UNIVERSITÉ DE FRANCHE-COMTÉ

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

STATISTICAL EXPLOITATION OF MEASUREMENTS

Practical exercises

MASTER

FUNDAMENTAL PHYSICS AND APPLICATIONS

Pr Fabrice DEVAUX

fabrice.devaux@univ-fcomte.fr

TABLE OF CONTENTS

GENERAL COMMENTS

The exercises directly connected to the master lectures. The work on these exercises must be included in a unique report whose evaluation will be a part of the individual final rating. All programs and results must be included in this report, that should also be written to show your conclusions.

The simplest way is to send this report by e-mail with the corresponding executable programs. A paper version of the report is also possible.

The recommended languages to handle these problems are Matlab or Octave. Note that Octave is a free software you may load at home, whose syntax is largely compatible with Matlab.

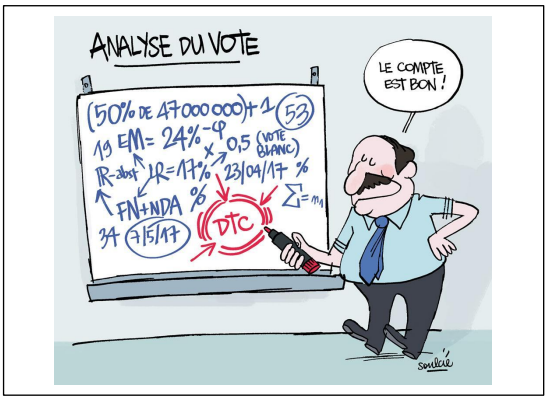
It is also possible to use Python, if you are autonomous with this language, with no need of assistance of the teacher.

TOPIC 1:
APPLICATION OF ESTIMATION TOOLS

1.1 UNCERTAINCY OF A POLITICAL POLL

1.1.1 Problem

A political journalist wrote a deep analysis about the reasons which leaded to a decrease of the popularity of Mr Dupont from 52% to 48%. The purpose of this exercise is to judge the interest of analyzing this variation of voting intention...



A pool in France is usually conducted by asking 1000 people. We admit as hypothesis that the true popularity of Mr Dupont is 50%. The goal of this exercise is to simulate a great number of polls, in order to assess the precision of a poll conducted on 1000 people.

1.1.2 Work to be done

- *Using an appropriate random number generator, build a program generating a sequence of 1000 random numbers with two possible equally probable values: 0 and 1 corresponding respectively to the answer given to the poll question, "Would you vote for Mr. Dupont?": NO (0), YES (1). How many 1 are obtained?*

- *Repeat at least 500 times this generation and plot the histogram[*] of the results. Do not forget the label of the axes and the title of the figure*
- *Order the results (use the SORT function) and plot the cumulative probability function of the 500 polls in order to find a confidence interval at 95% for the results. Conclusion?*
- *Compare with the theoretical result (see the exercises on estimation), are the results coherent?*
- *Repeat all the steps for a poll conducted on 4000 people. What conclusion do you draw about the accuracy of a poll?*

1.2 MEASUREMENT ERRORS AND AVERAGE

1.2.1 Problem

A period of one second is measured with an uncertainty of ± 0.02 seconds. The measurement is repeated 100 times. The purpose of this exercise is to estimate the uncertainty on the arithmetical average of 100 measurements. The random measurement error is supposed to be centered (null mean), Gaussian, independent from a measurement to another.

1.2.2 Work to be done

1. *If the uncertainty corresponds to a 95% confidence interval, which value has the standard deviation on the measurements?*
2. *Verify: use randn to "make" 1000 measurements with this standard deviation. Show the histogram and compare it with a normal distribution of the same standard deviation as the measurements. Using the cumulative distribution function of the data verify that uncertainty on measurements corresponds to the 95% confidence interval?*
3. *"Make" 100 measurements. What is the value of the arithmetical average (mean in matlab)? Repeat 100 measurements. Is the arithmetical average the same?*
4. *The arithmetical average is, just as the measurements, a random variable. To estimate its standard deviation, make 1000 arithmetical averages on 1000 series of 100 measurements. Show the histogram, estimate the standard deviation (std function) and the confidence interval from these 1000 averages. Conclusion?*

1.3 REPORT AND RECOMMENDATIONS

In your report, you must write down all the requested graphs (with axis labels, legends of the different curves and title) and accompany them with the requested

[*] You can use HIST function of Matlab (or Octave). Read carefully the syntax of the function to use it in an optimal way

comments. Any additional information or graphs that you deem relevant to include in your report will be appreciated.

