For students, teachers and enthusiasts

Probability, Statistics and Estimation

Propagation of Uncertainties in Experimental Measurement

Population and Sample

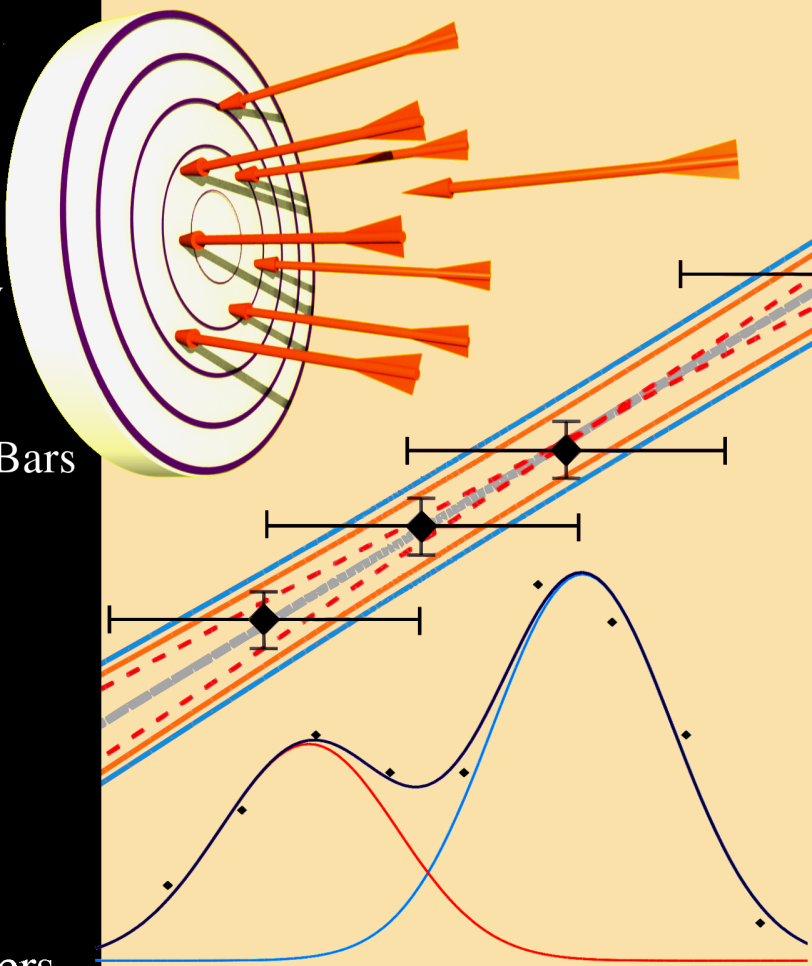
Sources of Uncertainty

Regression with Error Bars

Estimators

...

Exercises with Answers



Mathieu ROUAUD

Probability, Statistics and Estimation

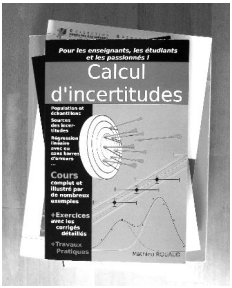
Propagation of Uncertainties in Experimental Measurement

Mathieu ROUAUD
Physics and Chemistry Lecturer in Preparatory Classes for the Engineering Schools. Associate Professor.
MSc Degree in Theoretical Physics.
Phd Student in Particle Physics.

To share the knowledge and provide access to the greatest number of reader, the book license is free, the digital book is free and to minimize the cost of the paper version, it is printed in black and white and on economic paper.

Complete Digital and Paper Books (June 2017)
with all the exercises corrected
on www.lulu.com

To contact the author:
ecrire@incertitudes.fr
Boudiguen 29310 Querrien
France



Date of publication: July 2013 Revision and translation: April 2017

French books: *Calcul d'incertitudes* and *Probabilités, statistiques et analyses multicritères* .

Foreword

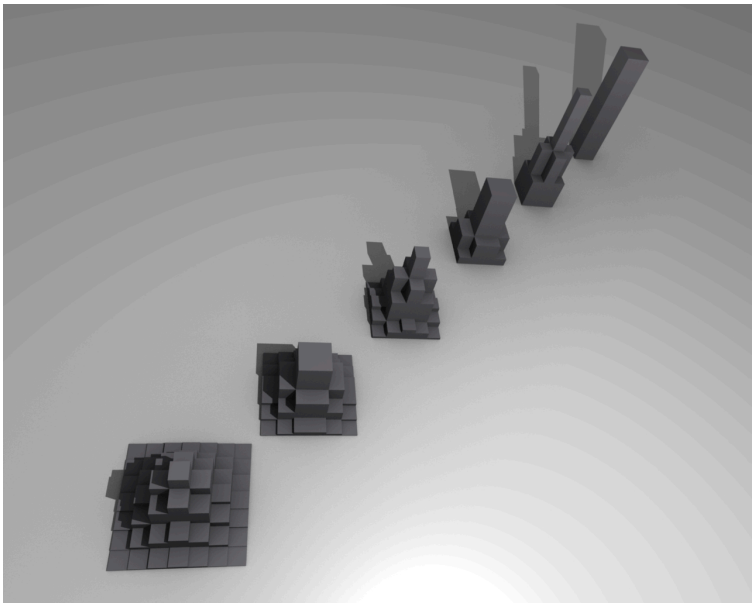
This handbook aims to be accessible to the major public. This work is meant to be used as a pedagogical tool and as a reference book. The following book represents personal reflections on the probabilistic nature of measures in science. In a classical curriculum these aspects are seldom, if not at all, dealt with. It is important that the experimental and practical foundations of science are complementary to the theoretical lectures. There is a scientific beauty that arises from the interaction between theory and experience.

While introducing the fundamental principles of statistics, this book explains how to determine uncertainties in different experimental situations. Many examples come from courses and practical work done in preparatory classes for the engineering schools.

I hope you enjoy reading!

Thanks to the readers who by their questions, comments and constructive criticisms make it possible to improve the book.

Thanks to the life and to all those who have come before me.



Contents

I. RANDOM VARIABLE

A. How to measure a quantity ?

In what follows, X is a random variable and $\{x_i\}$ a sample of n outcomes.

If we ask how many days are in a week, it is easy to answer. Now, consider different groups of students who have measured the thermal capacity of water[1] and have obtained the following values : {5100; 4230; 3750; 4560; 3980} J/K/kg. How can we estimate the water capacity in this situation? The answer will use a probabilistic approach.

B. The center of a distribution

How does one determine the most representative value of a sample? There are different ways to define the center of a distribution. For example, we have the mode, the median and the mean. The mode is the value which occurs with the greatest frequency in a data set. The median is the middle of a distribution that divides the sample into two parts whereby each half has an equal number of observations. The most commonly used measure of the

1 PHYSIQUE : The amount of energy needed to raise the temperature of one kilogram of mass by 1°C. That way, the water stores energy and can then return it by decreasing its temperature. Tables : c_{water} = 4180 Joules per Celsius degree and per kilogram.

center is the mean. The mean is the sum of the observed values divided by the number of observations :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_i + \ldots + x_n}{n} \quad \text{also} \quad \boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$ [2]

For the thermal capacity of water we have :

$$\bar{c} = \frac{5100 + 4230 + 3750 + 4560 + 3980}{5} = 4324\, J/K/kg$$

We have considered the arithmetic mean. We could have used the geometric mean:

$$\bar{x} = \sqrt[n]{\prod x_i}$$

For instance for two speeds of 20 m/s and 40 m/s, the geometric mean is $\sqrt{20\,m/s \cdot 40\,m/s} \simeq 28.3\,m/s$ whereas the arithmetic mean is of 30 m/s. The arithmetic mean is used more often globally due to its conveniently simpler calculation.

2 MATH : To simplify the writing of a sum, the Greek letter sigma is used as a shorthand and read as "the sum of all x i with i ranging from 1 to n".

C. The dispersion of a distribution

In addition to locating the center of the observed values we want to evaluate the extent of variation around the center. Two data sets may have the same mean but may be different with respect to variability. There are several ways to measure the spread of data. Firstly the range is the difference between the maximum and minimum values in the sample. The sample range of the variable is very easy to compute, however it is sensitive to extreme values that can be unrepresentative.

The sample standard deviation is preferred:

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

It is the most frequently used measure of variability.

For the thermal capacity of water we have :

$$s_c=\sqrt{\frac{(5100-4324)^2+(4230-4324)^2+(3750-4324)^2+(4560-4324)^2+(3980-4324)^2}{4}}$$

and $s_c \simeq 530\, J/K/kg$

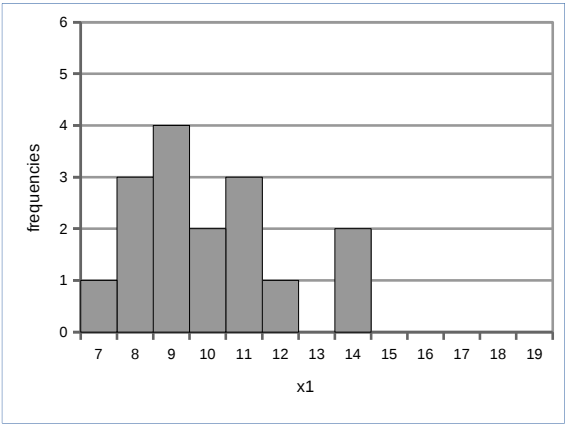
The mean deviation may also be used (see Exercise 1).

Using the standard deviation formula, dividing by n rather than $n-1$, will obtain the root mean square deviation (square root of average square deviation). The choice of the standard deviation is justified on page 130. Besides, the values of n are often large and the difference is small.

D. Examples of distributions

Case 1 :

	x_1
x_1^1	11
x_1^2	9
x_1^3	10
x_1^4	14
x_1^5	11
x_1^6	8
x_1^7	9
x_1^8	12
x_1^9	7
x_1^{10}	8
x_1^{11}	8
x_1^{12}	9
x_1^{13}	11
x_1^{14}	14
x_1^{15}	10
x_1^{16}	9



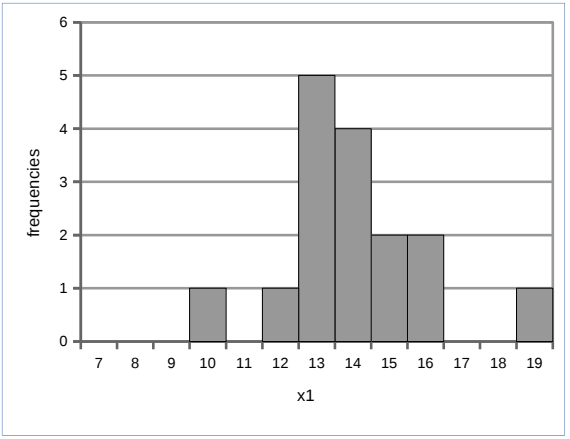
mean = 10 **standard deviation= 2.07**

mode= 9 range= 7

median= 9.5 root mean square deviation= 2.00

Case 2 :

	x_1
x_1^1	15
x_1^2	13
x_1^3	12
x_1^4	13
x_1^5	14
x_1^6	13
x_1^7	16
x_1^8	19
x_1^9	13
x_1^{10}	14
x_1^{11}	10
x_1^{12}	16
x_1^{13}	14
x_1^{14}	15
x_1^{15}	13
x_1^{16}	14



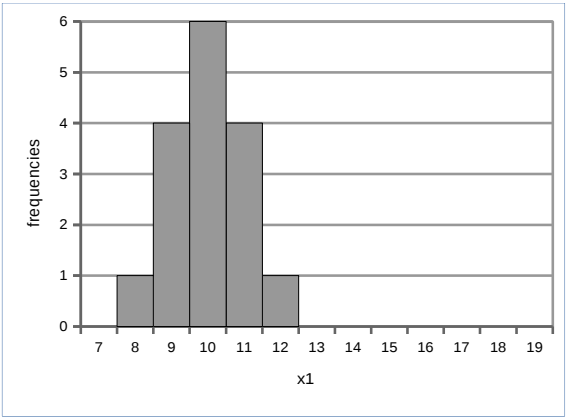
mean = 14 **standard deviation= 2.00**

mode= 13 range= 9

median= 14 root mean square deviation= 1.94

Case 3 :

	x_1
x_1^1	10
x_1^2	10
x_1^3	12
x_1^4	11
x_1^5	9
x_1^6	8
x_1^7	10
x_1^8	9
x_1^9	9
x_1^{10}	11
x_1^{11}	9
x_1^{12}	11
x_1^{13}	10
x_1^{14}	10
x_1^{15}	11
x_1^{16}	10



mean = 10 **standard deviation= 1.03**

mode= 10 range= 4

median= 10 root mean square deviation= 1.00

The mean is not always the most represented value (*case 1 and 2*) and in some cases does not appear at all. In *case 3* the histogram is symmetrical and illustrating that the median and mean are equal.

In the event that some values are represented several times, we determine the *frequency* f_i for each value x_i.

We have $n=\sum_{i=1}^{c} f_i$, where c gives the number of different values of x_i.

The mean and standard deviation then become:

$$\bar{x}=\frac{\sum_{i=1}^{c} f_i \cdot x_i}{n}=\sum_{i=1}^{c} \frac{f_i}{n} \cdot x_i \qquad s=\sqrt{\frac{\sum_{i=1}^{c} f_i (x_i-\bar{x})^2}{n-1}}$$

Sometimes the data can be grouped into class intervals. For example, if we measure the size of the inhabitants of a city, we can group all the inhabitants with a size between 160 cm and 170 cm within the same class interval. The number of observations in this class is the frequency and the middle of this class interval is the value assigned, here 165 cm (see Exercise 5).

The more the histogram is concentrated around the center, the more the standard deviation is small.

E. Central limit theorem

1) Population and samples

Consider a city of one million inhabitants. To survey the population we can interview a sample of only one thousand people drawn at random. Thanks to the statistical tools, from this $n=1000$ individuals sample, we can have information on the whole population. The larger the sample size is, the more accurate the results will be. Let $\bar{x}$ be the sample mean and s be the sample standard deviation. For the population we denote μ (Greek letter mu) the mean and σ (sigma) the standard deviation. The larger the sample, the more likely $\bar{x}$ and s are close to μ and σ respectively.

In the case of opinion polls, samples are around one thousand people. If we measure the size of one thousand inhabitants selected randomly from the population of a city of one million people, the average size of this sample is likely to be close to the average size of the entire population but has no reason to be equal.

Let us take the example of a coin toss. The coin is balanced and the outcomes are heads or tails. In this case the population is infinite and we can have an infinit number of measurements. Furthermore, the probabilities are known and we can determine the population features.

When the sample size becomes very large, it tends towards the population : $\mu = \lim_{n \to \infty} \bar{x}$ [3].

We introduce here the concept of probability :

$p_i = \lim_{n \to \infty} \frac{f_i}{n}$ where p_i is the probability of the outcome x_i.

With this formula (using the formula page 6) we find the population mean formula : $\mu = \sum p_i \cdot x_i$.

Also if we consider all the events possible, we have $\sum p_i = 1$ (1=100%).

The outcome *heads* is associated with $x_0=0$, and the outcome *tails* with $x_1=1$. The coin is balanced as $p_0=p_1=1/2=0,5=50\%$ and $\mu=p_0.x_0+p_1.x_1$.

Furthermore: $\sigma = \lim_{n \to \infty} s$ and with the formula for *s* page 6 we obtain $\sigma = \sqrt{\sum p_i \cdot (x_i - \mu)^2}$ (for *n* large, *n-1* is close to *n*).

Eventually : $\mu = 0,5$ and $\sigma = 0,5$.

3 MATH : Reads as «μ is equal to the limit of $\bar{x}$ when *n* tends to infinity».

Let us flip nine coins and collect a sample :

$$\{0; 1; 1; 0; 1; 1; 1; 0; 0\}.$$

Then we find $\bar{x} \simeq 0{,}56$ and $s \simeq 0{,}53$.

If this procedure is performed many times, each time we would have a different result for $\bar{x}$.

For example, if two other samples are taken:

$\{1; 1; 0; 1; 1; 0; 1; 1; 0\}$ then $\bar{x} \simeq 0{,}67$ and $s \simeq 0{,}50$

$\{0; 1; 0; 0; 0; 0; 1; 0; 0\}$ then $\bar{x} \simeq 0{,}22$ and $s \simeq 0{,}44$

What would be the distribution of these results as a whole? (Called the sampling distribution)

The values obtained for the samples are generally different from those of the population, but the larger the sample, the more likely it is that the values are closer to those of the population.

Case of a sample with n=50, where $\bar{x} \simeq 0{,}520$ and $s \simeq 0{,}505$:

$\{00100011111110010111110001001101101000110000011101\}$

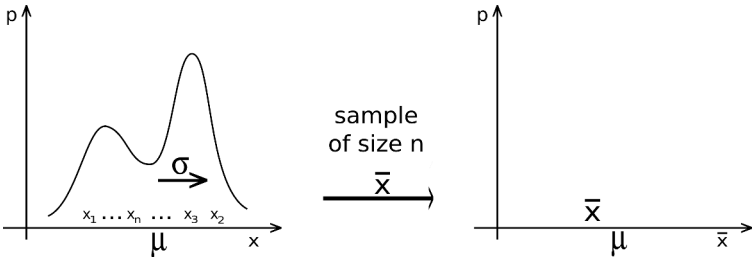
2) The central limit theorem

CENTRAL LIMIT THEOREM :

Within a population we collect random samples of size n. The mean of the sample $\bar{x}$ *varies around the mean of the population* μ *with a standard deviation equal to* $\sigma/\sqrt{n}$*, where* σ *is the standard deviation of the population.*
As n increases, the sampling distribution of $\bar{x}$ *is increasingly concentrated around* μ *and becomes closer and closer to a Gaussian distribution.*

We will describe in due course what a Gaussian distribution, also called normal distribution, is. For the moment we will simply consider a bell curve. This is a very important theorem. Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem.

This is illustrated through the following diagrams:



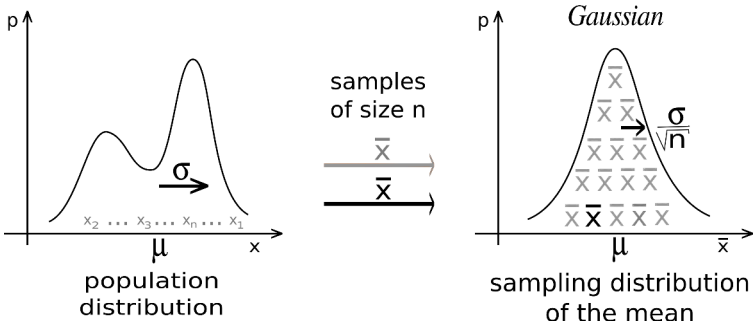
On the left we have the probability *p* of an event *x* (population distribution).

Hypothetically, for a city with a population of one million inhabitants, *p* could represent the probability that one have a given height *x*. If we could measure the height of all the inhabitants, we could exactly determine their average height μ and the standard deviation σ. However, it is pragmatically difficult, if not impossible, to measure a population of that size. Therefore a sample size of only a thousand inhabitants is taken to reduce the burden of labour. For this to be as representative of the whole population as possible, the thousand-person sample is picked randomly.

We obtain a thousand measures of height from x_1 to x_{1000}. From this sample of size $n=1000$ we calculate a mean $\bar{x}$ and a standard deviation *s*. We think that $\bar{x}$ is close to μ, but at the same time there is no reason for it to be equal to μ. We put this value of $\bar{x}$ on the right side of the figure on page 10.

We take a new random sample of a thousand people and a new value for $\bar{x}$.

We then repeat this operation a great number of times. We see on the right the distribution of the samples obtained:



3) Student's t-value and uncertainty

The central limit theorem applies to the limit of large numbers. In the particular case where the distribution of the population is normal we can apply it from *n* small thanks to the coefficients of Student *t*.

Prediction interval :

If μ and σ are known, the sampling distribution is also Gaussian and the expected statistical fluctuations are with a probability of *p*% between $\mu - t_\infty \sigma/\sqrt{n}$ and $\mu + t_\infty \sigma/\sqrt{n}$.

The t-values are read on page 239.

Confidence interval :

In the case of the calculation of uncertainties μ and σ are not known and we estimate them from the sample with $\bar{x}$ and *s*. Due to a small statistic, there is widening given by the Student's t-distribution:

$$\mu = \bar{x} \pm \boldsymbol{t} \cdot \frac{s}{\sqrt{n}}$$

The Student's *t*-value depends on the sample size *n* and on the confidence. If the confidence is 95%, we have 95 in 100 chance that μ is between $\bar{x} - t \cdot s/\sqrt{n}$ and $\bar{x} + t \cdot s/\sqrt{n}$.

We recognize here the notion of measurement uncertainty Δx [4]:

$$x = \bar{x} \pm \Delta x \quad \text{with} \quad \Delta x = t \cdot \frac{s}{\sqrt{n}}$$

Δx is also called absolute uncertainty and $\Delta x / |\bar{x}|$ relative uncertainty.

Let us take again the calorimetry experiment described on page 1. We want to know the thermal capacity of the water with a confidence of 95%. As is often the case in experimental sciences, we consider that the data follow a normal distribution, because by the influence of many independent factors on the value of the measured quantities, we still expect, under the central limit theorem, to have Gaussian fluctuations.

We find for four degrees of freedom (*ddl=n-1*) a Student's *t* of 2.78.

From where : $c = \bar{c} \pm t \cdot s_c / \sqrt{n} = 4320 \pm 660\ J/K/kg$ with 95% confidence.
Here following the dispersion of the values measured by the students : $\Delta c / \bar{c} \simeq 15\,\%$. The calorimetry measurements are usually imprecise. The expected value, known here, is well within the range :

$$3660 < 4180 < 4980$$

In experimental sciences we endeavor to quantify all natural phenomena. Yet, due to the very nature of the ex-

4 MATH : Reads "delta *x*".

perimental approach, the various parameters which make it possible to describe a experimental situation are not perfectly known. We do not have a simple numerical value associated with each characteristic, but an interval for a given confidence. Strictly speaking, any experimental value must associate its uncertainty with its confidence.

Exceptions:

▸ Large number samples: the size of the sample *n* is large enough to be able to directly apply the central limit theorem. The sampling distribution is normal regardless of the distribution of the population. We do not have to worry about Student's distribution which anyway identifies with a Gaussian distribution in this case.

▸ Small samples and normality of the data: we apply the Student law as before.

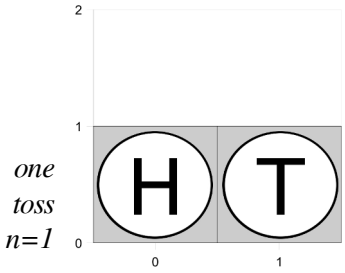
▸ Small samples with non-normality of data: For example, we find by computing the skewness and kurtosis of the data that they do not match a normal distribution. To counter this, it is necessary to perform a case by case study. For instance, when we have a uniform distribution of the population, the prediction interval given on page 12 works from n=2 (it is shown by using the data of the article *Measure with a ruler* p154) . However for a binomial distribution with parameters n=2 and p=0.24, the 50% prediction interval contains 0% of the values... A more complex case on page 146 shows for n = 12, in comparison with a numerical simulation, that the central limit theorem underestimates the confidence interval.

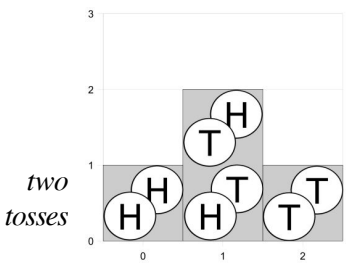
4) Examples

A large number of random factors will influence the measurement of a physical quantity; independent factors, which, whatever their natures will ultimately generate a Gaussian distribution of observations. Let us consider two examples, the toss of a coin and the roll of a six-sided die.

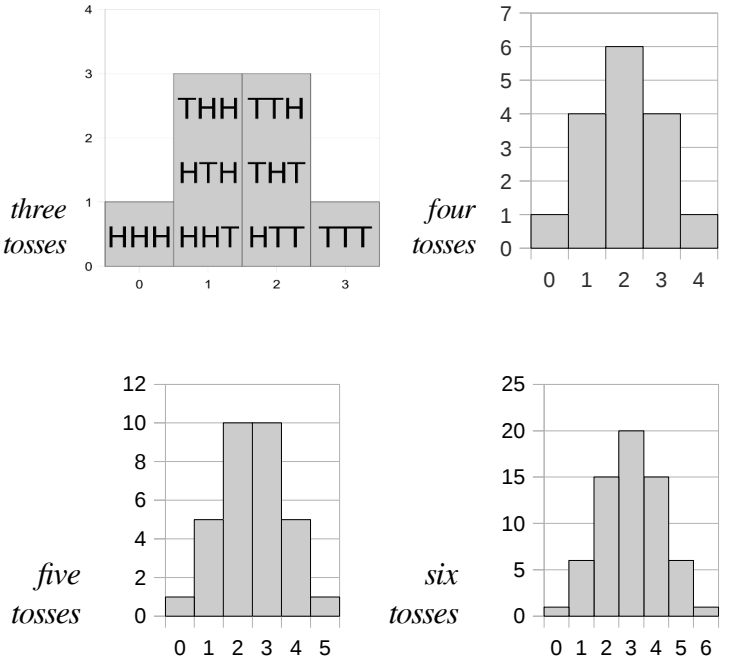
Consider the sample space for tossing a fair coin *n* times. We count the number of *tails*. For one toss, we have two outcomes possible, one with zero *tails* and one with one *tails*.

For two tosses, we have four outcomes possible, one with zero *tails* (H H), two with one *tails* (H T or T H) and one with two *tails* (T T). The more tosses that are thrown, the closer we get to a Gaussian distribution.



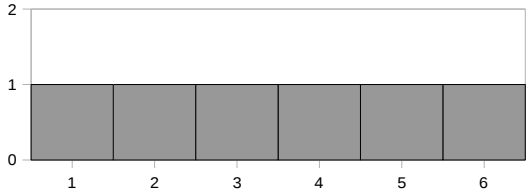


We can obtain the probability (the number of tails divided by the number of possibilities 2^n) as a function of the number of tails. For n = 1 we have the distribution of the population and following the sampling distributions for different values of *n*.



