The University of Franche-Comté

Master CompuPhys - Statistical exploitation of measurements

Application of estimation tools

Chiari Even
Legrand Maxime
M1 Compuphys 2022

Supervisor : Fabrice Devaux

October 24^{th} 2022

Contents

1 Introduction

The estimation tools are very useful in physics and other domains, in this practical work we'll be using multiple tools we have learned during the SEM lectures, such as :

- The normal law
- Cumulative probability function
- The confidence interval notion
- Mean and standard deviation values related to a law

2 Requirements and limits

The algorithm for this practical work was coded under *Python 3.8* and may not work properly with older versions. You may also consider having the following libraries installed :

- *numpy*
- *matplotlib*
- *scipy*

The algorithm for the first part was made for a probability of 0.5 and may not properly display histograms for other probabilities. Consider removing the $range$ argument inside the $hist$ function to avoid this issue.

3 Uncertainty of a political poll

We have to simulate a huge amount of polls in order to verify the accuracy of a poll. The probability of a vote in favor of Mr.Dupont is 0.5. It means each trial can be assimilated to a Bernoulli experiment. The results can be one or zero (one if the vote is yes):

- First we have to generate 1000 random values of zero and one.
- We will repeat this 500 times in a second time to plot an histogram.
- We will use the cumulative probability function (CDF) to visualize the 95 percent confidence interval.
- Finally we will change the amount of votes and see how it influences our results.

3.1 Case of a poll conducted on 1000 people

By generating 1000 random votes with a probability of 0.5 each, we have obtained 505 numbers of "yes" votes. It seems logical because we already demonstrated through the CLT that a huge amount of Bernoulli experiments tends to a Gaussian law of mean Nm. Now we want to repeat the 1000 votes 500 times and visualize the results through an histogram.

In order to justify if the poll is trustworthy or not, a good approach would be to plot the CDF associated to our configuration and to see if the affirmation *"a decrease of the popularity of Mr Dupont from 52% to 48%"* fits in the 95% confidence interval.

The CDF is obtained by sorting the amount of "yes" votes in an increasing order, and by summing the associated probabilities at each step.

Below you'll find the results obtained for the following configuration :

- 1000 people voting
- 500 number of repetitions
- 0.5 probability of a "yes" vote

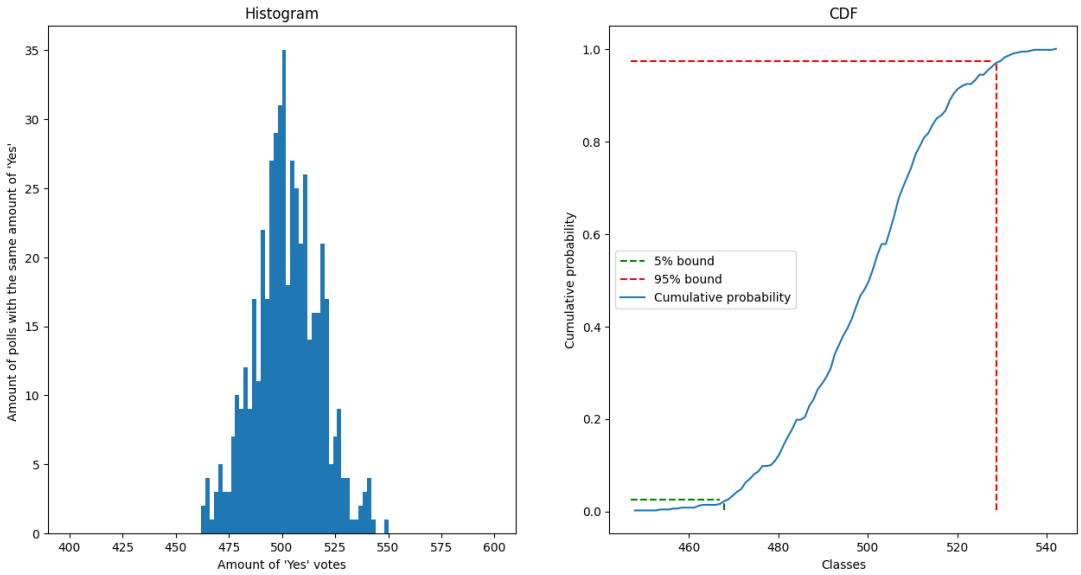

Figure 1: Histogram and CDF for 1000 votes repeated 500 times

52% corresponds to 520 "yes" votes, and 48% to 480. Both are in the 95% confidence interval so we can conclude that for 1000 people voting, the poll is trustworthy.