Matlab or Octave functions useful for these exercices (non exhaustive list): rand, hist, sort, <, <=, mean, std, randn, ones, plot, hold on, hold off. For more, type help followed by the name of the function and make trials. For example, what is the result of 2<3?

TOPIC 2:
$\chi^{(2)}$ TEST AND NONLINEAR FITTING OF PARIS-ALGER FLY DATA

2.1 $\chi^{(2)}$ TEST

2.1.1 Background

Two solutions are possible to perform a $\chi^{(2)}$ test:

1. *With Matlab use the standard function **chi2gof** on a vector of data (one element of the vector per data). Useful Matlab functions: **mean**, **std**, **cdf**, **icdf**, **cumsum***
2. *With other languages (Octave or Python), write a program which makes the χ^2 test as it is described in the Chapter 4 of the lecture (p 26-27)*

2.1.2 Problem

We want to use $\chi^{(2)}$ test on the data of the Paris-Alger fly. Data are given in the table 2.1.

T	1.90 - 1.95	1.95 - 2.00	2.00 - 2.05	2.05 - 2.10	2.10 - 2.15	2.15 - 2.20	2.20 - 2.25	2.25 - 2.30	2.30 - 2.35	2.35 - 2.40	2.40 - 2.45	2.45 - 2.50	2.50 - 2.55
N	19	19	39	48	87	94	104	92	57	44	28	26	13

Table 2.1: Number of flights (N) as a function of time slots (T) given in hours and hundredths of hours

2.1.3 Work to be done

1. *Built the vector of data corresponding to the duration of each flight. Duration of each flight will be the average time of a time slot. Calculate the mean and the standard deviation of data.*
2. *Plot the histogram and the cumulative distribution of data using time slots as bins. Compare these graphs with distribution and cumulative distribution function of a normal law of equivalent mean and standard deviation.*
3. *Compute D^2 of data.*
4. *What is the degree of freedom of the data? Make the $\chi^{(2)}$ test on data and say if data are compatible with the gaussian hypothesis at a confidence of 95%.*

2.2 NONLINEAR FIT OF DATA

2.2.1 Background

Two solutions are possible to perform nonlinear fit of data:

1. *With Matlab apply the standard function **nlinfit** on a vector of data to adjust the parameters of the model you have identified.*
2. *With other languages (Octave or Python), write a program which makes the Gauss-Newton method as it is described in the Chapter 5 of the lecture.*

2.2.2 Problem

We want to fit the data given in table 2.1 with a Gaussian function by optimizing its following parameters: amplitude, standard deviation and offset.

2.2.3 Work to be done

1. *Write a program implementing the Gauss-Newton method or use NLINFIT function (with Matlab).*
2. *Calculate the uncertainty of the optimized parameters and compare them to the initial parameters.*
3. *Draw the relevant curves representing the different results obtained. Comment on your results with respect to the χ^2 test performed on the data in section 2.1.*

2.3 REPORT AND RECOMMENDATIONS

In your report, you must write down all the requested graphs (with axis labels, legends of the different curves and title) and accompany them with the requested

comments. Any additional information or graphs that you deem relevant to include in your report will be appreciated.

TOPIC 3:
APPLICATION OF PRINCIPAL COMPONENT ANALYSIS (PCA) TO THE ANALYSIS OF SPORTS RESULTS

Tokyo 2021 Olympic Games decathlon podium: Damian Warner (gold medal); Kevin Mayer (silver medal); Ashley Moloney (bronze medal).

PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. We propose here to apply this method to the analysis of the sports results of the Decathlon at the Olympic Games in Tokyo in 2021.

3.1 DATA

The score of each competitor obtained in the different events are grouped in the following table (Table 3.1). The data of this table are contained in the Matlab or Octave data file *"DataDecathlonTokyo2021.mat"* which is downloadable on Moodle.

Ranking	Athlete	1500 m	100 m	400 m	110 m hurdles	Long jump	Shot put	High jump	discus throw	Pole vault	javelin throw
1	Warner	737	1066	934	1045	1123	777	822	843	880	790
2	Mayer	660	933	800	987	935	794	878	830	972	937
3	Moloney	685	1013	994	964	970	758	906	754	910	695
4	Scantling	709	935	897	971	886	826	794	776	941	876
5	Lepage	733	992	962	925	972	809	794	811	910	696
6	Ziemek	690	963	858	910	862	789	850	764	1004	744
7	Victor	593	935	851	870	871	814	822	865	880	913
8	Shkurenyov	715	876	862	920	957	787	794	809	941	752
9	Urena	759	938	909	958	886	727	850	740	880	675
10	Bastien	765	931	927	921	908	753	850	679	790	711
11	Erm	755	852	897	905	900	765	794	782	849	714
12	Wiesiolek	745	899	898	856	878	784	822	834	849	612
13	Zhuk	664	852	851	856	797	865	767	808	941	730
14	Kazmirek	629	841	901	882	930	757	822	720	849	795
15	Uibo	689	791	777	870	903	725	822	795	1067	598
16	Helcelet	651	847	842	930	852	789	767	775	790	761
17	Sykora	589	821	866	913	821	767	714	868	790	794
18	Dos Santos	604	956	847	901	905	736	822	663	790	656
19	Roe	633	892	772	794	821	727	767	836	849	772

Table 3.1: Results and ranking of athletes in the decathlon event of the Olympic Games of Tokyo 2021

3.2 METHODOLOGY

The method to be used is the one exposed in chapter 3 of the lecture. The sports events are the variables ($N = 10$) and the athletes' scores are the measures ($K = 19$) of these variables. Following the methodology explained in class, build a Matlab or Python program to calculate successively the following matrices:

- *The matrix of data:* M
- *The covariance matrix of data:* C
- *The correlation coefficient matrix :* R
- *The Karhunen-Loève transformed matrix of data:* Y
- *The* $N \times N$ *rotation matrix between orthonormal bases in the N-dimensional space:* Φ

In order to analyze these data, you must first reduce their dimensionality from $N = 10$ to $L = 3$ by isolating the 3 main components of the matrices Φ and Y. From there, you can start analyzing the data.

3.3 ANALYSIS OF DATA

By analyzing the available data answer the following questions.

1. Calculate the truncated data matrix M_{trunc} *and compare it with the original data matrix. Comment on your observations.*
2. Estimate the ratio of information losses when only 3 principal components are considered.
3. Plot the 3D representation[†] *of the correlation coefficients matrix R, and identify events with significant correlations.*
4. From the 3 main components of the matrix Φ*, plot a 3D graph*[‡] *of the events in the 3-dimensional space. Comment on your observations. Is it consistent with the correlation coefficients? Give meaning of positive or negative value of each of the 3 principal components.*
5. From the 3 main components of the matrix Φ*, plot a 3D graph*[§] *of the events in the 3-dimensional space. Comment on your observations. Is it consistent with the correlation coefficients?*
6. From the 3 main components of the matrix Y*, plot a 3D graph of the athletes in the 3-dimensional space. Comment on your observations. Is it consistent with the results given in table 1?*

[†] With Matlab, use ***bar3*** function and use the event's name to use the name of the sport as a scale for the axes.
[‡] With Matlab, use biplot function and use the event's names to label each point of the graph
[§] With Matlab, use biplot function and use the event's names to label each point of the graph

3.4 REPORT AND RECOMMENDATIONS

In your report, you must write down all the requested graphs and accompany them with the requested comments. Any additional information or graphs that you deem relevant to include in your report will be appreciated.