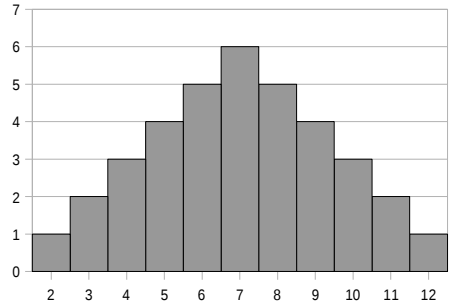
Similarly for the dice we enumerate the possibilities for their sum and the distribution also tend towards a Gaussian. For a single die, the sum simply corresponds to the value of the die. We have a possibility for each value:

one die :

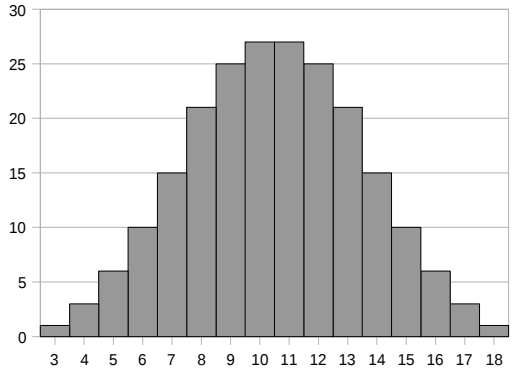


For two dice, there is only one possibility for the sum to be two: 1 for the first die and 1 for the second die. For the sum to be three, there are two possibilities: 1 and 2, or, 2 and 1. The most likely with two dice is to obtain a sum of 7: (1,6) (6,1) (2,5) (5,2) (3,4) (4,3).

two dice :

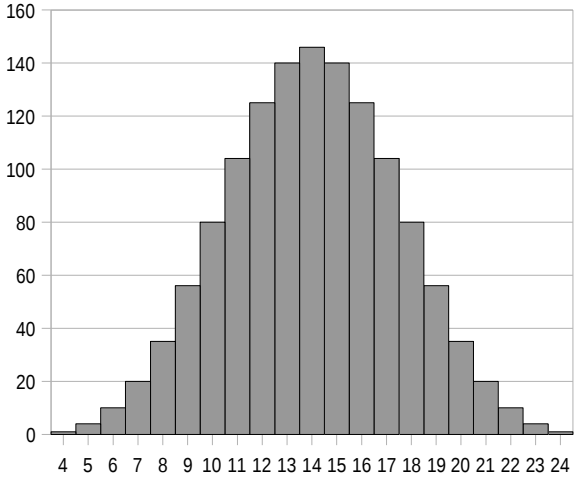


three dice :



For four dice, we already recognize the bell curve and the profile is clearly of the Gaussian type:

four dice :



On this last example, we check the validity of the central limit theorem.

The population mean is :

$$\mu_x = (1+2+3+4+5+6)/6 = 3.5$$

We verify that the sampling distribution mean is the same: $\mu_{\bar{x}} = 14/6 = 3.5$.

The population standard deviation is :

$\sigma_x = \sqrt{\sum p_i \cdot (x_i - \mu)^2}$ then

$$\sigma_x = \sqrt{1/6 \cdot \{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\}}$$

and $\sigma_x \simeq 1.71$

also for four dice : $\sigma_{\bar{x}}=\sigma_x/\sqrt{n}=\sigma_x/2\simeq 0.85$. Now, on the above curve, 40% from the maximum (explanation page 21), we have a deviation close to *3.5* (between *3* and *4*), and an average of $3{,}5/4\simeq 0.88$. It matches.

F. Gaussian distribution

1) Definition of continuous distribution

Some quantities are fundamentally continuous : time, space, temperature, et cetera. Time is like fluid, it does not jump from one value to another. A continuous random variable can take any value in an interval of numbers and has a continuum of possible values. On the other hand when we throw a six-sided die, it is impossible to say "I got 2.35!". It is a forbidden value, only integer values from one to six are allowed.

Thus, some probability distributions are discrete and others continuous. For a die we have : $p_1=...=p_6=1/6$ and $\sum_{i=1}^{n=6} p_i=1$. Now, if we are interested in the height of the inhabitants of a city, there is a continuum of possibilities, the distribution is continuous. We define the probability density function *p(x)* with *p(x)dx* the probability to be between *x* and *x+dx*. Where *dx* is a small variation, and *x+dx* is close to *x*. The continuous random variable probability model assigns probabilities to intervals of outcomes rather than to individual outcomes.

So, the probability that the event is realized on the set of possible values is 100%:

$$\int_{x=-\infty}^{+\infty} p(x)\,dx = 1$$

Mean and variance in the continuous case :

$$\mu = \int_{-\infty}^{+\infty} x \cdot p(x)\,dx \qquad V = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\,dx$$

2) Bell-shaped density curve

A continuous random variable X following normal distribution has two parameters: the mean μ and the standard deviation σ. Density function :

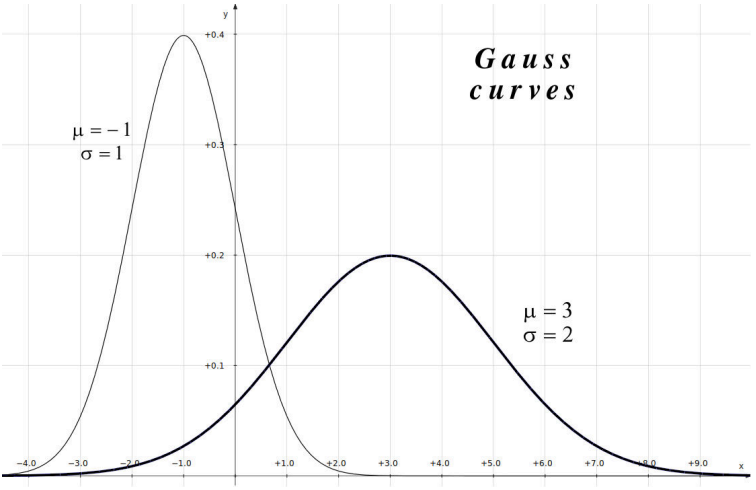
$$p(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In the mathematic tools section page 168 some demonstrations are performed.

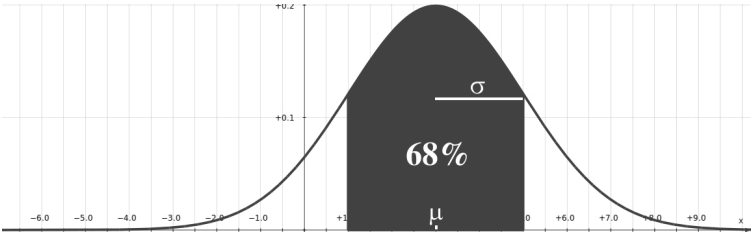
---

5 MATH : The mean is also called E(X), expectation of X. σ^2=V(X) is called variance of the random variable X. Properties : E(aX+b)=aE(X)+b , V(aX)=a^2V(X) and V(X)=E(X^2)-E(X)^2.

We have represented two cases on the following graph:



The total area under the curve is always 1. The probability concentrated within interval $[\mu - \sigma, \mu + \sigma]$ is 0.68 :

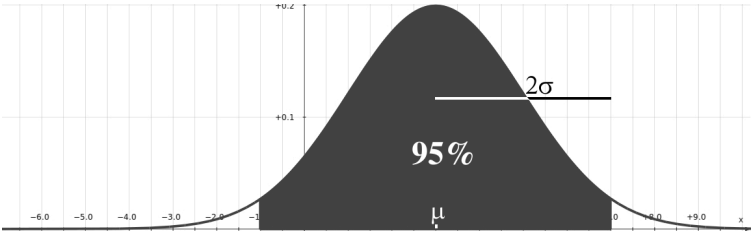


$$\int_{\mu-\sigma}^{\mu+\sigma} p(x)\,dx = 0.683... \simeq 68\,\%$$

We evaluate the standard deviation at *60%.*p_{max} :

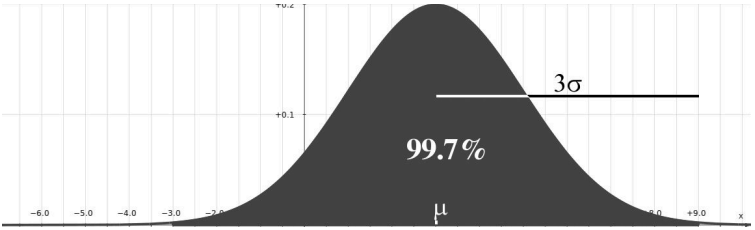
$$p(\mu \pm \sigma)/\,p_{max} = 1/\sqrt{e} \simeq 0.607$$

The probability concentrated within interval [μ – 2σ, μ + 2σ] is 0.95 :



$$\int_{\mu-2\sigma}^{\mu+2\sigma} p(x)dx = 0.954... \simeq 95\%$$

The probability concentrated within interval [μ – 3σ, μ + 3σ] is more than 0.99 :



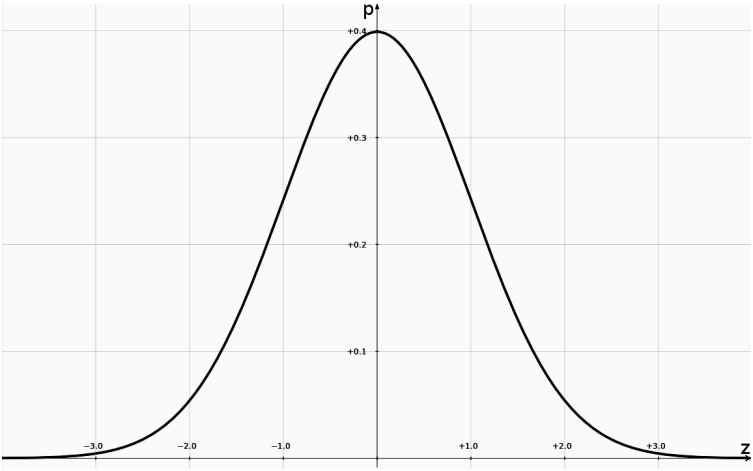
$$\int_{\mu-3\sigma}^{\mu+3\sigma} p(x)dx = 0.997... > 99\%$$

3) Standard normal distribution

A standard normal distribution Z is normally distributed with a mean $\mu = 0$ and a standard deviation $\sigma = 1$.

The distribution X is transformed into a distribution Z using the following two transformations : $x' = x - \mu$

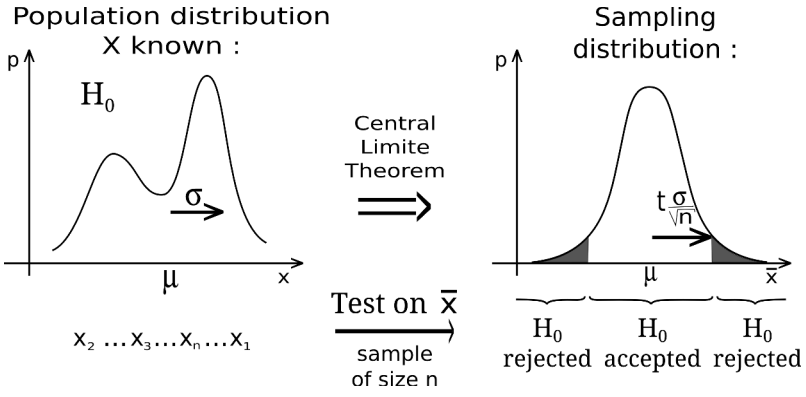
and $\boxed{z = \frac{x - \mu}{\sigma}}$ then : $p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$



G. Hypothesis test

Similarly to how to estimate the mean of an unknown probability distribution, the central limit theorem is also used for a hypothesis test. With a collected sample we have used the properties of the sampling distribution to determine a value and it's uncertainty. For the hypothesis tests, we proceed in the other direction: the law of probability is assumed known, so the sampling distribution is perfectly known and we take a sample to define a decision criterion allowing us to accept or reject the hypothesis.

Using the mean we test a hypothesis H_0. This is referred to as the null hypothesis. It is an assumption made on the probability distribution X. Let μ and σ be the mean and standard deviation of X. We take from the population a sample of size n large enough. The sample mean is $\bar{x}$. If $\bar{x}$ is between $\mu - t_\infty . \sigma/\sqrt{n}$ and $\mu + t_\infty . \sigma/\sqrt{n}$ then H_0 is accepted. However if $\bar{x}$ is outside of those values, the null hypothesis is rejected (two-tailed test).



We consider the coefficient t_∞ of a normal distribution for a p% confidence (or Student's t-value when $n \to \infty$).

We can also use other characteristic intervals of the sampling distribution. In general, the hypothesis H_0 implies a property A of the sampling distribution. Here, the involvement is not deterministic but statistical and decision-making proceeds differently.

Deterministic test case : $H_0 \Rightarrow A$

- If we observe *A* then H_0 cannot be rejected.
- If we do not observe *A* then H_0 is rejected[6].

Statistical test case :

In *p%* of cases : $H_0 \Rightarrow A$

In *(1-p)%* of cases : $H_0 \not\Rightarrow A$

- If we observe *A* then H_0 cannot be rejected with *p%* confidence.
- If we do not observe *A* then H_0 is rejected, with a risk to reject H_0 when it is true of (1-*p)%*.

This protocol makes it possible to make a decision, but at the same time it carries risks of making a mistake. We can reject the hypothesis when it is true or we can accept the hypothesis when it is false.

---

6 The contrapositive of an implication : if $P \Rightarrow Q$ then $\bar{Q} \Rightarrow \bar{P}$.

	Accept H_0	Reject H_0
H_0 is true	Right decision	Wrong decision (error of the first kind α)
H_0 is false	Wrong decision (error of the second kind β)	Right decision

The aim is to minimize the risks linked to α and β. If we accept H_0, then α (the probability of rejecting the null hypothesis when it is true) has to be important, and β (the probability of accepting the null hypothesis when it is false) has to be low. If, on the contrary, we reject H_0 it is because α is low and β important.

When H_0 is true we are able to compute the risk, and α is equal to 1-p. However if H_0 is false we cannot compute β, unless an alternative hypothesis H_1 is known.

For a standard test we fix 1-p in advance. For example we can consider the test statistically significant at the threshold of 5%, and, according to the result found reject or accept the hypothesis. Another method is to calculate the probabilities α and β which correspond to the value $\bar{x}$ found with our sample. Then we measure the credibility of H_0 and we choose whether or not to accept our hypothesis.

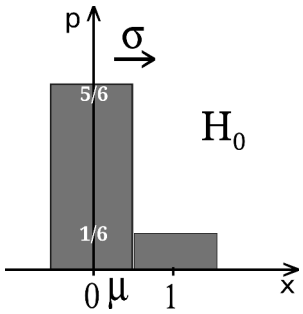
For example, let's imagine that we have several dice. All the dice are unbiased except one which has double the chances of falling on six. Unfortunately the rigged die is

mixed with the others and it does not bear any distinctive marks. We choose one die for a game night and we want to distinguish the biased die to be sure that the chosen die is well balanced.

The die is thrown 92 times :

3151353256243365313441354244652632465436616546
2241154636433165514444241456414316555146362534

For H_0 we define a discrete random variable *X*. When the outcome is six the value is 1. All other outcomes are recorded as 0 :



P(X=0)=5/6 and *P(X=1)=1/6*.

x_i	$x_0 = 0$	$x_1 = 1$
p_i	$p_0 = 5/6$	$p_1 = 1/6$

$$\mu = p_0 x_0 + p_1 x_1$$

$$\sigma^2 = p_0 (x_0 - \mu)^2 + p_1 (x_1 - \mu)^2$$

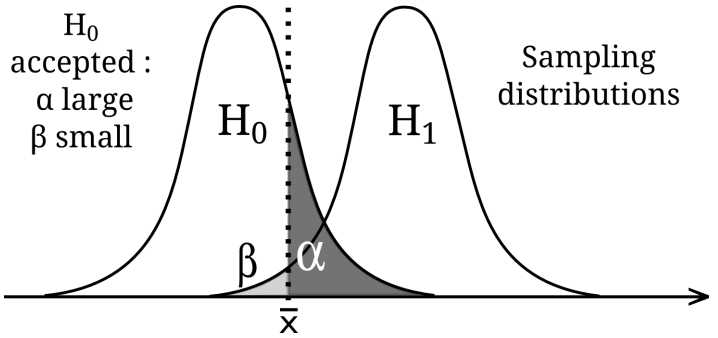
The mean of X is $\mu = 1/6 \simeq 0.167$ and the standard deviation is $\sigma = \sqrt{5}/6 \simeq 0.373$.

In our sample there are sixteen sixes and the sample mean is $\bar{x} = 16/92 \simeq 0.174$. Therefore $\bar{x} - \mu = t_\infty . \sigma / \sqrt{n}$ which gives:

$t_\infty = (\bar{x} - \mu)\sqrt{n}/\sigma \simeq 0.187$.

The right tail of the Gauss distribution for values greater than 0.187 has an area of 0.43, showing that $\alpha \simeq 43\%$ [7].

7 Here σ is known and n=92 is large enough to use the central limit



If H_0 is false then H_1 is true. If the alternative hypothesis H_1 was "the die is biased" (with no details about the way the die is loaded) we would have conducted a two-tailed test and for α we would have to consider the two tails of the distribution. This scenario only requires a one-tailed test: if H_0 were false, the probability of observing six would be doubled and we would observe greater values.

For H_1 we define the random variable Y with $P(Y=0)=2/3$ and $P(Y=1)=1/3$. The mean is $\mu' = 1/3 \simeq 0.333$ and the standard deviation is $\sigma' = \sqrt{2}/3 \simeq 0.471$.

Then $\bar{x} - \mu' = t_\infty' . \sigma' / \sqrt{n}$ and $t_\infty' = (\mu' - \bar{x})\sqrt{n}/\sigma' \simeq 3.24$.

The left tail of this distribution has an area of $\beta \simeq 0.06\%$.

We can therefore very comfortably accept the hypothesis that the chosen die is balanced. In the case of rejection we would have a 43% chance of making a mistake (we try to minimize this error first, classically it is only below the threshold of 5% that one begins to question the null hypothesis). With regard to the alternative hypothesis, there is

theorem.

less than a 1 out of 1000 chance that we considered the die balanced while the die is rigged (we also talk about the power of the test : $\eta=1-\beta$).

Note that we never calculate the probability that a hypothesis is true but the probability to reject the hypothesis while it is true (error of the first kind).

In the legal framework an error of the first kind is made if an innocent person is convicted and second kind if a guilty person is acquitted. The jury is asked to prove the guilt beyond a reasonable doubt and if the person is convicted α must be sufficiently small [vi]. We try to minimize the probability to condemn an innocent person. We do not directly consider the probability of being guilty, the person on trial is presumed innocent, a defendant is considered not guilty as long as his or her guilt is not proven (H_0 : "the defendant is not guilty").

Exercises page 39 treat different cases for this test.

H. Chi-squared test

The Chi-squared test is an another hypothesis test that is simple to use. It tests a null hypothesis stating that the frequency distribution of events in a sample is consistent with a theoretical distribution. We consider different disjoint events that have a total probability of 1.

number of observations :	O_1	O_2	...	O_j	...	O_c
expected frequency :	E_1	E_2	...	E_j	...	E_c

We compute the following sum :

$$\chi^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}$$

Next we have a table on page 240 to estimate the probability of rejecting the null hypothesis H_0 when it is true. According to the value of χ^2 and the number of degrees of freedom we determine whether the assumption is acceptable. The number of degrees of freedom is :

$$ddl = c - 1 \quad \text{(number of categories minus one)}$$

Let us illustrate with the experiments carried out by the botanist Mendel. He makes crosses between plants. He crosses peas with pink flowers. His theory implies that he must obtain 25% of peas with red flowers, 25% of peas with white flowers and 50% of peas with pink flowers. This result is derived from the random encounter of gametes.

Imagine that he observes one thousand flowers with the following values: 27% white, 24% red and 49% roses. Should he continue to believe in his hypothesis?

Observed numbers :	270	240	490
Theoretical frequencies :	250	250	500

then

$$\chi^2 = \frac{(270-250)^2}{250} + \frac{(240-250)^2}{250} + \frac{(490-500)^2}{500} \simeq 2.2$$

and $ddl = 3-1 = 2$.

According to the table, there is more than a 30% chance that the assumption will be rejected when it is true.
We then decide to accept the hypothesis. In general, we take a critical probability α of 5%, below which it is envisaged to reject the hypothesis.

The test is easily generalized for a table :

Experimental frequencies : Theoretical frequencies :

$$\begin{pmatrix} O_{11} & O_{12} & \dots & O_{1j} & \dots & O_{1c} \\ O_{21} & O_{22} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ O_{i1} & \dots & \dots & O_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ O_{r1} & \dots & \dots & \dots & \dots & O_{rc} \end{pmatrix} \quad \begin{pmatrix} E_{11} & E_{12} & \dots & E_{1j} & \dots & E_{1c} \\ E_{21} & E_{22} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ E_{i1} & \dots & \dots & E_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ E_{r1} & \dots & \dots & \dots & \dots & E_{rc} \end{pmatrix}$$

The *ddl* is dependent on the number of columns *c* and of rows *r* :

$$ddl = (c-1)(r-1)$$

We compute the χ^2 with a similar formula :

$$\chi^2 = \sum_{(i,j)} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Moreover, we use the same table to determine the validity of the hypothesis.

I. The sources of the uncertainties

A random variable has a small uncertainty if the measurement is precise, accurate and the acquisition system has a good resolution.

Accuracy is ensured by the absence of systematic errors. There could be a bias that makes the measurement inaccurate (even if the dispersion is low). Reading errors, absence of systematic control and corrections of influential factors, hypotheses in the modeling, etc. All biases must be identified and estimated in order to be added to the dispersion, so the system becomes accurate.

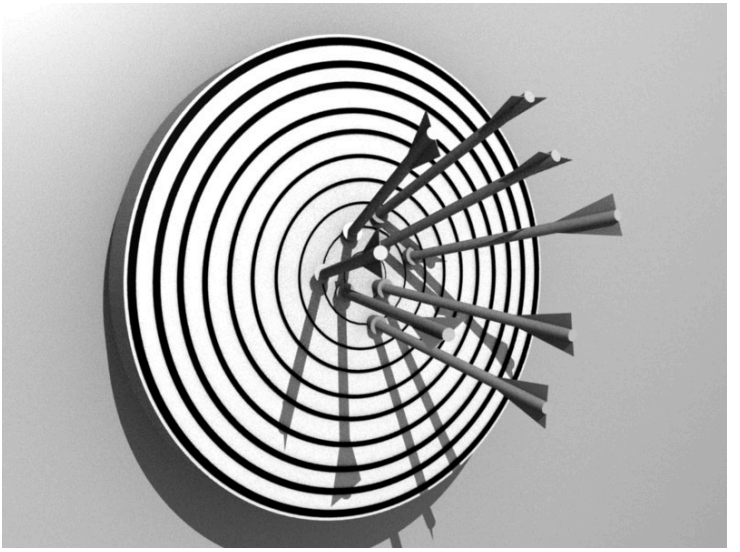
Precision pertains to the repeatability and reproducibility of the measurements. The values of a precise system have low variability. The dispersion may be caused by accidental errors or by a random physical phenomenon (such as radioactivity). The experimenters by their own work, conscientious and according to a well defined and rational protocol, can minimize dispersion. The sources can be countless, but we will try to identify a maximum of sources in order to evaluate them.

The resolution of a sensor depends on the distance between the graduation marks, the type of vernier or the number of digits on the display screen. Sometimes other factors have to be added to the uncertainty due to discretization. You have to refer to the technical datasheet, the instruction guide or contact the manufacturer for a good knowledge of your measuring tool. Calibration of measuring instruments

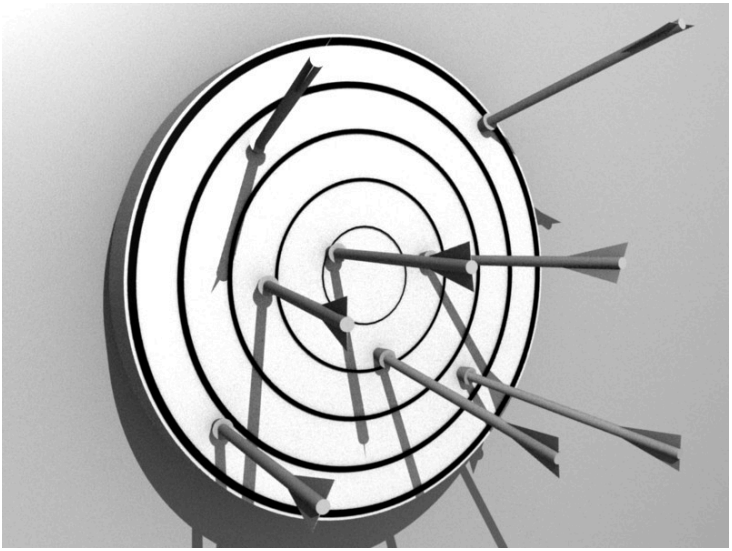
can also be performed with a high-precision apparatus that is used as a reference.

The influence of these different sources of uncertainty can be illustrated by a target and arrows. The center of the target corresponds to the quantity to be measured and the arrows represent the different measurements. If the arrows as a whole are not correctly centered, the accuracy is not assured. The tightening of arrows represents precision. The distance between the circles on the target indicates the resolution. The value noted is that of the circle whose arrow is closest. The experimenter sees the arrows and the circles, however he does not know where is the center of the target. He holds the bow and his desire to be closer to the center of the target shows the quality and rigor of his work.

