above, and what is the next parameter that can be estimated once you "clean" your data using the knowledge you already have. In this phase work on paper, write down equations and the functions of the random variables that you can measure on the dataset, if you use all the theoric tools we described in class, you can end up with a very simple solution where each parameter can be derived by a single measure plus the knowledge of the parameters already estimated;
- Repeat for every parameter;

To write your report use the LaTeX template we give you and do not write more than 3 pages. Deliver the PDF file of the report and the R, Matlab, or python script you used for processing and plotting as a single .zip or .tar file through Classroom; if the script does not run on a standard Linux box, we simply notify you that it does not work, and we will not attempt correction. Keep your code **CLEAN, ORGANIZED, and COMMENTED**. Do not send us your source code with unused portions commented out, blocks of code with no comments, or with monolithic pieces of code. Split your code in functions. Do not include the data set in the zip file, we already have it! The script MUST work assuming that the dataset is in the same folder of the script. DO NOT use absolute folders like

```
ds <- read.csv('/home/john.doe/Documents/spe/doe.csv')
```

but rather

```
ds <- read.csv('./doe.csv')
```

The deadline to have a correct-and-redo chance for this assignment is December 14 . . . 2016!!. If you deliver the assignment within this date, we will correct it and give you the chance to refine it before the Christmas vacations, otherwise we will consider the work "as is" before the oral discussion is agreed upon. However, if the quality of the delivery is unacceptable (e.g., no methodology is described, plots are meaningless and not explained, etc.) we will not correct it, but simply reject it, so you lose the privilege of a pre-correction.

If you have some doubts, just write us an email or ask in class.

Have Fun!