3.2 Case of 4000 voters

Now we want determine the same thing for 4000 people voting and 500 number of repetitions, by repeating the steps described in 3.1, we obtain the following results :

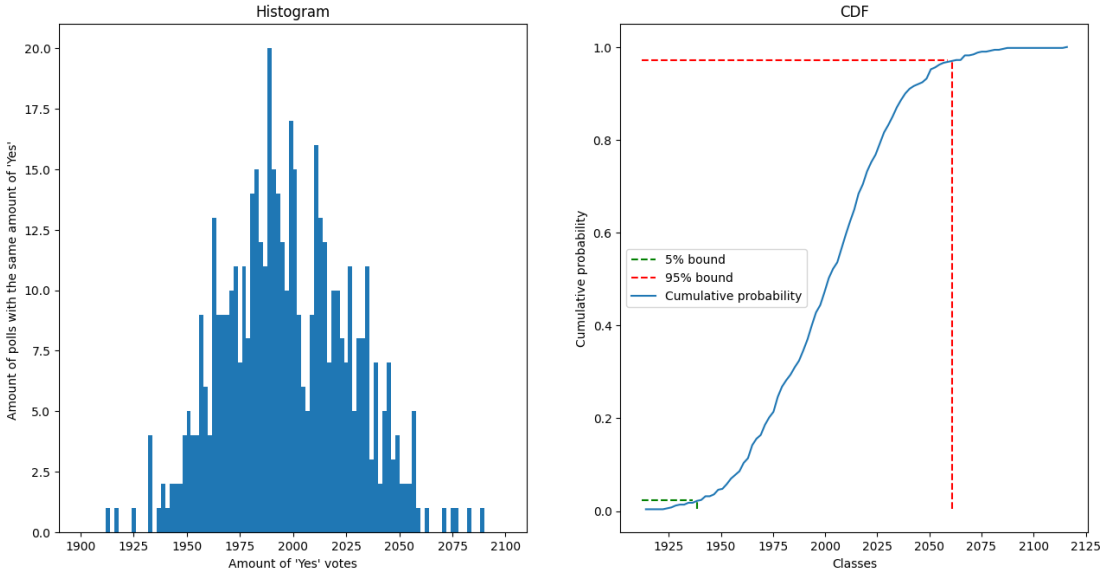

Figure 2: Histogram and CDF for 4000 votes repeated 500 times

We can see that the histogram is more spread out than the previous one for 4000 people voting, this is expected because with more people voting we can have more various numbers of "yes" votes.
Concerning the cumulative probability, for 52% we have 2080 "yes" votes and for 48% we have 1920. We can now conclude that the poll isn't trustworthy anymore due to the fact we are out of the 95% confidence interval.

In a more general case, the everyday life polls are defined in a way not to be contradicted with the safe argument *"It's on 1000 people"*, thus the demonstration above shows a really disturbing phenomenon.

4 Measurement errors and average

In this part we will study how a measurement error evolves regarding the amount of times it's performed. The studied case is very concrete and is almost always leaded in industrial processes : *"A period of one second is measured with an uncertainty of ± 0.02 seconds. The measurement is repeated 100 times"*. Thus we will be estimating the following behaviors :

- The influence of standard deviation
- If our uncertainly is in the confidence interval when we compare the data with the expected theoretical values
- The standard deviation of the mean of our data when we perform 1000 times 100 measurements

4.1 Uncertainty on 1000 measurements

First one needs to determine the value of the standard deviation on 1000 measurements. This can be obtained through the following relation :

$$\begin{aligned} \mu_{exp} \pm 0.02 &= \mu_{theo} \pm 1.96\sigma_{theo} \\ \sigma_{theo} &= 0.02/1.96 \end{aligned}$$

With $\mu_{theo} = 1$ the theoretical mean.

Below you'll find the histogram and CDF graphs for 1000 measurements with the σ_{theo} STD, the mean of 1, and the Normal law distribution fit with respect to the mean μ_{theo} and σ_{theo}.