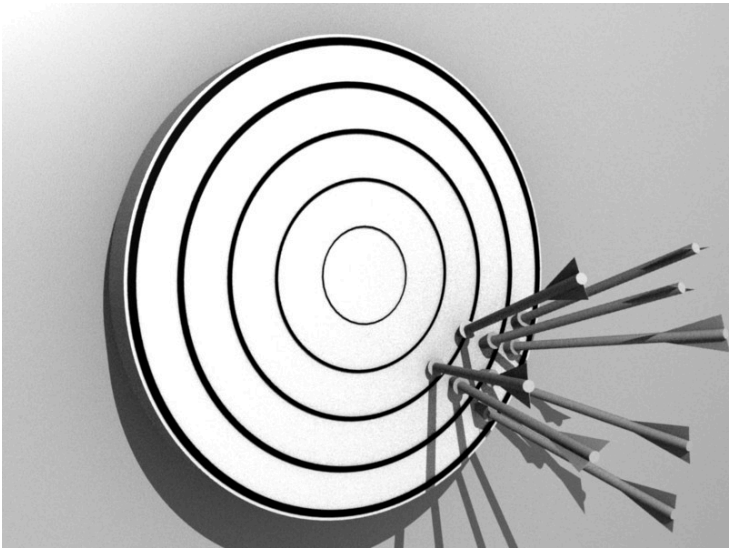
Measure accurate, precise and with a good resolution :



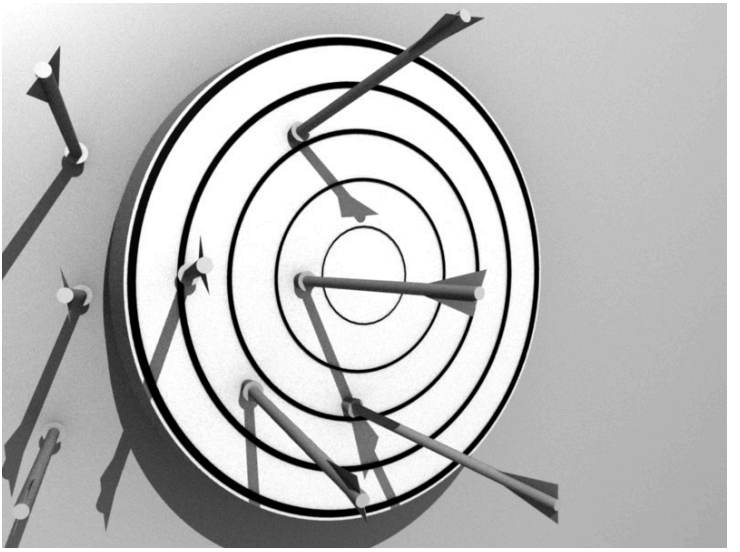
Measure accurate, with a poor precision and a low resolution :



Measure precise but with a bias and a low resolution :



Measure biased, with low precision and resolution :



The full standard deviation will be determined from the deviations of each source by adding the squares (due to the propagation of uncertainties explained in Chapter 2):

$$\sigma = \sqrt{{\sigma_1}^2 + {\sigma_2}^2 + {\sigma_3}^2 + \ldots}$$

J. Exercises

Exercise 1 : Ages *Answers p171*

Students of a class have the following ages : {18; 20; 18; 19; 18; 18; 18; 17; 18; 19; 17; 19; 17; 21; 18}. Determine the mode, median, arithmetic mean, geometric mean, range, standard deviation, root mean square deviation, and mean deviation[8].

Exercise 2 : Card game *Answers p171*

We play with a 32-card standard deck. We randomly draw five cards.
a) Determine the probability of a four aces hand.
b) Determine the probability of having a flush (a flush is a poker hand containing five cards all of the same suit - hearts, spades, diamonds or clubs).

Exercise 3 : Gravity field *Answers p171*

Students measure the Earth's gravitational field strength g. The students measure the following values at the laboratory : 6,20 ; 8,35 ; 13,00 ; 8,37 ; 8,54 ; 9,67 ; 9,75 ; 10,66 (m/s²).
a) What comments can be made about these results?
b) Calculate the mean and standard deviation.
c) What is the mean uncertainty (95% confidence)? Is the result consistent with the expected value?
d) A ninth student performs a new measurement under the same experimental conditions. Evaluate the probability that the student will get a result between 8 and 12 m/s².

8 *Absolute deviation mean* $= \left(\sum \left|x_i - \bar{x}\right|\right) / n = \left(\sum \sqrt{\left(x_i - \bar{x}\right)^2}\right) / n$

Exercise 4 : Elevator *Answers p172*

The maximum load of an elevator is 300 kg, and the total mass of the occupants is 280 ± 10 kg at σ. What is the probability of being overloaded?

Exercise 5 : Assignment *Answers p172*

The following table represents Students' assignment grades (ranking on a 20 point scale) :

10	5	13	7	6	9	5	5	10	15	5	3
15	12	11	1	3	13	11	10	2	7	2	8
2	15	4	11	11	5	8	12	10	18	6	

a) Calculate the mean and standard deviation.

b) Make a graph with the grades on the x-axis and the frequencies on the y-axis.

c) Make another diagram with the following class intervals : [0, 1, 2], [3, 4, 5] ... , [18, 19, 20].
Which bar chart do you prefer?

Exercise 6 : Yahtzee *Answers p173*

We play this game with five six-sided dice.

1) The five dice are drawn. What is the probability of having Yahtzee (all five dice the same).

2) What is the probability of having a sum smaller than ten?

3) We do a series of throws and we get the following sums: 18, 15, 17, 22, 16, 12, 14, 22, 23, 14, 23, 14, 18, 21, 12, 15, 18, 13, 15, 18, 17, 15, 17, 21, 25, 16, 8, 15, 15, 13.

a) Calculate the mean and standard deviation.
b) What mean do you estimate with 95% confidence? Is it consistent with the theoretical value?
c) Make a graph with the values and their frequencies.
d) If we roll the dice again, what is the probability of the result being higher than 24?

Exercise 7 : Elastic bands *Answers p174*

An elastic bands manufacturer indicates that among a thousand elastics sold, on average, ten are not functional. A buyer wants to test the delivered batches before accepting the delivery. He decides to refuse the delivery if the number of damaged elastic bands is too high and wants to have a less than 1% chance of making a mistake by refuting the manufacturer's indication. The buyer picks *n* elastics randomly. How many damaged elastics should the delivery contain for the client to refuse the delivery? Indicate this number for three different cases: a sample of 1000, 200 and 50 elastic bands.

Exercise 8 : Testing an insulating panel
Answers p175

A manufacturer specifies a thermal conductivity of 0.039 W/m/K for a insulation board. The value is certified within ±5%. You want to check if this is true. To do so, you take ten panels at random and measure their respective conductivity ($mW.m^{-1}.K^{-1}$) :

39.1	38.8	39.5	39.2	38.9	39.1	39.2	41.1	38.6	39.3

Are the values in agreement with those announced by the manufacturer (95% confidence on the given margin is consider)? Could he, according to your results, announce another margin?

Exercise 9 : Coins *Answers p175*

We perform a large number of coin tosses to test if the probabilities of landing heads or tails are equal. We carry out the experiment with three different coins. Are they balanced? (Answers with 95% confidence)

1) 42 tails and 58 heads.
2) 510 tails and 490 heads.
3) 420 tails and 580 heads.

Exercise 10 : Parity *Answers p176*

Is gender equality respected in both chambers and The Supreme Court?

	Male	Female
Parliament	470	107
Senate	272	76
The Supreme Court	10	2

Exercise 11 : Births *Answers p176*

Let us test the following hypothesis : births in Sweden are distributed uniformly throughout the year. Suppose we have a random sample of 88 births. The results are grouped according to seasons of variable length : 27 births in the spring (April - June), 20 in the summer (July / August), 8 in the fall (September / October) and 33 in the winter (November - March).

At a 5% critical level, can the hypothesis be rejected?

Now, we collect a very large sample : 23385, 14978, 14106 and 35804.

What is the conclusion?

Theory

Exercise 12 : Gaussian distributions in the plane and the space

Answers p176

One-dimensional Gaussian distribution :

Let $p(x)$ be the probability density function of the standard normal distribution.

1- Calculate and compare the means of x and |x|. Which is equivalent to the mean distance from the origin? What do you think about σ_x and $\sigma_{|x|}$?

2- Do a numerical calculation of $P(|x|\leqslant 1)$, $P(|x|\leqslant 2)$ and $P(|x|\leqslant 3)$.

Two-dimensional Gaussian distribution :

Let $p(x,y)$ be a two-dimensional standard normal density with $p(x,y)=p(x)\,p(y)$. Where $p(x)$ and $p(y)$ are one-dimensional standard normal densities.

Hints:

Let consider a multiple integral of a two-variable continuous function. If we can separate the variables and the limits of integration are independent of x and y :

$$\iint f(x,y)\,dxdy=\iint f(x)\,f(y)\,dxdy=\int f(x)\,dx\int f(y)\,dy$$

Converting between polar and Cartesian coordinates : $\rho^2=x^2+y^2$ and $dxdy=2\pi\rho\ d\rho$ (rotational symmetry)

1- What is the expression of $p(x,y)$? Show that $p(x,y)$ satisfies the two necessary conditions for a probability distribution.

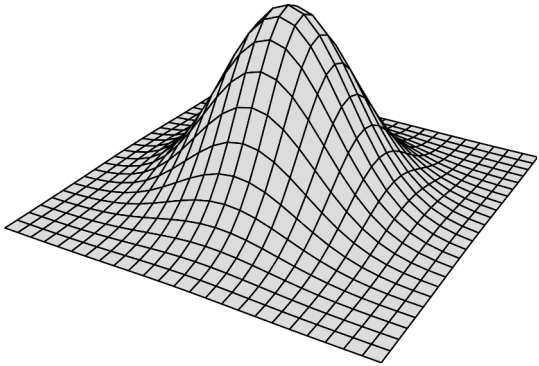
2- By introducing the polar coordinates verify that the probability on all the plane is one. You will express $p(\rho)$ define as:

$$\iint p(x,y)dx\,dy = \int p(\rho)\ d\rho\ .$$

$p(\rho)$ is the density function with respect to ρ.
$p(\rho)\ d\rho$ corresponds to the probability of an event being between ρ and $\rho+d\rho$.

What is the value of the mean $\bar{\rho}$ of the distance ρ from the point of origin? What is the standard deviation σ_ρ for this distribution?

3- Calculate $P(\rho \leqslant \sigma_\rho)$, $P(\rho \leqslant 2\sigma_\rho)$ and $P(\rho \leqslant 3\sigma_\rho)$.



Three-dimensional Gaussian distribution :

Let $p(x,y,z)$ be a three-dimensional standard normal density with $p(x,y,z)=p(x)\,p(y)\,p(z)$. Where $p(x)$, $p(y)$ and $p(z)$ are one-dimensional standard normal densities.

Hints:

Same properties for the multiple integral than in two dimensions.

Converting between spherical and Cartesian coordinates : $r^2 = x^2 + y^2 + z^2$ and $dx\,dy\,dz = 4\pi r^2\,dr$ (spherical symmetry)

1- What is the expression of $p(x, y, z)$? Determine $p(r)$ define as : $\iiint p(x, y, z)\,dx\,dy\,dz = \int p(r)\,dr$.
$p(r)$ is the density function with respect to *r*.
$p(r)\,dr$ corresponds to the probability of an event being between *r* and *r*+*dr*.

2- Verify that the probability on all the space is one.
3- What is the value of the mean $\bar{r}$ and the standard deviation σ_r ?

4- Calculate $P(R \leqslant \sigma_r)$, $P(R \leqslant 2\sigma_r)$ and $P(R \leqslant 3\sigma_r)$.

5- Compare the three cases.

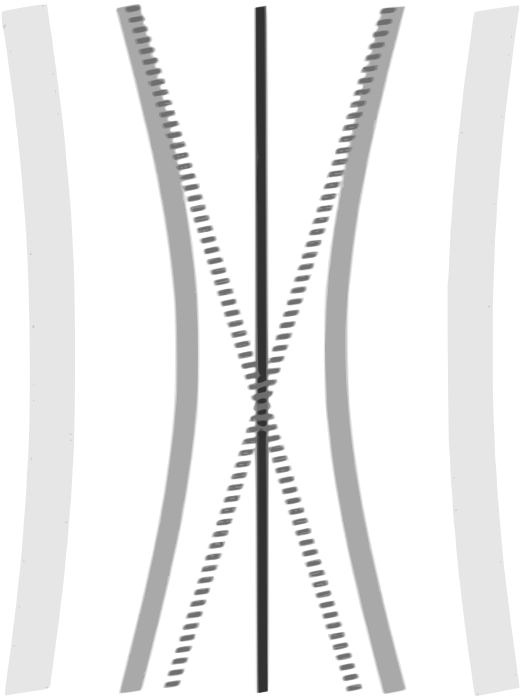
Gaussian	1D	2D	3D
Distance from the origin	$\lvert x \rvert$ [9]	ρ	r
Mean	$\sqrt{\frac{2}{\pi}}$	$\sqrt{\frac{\pi}{2}}$	$2\sqrt{\frac{2}{\pi}}$
Standard deviations	1	$\sqrt{2}$	$\sqrt{3}$
P (σ)	68.3%	1-1/e=63.2%	60.8%
P (2σ)	95.4%	98.2%	99.3%
P (3σ)	99.7%	99.988%	99.9994%

To check your calculations on a spreadsheet you can use the following functions:

On OpenOffice :

- sum of a selected area * (ex.: B43:B53) : =SOMME(*)
- value set to the cell B3 : B3
- =MOYENNE(*)
- squared value : *^2
- square root : *^(1/2)
- =ECARTYPE(*)
- Student's t-value 95% confidence and n=20 : =LOI.STUDENT.INVERSE(0,05;19)
- =TEST.KHIDEUX(*;**); * : experimental frequencies (ex.: B71:E72), ** : theoretical frequencies.

9 MATH : We may be surprised at the difference of the one-dimensional expression with the absolute value, it is only a question of definitions in cylindrical and spherical coordinates. For example, $\rho \in [0;+\infty[$ and $\theta \in [0;2\pi[$, but we could also take $\rho \in]-\infty;+\infty[$ and $\theta \in [0;\pi[$, then the mean of ρ would be zero and we would have considered its absolute value.



II. CORRELATION AND INDEPENDENCE

In the previous chapter we looked at a singular random variable X with n number of outcomes $\{x_i\}$. Now we have several random quantities and a new index to distinguish them : X_j and its observations $\{x_{jk}\}$. X_j is the j^{th} quantity and x_{jk} is the k^{th} observation of this quantity. We are interested in the interactions between these different quantities.

To illustrate, we consider a sample of four individuals with three characteristics, of height X_1, weight X_2 and of birth month X_3. A priori, we expect a correlation between height and weight : generally taller people also have heavier body mass (positive correlation). On the other hand, we can think that birth months have no effect on weight and height (X_3 uncorrelated with X_1 and X_2).

A. Correlation coefficient

The sample correlation coefficient r is used to identify a linear relationship between two variables X_i and X_j :

$$r_{ij} = \frac{\sum_k [(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)]}{\sqrt{\sum_k [(x_{ik} - \bar{x}_i)^2]} \cdot \sqrt{\sum_k [(x_{jk} - \bar{x}_j)^2]}}$$

r varies between *-1* and *+1*. If $|r|=1$ the variables are perfectly correlated : $r=1$ is verified in the case of a perfect increasing linear relationship, and $r=-1$ in the case of a perfect decreasing linear correlation. If $r=0$, the variables are uncorrelated and independents.

Calculus of r_{12} ,r_{13} and r_{23} :

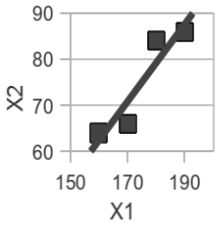
	X_1 *(cm)*	X_2 *(kg)*	X_3	$x_1-\bar{x}_1$	$x_2-\bar{x}_2$	$x_3-\bar{x}_3$	$(x_1-\bar{x}_1)^2$
1	160	64	4	-15	-11	-2	225
2	170	66	8	-5	-9	2	25
3	180	84	9	5	9	3	25
4	190	86	3	15	11	-3	225
$\bar{x}$	175	75	6			Σ=	500

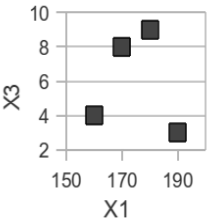
and :

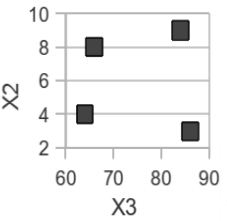
$(x_2-\bar{x}_2)^2$	$(x_3-\bar{x}_3)^2$	$(x_1-\bar{x}_1)$ $.(x_2-\bar{x}_2)$	$(x_1-\bar{x}_1)$ $.(x_3-\bar{x}_3)$	$(x_2-\bar{x}_2)$ $.(x_3-\bar{x}_3)$
121	4	165	30	22
81	4	45	-10	-18
81	9	45	15	27
121	9	165	-45	-33
404	26	420	-10	-2

then : $r_{12}=\frac{420}{\sqrt{500}\sqrt{404}}\simeq 0.93$, $r_{13}\simeq -0.09$ and $r_{23}\simeq -0.02$.

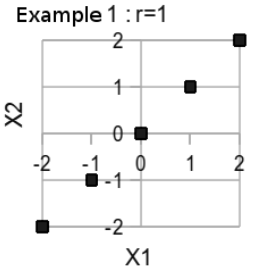
r_{12} is close to +1, so we have a strong positive correlation. r_{13} and r_{23} are close to zero and therefore: X_3 is independent of X_1 and X_2 :

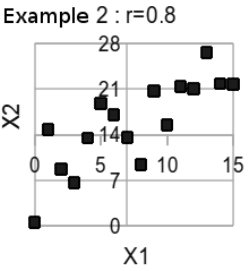


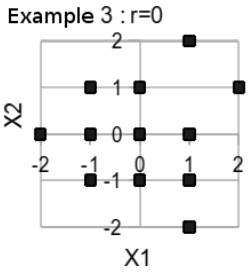


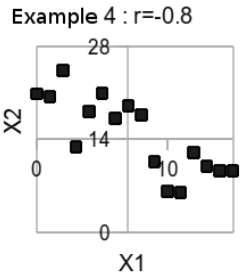


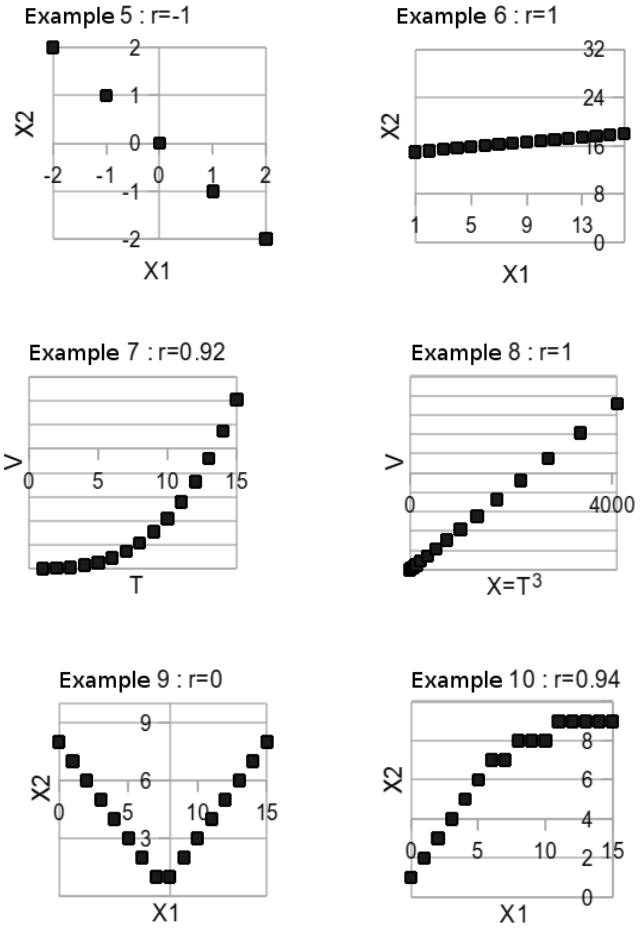
Examples of data sets :











Examples 7, 9 and 10 illustrate a strong correlation between two variables. Yet the correlation coefficient is not as close to -1 or +1 as we could imagine, it is even zero in example 9. This is due to the fact that the correlations are not linear.

There may be saturation phenomena (example 10) or

threshold effect (a product may be beneficial at low dose and harmful at higher doses - example 9). To prevent these phenomena from misrepresenting the data, it is important to carefully consider the relevance of the variables chosen before starting a statistical study.
Another example: if we study the volume V of different objects according to their size T, we find a positive correlation. But the correlation will be much stronger between V and $X = T^3$ (graphs 7 and 8).

The fact that two variables are correlated does not necessarily imply a causal relationship between the two variables (the variation of one variable leads to the variation of the other). Rather than the variables effecting each other, the change may be attributable to a common external cause.
For example, it can be assumed that there is a correlation between the consumption of sunscreen and of ice cream. There is obviously no causal link between the two but a common cause i.e. the weather.
A physics study can show a causality, not a statistical one.

B. Propagation of uncertainties formula

Consider a bucket filled with a million grains of sand. The mass of a grain is 10 mg with an uncertainty of 1mg. What is the mass of sand contained in the bucket?

1) Propagation of standard deviations formula

As a general approach, let f be a function of p independent variables:

$$f(x_1, x_2, \ldots, x_j, \ldots, x_p)$$

Each of these random variables is associated a mean value $\bar{x}_j$ and a standard deviation σ_j.

What are the values of f and σ_f ?

Statistics give the answer and demonstrates the propagation formula of the standard deviations:

$$\sigma_f^2 = \sum_{j=1}^{p} \left[\left(\frac{\partial f}{\partial x_j} \right)^2 \sigma_j^2 \right] \qquad [10]$$

10 We obtained the variance formula by replacing σ^2 by V.

2) Uncertainty calculations

For the uncertainties (defined on page 13) we also have a propagation formula:

$$\Delta f^2 = \sum_{j=1}^{p} \left[\left(\frac{\partial f}{\partial x_j} \right)^2 \Delta x_j^{\,2} \right]$$

The uncertainty propagation formula is not as exact as for standard deviations, but this formula is very practical and often very close to the exact result.

Considering our bucket: $M(m_1, m_2, \ldots, m_j, \ldots, m_p)$

with $M = \sum_{j=1}^{p} m_j$

where we call M the total mass of sand in the bucket, m_j the mass of each grain and p the number of grains.

$$\Delta M^2 = \sum_{j=1}^{p} \left(\partial M / \partial m_j \right)^2 \Delta m_j^{\,2}$$

$$\partial M / \partial m_j = \partial m_1 / \partial m_j + \ldots + \partial m_j / \partial m_j + \ldots + \partial m_p / \partial m_j$$

$$\partial M / \partial m_j = 0 + \ldots + 1 + \ldots + 0 = 1$$

(Calculations of partial derivatives are explained on page 167)

then $\Delta M^2 = \left(\sum_{j=1}^{p} 1^2 \right) \Delta m^2$

and $\Delta M^2 = p \cdot \Delta m^2$ with $\Delta m = \Delta m_j$ whatever *j*.

Finally : $\Delta M = \sqrt{p} \cdot \Delta m = \sqrt{1000000} \times 0.001\, g$.

The bucket weighs 10 kilograms with an accuracy to the gram. The precision on the mass of the bucket is thus 0.01%. Naively, we might have thought that the overall uncertainty on the bucket mass was the sum of the uncertainties of each grain. We would then have an absolute uncertainty of 1 kilogram and a relative of 10%, which is very different from reality and would ignore the compensations.

Here the propagation formula is very precise because we have a very large number of grains. It is even exact, from the small numbers, if the distribution of the mass of the grains is Gaussian[11].

11 MATH : A linear combination of Gaussian quantities is itself Gaussian (applies here to a sum). And in the propagation of uncertainties formula, if f and the x_i have the same kind of probability distribution, the formula is exact like this one with the standard deviations.

In practice, there are multiple common ways to calculate uncertainties' propagation, depending on the situation, given below:

- For sums or differences we add absolute uncertainties squared:

$$\Delta f^2 = \sum_{j=1}^{p} \Delta x_j^{\,2}$$

For example if $d=x_2-x_1$ with $\Delta x_2=\Delta x_1=1cm$ then $\Delta d \simeq 1.4\,cm$.

- For products or quotients we add relative uncertainties squared:

$$\left(\frac{\Delta f}{f}\right)^2 = \sum_{j=1}^{p} \left(\frac{\Delta x_j}{x_j}\right)^2$$

For example if $R=U/I$ with U and I with a precision of 1% then R is known with a precision of 1.4%.

In more complex cases, the partial derivative calculation must be performed explicitly.

Alternatively, using a random number generator or

uncorrelated packets approximately Gaussian, it is possible to do a numerical calculation.
The latter method is illustrated in Exercise 2 of this chapter. A spreadsheet can do the calculations automatically (for example see on sheet 4 of the file IncertitudesLibres on www.incertitudes.fr).

There are also methods which give general ideas on uncertainty. They develop a general impression at the risk of reliability.

For example as we add the uncertainties squared, we can anticipate that the largest uncertainty will quickly prevail over the others. Consider the example of $R=U/I$, if U is known with an uncertainty of 1% and I of 0,1% then R is known with an uncertainty of 1,005%≃1%, we can ignore the uncertainty of I.

For addition and subtraction, it is sometimes considered that the parameter with the last significant figure the less precise indicates the precision of the last significant figure of the result.
Yet on our example calculation of the mass of the sand-filled bucket, it does not work. Since the mass of a grain is m = 10 mg but the mass of the bucket M is known to the

gram and not to the milligram!

For multiplication and division, it is sometimes considered that the parameter with the lowest number of significant figures indicates the result's number of significant digits, but here too one must be careful.
As a simple illustration, if $H=2h$ with $h=5,00m$ (h known within a *cm*), what is the value of H ? According to the rule below, H would be known with three significant digits: $H=10,0m$. H would only be known to *10cm*, it goes without saying that it is closer the *cm* ...

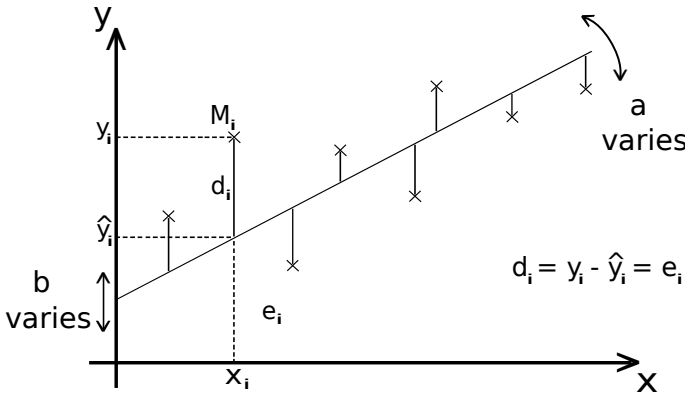
While these tricks serve to aid the calculations, they are not without their pitfalls and must used with caution.

C. Linear regression

We have two correlated random variables, *X* and *Y*, and are in need of an approach for modeling the relationship between them. The outcomes generate a cloud of data in the plane *y(x)* and we want to determine the best affine function $y = ax + b$ that fit the observations. For example, what is the most appropriate relationship between the height X and the weight Y in our initial example?

What are the uncertainties Δa and Δb ?

1) Principle and formulas



We choose the least squares method: this method minimizes the sum of squared residuals.

The set of dots is denoted $M_i(x_i, y_i)$. For x_i given, the estimated value of y is: $\hat{y}_i = a x_i + b$.

We have to minimize the following quantity:

$$\sum d^2 = \sum_i (y_i - \hat{y}_i)^2$$

We differentiate this quantity with respect to a and b, and we set the derivatives equal to 0. We then have the best straight line and can obtain the following equations:

$$\sum_i (y_i - a x_i - b) x_i = 0 \quad \text{and} \quad \sum_i (y_i - a x_i - b) = 0 .$$

This is equivalent to

$$\boxed{a = \frac{\overline{x y} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}} \quad \text{and} \quad \boxed{b = \bar{y} - a\,\bar{x}} .$$

We call e_i the residuals: $y_i = \hat{y}_i + e_i$.

we find the following different standard deviations [12]:

- for the residuals $s_r = \sqrt{\dfrac{\sum_i e_i^2}{n-2}}$
- for the slope $s_a = \dfrac{s_r}{\sqrt{\sum_i (x_i - \bar{x})^2}}$
- for the y-intercept $s_b = s_r \sqrt{\dfrac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$

12 Demonstrations p99.

Then $\boxed{\Delta a = t_{n-2}\, s_a}$ and $\boxed{\Delta b = t_{n-2}\, s_b}$.

t_{n-2} : Student's t-value for *n-2* degrees of freedom.

You are now able to carry out all calculations.
Let's do the calculations for weight in relation to height:

$$\overline{xy} = (160\text{x}64 + 170\text{x}66 + 180\text{x}84 + 190\text{x}86)/4$$

$$\overline{x^2} = (160^2 + 170^2 + 180^2 + 190^2)/4$$

$a = (13230 - 175\,x\,75)/(30750 - 175^2) = 0.84$ and
$b = 75 - 0.84\,x\,175 = -72$

$$s_r = \sqrt{[(64 - (0.84\,x\,160 - 72))^2 + (-4.8)^2 + 4.8^2 + (-1.6)^2]/2} \simeq 5.06$$

$$s_a \simeq 5.06/\sqrt{(160 - 175)^2 + (-5)^2 + 5^2 + 15^2} \simeq 0.226$$

$\Delta a \simeq 2.92\,x\,0.226 \simeq 0.66$ with a 90% confidence

$$s_b \simeq 5.06\sqrt{(160^2 + 170^2 + 180^2 + 190^2)/[4(15^2 + 5^2 + 25 + 225)]} \simeq 39.7$$

$\Delta b \simeq 2.92\,x\,39.7 \simeq 116$ with a 90% confidence

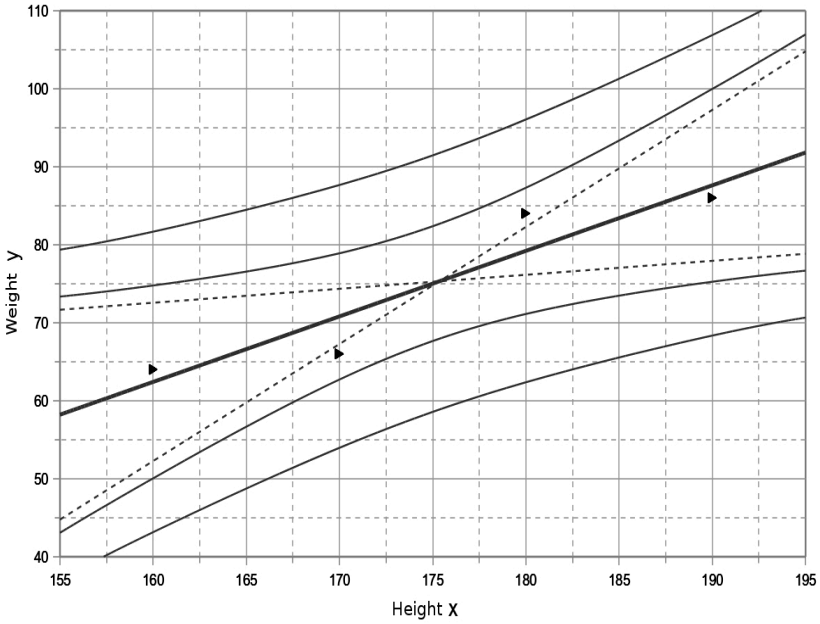
then: $Height = (0.84 \pm 0.66)\,Weight - (72 \pm 116)$

with a 90% confidence.

Here, the formula is very imprecise, which is unsurprising given the small number of points and the dispersion of the data. However, the method of calculation is now explicit and comprehensible.

In the graph that follows we have:

- In the middle, the interpolated straight line represents the best balance between the points above and below this line.



- The dotted lines represent the two extreme lines ($y = a_{min}\,x + b_{max}$ and $y = a_{max}\,x + b_{min}$).

- The first curves represent the estimated values for y. It is the mean confidence interval of y_o corresponding to x_o :

$$\Delta y_o = t_{n-2}\, s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

For example if $x_o = 175\ cm$ we can calculate $y_o = 75.0 \pm 7.4\ kg$. Also we can obtain an estimation out of the interval, for example if $x_o = 195\ cm$ we obtain $y_o = 92 \pm 15\ kg$.

- The outer curves represent a prediction for a new measurement. Prediction interval for an observation y_o :

$$\Delta y_o = t_{n-2}\, s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2} + 1}$$

For example, if the height equals 175 cm there is a 90% chance of their mass being between 58 and 92 kg (generally 90% of the data are inside the curves and 10% outside).

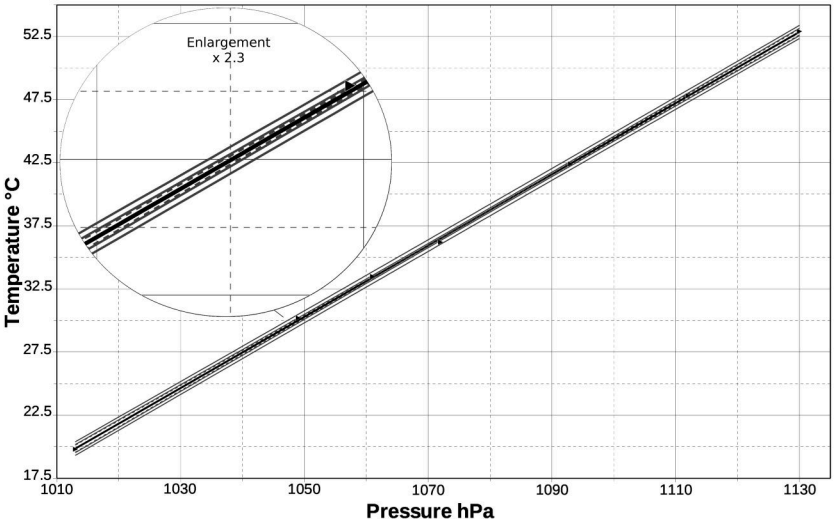
2) Absolute zero measurement

We study a gas enclosed in a rigid container of constant volume. We have sensors to measure its temperature and pressure. Initially the gas is at ambient temperature and pressure. Then we immerse the whole container in hot water and measure the changes over time[1] :

1 *This experiment was realized on Tuesday 17 October 2006 in Bourges by M. ROUAUD and O. LEROY at the Lycée Alain-Fournier (France).*

time	10h 15	10h 30	10h 40	10h 55	not noted	not noted	12h
temperature Θ (°C)	19.8	52.9	47.8	42.4	36.2	33.5	30.2
pressure P (hPa)	1013	1130	1112	1093	1072	1061	1049

We assume that the gas obeys the ideal gas law $PV = nRT = nR(\Theta - \Theta_{0K})$. Plotting *Θ(P)* we can obtain a temperature of absolute zero : the y-intercept give Θ_{0K}.



The regression is good (r=0.99991) but the measured points are far from absolute zero. By extension we obtain with a 95% confidence:

$$\Theta_{0K} = -266.0 \pm 4.8\,°C$$

We know that absolute zero is -273.15 °C, therefore this is not consistent. We can therefore assume that there is a bias and that we have not considered all the sources of uncertainties.

3) Regression with uncertainties on the data

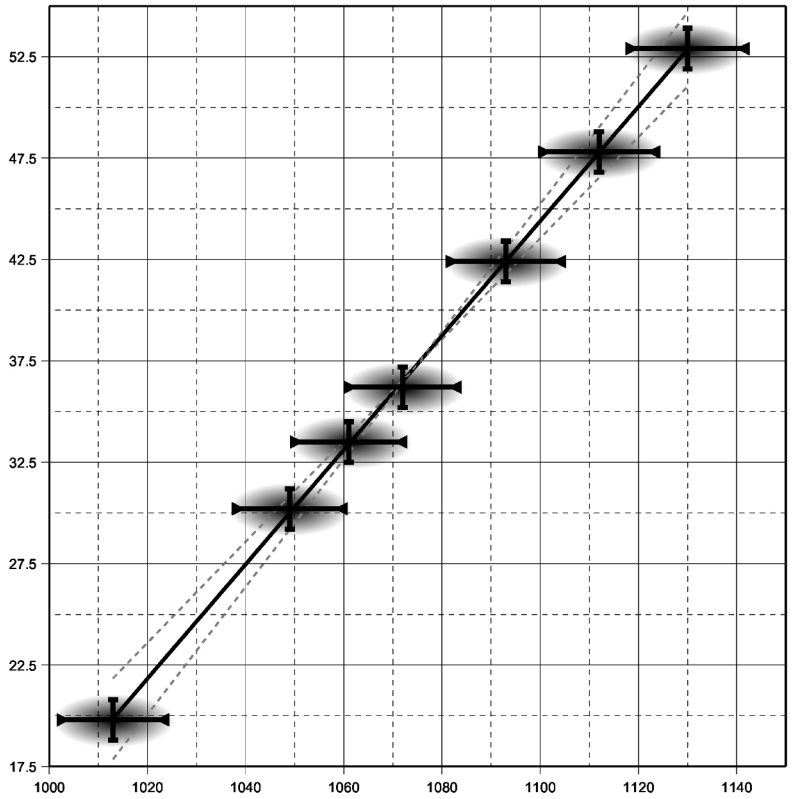
The measuring instruments are not perfect and the manuals indicate the precision for each. The uncertainties are 1% for the pressure sensor and 1°C for the thermometer. We thus have an uncertainty on x and y :

$$M_i(x_i \pm \Delta x_i, y_i \pm \Delta y_i)$$

Now both the dependent and independent variables have uncertainties which contribute to the weighting factors. Let w_i be the weights. In an experimental situation the weights are frequently not equal. The linear least-squared fit requires the minimization of :

$$\sum_i w_i e_i^2 \qquad \text{with} \qquad w_i = \frac{1}{(\Delta y_i)^2 + (a \Delta x_i)^2}$$

The problem is solved iteratively because the weights depend on the slope. We initially put an estimated value of a, then the value of a obtained replaces it until the equality of the values.



We obtain: $\Theta_{0K} = -266 \pm 35\,°C$ with the same confidence on the uncertainties of x_i and y_i. The value is now correct. The main sources of uncertainties seem to be included.

We could also consider the modeling uncertainty induced by the ideal gas hypothesis, but under the experimental conditions of this experiment, the model provides a good approximation. This source of uncertainty is negligible compared to the others uncertainties considered here. The use of a real gas model (like Van der Waals equation of state) would demonstrate this.

Formulas [i] :

$$S^2 = \sum_i w_i [y_i - (a x_i + b)]^2$$

$$\frac{\partial S}{\partial b}^2 = 0 \quad \text{and} \quad \frac{\partial S}{\partial a}^2 = 0$$

leads to

$$b = \frac{\sum w_i y_i \sum w_i x_i^2 - \sum w_i x_i \sum w_i x_i y_i}{\Delta}$$

and

$$a = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\Delta}$$

with

$$\Delta = \sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2$$

then

$$\Delta b = \sqrt{\frac{\sum w_i x_i^2}{\Delta}}$$

and

$$\Delta a = \sqrt{\frac{\sum w_i}{\Delta}}$$

4) Linearization

In many cases we can go back to a linear model (i.e. $y = a\,x + b$). Here are some examples below:

$y = \alpha\, x^{\beta}$ $x, y, \alpha > 0$	$y' = a\,x' + b$ with $y' = \ln(y) \quad x' = \ln(x)$ and $\beta = a \quad \alpha = e^{b}$ and $\Delta\beta = \Delta a \quad \Delta\alpha = \alpha\,\Delta b$
$y = \alpha\, e^{\beta x} \quad x, \alpha > 0$	$y' = \ln(y) \quad x' = x$
$y = \frac{1}{\alpha + \beta x}$	$y' = \frac{1}{y}$
$y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$ (logistic distribution)	$y' = \ln\left(\frac{y}{1-y}\right) \quad y' = \alpha + \beta\, x$
$y = 1 - \left(\frac{\alpha}{x}\right)^{\beta}$ (Pareto distribution)	$y' = \ln(1-y) \quad x' = \ln(x)$ $\beta = -a \quad \alpha = e^{-\frac{b}{a}}$
$y = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$ (Weibull distribution)	$y' = \ln\left(\ln\left(\frac{1}{1-y}\right)\right)$ $x' = \ln(x) \quad \beta = a \quad \alpha = e^{-\frac{b}{a}}$
$y = \alpha + e^{\beta x}$	no linear model
$y = \frac{\alpha\, x}{\beta + x}$	no linear model
$y = \alpha + \beta\, x + \gamma\, x^{2}$	no linear model as *y(x)*

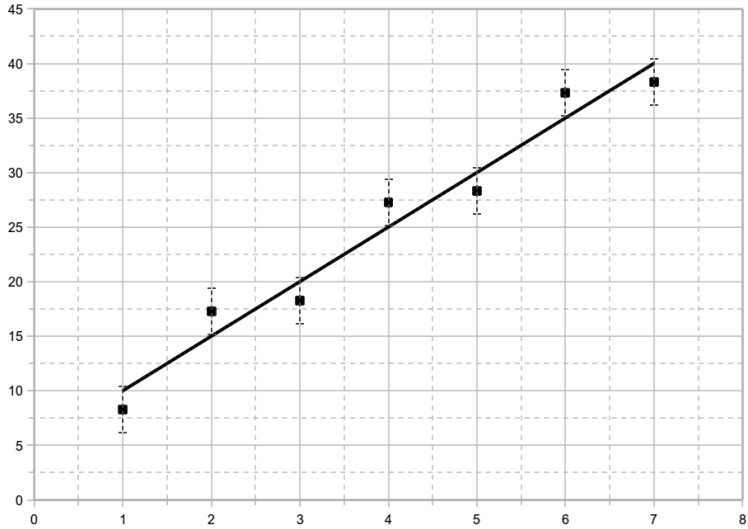
5) Comparison of methods

In all the regression methods we consider, we are interested in the variability of y to x fixed (the opposite approach would yield equivalent results).

For a given x_i we have a y_i and its standard deviation σ_{yi}.

a) Summary

1- Simple linear regression



y_1=8.3 y_1=17.3 y_1=18.3 y_1=27.3 y_1=28.3 y_1=37.3 y_1=38.3

Case 1 : a=5 b=5 s_r=2.34 s_a=0.443 and s_b=1.98

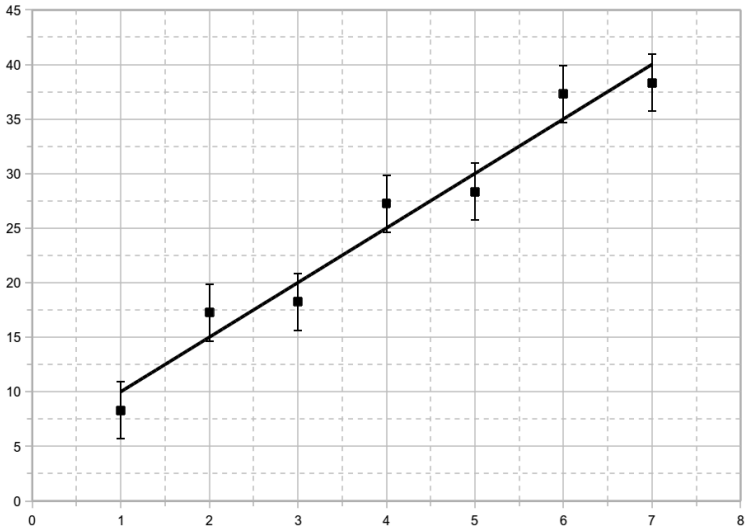
Simple regression does not mean that the data has no uncertainties. The uncertainties are unknown and we estimate them with the data itself. The uncertainties are considered

constant whatever regardless of y_i. The uncertainty corresponds to the standard deviation of y_i with respect to the estimated line :

$$s_{y_i} = s_y = s_r = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$$

$$s_a = \frac{s_r}{\sqrt{\sum_i (x_i - \bar{x})^2}} \qquad s_b = \sqrt{\frac{\overline{x^2}}{\sum (x_i - \bar{x})^2}}\, s_r$$

2- Regression with constant standard deviation s_y:



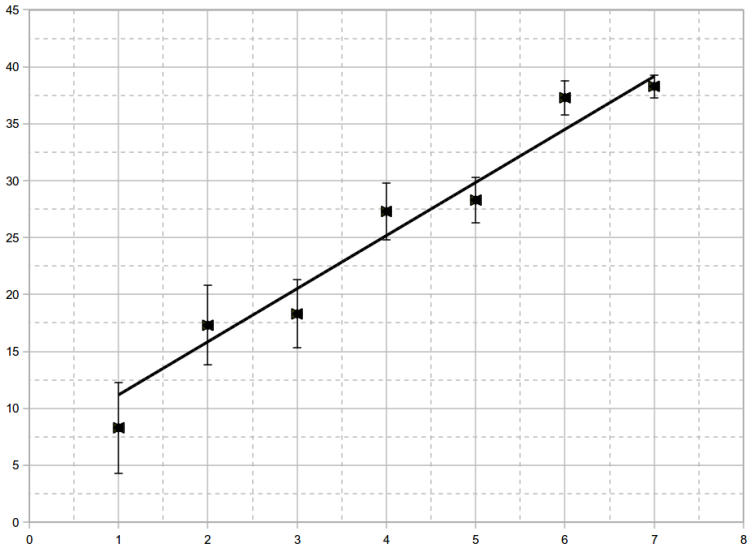
Case 2 : a=5 b=5 s_y=2.55 s_a=0.482 s_b=2.16

In this case the s_{yi} are equal and known : $s_{y_i} = s_y$ and

$$s_a = \frac{s_y}{\sqrt{\sum_i (x_i - \bar{x})^2}} \qquad s_b = \sqrt{\frac{\overline{x^2}}{\sum (x_i - \bar{x})^2}}\, s_y$$

If the straight line does not pass through the error bars, it can be assumed that all sources of uncertainty have not been calculated. It is necessary either to integrate them into s_y, or to apply the previous method.

3- Regression with standard deviation s_y:



$s_{y1}=4$ $s_{y2}=3.5$ $s_{y3}=3$ $s_{y4}=2.5$ $s_{y5}=2$ $s_{y6}=1.5$ $s_{y7}=1$

Case 3 : $a=4.67$ $b=6.51$ $s_a=0.482$ $s_b=2.16$

The s_{yi} are known. We can apply the propagation formula of the standard deviations:

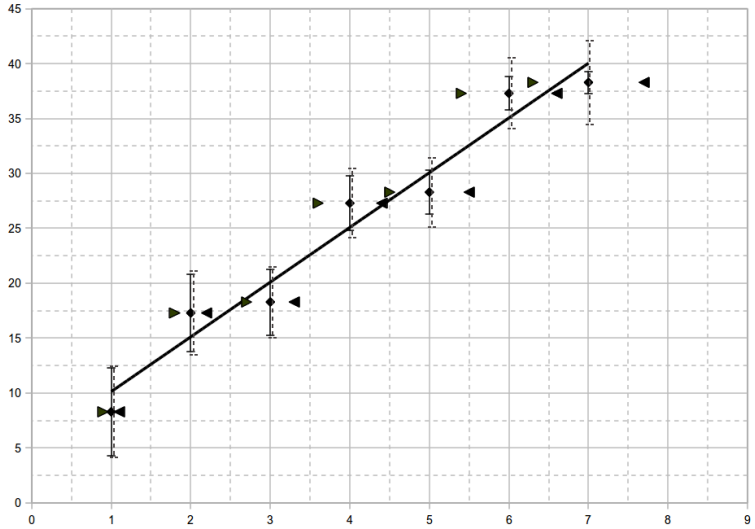
$$s_a^2=\sum_i \left(\frac{\partial a}{\partial y_i}\right)^2 s_{y_i}^2 \quad \text{and} \quad s_b^2=\sum_i \left(\frac{\partial b}{\partial y_i}\right)^2 s_{y_i}^2$$

The formulas' results are exact and help us to find the expressions of the previous cases. Also in this case we can

use the following estimates:

$$\Delta a = \frac{1}{\sqrt{\sum w_i}} \frac{1}{\sqrt{\overline{x^2} - \bar{x}^2}} \qquad \Delta b = \frac{1}{\sqrt{\sum w_i}} \sqrt{\frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}} \qquad w_i = \frac{1}{\sigma_{y_i}^2}$$

4- Regression with standard deviation s_y and s_x



$s_{y1}=4 \quad s_{y2}=3.5 \quad s_{y3}=3 \quad s_{y4}=2.5 \quad s_{y5}=2 \quad s_{y6}=1.5 \quad s_{y7}=1$

$s_{x1}=0.1 \quad s_{x2}=0.2 \quad s_{x3}=0.3 \quad s_{x4}=0.4 \quad s_{x5}=0.5 \quad s_{x6}=0.6 \quad s_{x7}=0.7$

$s_1=4.0 \quad s_2=3.6 \quad s_3=3.4 \quad s_4=3.2 \quad s_5=3.2 \quad s_6=3.4 \quad s_7=3.6$

Case 4 : $a=4.98 \quad b=5.14 \quad s_a=0.695 \quad s_b=3.16$

But if x_i is assumed fixed, we transfer the dispersion

on y_i : $\qquad s_{y_i\,Total}^2 = s_i^2 = s_{y_i}^2 + a^2 s_{x_i}^2$

Everything happens as if only y_i had standard deviations s_i.

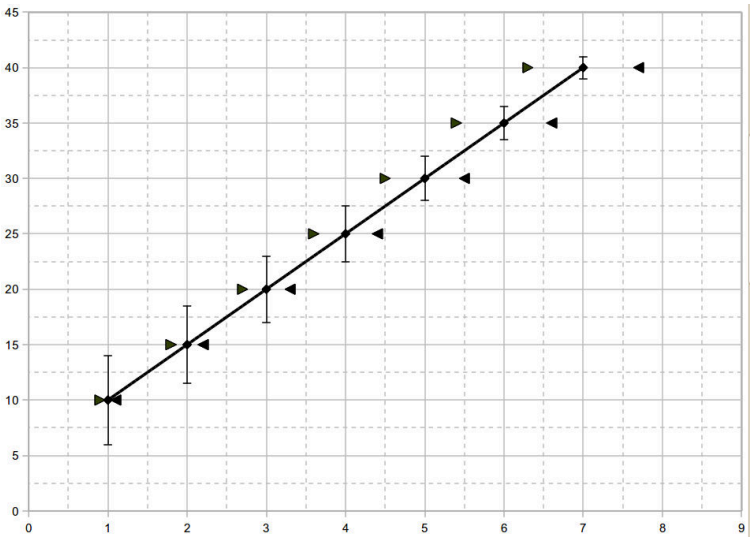
Hence the formulas :

$$s_a^2 = \sum_i \left(\frac{\partial a}{\partial y_i} \right)^2 s_i^2 \quad \text{and} \quad s_b^2 = \sum_i \left(\frac{\partial b}{\partial y_i} \right)^2 s_i^2$$

The derivatives are difficult to calculate (the weights depend on a), but we can easily evaluate them numerically. Also we commonly use the following estimates:

$$\Delta a = \sqrt{\frac{\sum w_i}{\Delta}} \quad \Delta b = \sqrt{\frac{\sum w_i x_i^2}{\Delta}} \quad w_i = \frac{1}{\sigma_{y_i}^2 + a^2 \sigma_{x_i}^2}$$

b) Discussion



Case 5 : a=5 b=5 s_a=0.696 s_b=3.16

Uncertainties in this case can come from measuring instruments. The dispersion of the first case on page 69 can come from accidental errors linked to the experiment or from a

fundamentally random phenomenon.

We would like that the fourth case on page 72 includes all the sources of uncertainty of the cases 1 and 5 :

$$s_{a_1} \simeq 0.443 \quad \text{and} \quad s_{a_5} \simeq 0.696 \quad \text{but} \quad s_{a_4} \simeq 0.695$$

Using the conventional formulas, we get the impression that the dispersion around the regression line is not included. In order to get the correct dispersion around the regression line, we perform the direct calculation with the propagation formula and I propose the small variations method:

$$s_a^{\,2} = \sum_i \left(\frac{\partial a}{\partial y_i} \right)^2 s_i^{\,2} \quad \text{and} \quad \left(\frac{\partial a}{\partial y_j} \right) = \lim_{\Delta y_j \to 0} \left(\frac{\Delta a}{\Delta y_j} \right)$$

Where $y_{i \neq j}$ are kept constant: $y_j + \Delta y_j \to a + \Delta a$

Δy_j is a small variation with respect of y_j , if Δy_j becomes smaller, $\frac{\Delta a}{\Delta y_j}$ stays constant (definition of the derivative).

We try to find the result for $s_{a\,5}$: we replace $y_1 = 10$ with $y_1 = 10.001$ and then from $a = 5$ we have after iteration $a = 4.999907$ then $\frac{\Delta a}{\Delta y_1} \simeq -0.093$.