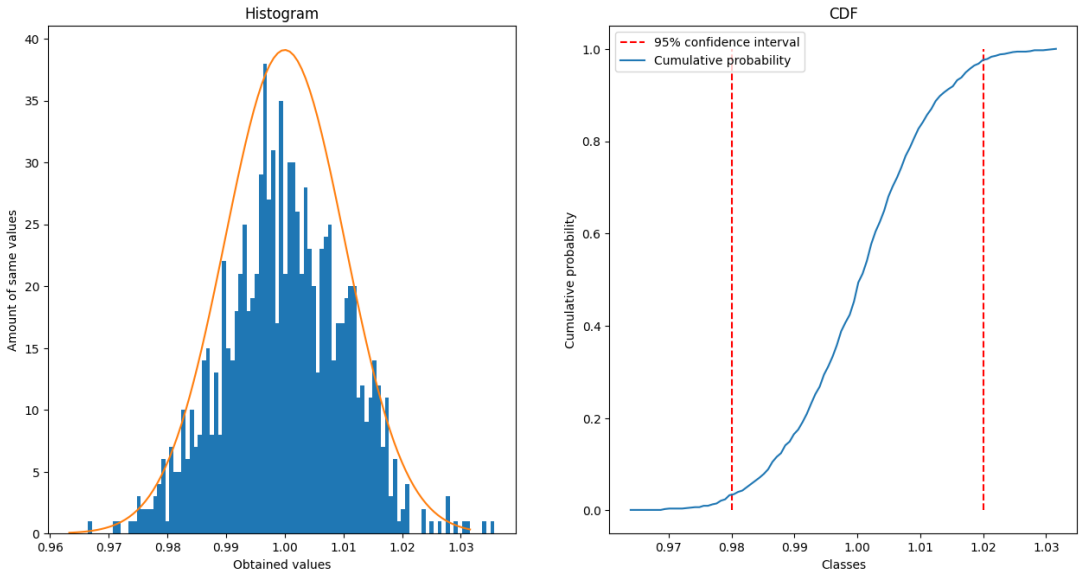

Figure 3: Histogram and CDF for 1000 measurements

By printing the experimental and the theoretical values for the mean and STD, one can notice that they agree :

```
experimental  0.979183396645173  (std)  1.0188903772002977  (mean)
theoretical   0.9897959183673469 (std)  1.010204081632653   (mean)
```

Note : These values tend to be the same for an infinite amount of measurements.

4.2 Uncertainty on the average of measurements

For this part, we're interested in making 100 measurements which we'll repeat 1000 times and compute the mean for each series. One can first wonder about the nature of the obtained mean. By "making" 100 measurements twice, we obtain the following values :

```
1st mean :     0.9997870590675716
2nd mean :     1.0002390978377904
```

We can notice that the mean isn't neither the same for the two trials, nor the expected value of 1. It comes from the fact that the measurements are randomized which also makes their average random.

4.3 Uncertainty on the STD of measurements

Now that we know that the average is also a random variable, let's get into the behavior of the STD of these averages. With the same configuration than in 4.2, one can plot the histogram of the obtained averages, and then the CDF associated to them :

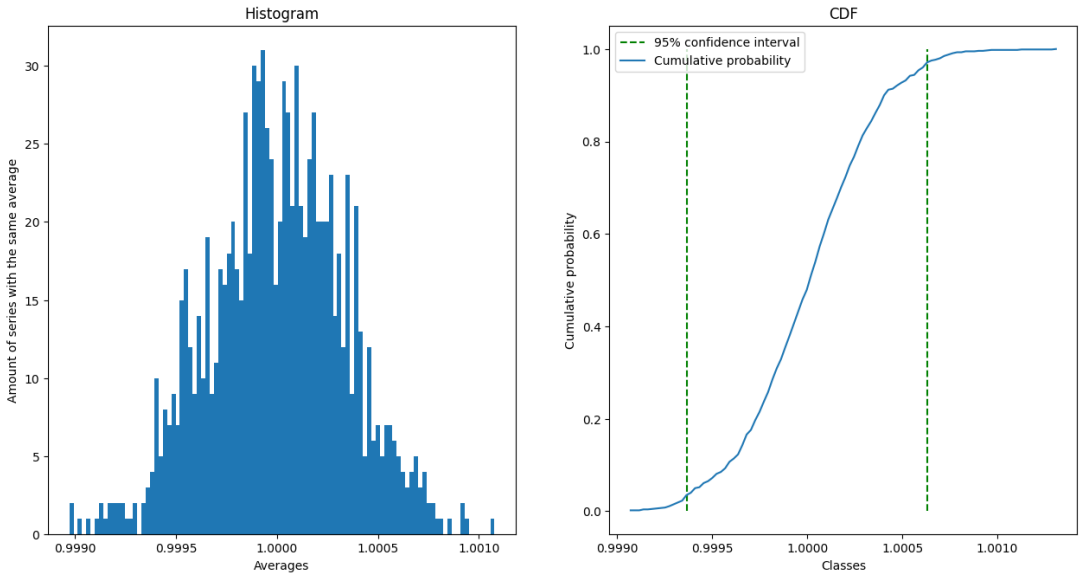

Figure 4: Histogram and CDF of averages for 1000 repetitions of 100 measurements

The STD associated to these measurements is $\sigma_{mes} = 0.0003221$ for a theoretical STD $\sigma_{theo} = 0.0003227$. We can thus define the 95% confidence interval such as :

$$\begin{aligned} \mu - 1.96 * \frac{\sigma_{theo}}{\sqrt{N_{mes}}} \leq \quad 1 \quad \leq \mu + 1.96 * \frac{\sigma_{theo}}{\sqrt{N_{mes}}} \\ 0.9993 \leq \quad 1 \quad \leq 1.0006 \end{aligned}$$

We can conclude that the more the measurements are repeated, the lower will be the confidence interval as it evolves with $1/\sqrt{N_{mes}}$. In addition to this, the obtained experimental STD draws an interval contained in the 95% one, so the measurement process is trustworthy.