We return $y_1 = 10$ and repeat the procedure for y_2 by replacing it with *15.001*. We obtain the following results:

	j=1	*j=2*	*j=3*	*j=4*	*j=5*	*j=6*	*j=7*
$\frac{\partial a}{\partial y_j}$	-0.093	-0.078	-0.049	-0.006	0.041	0.080	0.105

We find $s_{a_5} \simeq 0.696$, the same result than the previous method.

Let's do the same for $s_{a\,4}$:

	j=1	*j=2*	*j=3*	*j=4*	*j=5*	*j=6*	*j=7*
$\frac{\partial a}{\partial y_j}$	-0.096	-0.080	-0.050	-0.007	0.081	0.122	0.147

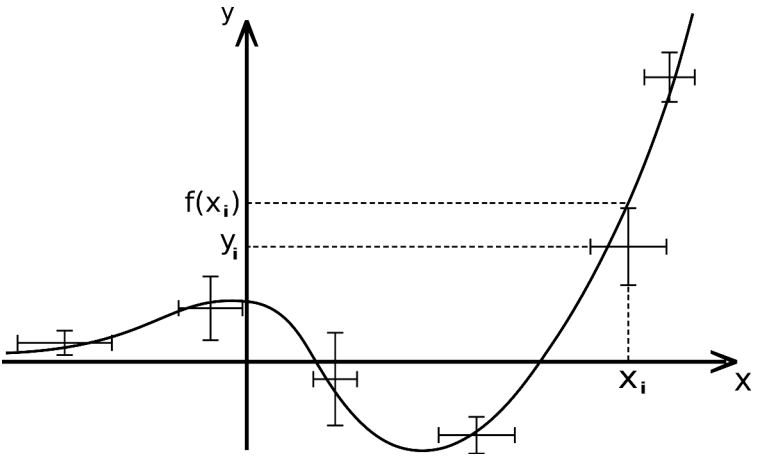
We then find that $s_{a_4} \simeq 0.786$. The result is significantly different from the classical estimate and now seems to incorporate all the sources of uncertainty.

In the exercise *Standard deviation proportional to y* on page 100 we study a particular case where we carry out the direct analytical calculation. The comparison can thus be extended.

D. Nonlinear regression

We generalize the weighted least squares method in the nonlinear case. The function is nonlinear with respect to x and also nonlinear with respect to the parameters. Although multiple regression has similar developments, it is not dealt with in this book [x].

1) Principle



We compare y_i to the value $f(x_i)$ given by the function sought: $y_i - f(x_i)$. The weight assigned to $(y_i - f(x_i))^2$ is inversely proportional to the variance of $y_i - f(x_i)$.

The quantities x_i and y_i are independent[13] from where:

13 Talking about independence between two variables as we look for a functional relationship between them may seem illogical. We refer here to the experimental determination of each quantity which is independent (in the sense of uncorrelated uncertainties).

$$V(y_i - f(x_i)) = V(y_i) + V(f(x_i))$$

By applying the variance propagation formula again:

$$V(f(x_i)) = f'(x_i)^2 V(x_i)$$

into $$S^2 = \sum_i w_i (y_i - f(x_i))^2$$

with $$w_i = \frac{1}{\sigma_{y_i}^2 + f'(x_i)^2 \sigma_{x_i}^2}$$ [14]

Then $\frac{\partial S^2}{\partial a_k} = 0$ allows to determine the parameters a_k of our function (by an analytical computation or a numerical resolution).
Each time we can return to a system of linear equations of the form $HA=B$, with H a square matrix and A the vector associated with the parameters. Then $A=H^{-1}B$, where H^{-1} is the inverse matrix of H.

The parameters uncertainties are given by the diagonal terms of the matrix H^{-1} :
$\sigma_{a_k}^2 = (H^{-1})_{kk} \sigma_r^2$ where σ_r^2 is the residual variance with respect to the estimated curve.

When there are no uncertainties on the data, the standard deviation of the residuals with p parameters are written:

14 S^2 is also called χ^2. If we assume the distributions Gaussian, the standard deviations can be replaced by the uncertainties using Student's t-values.

$$s_r = \sqrt{\frac{\sum (y_i - f(x_i))^2}{n - p}}$$

When w_i depends on the parameters we iterate the method until we can consider the weights to be constant. If we know the standard deviations of the data, the standard deviations of the parameters can be computed with the propagation formula, or they can be estimated using the same procedure used for linear regression with error bars on page 67.

2) Polynomial regression

In that case:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m = \sum_{i=0}^{m} a_i x^i$$

This function is non-linear with respect to x and linear with respect to the parameters.

Let us illustrate with the example of Cauchy's equation. This equation explains the phenomenon of light dispersion. It is an empirical relationship between the refractive index and wavelength:

$$n(\lambda) = a_0 + \frac{a_1}{\lambda^2} + \frac{a_2}{\lambda^4}$$

with the following data:

$\lambda(\mu m)$	0.6157	0.5892	0.5685	0.5152	0.4981
n	1.71276	1.71578	1.71852	1.72716	1.73060

The uncertainties on λ and n are initially neglected. What are the values and the uncertainties of a_0, a_1 and a_2?

We have the following parabolic regression:

$$f(x)=a_0+a_1 x+a_2 x^2 \quad \text{with} \quad f=n, \quad x=1/\lambda^2 .$$

$$S^2=\sum (y_i-a_0-a_1 x_i-a_2 x_i^2)^2 \quad \text{and} \quad \partial S^2/\partial a_k=0$$

we obtain:
$$\begin{cases} \bar{y}-a_o-a_1\bar{x}-a_2\overline{x^2}=0 \\ \overline{xy}-a_o\bar{x}-a_1\overline{x^2}-a_2\overline{x^3}=0 \\ \overline{x^2y}-a_o\overline{x^2}-a_1\overline{x^3}-a_2\overline{x^4}=0 \end{cases}$$

then $HA=B$ with:

$$H=\begin{pmatrix} 1 & \bar{x} & \overline{x^2} \\ \bar{x} & \overline{x^2} & \overline{x^3} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} \end{pmatrix}, \quad A=\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \quad \text{and} \quad B=\begin{pmatrix} \bar{y} \\ \overline{xy} \\ \overline{x^2y} \end{pmatrix}.$$

With a spreadsheet we can easily do the calculations:

$$H\simeq\begin{pmatrix} 1 & 3.3 & 11 \\ 3.3 & 11 & 38 \\ 11 & 38 & 135 \end{pmatrix}, \quad H^{-1}\simeq\begin{pmatrix} 4150 & -2530 & 376 \\ -2530 & 1546 & -230 \\ 376 & -230 & 34.3 \end{pmatrix},$$

$$B\simeq\begin{pmatrix} 1.7 \\ 5.7 \\ 19 \end{pmatrix} \quad \text{then} \quad A=H^{-1}B\simeq\begin{pmatrix} 1.68129 \\ 0.01135 \\ 0.00022 \end{pmatrix}.$$

For the uncertainties:

$$\Delta a_k=\sqrt{(H^{-1})_{kk}}\,t_{n-3}\,s_r \quad \text{with} \quad s_r=\sqrt{\frac{\sum (y_i-f(x_i))^2}{n-3}}$$

$s_r=1.87\times10^{-5}$ and with 95% confidence $t_{n-3}=4.30$

then

$a_0=1.6813\pm0.0017$, $a_1=(1.13\pm0.10)\times10^{-2}\ \mu m^2$

and $a_2=(2.2\pm1.5)\times10^{-4}\ \mu m^4$

Now, if we take the uncertainty $\Delta n=0.00004$ with 95% confidence, we have an uncertainty on y but not on x. Therefore the weights are constant: $w_i=w=1/\Delta n^2$. We obtain the same system of equations and we have the same results for H, H^{-1} B and A.

Here, in a similar way to the linear regression with error bars, we can estimate the dispersion of the residues by $1/\sqrt{w_i}$:

$$\Delta a_k=\sqrt{\frac{(H^{-1})_{kk}}{\sum w_i}} \qquad \text{then}$$

$a_0=1.6813\pm0.0012$, $a_1=(1.14\pm0.07)\times10^{-2}\ \mu m^2$

and $a_2=(2.2\pm1.0)\times10^{-4}\ \mu m^4$

With all the uncertainties, $\Delta n=0.00004$, $\Delta\lambda=0.005\mu m$ and $\Delta x=2\Delta\lambda/\lambda^3$, the weights depend on the observations. We nevertheless arrive at the same system of equations by considering the weights on an iteration constant, and we compute H, H^{-1}, B, and A with estimated parameters.

To calculate the means, we use the expression of the following weight:

$$w_i=\frac{1}{\Delta n^2+(a_1+2a_2x_i)^2\Delta x_i^{\,2}}$$

We thus obtain a first expression of the parameters. To iterate, we replace the previously used estimated parameters with these new parameters. We iterate as much as necessary to obtain a consistent value. Convergence is often very rapid:

Iteration	a_0	a_1	a_2
Estimated	1.5	0.005	0.0001
1	1.681282848208	0.011350379724	0.000219632215
2	1.681269875466	0.011358254795	0.000218466771
...			

Also here we consider: $\Delta a_k = \sqrt{\frac{(H^{-1})_{kk}}{\sum w_i}}$ then

$a_0 = 1.6813 \pm 0.0013$, $a_1 = (1.14 \pm 0.08) \times 10^{-2}\ \mu m^2$
and $a_2 = (2.2 \pm 1.2) \times 10^{-4}\ \mu m^4$

3) Nonlinear regression

We will start from p parameters a_k estimated and use an iteration method. On each iteration, the weights will be considered constant and the function will be linearized for each of n points on the set of parameters.
The function depends on the x_i and the parameters a_k.

Then we have $f_i = f(x_i ; a_k)$.

$$S^2=\sum_i w_i(y_i-f_i)^2 \quad \text{and} \quad \frac{\partial S^2}{\partial a_k}=-2\sum_{i=1}^{n} w_i\frac{\partial f}{\partial a_k}(y_i-f_i)=0$$

The first estimated parameters are noted $a_{0\,k}$. Following parameters will be noted $a_{j\,k}$ for the *jth* iteration. We will carry out a linearization with a small variation $\delta a_{0\,k}$ around $a_{0\,k}$ [15].

Let $\vec{a}=(a_{1,}\dots,a_k,\dots,a_p)$ and $\overrightarrow{\delta a}=(\delta a_{1,}\dots,\delta a_k,\dots,\delta a_p)$:

$$f(x_i;\vec{a}_0+\overrightarrow{\delta a_0})=f(x_i;\vec{a}_0)+\sum_{l=1}^{p}\delta a_{0l}\left(\frac{\partial f(x_i;\vec{a})}{\partial a_l}\right)_{a_{0l}}$$

or by reducing the notations :

$$f_i=f_{0,i}+\sum_l\left(\frac{\partial f}{\partial a_l}\right)_{0,i}\delta a_{0l}$$

$$\text{then} \quad \sum_i w_i\frac{\partial f}{\partial a_k}(y_i-f_{0,i}-\sum_l\left(\frac{\partial f}{\partial a_l}\right)_{0,i}\delta a_{0l})=0$$

$$\text{and} \quad \sum_i w_i\frac{\partial f}{\partial a_k}(y_i-f_{0,i})=\sum_{i,l} w_i\frac{\partial f}{\partial a_k}\frac{\partial f}{\partial a_l}\delta a_{0l}\,.$$

$$\text{We set} \quad H_{k,l}=\sum_i w_i\left(\frac{\partial f}{\partial a_k}\right)_i\left(\frac{\partial f}{\partial a_l}\right)_i=H_{l,k}\ ,$$

$$B_k=\sum_i w_i\frac{\partial f}{\partial a_k}(y_i-f_{0,i}) \quad \text{and} \quad A_l=\delta a_{0l}\,.$$

from where again $HA=B$ and $A=H^{-1}B$. We iterate until

15 MATH : We generalize the notion of derivative by adding the variations according to all the parameters:

$$f(x_0+\epsilon)\simeq f(x_0)+\epsilon(f'(x))_{x_0}$$

variation on the parameters is negligible and the values converge.
Here also, we use: $\sigma_{a_k}^2 = (H^{-1})_{kk}\,\sigma_r^2$.

Let us illustrate with a biological experiment, in which the relation between the concentration of the substrate [S] and the reaction rate in an enzymatic reaction is studied from data reported in the following table [ix] :

i	1	2	3	4	5	6	7
[S]	0.038	0.194	0.425	0.626	1.253	2.500	3.740
v	0.050	0.127	0.094	0.2122	0.2729	0.2665	0.3317

The model can be written as:

$$y = \frac{\alpha x}{\beta + x} \quad \text{where } v=y \text{ and } [S]=x.$$

We start with the estimations $\alpha_0=0.9$ and $\beta_0=0.2$

$$\frac{\partial f}{\partial \alpha} = \frac{x}{\beta + x} \;, \quad \frac{\partial f}{\partial \beta} = -\frac{\alpha x}{(\beta + x)^2} \;, \quad H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \;,$$

$$A = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

In the absence of given uncertainties on x_i and y_i :

$$H_{11} = \sum_{i=1}^{7} \left(\frac{\partial f}{\partial \alpha}\right)_{\alpha_0 i} \left(\frac{\partial f}{\partial \alpha}\right)_{\alpha_0 i} = \sum_i \left(\frac{\partial f}{\partial \alpha}\right)_{\alpha_0, i}^2 = \sum_i \left(\frac{x_i}{\beta_0 + x_i}\right)^2 ,$$

$$H_{12}=H_{21}=\sum_i \frac{\alpha_0 x_i^2}{(\beta_0+x_i)^3} \text{ and } H_{22}=\sum_i \frac{\alpha_0^2 x_i^2}{(\beta_0+x_i)^4}.$$

$$B_1=\sum_i \frac{x_i}{\beta_0+x_i}(y_i-f_{0,i}) \text{ and } B_2=-\sum_i \frac{\alpha_0 x_i}{(\beta_0+x_i)^2}(y_i-f_{0,i})$$

$$\text{with } f_{0,i}=\frac{\alpha_0 x_i}{\beta_0+x_i}$$

We can now put everything into a spreadsheet, which produces:

$$H_{11}\simeq 3.81\ ,\quad H_{12}=H_{21}\simeq -2.89\ ,\quad H_{22}\simeq 3.70\ ,$$

$$B_1\simeq -2.33 \text{ et } B_2\simeq 1.86\ .$$

and also

$$H^{-1}\simeq\begin{pmatrix}0.649 & 0.508\\ 0.508 & 0.668\end{pmatrix} \text{ then } A=H^{-1}B\simeq\begin{pmatrix}-0.567\\ 0.0602\end{pmatrix}.$$

From $\alpha_1=\alpha_0+\delta\alpha_0$ and $\beta_1=\beta_0+\delta\beta_0$ we have the new estimated parameters:

$$\alpha_1\simeq 0.9-0.567\simeq 0.333 \text{ and } \beta_1\simeq 0.2+0.06\simeq 0.260$$

We repeat the calculation from the start, this time using these values instead of α_0 and β_0 to iterate. We calculate new matrices and vectors H, H^{-1}, B and A, and obtain α_2

and β_2.
The results are shown in the following table:

Iterat°	α	β	$\delta\alpha$	$\delta\beta$	S^2
	0.9	0.2	-0.57	0.060	1.4454965815
1	0.333	0.26	0.0101	0.166	0.0150720753
2	0.343	0.43	0.0150	0.103	0.0084583228
3	0.358	0.53	0.0040	0.024	0.0078643240
4	0.3614	0.554	0.0004	0.0024	0.0078441826
5	0.36180	0.5561	0.00003	0.00018	0.0078440067
6	0.36183	0.5563	0.000002	0.000013	0.0078440057

After enough iterations: $H^{-1} \simeq \begin{pmatrix} 1.52 & 6.34 \\ 6.34 & 36.2 \end{pmatrix}$

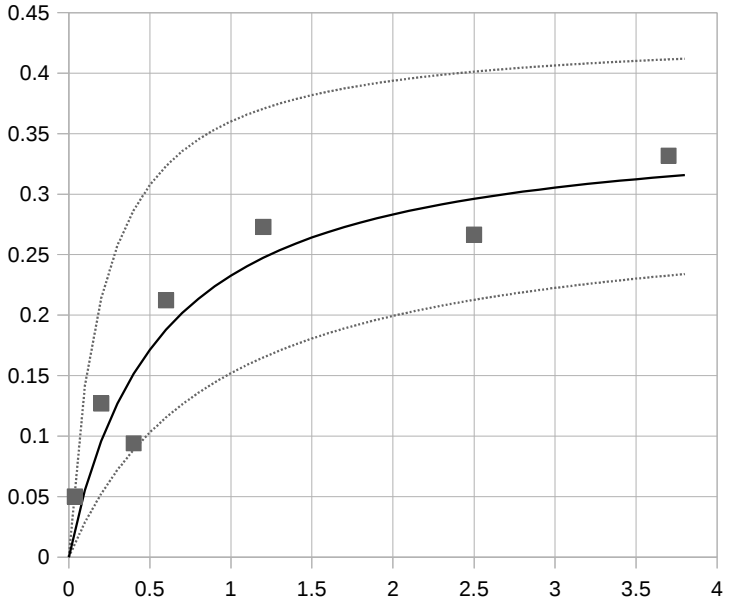
Calculate the uncertainties on the parameters:

$$s_r = \sqrt{\frac{S^2}{n-2}} \quad , \quad \Delta\alpha = \sqrt{(H^{-1})_{11}}\, t_{n-2}\, s_r \simeq \sqrt{1.52}.1.48.\sqrt{\frac{0.00784}{5}}$$

then $\Delta\alpha \simeq 0.07$, $\Delta\beta \simeq 0.35$.

Eventually: $\alpha = 0.36 \pm 0.07$ and $\beta = 0.56 \pm 0.35$
with 80% confidence.

The following graph shows the reaction rate as a function of substrate concentration. The squares are the experimental points, the solid line is the optimal curve and the dotted lines are the two extreme curves $f_{\alpha_{max},\beta_{min}}$ *and* $f_{\alpha_{min},\beta_{max}}$ *.*



To consider the uncertainties on the data we need to add weights. These are considered constant on each iteration. It will be necessary to calculate the derivative of f with respect to x which is present in the expression of the weight.

For standard deviations on the set of data we can calculate the standard deviations on the parameters using methods described on page 69 for the linear regression.

E. Exercises

Exercise 1 : Correlations *Answers p179*

1- We carry out nine experiments to obtain in each case three realizations of the quantities X_1, X_2 and X_3:

i=	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
X_1	$x_1^1 = -1$	-1	-1	0	0	0	1	1	1
X_2	$x_2^1 = 1$	0	-1	1	0	-1	1	0	-1
X_3	$x_3^1 = -1$	0	1	-1	0	1	-1	0	1

a) Determine the arithmetic means and standard deviations of these three quantities from these samples.

b) Plot X_2 as a function of X_1. The same for $X_3(X_1)$ and $X_3(X_2)$.

c) Calculate the correlation coefficients r_{12}, r_{13} and r_{23}. What comments can be we made from the results?

2- Same for the following data:

X_1	0	1	-1	2	0	-1	0	-2	1
X_2	1	2	-1	2	0	-2	-1	-2	1

3- Same for the following data:

X_1	-1	2	-2	0	-2	2	1
X_2	-1	0	2	-2	0	2	-1

Exercise 2 : Volumes *Answers p180*

We fill four beakers with 100 mL of water each with a pipette. To test the pipette and know the precise quantity of water, we weigh to the nearest decigram and obtain the following results for the different beakers in mL:

$$V_1=\{100.1 , 100.0 , 99.9 , 100.0\}$$

1- Calculate the mean and the standard deviation of V_1. Estimate the precision of the pipette with a 95% confidence.

We now fill two beakers and gather the contents of the two into one:

$$V = V_1 + V_2.$$

In the same order as V_1, we obtain the following measurements for V_2:

$$V_2=\{100.0 , 100.1 , 100.0 , 99.9\}$$

For example, for the third measurement, V_1=99.9 mL and V_2=100.0 mL.

2- Show that V_1 and V_2 are independent quantities.

3- Calculate the mean of V, its standard deviation and its uncertainty ΔV with 95% confidence.

4- Could you find this result with the uncertainty propagation formula?

(To improve the test it would take more measurements, but the principle remains the same, and the results remain valid because we have used the Student, considered decorrelated data and globally Gaussian packages. We should also take into account the uncertainties on the measures - resolution - in addition to their dispersion.)

Exercise 3 : Trees *Answers p181*

We want to measure the distance d between two trees. For this we have a stick of length one meter. From one tree to the other, we place the stick end to end a hundred times. For each displacement, we estimate an uncertainty of 1 cm.

What is the uncertainty estimated on the value of d?

Exercise 4 : The two-position method
Answers p182

We measure the focal length f' of a convex lens using the two-position method (also called Bessel method or displacement method).
An illuminated object is set up in front of a lens and a focused image forms on a screen. The distance D between the object and the screen is fixed. When $D>4f'$, there are two lens positions where the image is sharp. The distance between these two positions is denoted d. We then have the focal length of the lens by the relation $f'=(D^2-d^2)/4D$. We measure D=2000±10 mm and d=536±20 mm.

What is the uncertainty on f' ?

Exercise 5 : Refractive index *Answers p184*

We want to measure the index of refraction n_2 of a window glass. We perform the experiment of the refraction of a laser beam. According to Snell's law of refraction $n_1.\sin(i_1)=n_2.\sin(i_2)$, where n_i are the indices of the materials and i_i the angles of incidence and refraction. We get $n_1=n_{air}=1$, $i_1=30\pm1°$ and $i_2=20\pm2°$.

Determine n_2 with its uncertainty.
Exercise 6 : Cauchy's equation *Answers p185*

We want to measure the variation of the light index n as a function of the wavelength λ in a transparent material (phenomenon of dispersion). To carry out the experiment we need a prism, a sodium lamp and a goniometer. According the theory, in the visible spectrum, the index variation $n(\lambda)$ follows Cauchy's equation:

$$n(\lambda)=A+\frac{B}{\lambda^2}$$

The sodium line spectrum is known. For each line of wavelength λ_i, the corresponding index n is calculated using the formula of the prism:

$$n_i=\frac{\sin\left(\frac{A+D_{m,i}}{2}\right)}{\sin\left(\frac{A}{2}\right)}$$

D_m is the minimal deviation angle. $A=60°$ is the internal angle of the prism. These two angles are measured within 2' (1'=arc-minute and 1°=60').

We obtain the following values:

$\lambda(nm)$	615.7	589.2	568.5	515.2	498.1
Color	red	yellow	green-yellow	green	blue-green
D_m	57° 49.5'	58° 9'	58° 28'	59° 26.5'	59° 50'
n	1.71276	1.71568	1.71852	1.72716	1.73060

1- Determine the uncertainty for n (Δn is assumed constant).
2- Using Cauchy's equation find A, B and their respective uncertainties.

What is the value of the regression coefficient *r*?

3- We hypothesize that plotting *n* as a function of $1/\lambda$ or $1/\lambda^3$, will produce a better alignment of the points. We want to verify that the variation in $1/\lambda^2$ is indeed the best of the polynomial relations. For this we take the form:

$$n(\lambda)=A+B.\lambda^{\alpha}$$

Propose a method for determining A, B and α.
We can verify the model because we have α with its uncertainty.

Exercise 7 : Wall *Answers p187*

There is a wall with area S=72 m². The outside temperature is 6°C and the internal temperature is maintained at 18°C. This wall is 50 cm thick and consists of e_p=40 cm of compressed straw (thermal conductivity λ_p=45 mW/K/m) and e_e=10 cm of coating (λ_e=200 mW/K/m). The values for λ, thicknesses, and the temperature are rounded to within 10%, the nearest cm and the nearest half degree respectively.

1- Determine the thermal resistance with its uncertainty of the straw for this wall (R.λ.S=e)

2- Repeat for the coating.

3- Taking into account that thermal resistances associate like electrical resistors in series, determine the total thermal resistance of the wall with its uncertainty.

4- What should be the minimum heating power of the house to compensate for the losses by the walls? (ΔT=R.Φ)

Exercise 8 : Insulation and inertia *Answers p188*

In a low-energy house, the thermal resistances e/λ are measured per square meter. The thermal resistances are 8 m².K/W for the roof, 4 m².K/W for walls and floor, and 1 m².K/W for door and window frames. The R-values are known to the nearest 10%. The house area for the floor, the roof, the walls and the frames are 36 m², 54 m² , 82 m² and 8 m² respectively.

1- The equivalent resistances of the roof, the walls, the floor and frames are in parallel. Determine the total thermal resistance (in K/W) with its uncertainty.

Outdoor and indoor temperatures are constant.

2- What should be the minimum heating power of the house to keep the indoor temperature constant while offsetting the losses?

3- We switch off the heating and measure the temperature over time to obtain the following results:

t in hours	0	1	2	4	5	6	8	9	10
T in °C	18	16	14	12	11	10	9	9	8

Explain why the decay cannot be linear. We consider an exponential decay: $T(t)=a.\exp(-t/\tau)+b$. Calculate b, a and τ, with their uncertainties.

4- The house is insulated from the outside. The lost heat flux corresponds to a decrease in the energy stored in the house. This inertia is due to the thermal capacity C of the materials (J/K).

a) By reasoning on an infinitesimal time interval dt, find the differential equation verified by T(t) and the expression of question 3.

b) What is the relationship between τ, R and C?

Determine C and its uncertainty.

In question 3 we could also take into account measurement uncertainties: time can be considered as perfectly known and temperature is measured using a thermometer with graduations all the Celsius degrees.
For simplicity we considered that the outside temperature remains constant (to account for day/night variations we would consider sinusoidal variations and a harmonic approach).

Exercise 9 : Yield *Answers p191*

Specific quantities of fertilizer are spread on fields and we obtain the following yields:

Fertilizer (kg/ha)	100	200	300	400	500	600	700
Yield (Quintal/ha)	41	44	53	63	66	65	78

1- Determine the regression line that passes through the scatterplot. Slope, intercept and uncertainties with a confidence of 95%.

2- For 550 kg/ha of fertilizer, estimate the yield.

3- Repeat the calculation without fertilizer.

4- If a farmer spreads 250 kg/ha of fertilizer, what is the probability that he will get 40 to 60 quintals of grain?

Exercise 10 : Study of a battery *Answers p193*

To determine the open-circuit voltage E and the internal resistor r, we measure for the battery different values of U and I with a voltmeter and an ammeter (U = E - r.I) :

range for U :	unit : V accuracy ±0.05% ±0.003									
U (V)	4.731	4.731	4.730	4.728	4.724	4.724	4.722	4.721	4.719	4.716
range for I :	unit : µA accuracy ±0.2% ±0.03			unit : mA accuracy ±0.2% ±0.0003						
I	92.83	115.45	152.65	0.2352	0.4686	0.5200	0.5841	0.6661	0.7750	0.9264

1- Without error bars: determine E±ΔE and r±Δr.

2- Repeat including the uncertainties on U and I indicated in the instructions of the multimeter manufacturer.

Exercise 11: Thin lens formula *Answers p195*

We want to measure the focal length f' of a converging lens. At point O, the lens forms a sharp image at A' of an object at A. We measure OA, OA' and their uncertainties (are included all the sources of uncertainty: geometry, focusing and modeling). We consider that the lens verifies the thin lens formula:

$$1/OA' + 1/OA = 1/f'.$$

Determine f' using a linear regression. Table on the next page.

Experimental Data (mm) :			
OA	ΔOA	OA'	ΔOA'
635	5	*150*	*15*
530	5	*160*	*17*
496	5	*164*	*15*
440	5	*172*	*18*
350	5	*191*	*20*
280	5	*214*	*25*
210	5	*292*	*28*
150	5	*730*	*102*

Theory

Exercise 12 :

For simple regression, show that:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = n(\overline{x^2} - \bar{x}^2)$$

Exercise 13 :

For simple regression, show that we can also write:

$$s_b = s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \ .$$

Exercise 14 :

For simple regression, show that: $\Delta b = \sqrt{\overline{x^2}}\, \Delta a$.

Exercise 15 : Asymptotes *Answers p198*

For the simple regression, show that, above and below the regression line, extreme line, confidence and prediction curves, have the same asymptote when x_o becomes large.

Exercise 16 : Confidence Interval and Prediction Interval for Linear Regression with error bars
Answers p198

Regression with error bars

1- Give the expressions of $\overline{x}$ and $\overline{x^2}$ using the w_i weights. Could you find a new expression of Δ with $\overline{x}$ and $\overline{x^2}$ (p67) ?

Analogy

2- From the confidence and prediction intervals for simple regression (p62 and following), use an analogy to determine the following formulas:

Confidence: Prediction:

$$\Delta y_o = \frac{1}{\sqrt{n\sum w_i}}\sqrt{1+\frac{(x_o-\overline{x})^2}{\overline{x^2}-\overline{x}^2}} \qquad \Delta y_o = \frac{1}{\sqrt{n\sum w_i}}\sqrt{1+n+\frac{(x_o-\overline{x})^2}{\overline{x^2}-\overline{x}^2}}$$

3- Determine the y-distances to the regression line for extreme lines, confidence curves, and prediction curves when x_o becomes large.
In a previous exercise we showed that for the linear regression, the asymptotes are the same. By analogy what should we set so that it is the same in regression with bars of errors?

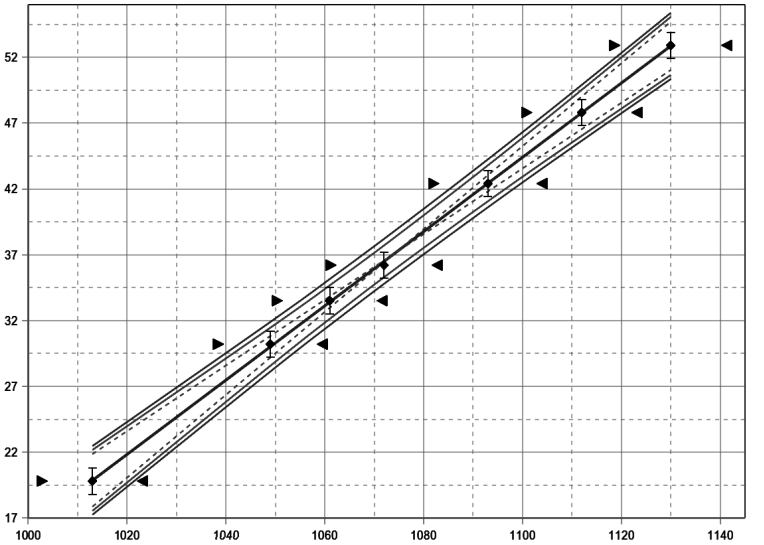
Show that we then have the following formulas:

Confidence: Prediction:

$$\Delta y_o = \frac{1}{\sqrt{\sum w_i}} \sqrt{1 + \frac{(x_o - \overline{x})^2}{\overline{x^2} - \overline{x}^2}} \qquad \Delta y_o = \frac{1}{\sqrt{\sum w_i}} \sqrt{2 + \frac{(x_o - \overline{x})^2}{\overline{x^2} - \overline{x}^2}}$$

The formulas obtained by analogy are thus empirical. Yet, while they seem experimentally coherent, they require confirmation by theoretical demonstration.

Confidence and prediction curves for the absolute zero experiment: T (°C) as a function of P (hPa)



Exercise 17 : Other expressions *Answers p200*

For regression with error bars give the expression of a, b, Δa and Δb as a function of $\overline{x}$, $\overline{y}$, $\overline{xy}$, $\overline{x^2}$ and $\overline{y^2}$. Compare with simple regression.

Exercise 18 : Least squares method

Answers p201

Demonstrate by the least squares method the expressions of a and b:

1- For simple linear regression.

2- For linear regression with error bars. The Δx_i and Δy_i are considered small with respect to x_i and y_i.

Proof of the expressions of Δa and Δb for the simple regression:

Method 1 :

1- Show that $a=\sum_i p_i y_i$ with $p_i=\frac{x_i-\bar{x}}{\sum (x_i-\bar{x})^2}$.

2- Deduce from this formula of a its variance $V(a)$ [16].

Method 2 :

Use the propagation of standard deviation formula.

For simple linear regression, is it possible to find a, b, Δa and Δb using the generalized regression matrix method?

Exercise 19 : Expectation of a

Answers p203

For linear regression, we denote α and β as the parameters of the population: $E(y_i)=\alpha x_i+\beta$.
a and b are the estimated parameters from a sample: $\hat{y}_i=a x_i+b$
Show that we have an unbiased estimator for α, so $E(a)=\alpha$.

16 MATH : E(X+Y) = E(X) + E(Y). If X and Y are two independent variables: V(X+Y) = V(X) + V(Y).

Exercise 20 : Standard deviations proportional to y *Answers p203*

We consider the particular theoretical case where the standard deviations of the linear regression are proportional to y : $\sigma_{y_i}=k\,y_i$. This case, where the relative uncertainties are constant, is experimentally common. We are interested in the slope a.

1- Show that: $$a=\frac{\sum\frac{1}{y^2}\sum\frac{x}{y}-\sum\frac{x}{y^2}\sum\frac{1}{y}}{\sum\frac{1}{y^2}\sum\frac{x^2}{y^2}-\left(\sum\frac{x}{y^2}\right)^2}$$

2- Express $\frac{\partial a}{\partial y_j}$ (long calculation).

3- Calculate s_a using the expressions found for the following two datasets:

x_i	1	2	3	4	5	6	7
1 : y_i	10	15	20	25	30	35	40
2 : y_i	8.286	17.286	18.286	27.286	28.286	37.286	38.286

(We let $k=0.1$).

Find these results again with the numerical method (evaluation of the derivatives with the small variations method).

Compare the values obtained by the classical method.

Exercise 21 : Interpretation of the expression of w_i *Answers p205*

Graphically justify the position of a in the expression of w_i.

Exercise 22 : Decomposition into Gaussians

Answers p206

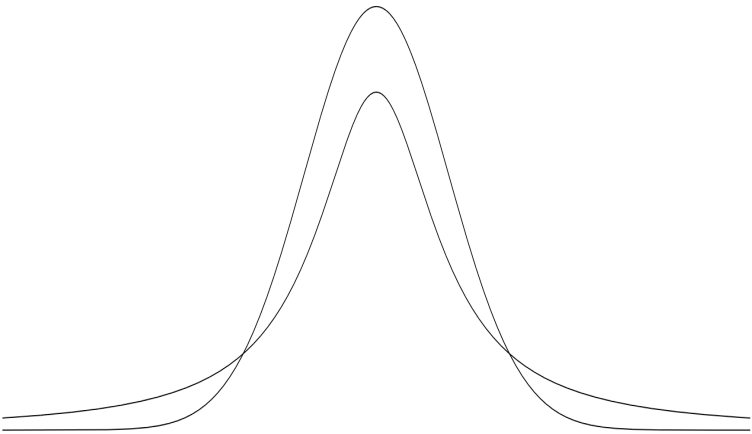
A factory manufactures nails. With a caliper, we measure the size of 48 nails:

Size	59.97	59.98	59.99	60.00	60.01	60.02	60.03	60.04	60.05
Quantity	2	4	6	5	5	10	9	6	1

Two machines manufacture the nails. The nails made by one of the machines are not exactly the same size. We assume that the size is distributed according to a Gaussian distribution. Determine the parameters of the Gaussian distributions of the nails manufactured by each machine (maximum, mean and standard deviation). How many nails have each machine produced?

To check your correlation calculations you can use the spreadsheet function of OOo :
COEFFICIENT.CORRELATION(*;**).
Courbes : Insertion>Diagramme...>XY(Dispersion)>etc.
For matrix calculations, inversion of matrices : INVERSEMAT(*), product of matrices : PRODUITMAT(*;**).

You can use the file IncertitudesLibresOOo32.ods on the website www.incertitudes.fr to realize regressions. Sheet 2, simple regression and sheet 3 with error bars.



III. PROBABILITY DISTRIBUTIONS

We list different laws of discrete and continuous probabilities. We verify they are distributions of probability and we give their expectation $\mu = E(X)$ and variance $V = E[(X-\mu)^2]$.

The variance can also be calculated with the formula $V = E[X^2] - E[X]^2$ and then we have the standard deviation $\sigma = \sqrt{V}$.

We can calculate other moments that allow us to determine, for example, the symmetry and the flatness of the probability distribution.

Moments: $\mu_k = E(X^k)$

Normalized moments are instructive because they characterize the form of the distribution:

$$\beta_{k-2} = E\left[\left(\frac{X-\mu}{\sigma}\right)^k\right] \quad \text{or} \quad \beta_{k-2} = \frac{\mu_k}{\sigma^k}$$

β_1 : Skewness (third standardized moment)

β_2 : Kurtosis

We also consider the sum of independent random variables: $Z = X + Y$.

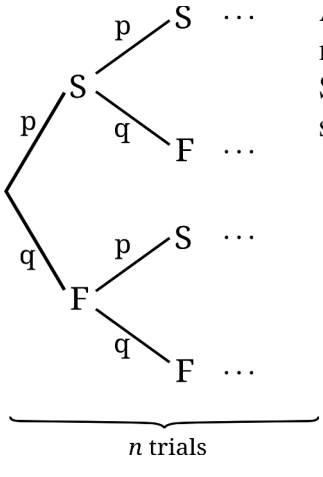
A. Discrete Random Variables

1) Binomial distribution

We consider the number of successes in a sequence of *n* identical and independent experiments. Each trial has two outcomes named success and failure (Bernoulli trial). Each success occurs with the probability $p=P(S)$. The two parameters are *n* and *p* and we write $\mathcal{B}(n,p)$. We want to determine the probability to obtain exactly *k* successes out of *n* trials. A path containing *k* successes has *n-k* failures and its probability is $p^k q^{n-k}$, where $q=1-p$ is the probability of a failure.

Then we have to count the number of paths where *k* successes occur, there are different ways of distributing *k* successes in a sequence of *n* trials. *n* choice for the position of the first success, *n-1* for the second and *n+1-k* for the *k*th success:

$n(n-1)...(n+1-k)=n!/(n-k)!$ possibilities[17].



After we divide by *k!* to remove multiple counts (for example S_1S_2F and S_2S_1F correspond to the same path). From where:

$$P(X=k) = \binom{n}{k} p^k q^{n-k}$$

with $\binom{n}{k}=\frac{n!}{(n-k)!k!}$

(often read aloud as « n choose k »)

and $\sum_{k=0}^{n} P(X=k)=1$

17 *n!*, is said « n factorial » with $n!=1\times2\times3...\times n$.

we show that $E(X)=np$ and $V(X)=npq$.

The sum of two independent binomial laws of the same p is also a binomial law. The sum of a $\mathcal{B}_x(n_1,p)$ and a $\mathcal{B}_y(n_2,p)$ is a $\mathcal{B}_z(n_1+n_2,p)$.

To determine the distribution of Z sum of X and Y we use the following property for discrete distributions:

$$P(Z=k)=\sum_i P([X=i]\cap[Y=k-i])$$

2) Geometric distribution

We consider a random experiment with exactly two possible outcomes : S="success" and F="failure" ; $p=P(S)$ and $q=1-p=P(F)$. We repeat independent trials until the first success. Let X be a random variable of the number of trials needed to get one success. We denote the distribution by $\mathcal{G}(p)$.

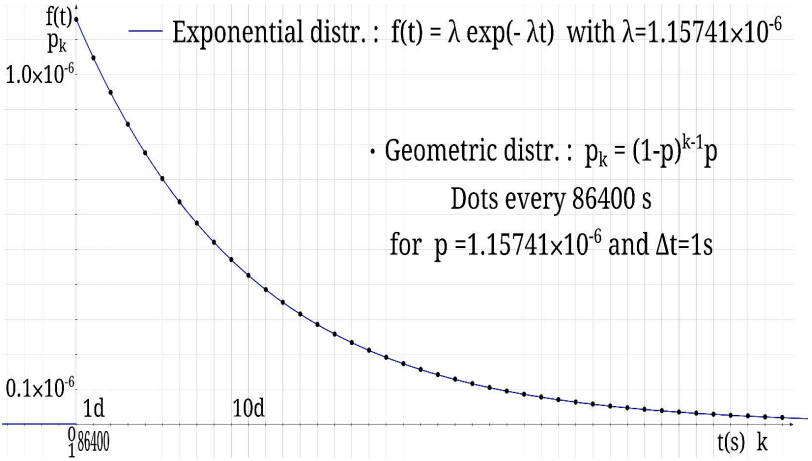
$$P(X=k)=q^{k-1}p \quad , \quad \sum_{k=1}^{\infty} P(X=k)=1$$

$$E(X)=1/p \quad \text{and} \quad V(X)=q/p^2$$

• Example: We have a box with two white balls and a black ball inside. The balls are indistinguishable to the touch. We draw a ball, if we get a black ball we stop, otherwise we return the ball to the box and we repeat. What is the probability of drawing the first black ball in the third draw?

Answer: $P(X=3)=(2/3)^2 1/3=4/27$.

• Application: This distribution can be used like a discrete model of lifetime. For example, consider a cell which every day has a chance in ten to die. If it did not die the next day, its probability of dying the next day did not change: there is no aging[18]. Every day the cell can escape death and it can live forever. Rather than making a decision every day, we can decide his death every hour with a probability q' of survival, like this $q'^{24}=q$ and $q'\simeq 0.996$. Or even a choice every second, we then have different geometric distributions witch modelize the same reality with $p\to 0$ and $q\to 1$.



From a discrete distribution we reach a continuous one. The elapsed time since the beginning of the experiment is $t=k\Delta t$.

Let's look at the cumulative distribution function:

18 The geometric distribution has a remarkable property, which is known as the memoryless property.

$$P(X \leqslant k) = \sum q^{k-1} p = \sum e^{\ln(1-p).(k-1)} p \simeq \sum p e^{-pk}$$

We have used a Taylor series, indeed $t \gg \Delta t$ and $k \gg 1$.

Then $P(X \leqslant k) \to \int f(t)\,dt$, $\sum p e^{-\frac{p}{\Delta t}t} \to \int \lambda e^{-\lambda t}\,dt$ so, if we look at the limit, the geometric distribution becomes the exponential one with $\lambda = p/\Delta t$. Here for our cell λ is about 1.16×10^{-6} per second.

The sum of two independent geometric random variables $\mathcal{G}(p)$ is a negative binomial distribution $\mathcal{BN}(2,p)$, rank for obtaining two successes.

For a binomial distribution the number of trials is fixed and we look at the number of successes, in the case of the negative binomial distribution it is the opposite we are interested in the number of trials necessary to achieve a number of successes fixed in advance. $\mathcal{BN}(r,p)$ is the probability distribution of the rank of the r-th success[19].

Then $\mathcal{G}(p) = \mathcal{BN}(1,p)$.

3) Poisson distribution

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval. The events occur with a known frequency independent of the

19 There is also another definition: number of failures preceding the r-th success. The relations remain true by redefining the geometric distribution in the same way.

time elapsed since the previous events[20]. If λ is the number of times the event occurs on average over a given interval then the probability that the event occurs k times over this time interval is:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \quad , \quad \sum_{k=0}^{\infty} P(X=k)=1$$

$$E(X)=\lambda \quad \text{and} \quad V(X)=\lambda$$

- Example: During the Perseids meteor shower of August 1911 the hourly rate of shooting stars was 50. What was the probability of seeing exactly 7 meteors in 12 minutes?

  Answer:

  $$\lambda=\frac{12}{60}50=10 \quad \text{and} \quad P(X=7)=\frac{10^7}{7!}e^{-10}\simeq 9\%$$

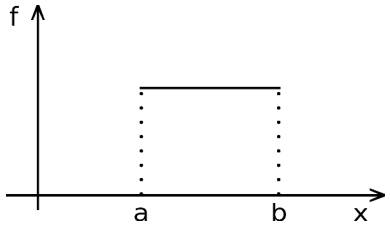
The sum of two independent Poisson distributions $\mathcal{P}_X(\lambda_1)$ and $\mathcal{P}_Y(\lambda_2)$ is a Poisson distribution $\mathcal{P}_Z(\lambda_1+\lambda_2)$.

20 or space traveled.

B. Continuous Random Variables

1) Uniform distribution

$$\begin{cases} x<a: & f(x)=0 \\ a\leqslant x\leqslant b: & f(x)=\dfrac{1}{b-a} \\ x>b: & f(x)=0 \end{cases}$$



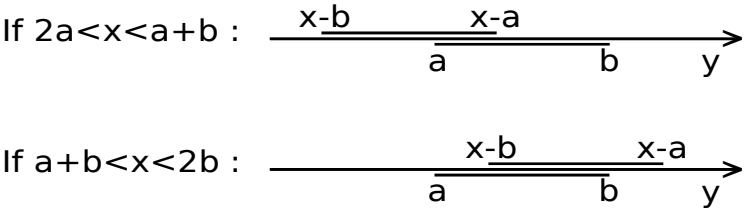
$$E(X)=\frac{a+b}{2} \quad \text{and} \quad V(X)=\frac{(b-a)^2}{12}$$

To determine the distribution of Z sum of two continuous independent random variables X and Y we use a convolution:

$$f_Z(x)=\int_{-\infty}^{+\infty} f_X(y)f_Y(x-y)\,dy$$

Consider the sum of two independent uniform distributions $\mathcal{U}(a,b)$. The integrand is nonzero if:

$$a<y<b \quad \text{and} \quad a<x-y<b \quad \text{then} \quad x-b<y<x-a$$

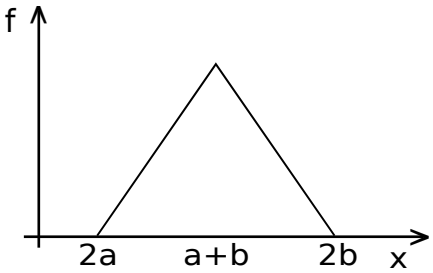


$$\text{If} \quad 2a<x<a+b$$

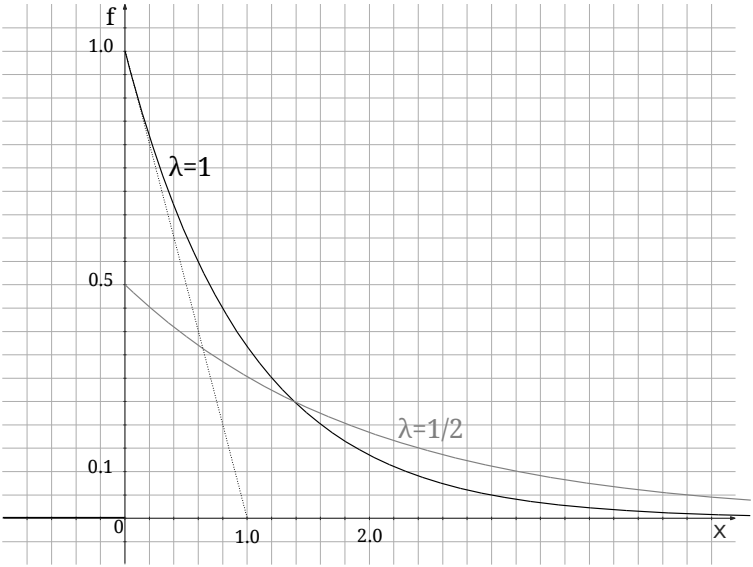
$$\text{then} \quad f_Z(x)=\int_a^{x-a}\frac{1}{(b-a)^2}\,dy=\frac{x-2a}{(b-a)^2}$$

$$\begin{cases} x<2a: \ f_Z(x)=0 \\ 2a\leqslant x<a+b: f_Z(x)=\dfrac{x-2a}{(b-a)^2} \\ a+b\leqslant x\leqslant 2b: f_Z(x)=\dfrac{2b-x}{(b-a)^2} \\ x>2b: \ f_Z(x)=0 \end{cases}$$

We obtain a triangular distribution:



2) Exponential distribution



$$\begin{cases} t \geqslant 0: \ f(t) = \lambda \, e^{-\lambda t} \\ t < 0: \ f(t) = 0 \end{cases} \qquad E(T) = \frac{1}{\lambda} \quad \text{et} \quad V(T) = \frac{1}{\lambda^2}$$

The exponential distribution satisfies the memoryless property: $P(T > b) = P_{T>a}(T > a+b)$. This distribution is widely used to model waiting times. It is also used to model lifetime without aging : like the lifetime for a particle decay.

The mean life $E(T)$ and the half-life $t_{1/2}$, $P(T > t_{1/2}) = 0.5$, are two different times.

The distribution of the sum of two independent exponential distributions is not an exponential distribution.

3) Normal distribution

The normal or Gaussian distribution has previously been described page 20.

The sum of two independent Gaussian distributions $\mathcal{N}_X(\mu_1,\sigma_1^2)$ and $\mathcal{N}_Y(\mu_2,\sigma_2^2)$ is the Gaussian distribution $\mathcal{N}_Z(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

4) Student's t-distribution

The t-distribution is express with the number of degrees of freedom k and the gamma function (function described in the mathematical tools).

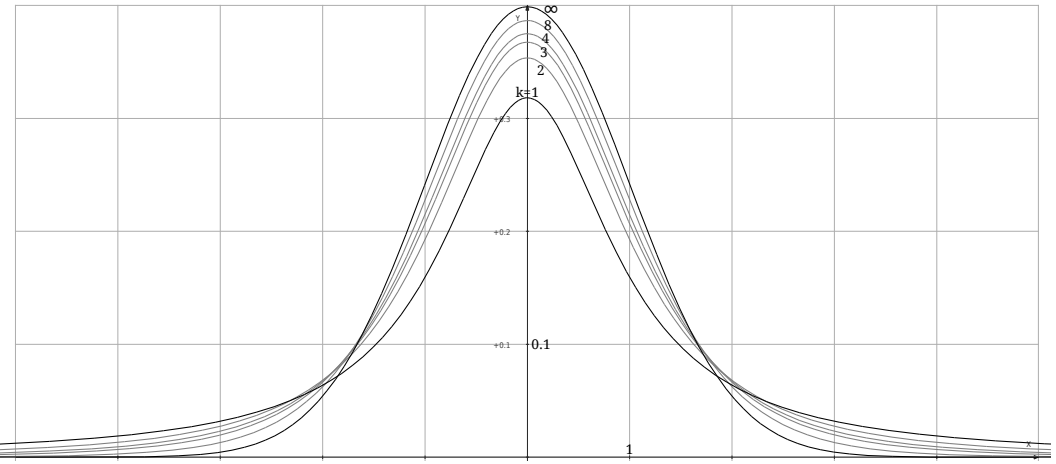
$$\text{For} \quad k \geq 1 \quad , \quad f_k(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}}$$

As the number of degrees of freedom grows, the t-distribution approaches the normal distribution $\mathcal{N}(0,1)$:

$$\lim_{k \to +\infty} f_k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\text{Variance:} \quad V_k = \frac{k}{k-2} \quad \text{if} \quad k \geq 3 .$$

$$\text{Kurtosis:} \quad \beta_k = 3\frac{k-2}{k-4} \quad \text{if} \quad k \geq 5 .$$



In exercises the expressions of the first Students are expressed, we show that $\int_{-\infty}^{+\infty} f_k(x)\,dx = 1$, we compute the expression of the variance and finally we show on examples that the sum of two independent Students is not a Student.

5) Chi-squared distribution

Let consider k independent normal distributions $\mathcal{N}(0,1)$: T_1, T_2,… and T_k. The sum $X_k = T_1^2 + T_2^2 + ... + T_k^2$ follows a χ^2-distribution with k degrees of freedom.

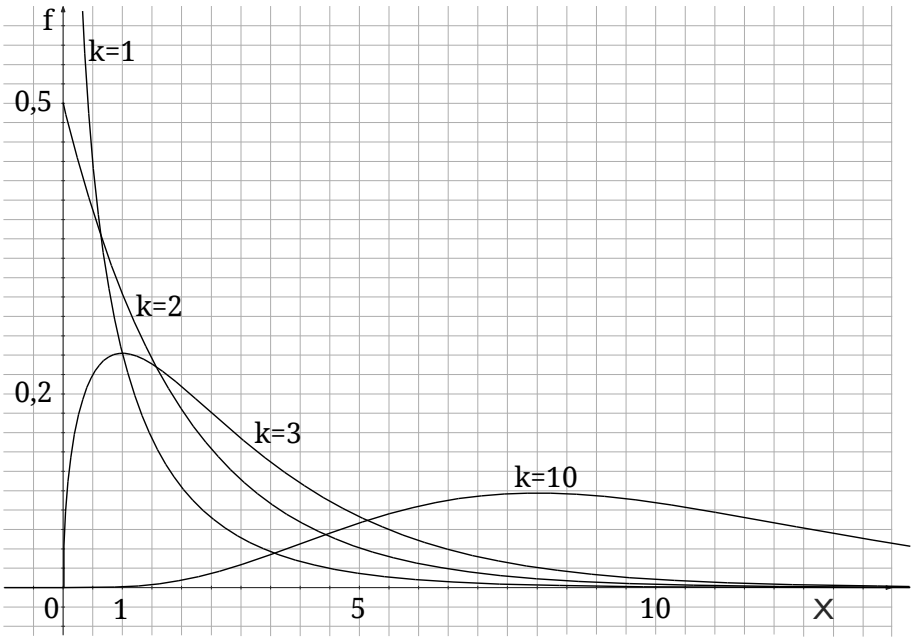
$$E(X_k) = \sum_{i=1}^{k} E(T_i^2) = k\left(V(T) + E(T)^2\right) = k$$

For $k \geq 1$ and $x \geq 0$, $f_k(x) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$

Expectation : $E_k = k$ Variance : $V_k = 2k$

Skewness : $\beta_{1,k} = \sqrt{8/k}$

Kurtosis : $\beta_{2,k} = 3 + 12/k$



The chi-squared distribution converges to a normal distribution for large k. Normal distribution of expectation k and variance $2k$.

C. Function of a continuous distribution

We have a random variable Y defined as a function of a continuous random variable X: $Y=\varphi(X)$.

We know the law of X and we want to determine the law of Y. We use the cumulative distribution functions F and then we consider the derivative of F_X to get the probability density function f.

Cumulative distribution function of X :

$$F_X(x)=P(X\leq x)=\int_{-\infty}^{x} f_X(x)\,dx$$

We consider the case where $\varphi(x)$ is a strictly monotonic function. Then Y is a continuous distribution.

Cumulative distribution function of Y :

$$F_Y(y)=P(Y\leq y) \quad \text{and} \quad f_Y(y)=\frac{d\,F_Y(y)}{dy}$$

We try to express $F_Y(y)$ with F_X.

1) Case where $\varphi(x)=\ln(x)$: $x>0$ and $Y=\ln X$

$$F_Y(y)=P(Y\leq y)=P(\ln X\leq y)=P(e^{\ln X}\leq e^y)=P(X\leq e^y)$$

We have applied the inverse function $\varphi^{-1}(x)=e^x$.

The exponential function is strictly increasing, so the direction of the inequality has been preserved.

so $\quad F_Y(y)=F_X(e^y) \quad$ and $\quad f_Y(y)=e^y f_X(e^y)$

• Example: X is a uniform distribution $\mathcal{U}(1;2)$. What probability distribution is Y?

We have $f_X(x)=\begin{cases} 0 & x\leqslant 1 \\ 1 & 1<x<2 \\ 0 & x\geqslant 2 \end{cases}$ then if $e^y\leq 1$, $y\leq 0$ and $f_Y(y)=0$. We continue like this for the two other cases and we obtain the distribution of Y:

$$f_Y(y)=\begin{cases} 0 & y\leqslant 0 \\ e^y & 0<y<\ln 2 \\ 0 & y\geqslant \ln 2 \end{cases}$$

2) Case where φ(x)=ax+b: $a\neq 0$ and $Y=aX+b$

If a>0 :

$$F_Y(y)=P(Y\leq y)=P(aX+b\leq y)=P\left(X\leq \frac{y-b}{a}\right)$$

A linear function is strictly increasing for *a* strictly positive, the direction of the inequality is not changed.

Then $F_Y(y)=F_X\left(\frac{y-b}{a}\right)$ and $f_Y(y)=\frac{1}{a}f_X\left(\frac{y-b}{a}\right)$

If a<0 :

$$F_Y(y)=P\left(X\geq \frac{y-b}{a}\right)=1-P\left(X<\frac{y-b}{a}\right)=1-F_Y\left(\frac{y-b}{a}\right)$$

and $f_Y(y)=-\frac{1}{a}f_X\left(\frac{y-b}{a}\right)$ then $f_Y(y)=\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$

• Example 1: X is a uniform distribution $\mathcal{U}(0,1)$. What probability distribution is Y?

We have $f_X(x)=\begin{cases} 0 & x\leqslant 0 \\ 1 & 0<x<1 \\ 0 & x\geqslant 1 \end{cases}$, so if $\frac{y-b}{a}\leq 0$ and $a>0$ then $y\leq b$ and $f_Y(y)=0$. We continue like this for the two other cases and we obtain the distribution of Y:

$$f_Y(y)=\begin{cases} 0 & y\leqslant b \\ \frac{1}{a} & b<y<a+b \\ 0 & y\geqslant a+b \end{cases}$$

$U(0,1)\rightarrow U(b,a+b)$

If φ(x)=(b-a)x+a and a<b:

$U(0,1)\rightarrow U(a,b)$

• Example 2: X is a Gaussian distribution $\mathcal{N}(0,1)$. We can find again a $\mathcal{N}(\mu,\sigma)$ distribution with Y= σX+μ. In general, by applying an linear function we obtain a distribution of the same kind.

3) Case where φ(x)=x²: Y=X^2 and y>0

$$F_Y(y)=P(X^2\leq y)=P(-\sqrt{y}\leq X\leq\sqrt{y})=F_Y(\sqrt{y})-F_Y(-\sqrt{y})$$

and $f_Y(y)=\frac{1}{2\sqrt{y}}(f_X(\sqrt{y})+f_X(-\sqrt{y}))$ if y>0 else zero.

4) Case where $\varphi(x)=e^x$: $Y=e^X$ and $y>0$

$$F_Y(y)=P(e^X\leq y)=P(X\leq \ln y)=F_Y(\ln y)$$

and $f_Y(y)=\frac{1}{y}f_X(\ln y)$ if y>0 else zero.

D. Numerical simulation

We simulate continuous and discrete probability distributions using computers. For this purpose we use uniform distributions created by random number generation algorithms:

Continuous uniform distributions U(0,1) : Ran# on a pocket calculator, ALEA() in the LibreOffice spreadsheet, etc.

Discrete uniform distributions U(i,j) : for example, rand(i,j) in PHP language. rand(0,999)/1000 simulates a uniform continuous law discretized to the thousandth.

• Inverse transform sampling : the inverse transformation method takes uniform samples between 0 and 1 from U. We express a X distribution as a function of U by inversion of the cumulative distribution function F_X. With F strictly increasing: $X=F^{-1}(U)$.

Case of an exponential distribution: $f_X(x)=\lambda e^{-\lambda x}$ if x>0

else zero, then $F_X(x)=\int_{-\infty}^{x} f_X(x)\,dx=1-e^{-\lambda x}=y$ if x>0 else zero. So $x=-\ln\frac{(1-y)}{\lambda}$ and finally we simplify, knowing that *1-U* and *U* have the same distribution.

Simulation of an exponential distribution: $X=-\frac{\ln U}{\lambda}$

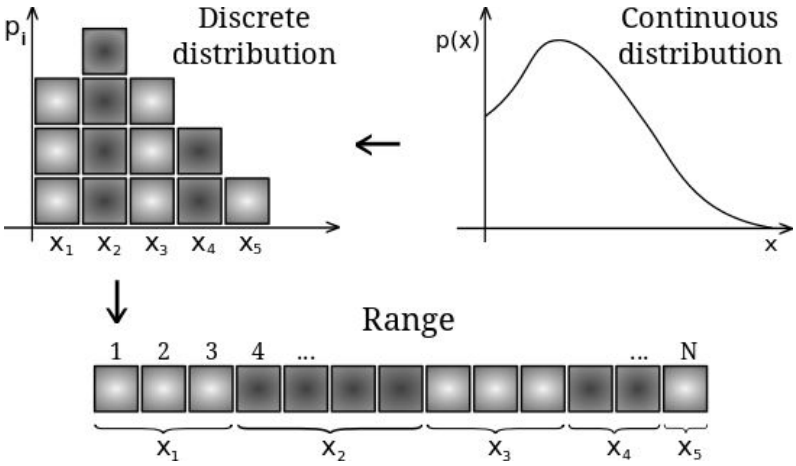
• Simulation of two independent normal $\mathcal{N}(0;1)$ distributions X_1 and X_2 from two independent uniform $U(0,1)$ distributions U_1 and U_2:

$$X_1=\sqrt{-2\ln U_1}\cos(2\pi U_2)$$
$$X_2=\sqrt{-2\ln U_1}\sin(2\pi U_2)$$

This is the Box-Muller transform. For a Gaussian there is no direct formulation of the cumulative distribution function F. Thus F^{-1} has no simple analytic expression.

• Method for any X distribution:

If we have a continuous distribution we obtain a discrete distribution using class intervals for *x*. We then have a histogram and each bar of height p_i is also discretized in units. We get an outcome of the distribution X drawing randomly and equiprobably a unit of the histogram. We get by putting the bars end to end a full range of units. We use a discrete uniform distribution U(1,N) where N is the total number of units in the histogram. The generated value is compared to its position on the range and we get a x_i observation.



• There are many other methods that use the different properties of the probability distributions. For example, by simulating the Bernoulli distribution, we obtain, by sum, a binomial distribution which itself allows us to simulate a normal distribution.

E. Exercises

Exercise 1: Binomial distribution *Answers p210*

Check that the binomial distribution defines a probability distribution. Calculate the expectation and the variance as a function of the parameters n and p.

Exercise 2: Sum of binomial distributions
Answers p211

Show that the sum of independent binomial distributions with the same parameters p is itself a binomial distribution whose parameters will be determined.

Exercise 3: Geometric distribution *Answers p211*

Check that the geometric distribution defines a probability distribution.
Calculate the expectation and the variance as a function of the parameter p.
Show that the sum of two independent geometric distributions with the same parameter p is a negative binomial distribution $\mathcal{NB}(2,p)$.

Exercise 4: First successes *Answers p212*

1- Let consider a balanced coin. What is the probability that the first tail will appear on the fifth toss? Knowing that the first tail has not yet appeared on the third throw, what it is the probability that it will appear for the first time at the eighth toss?

2- Let consider a balanced die. On average after how many tosses appears the first six? What is the probability that the first six will appear in the first six throws?

Exercise 5: Poisson distribution *Answers p212*

Check that the Poisson distribution defines a probability distribution.
Calculate the expectation and the variance as a function of the parameter λ.
Determine the sum of two Poisson distributions.

Exercise 6: Uniform distribution *Answers p213*

Calculate the variance for a continuous uniform distribution $\mathcal{U}(a,b)$.

Exercise 7: Exponential distribution

Answers on page 214

Check that the exponential distribution defines a probability distribution.
Calculate the expectation and the variance as a function of the parameter λ.
Determine the sum of two exponential distributions.

Exercise 8: Sum of Gaussians *Answers p215*

1- Determine the sum of two standard normal distributions $\mathcal{N}(0,1)$.
2- Determine the sum of two normal distributions.

Exercise 9: First Students *Answers on page 215*

1- Give the expressions of the first Student functions.
2- Give the polynomial degree in denominator, the variance, the kurtosis and the y-intercept.
3- Give the expression of the Student for k=9, centered and with a variance equals to 14/5.

Exercise 10: Student's t-distribution

Answers on page 216

Whatever k, show that the Student distribution corresponds to a probability distribution.

We can use the integral $I_k = \int_{-\infty}^{+\infty} \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}} dx$ and carry out an integration by substitution with $u = \frac{x}{\sqrt{k}}$.

Exercise 11: Variance of Student's t-distribution

Answers on page 218

Determine the variance of a t-distribution.

Exercise 12: Sum of Student's t-distributions

Answers on page 218

Using an example show that the sum of two independent t-distributions is not a t-distribution.

Exercise 13: Chi-squared distribution

Answers on page 221

Give the expressions of the first chi-square density functions in the cases where k=1, k=2, k=3 and k=10.

Exercise 14: Product of distributions

Answers on page 221

Let X and Y be two independent distributions.

1. Propose a general method for determining the probability distribution of Z=XY.
2. We now consider the case where X and Y are two

independent continuous uniform distributions $\mathcal{U}(1,2)$.

a. Determine the analytic expression of the Z distribution.

b. Find the shape of probability density function of Z with a numerical simulation of the product on a spreadsheet for n=10,000.

Exercise 15: Sum of exponentials *Answers p222*

Let X_i be n independent exponential distributions with the same parameters λ. Let S_n be the distribution of the sum: $S_n=X_1+X_2+...+X_n$. We have too: $M_n=S_n/n$.

1. Determine the probability distribution S_2.

2. Determine the probability distribution S_n.

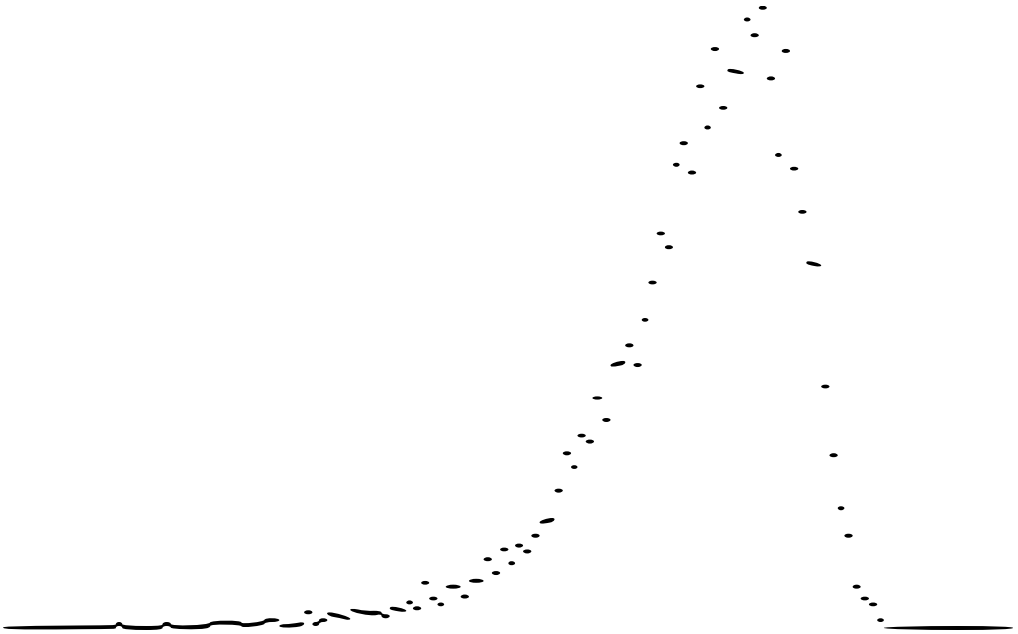
3. Determine the probability distribution M_n.

Exercise 16: Inverse distribution *Answers p223*

Let X be a random variable with strictly positive support.

1. Determine the probability density function of Y=1/X.
2. a. Find T_n the inverse distribution of M_n defined in the previous exercise.
   b. Find the inverse distribution of the Cauchy distribution:

   Cauchy distribution X: $f_x=\frac{1}{\pi}\frac{1}{1+x^2}$ for all x.



IV. ESTIMATORS

We have a random variable X that follows a known distribution with parameters that are unknown.
We have a n sample $(X_1,X_2,\ldots,X_n)$ of the random variable X from a population. We want to have a method to estimate at best the different parameters that define our distribution (inferential statistics). Our parameters are denoted by the letter θ and we call T_n our estimator of θ.

At first we use the sample to give a point estimate of θ, that has a high probability of being close to the parameter, next we provide an interval estimate of θ that has a high probability of containing θ.

A. Properties of an Estimator

1) Bias

The bias of an estimator T_n is the expectation $E(T_n-\theta)$.

Then : $b_{T_n}(\theta)=E(T_n)-\theta$

The estimator is unbiased if $b_n=0$ so $E(T_n)=\theta$.
It is best to have an unbiased estimator if not asymptotically unbiased: $\lim_{n\to+\infty} b_n=0$.

2) Mean Square Error

We can compare different estimators with their mean square errors. The mean square error of T_n is defined as the expectation $E[(T_n-\theta)^2]$ and we show that:

$$r_{T_n}(\theta)=V(T_n)+b^2$$

Indeed $E[(T_n-\theta)^2]=E(T_n^2)-2\theta E(T_n)+\theta^2$ (linearity of the expectation) and eventually after simplifications of $r=V(T_n)+E(T_n)^2-2\theta(b+\theta)+\theta^2$ we find the previous expression. For an unbiased estimator its mean square error is equal to its variance. It is useful to choose an estimator whose mean square error tends to zero with the sample size and the faster is the convergence the better is the estimator.

• Example 1: $\bar{X}_n$ is the sample average regarding the sample $(X_1,X_2,\ldots,X_n)$ defined as:

$$T_n=\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Demonstrate that $\bar{X}_n$ is an estimator of the expectation $\theta=m=E(X)$. Study its properties: bias, convergence.

$E(T_n)=1/n\sum E(X_i)$ using linearity of the expectation then $E(T_n)=1/n.nm=m$ and $b=0$.

$V(T_n)=1/n^2\sum V(X_i)$ (variance of a sum with independent variables) then $r=1/n^2.n\sigma^2$ and $r=\sigma^2/n$.

The sample average is a good estimator of the expectation of a random variable. The estimator is unbiased and the

mean square error tends to zero when *n* tends to infinity.

• Example 2: We are looking for an estimator of the variance of a random variable, we propose the following two estimators:

$$S_n^{\,2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n} \quad \text{and} \quad R_n^{\,2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n-1}$$

Which, in your opinion, is the best estimator of σ^2 ?

$S_n^{\,2}=1/n\sum(X_i^{\,2}-2X_i\bar{X}_n+\bar{X}_n^{\,2})$ and
$S_n^{\,2}=1/n\sum X_i^{\,2}-2T_n\sum X_i/n+T_n^2=1/n\sum X_i^{\,2}-T_n^2$
then $E(S_n^{\,2})=E(X^2)-E(T_n^2)=\sigma^2+m^2-V(T_n)-E(T_n)^2$
and $E(S_n^{\,2})=\sigma^2-\frac{\sigma^2}{n}$ so $E(S_n^{\,2})=\frac{n-1}{n}\sigma^2$ and $b_{S_n^2}=-\frac{\sigma^2}{n}$
.

By a quite similar calculation we find $b_{R_n^2}=0$.

This quantities are two estimators of the variance because they are unbiased or asymptotically unbiased.
$R_n^{\,2}$ is the best estimator because it has no bias.

B. Construction of estimators

We want to have a general method to find the appropriate estimators to estimate the parameters of a distribution.

1) Method of Moments

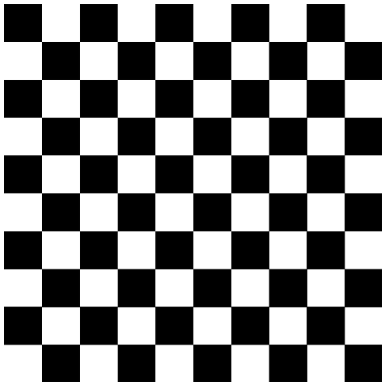
We identify the moments of the population with those of the sample. We consider as many moments as we have parameters starting with the first moment. As we will see on examples this method provides us with estimators but does not guarantee us that these are the best in terms of bias and mean square error.

Theoretical moments: $m_k = E(X^k)$, $k \in \mathbb{N}^*$

Sample moments: $\overline{X_n^{\,k}} = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,k}$

We can also consider the moments centered or completely normalized, the approach remains the same. The first n-sample moment is the sample average and it corresponds to an excellent estimator as shown before. On the other hand, the centered second moment has a bias, as we have shown before we should divide by *n-1* instead of *n* to have a unbiased estimator.

• Example 1: We have a checkerboard of 100 squares and 200 seeds. Every second we randomly place a seed in one of the squares. At the end of the experiment we count the number of seeds in each square and we count the number of squares that contain no seed, one seed, two seeds and so on. Let X be the random variable for the numbers of seeds per square. We obtain the following distribution:



k	0	1	2	3	4	5	6	7	8
n	12	28	34	10	8	7	0	1	0

We assume that this distribution follows a Poisson distribution. Deduce the value of the parameter λ.
The parameter of this distribution is equal to the expectation therefore according to the theorem of moments, λ is estimated by the sample average:

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{p} n_k x_k \quad \text{then} \quad \lambda = 2$$ [21].

This result is consistent with our model, indeed let us name n the number of squares and N the number of seeds. We have a uniform random distribution. We could, for example, use two balanced ten faces dice of different colors for the horizontal and vertical position. The frequency of

21 Using the second moment we will find $\lambda = 2.06$.

the event per second for each square is *1/n*. For a Poisson distribution at any instant the event can occur, here the time is discretized, nevertheless the approximation of a continuous time is correct because one second is a small duration compared to that of 200 seconds of the experiment. So we can take $\lambda = N/n$. We tend to a Poisson distribution when the number of squares and seeds tend to infinity.

• Example 2: We assume that the lifespan of a glass follows an exponential distribution. We observe in months the following lifetimes: 7, 5, 37, 35, 17, 9, 6, 13, 4 and 8 months. Determine thee parameter λ of the distribution.

For an exponential distribution $E(T) = \frac{1}{\lambda}$, hence by using the first moment: $\lambda = \frac{n}{\sum_{i=1}^{n} t_i}$ and $\lambda = \frac{10}{141} \simeq 0.071$.

We thus have a point estimate of the parameter λ and the life expectancy is about 14 months.

We will show in an exercise that this estimator T_n has a bias: $E(T_n) = \frac{n}{n-1}\lambda$ and $b_{T_n} = \frac{\lambda}{n-1}$.

This bias and its mean square error tend to zero:

$$r_{T_n} = \lambda^2 \frac{(n+2)}{(n-1)(n-2)}.$$

We construct from this estimator a new estimator W_n without bias: $W_n = \frac{n-1}{n} T_n = \frac{n-1}{\sum X_i}$.

We show that $E(W_n) = \lambda$ and $r_{W_n} = V(W_n) = \lambda^2/(n-2)$.

This estimator is better than the previous: zero bias and

lower risk (mean square error). A new estimate of the parameter gives $\lambda \simeq 0.064$ and the life expectancy is about 16 months.

• Example 3: We consider that the mass of the apples produced by a tree follow a normal distribution. We randomly draw and measure the following masses in grams:

158	131	146	158	125	153	166	121
127	123	195	149	124	153	123	129

Determine the mean and standard deviation of the distribution.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} m_i \simeq 142.5 \ , \quad \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} m_i^{\,2} \ \text{ and } \ \sigma \simeq 20.2 \ .$$

• Example 4: The planks produced by a sawmill have a length which follows a uniform distribution $\mathcal{U}(a,b)$.
We measure the lengths in millimeters of 8 planks drawn at random: 2017, 1987, 2018, 2014, 2003, 1985, 2013 and 1981. Estimate a and b.

$$\text{We have} \quad E(X) = \frac{a+b}{2} = \overline{X_n^{\,1}}$$

$$\text{and} \quad E(X^2) = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4} = \frac{a^2+ab+b^2}{3} = \overline{X_n^{\,2}}$$

$$\text{then} \quad a+b = 2\,\overline{X_n^{\,1}} \quad \text{and} \quad ab = 4\,\overline{X_n^{\,1}}^{\,2} - 3\,\overline{X_n^{\,2}}$$

$$\text{Eventually} \quad a \simeq 1977 \quad \text{and} \quad b \simeq 2028$$

2) Method of Maximum Likelihood

Let X be a discrete or continuous random variable. We want to estimate a unknown parameter θ from a set of observations $\{x_i\}$.

We define a function f(x,θ) such that:

$$f(x,\theta) = \begin{cases} P_\theta(X=x) & \textit{for a discrete variable} \\ & \text{or} \\ f_\theta(x) & \textit{for a continuous variable} \end{cases}$$

We define the likelihood function *L.* This is a function of θ, determine by the numbers x_1, x_2, . . . , x_n:

$$L(x_1,x_2,\ldots,x_i,\ldots,x_n,\theta)=f(x_1,\theta)\times f(x_2,\theta)\times\ldots\times f(x_n,\theta)$$

also we can simply write:

$$L(x_1,x_2,\ldots,x_i,\ldots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

The value of the parameter for which the product of the probabilities, or probability densities, taken at the different points of the sample, is maximum, is considered the most probable value. The maximum likelihood estimate of θ is the value that maximizes the likelihood function $L(\theta)$.

The maximum likelihood principle is simple and the method is easy to implement:

$$\frac{\partial L(x_i,\theta)}{\partial\theta} = 0 \quad \text{and} \quad \frac{\partial^2 L(x_i,\theta)}{\partial\theta^2} < 0$$

Usually on can find the maximum by differentiating the likelihood function $L(\theta)$. However the calculation of the derivative may be tedious, it is why we prefer to consider the logarithm of $L(\theta)$. The logarithm is a increasing function and we can consider the extreme values of $ln(L(\theta))$ instead of $L(\theta)$.

• Example 1: Let us take again the case of a Poisson distribution and determine the estimator of λ by the method of maximum likelihood.

$$P_\lambda(X=x)=\frac{\lambda^x}{x!}e^{-\lambda} \quad \text{and} \quad L=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}$$

$$\ln L=-n\lambda+\sum_{i=1}^{n}\ln\frac{\lambda^{x_i}}{x_i!} \quad \text{and} \quad \frac{\partial \ln L}{\partial \lambda}=-n+\sum_{i=1}^{n}\frac{x_i}{\lambda}=0$$

$$\text{then} \quad \lambda=\frac{1}{n}\sum_{i=1}^{n}x_i$$

We find the same estimator as by the previous method.

• Example 2: Let us take again the case of an exponential distribution and determine the estimator of λ by the method of maximum likelihood.

$$f_\lambda(t)=\lambda e^{-\lambda t} \quad \text{and} \quad L=\prod_{i=1}^{n}\lambda e^{-\lambda t_i}=\lambda^n e^{-\lambda\sum_{i=1}^{n}t_i}$$

$$\ln L = n\ln\lambda - \sum_{i=1}^{n} \ln\frac{\lambda^{x_i}}{x_i!} \quad \text{and} \quad \frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i$$

$$\text{then} \quad \lambda = \frac{n}{\sum_{i=1}^{n} t_i}$$

The same estimator as by the previous method.

• Example 3: Let us take again the case of a Gaussian distribution and determine the estimators of μ and σ by the method of maximum likelihood.

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{and} \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$\text{so} \quad \ln L = -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0 \quad \text{and} \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

We obtain the same estimators as by the moment theorem and we have again an estimator of the variance biased.

• Example 4: Let consider now the uniform distribution.

$$L(x_i,a,b)=\prod_{i=1}^{n} f(x_i,a,b)=\begin{cases}\dfrac{1}{(b-a)^n} & if \quad \{x_i\}\subset[a,b] \\ 0 & else\end{cases}$$

For a fixed, smaller is b bigger is the likelihood, then $b=max(\{x_i\})$.
For b fixed, bigger is a bigger is the likelihood, then $a=min(\{x_i\})$.

We have here very different estimators than by the method of moments, for the example of the planks we have the following estimates: $a=1981$ and $b=2018$.

These estimators are biased, for example b is necessarily smaller than the theoretical value, but, contrary to the method of moments we are assured that all the x_i belong to [a, b].

C. Interval estimate

We now have effective tools to determine the values of the parameters of a distribution. We have determined point estimates and now we want to determine a confidence interval.

The central limit theorem allows, from a large sample, to estimate the mean of a probability distribution with a confidence interval.

But what about the other parameters different from the mean? For example, what are the uncertainties on the parameter λ of an exponential distribution, or the bounds a and b for a uniform distribution?

We consider an unbiased estimator and if the estimator is biased we create a new estimator by removing the bias.

We use three different methods. The integral method which requires to determine the full probability distribution of the estimator. A second method, simpler, provides an inequality that overestimates our uncertainty, but only requires knowledge of the variance of the estimator. And finally a third by numerical simulation.

For the second method we use the Chebyshev's inequality. This inequality can be applied to any probability distribution in which the mean and variance are defined. We do not have to know the aspects of the estimator distribution, only its variance, but the inequality generally gives a poor bound:

$$\forall \epsilon > 0, \quad P(|X - E(X)| \geq \epsilon) \leq \frac{V(X)}{\epsilon^2}$$

• Example 1: Consider again the checkerboard and the Poisson distribution with parameter λ. This parameter is estimated by the mean and we can therefore, in the case of large numbers, use the central limit theorem:

$$\lambda = \lambda_m \pm t_\infty \sigma / \sqrt{n}$$

Let us estimate the variance:

k	0	1	2	3	4	5	6	7	8
$k-\lambda_m$	-2	-1	0	1	2	3	4	5	6
$(k-\lambda_m)^2$	4	1	0	1	4	9	16	25	36
n	12	28	34	10	8	7	0	1	0

$\sigma_\lambda^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\sigma_\lambda \simeq 1.44$ and $\lambda = 2.0 \pm 0.3$ with 95% confidence.

Here the sample size is not sufficient for this method but we have illustrated the general method for large samples. We probably have underestimated the width of the interval.

• Second example: The waiting time of a train follows a uniform distribution $\mathcal{U}(0,a)$. We observe the following waiting times 3, 12, 7, 17, 8, 14, 2, 5, 10, 14, 15 and 11 minutes. What is the value of a ?

What is the bias, the variance of the estimator T_n and the uncertainty on a at 90% confidence?

a) Let us first use the method of maximum likelihood which gives us the estimator T_n , the distribution of the maximum of a number n of independent, identically distributed variables: $T_n = max(\{X_i\})$

a is here estimated at 17 minutes.

The distribution of the maximum is such that:

$$P(T_n \le x) = P([X_1 \le x] \cap [X_2 \le x] \cap ... \cap [X_n \le x])$$

the variables are independents, then:

$$P(T_n \le x) = \prod_{i=1}^{n} P(X_i \le x)$$

Cumulative distribution function of X: $F_X(x) = P(X \le x)$

$$F_{T_n}(x) = \prod_{i=1}^{n} F_{X_i}(x) = \left(\frac{x}{a}\right)^n \text{ for } 0 \le x \le a$$

$$F_{T_n}(x) = 0 \text{ if } x < 0\,, \; F_{T_n}(x) = 1 \text{ if } x > a$$

So we obtain the density of T_n : $f_{T_n}(x) = \frac{dF_{T_n}(x)}{dx} = n\frac{x^{n-1}}{a^n}$

for $0 \le x \le a$ and $f_{T_n}(x) = 0$ else.

Expectation: $E(T_n) = \int x f_{T_n}(x)\,dx = \frac{n}{a^n}\int_0^a x^n\,dx = \frac{n}{n+1}a$

The estimator T_n is biased: $b_{T_n} = \frac{n}{n+1}a - a = -\frac{a}{n+1}$

a is underestimated and the bias is asymptotically zero.

Variance: $E(X^2)=\int x^2 f_{T_n}(x)dx=\frac{n}{a^n}\int_0^a x^{n+1}dx=\frac{n}{n+2}a^2$

$$V(T_n)=E(X^2)-E(X)^2=\frac{n}{n+2}a^2-\frac{n^2}{(n+1)^2}a^2=\frac{na^2}{(n+2)(n+1)^2}$$

It is better to take an unbiased estimator, for this we remove the bias and we get W_n:

$$W_n=\frac{n+1}{n}T_n \quad \text{with} \quad b_{W_n}=0 \quad \text{and} \quad V(W_n)=\frac{a^2}{n(n+2)}$$

New estimate of a with W_n : $a\simeq 18.4$ minutes.

For the uncertainty Δa we will compare three methods.

-> With determine a confidence interval with the Chebyshev's inequality:

We apply the inequality to W_n : $P(|W_n-a|\geq\epsilon)\leq\frac{V(W_n)}{\epsilon^2}$

so $P(|W_n-a|\leq\epsilon)=P(-\epsilon\leq W_n-a\leq\epsilon)$

and $P(\epsilon\geq a-W_n\geq-\epsilon)=P(U_n+\epsilon\geq a\geq W_n-\epsilon)$

we set $\epsilon=\sqrt{\frac{V(W_n)}{\alpha}}$ and we have:

$$1-P(W_n-\epsilon\leq a\leq W_n+\epsilon)\leq\alpha$$

Eventually:

$$P\left(W_n-\sqrt{\frac{V(W_n)}{\alpha}}\leq a\leq W_n+\sqrt{\frac{V(W_n)}{\alpha}}\right)\geq 1-\alpha$$

In the 90% confidence case $\alpha = 0.1$ and for our sample $V(W_{12}) = \frac{18.4^2}{12 \times 14} \simeq 2.0$ and $\Delta a = \sqrt{\frac{V(W_n)}{\alpha}} \simeq 4.5$.

Then $13.9 \leq a \leq 22.9$ and $a \simeq 18.4 \pm 4.5$ minutes.

-> Let us determine a confidence interval with an integral calculation on the probability density of the estimator.

We determine the probability density of W_n:

$$P(W_n \leq x) = P\left(\frac{n}{n+1} W_n \leq \frac{n}{n+1} x\right) = P\left(T_n \leq \frac{n}{n+1} x\right)$$

$$F_{W_n}(x) = F_{T_n}\left(\frac{n}{n+1} x\right) = \left(\frac{n}{n+1} \frac{x}{a}\right)^n \text{ if } 0 \leq x \leq \frac{n+1}{n} a .$$

$$f_{W_n}(x) = \left(\frac{n}{a(n+1)}\right)^n n x^{n-1} \text{ if } 0 \leq x \leq \frac{n+1}{n} a$$

and $f_{W_n}(x) = 0$ else.

We have an asymmetric probability distribution whose maximum corresponds to the upper bound. For a 90% confidence we remove the 10% of the left distribution tail from the confidence interval.

Thus we define a_{max} and a_{min} such as:

Upper bound: $a_{max} = \frac{n+1}{n} a$

so $a_{max}=\left(\frac{13}{12}\right)^2\times 17\simeq 19.95$ and $a_{max}\simeq 20.0$ minutes.

Lower bound: $\int_{a_{min}}^{a_{max}} f_{W_n}(x)\,dx=1-\alpha$

By numerical calculation with $a\simeq 18.4$ and *n=12,*

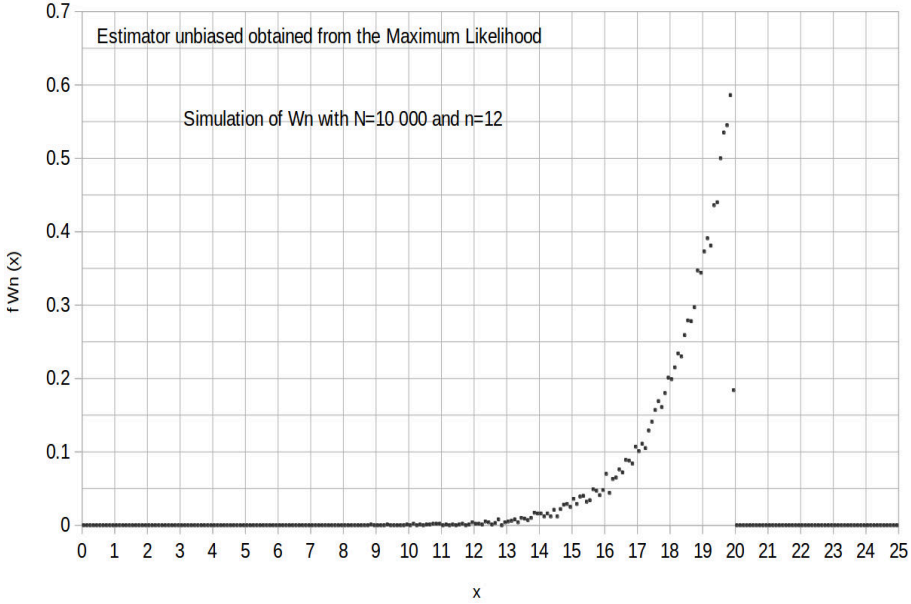
$$\int_{a_{min}}^{20} f_{W_{12}}(x)\,dx=0.9 \text{ with } a_{min}\simeq 16.4 \text{ minutes.}$$

In conclusion, $a\simeq 18.4\,{}^{+1.6}_{-2.0}$ minutes

with 90% confidence.

The interval found is asymmetrical. The bounds are included into those of the Chebyshev's inequality and we have here a more precise estimate of *a*.

-> Let us now perform a numerical simulation on a spreadsheet. The ALEA() function of the spreadsheet provides real randomly and uniformly distributed between 0 and 1. Then we multiply by the point estimate of *a* to obtain the distributions $X_i=U_i(0,a)$. We generate 10,000 samples of size 12. We place the maximum of each of these samples on a graph and thus have the sampling distribution of T_n:



We thus find the same results as by the preceding method (file : www.incertitudes.fr/livre/Train.ods).

a) We now use the estimator of the moment theorem: $T_n = 2\overline{X_n}$. Here, for large n, we can use the central limit theorem because an affine function of a distribution gives a new distribution of the same form. We estimate the mean with the Gaussian law:

$$\bar{x} = \frac{3+12+...+11}{12} \simeq 9.83 \quad \text{and} \quad s = \sqrt{\frac{(3-\bar{x})^2+...}{11}} \simeq 4.88$$

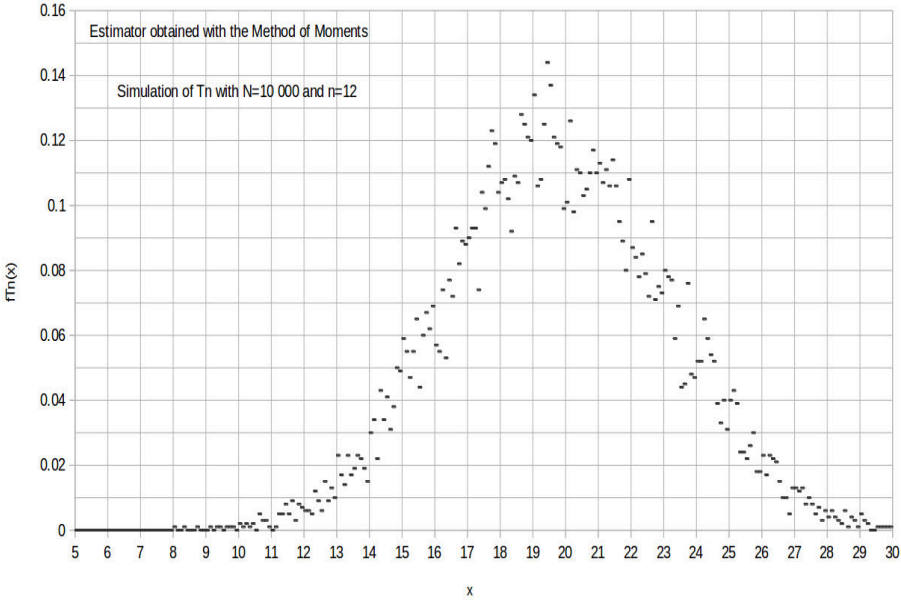
$$m = \bar{x} \pm t_\infty s/\sqrt{n} = 9.8 \pm 2.3 \text{ with 90\% confidence}$$

Then $a = 19.7 \pm 4.6$ minutes with 90% confidence.

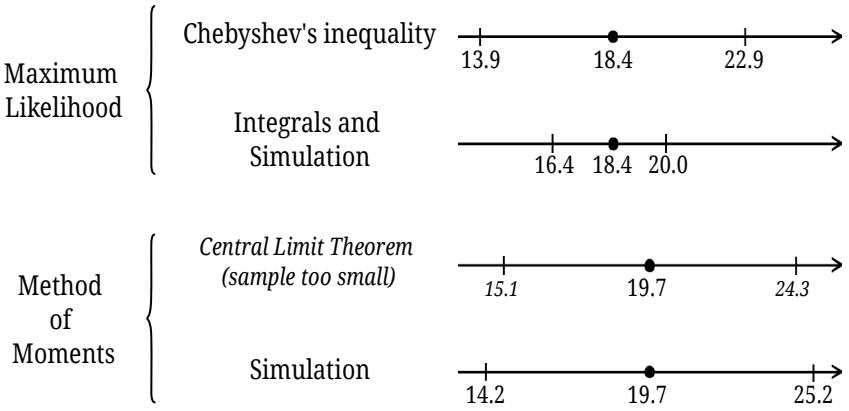
In this case the value of n is only 12 and the distribution of the population is not normal, we want nevertheless to show how one would proceed for n large. As we will see with

numerical simulation, the interval is thus underestimated.

Indeed by carrying out a numerical simulation with this estimator, $a = 19.7 \pm 5.5$ minutes with 90% confidence:

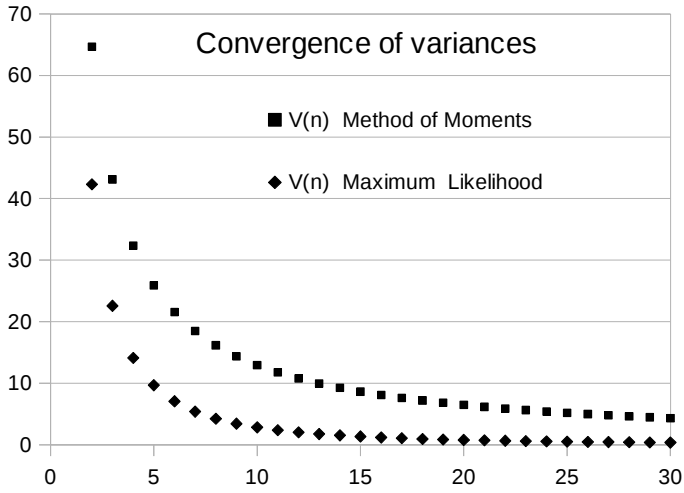


Comparison of the results of the different methods with 90% confidence:



In conclusion, the maximum likelihood estimator converges much faster than that of the moment theorem and we prefer this first method. The variance converges to $1/n^2$ instead of $1/n$:

$$\left[V\left(W_n\right)\right]_{ML}=\frac{a^2}{n\left(n+2\right)} \quad \text{and} \quad \left[V\left(T_n\right)\right]_{MM}=\frac{a^2}{3n}$$



D. Exercises

Exercise 1 : Estimators of the mean

Answers p224

Let consider the (X_1, X_2, X_3). X_1, X_2 and X_3 are three independent variables with the same distribution, expectation m and variance σ^2.
Compare the following three proposed estimators to estimate the mean m of the sample:
$A_3=(X_1+X_2+X_3)/3$, $B_3=(X_1+2X_2+3X_3)/6$ and
$C_3=(X_1+2X_2+X_3)/3$.

Exercise 2 : Homokinetic Beam

Answers p224

We consider a homokinetic beam of C^+ ionized carbon atoms. We measure the momentum and the kinetic energy of 100 atoms of the beam. The beam is considered perfectly unidirectional.
The total momentum magnitude and kinetic energy are:

$$p=\sum_{i=1}^{100} m v_i = 2.418\times10^{-21} \text{ kg.m/s and}$$

$$E_k=\sum_{i=1}^{100} \frac{1}{2} m v_i^2 = 1.518\times10^{-18} \text{ J with m}=1.993\times10^{-26} \text{ kg.}$$

Let V be the probability distribution of the speed of the ions. Determine the average speed v_m, its variance σ_V^2 and the uncertainty Δv with 95% confidence, using the sample taken and the appropriate estimators.

Exercise 3 : Two estimators *Answers p225*

Let X be the following discrete random variable:

Values	0	1	2
Probabilities	3θ	θ	1 - 4θ

1. What values of θ define a valid probability distribution?

2. Calculate E(X) and V(X).

3. Let $(X_1, X_2,..., X_n)$ be a sample of X.

We have $\overline{X_n}=\frac{1}{n}\sum_{i=1}^{n} X_i$ and $T_n=a\overline{X_n}+b$.

Determine a and b for which T_n is an estimator without bias for θ.

Determine $V(T_n)$.

4. We now consider the random variable Y_i defined for all $i\in[\![0;n]\!]$ by $Y_i=1$ if $X_i=1$ and $Y_i=0$ otherwise.

Let $Z_n=\sum_{i=1}^{n} Y_i$. Determine $E(Z_n)$.

Show that $U_n=\frac{Z_n}{n}$ is an unbiased estimator of θ. Determine $V(U_n)$.

5. We make estimates of θ with the following realizations:

Values	0	1	2
Frequencies	31	12	57

Estimate θ. Which estimator do you prefer?

Exercise 4 : Ballot boxes *Answers p225*

Two identical urns contain the same proportion p of black balls. In the first ballot box we draw with remplacement a sample of size n_1 and we note P_1 the proportion of black balls in this sample. We perform the same experiment for the second urn.

We define $T=\frac{P_1+P_2}{2}$ and $U=x\,P_1+(1-x)P_2$

with $x \in]0,1[$

Show that T and U are two estimators of p.
Which one is the best ?

Determine the value of x to have the optimum estimator.

Exercise 5 : Continuous variable *Answers p226*

We consider the continuous variable X with the following density:

Where a is the parameter we want to estimate ($a>1$).

$$f(x)=\begin{cases}\dfrac{a}{x^{a+1}} & \text{if } x>1\\ 0 & \text{otherwise}\end{cases}$$

1. Check that f defines a valid probability density.

2. Calculate E(X).

3. Determine estimators of a by the method of maximum likelihood and by the method of moments.

4. Give point estimates of a for the following observations:

1.16 / 1.80 / 1.04 / 3.40 / 1.22 / 1.06 / 1.35 / 1.10.

5. Can we calculate V(X)?

6. Perform a numerical simulation of the law of X. What can we guess about the biases and convergences of the estimators found?

Exercise 6 : Linear density *Answers p229*

Consider the following continuous random variable X:

$$f(x)=\begin{cases} a\,x+b & if \quad 0\le x\le 1 \\ \\ 0 & otherwise \end{cases}$$

1. Express b as a function of a such that f defines a probability distribution.

2. Calculate E(X) and V(X).

3. Determine an estimator T_n of a by the method of moments. Discuss the properties of this estimator.

4. We draw a sample:

0.81	0.67	0.72	0.41	0.93	0.55	0.28	0.09	0.89

Determine a point estimate of a. How would we get an interval estimate?

Exercise 7 : Estimators for the exponential law

Answers p230

The exercise of the previous chapter on page 125 provides the expression of the probability density of the estimator T_n of λ obtained in the course:

$$f_{T_n}(x)=\frac{n^n\lambda^n}{(n-1)!\,x^{n+1}}e^{-\frac{n\lambda}{x}}$$ if x>0 and zero if not

1. Determine the expectation, bias, variance and the mean square error of T_n.

2. Let $W_n=\frac{n-1}{n}T_n$. Determine the expectation, bias, variance and the mean square error of W_n.

3. Which estimator would you recommend for λ?

Exercise 8 : Decays *Answers p231*

The law of probability X of particle decay as a function of time follows an exponential distribution with parameter λ. We measure the lifetimes in microseconds of a sample of ten particles and we want to deduce a point estimate and an interval estimate of λ. We will use different methods and comment on them.

2.2	3.32	1.93	0.88	4.36
7.1	0.65	0.53	4.69	0.31

1. With the central limit theorem could you estimate the expectation m of X and its uncertainty with a confidence of 90% (m : mean lifetime)? With the propagation of uncertainties formula we then could find the value of λ with its uncertainty. What do you think about this estimate of λ?

2. Let T_n be the estimator of λ found during the lesson. Like we shown previously this one is biased and then we use $W_n = \frac{n-1}{n} T_n$.

Determine by an integral calculus the uncertainty on λ with 90% confidence.

3. Find now this result by numerical simulation.

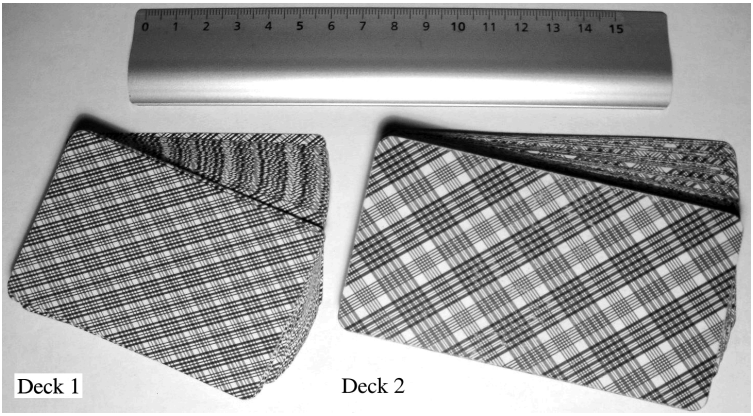
V. Measure with a ruler

Article published in the BUP [ii].

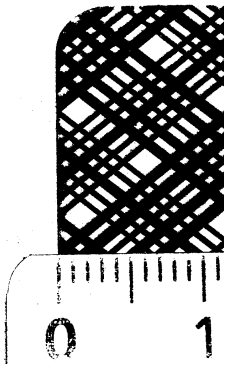
ABSTRACT
The measurement of a physical quantity by an acquisition system induces because its resolution a discretization error. We are here concerned with measuring a length with a graduated ruler. This type of measure leads us to consider a uniform continuous probability distribution. We then use a convolution to determine the uncertainty with its confidence of a sum of lengths. Finally, we generalize to the general case of the calculation of uncertainties for independent random variables using the error propagation formula.

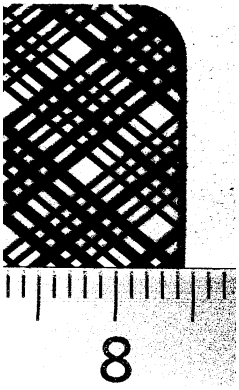
INTRODUCTION

We want to measure lengths and evaluate uncertainties as exactly as possible. Uncertainties about the measured values and their sums. We have a ruler of 15cm graduated to the millimeter and two sets of cards. The ruler is assumed to be perfect and the cards of each game identical.



1. MEASURE OF THE LENGTH OF ONE CARD



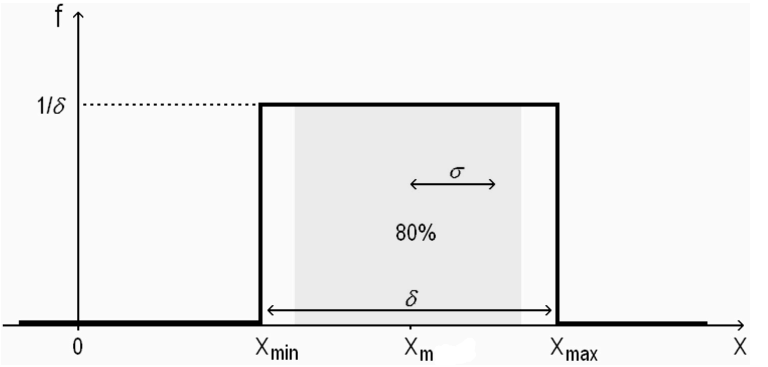


We place the graduation zero on the left edge of the card. On the right edge we consider the graduation closest to the edge. The experimenter does not read between the graduations. The thickness of the lines which delimit a graduation is considered negligible compared with the width of this graduation. We get thus for the deck 1:

$$x_1 = 8.4 \pm 0.05\, cm\,.$$

Concerning the cards of the second deck:

$$x_2 = 11.2 \pm 0.05\,cm.$$



We accept a loss of information due to the resolution $\delta = 1\,mm$ of the ruler. When we use these data later, all values between $x_{\min} = x_m - \delta/2$ and $x_{\min} = x_m - \delta/2$ are equally probable. We have to consider a continuous and uniform random variable X. x is a realization of X. This distribution of probability has a range $E = x_{\max} - x_{\min}$ and its density $f(x)$ verify:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

The probability to be between x and $x + dx$ is $f(x)\,dx$. The result is necessarily between $x_{\min}$ and $x_{\max}$: for example $x_1 = 8.4 \pm 0.05\,cm$ with 100% of confidence, but $x_1 = 8.4 \pm 0.04\,cm$ with a probability of 80%.

To characterize the spreading of a distribution we consider the range E and the standard deviation σ whose definition for a continuous law is:

$$V=\sigma^2=\int (x-x_m)^2 f(x)\,dx\ ,$$

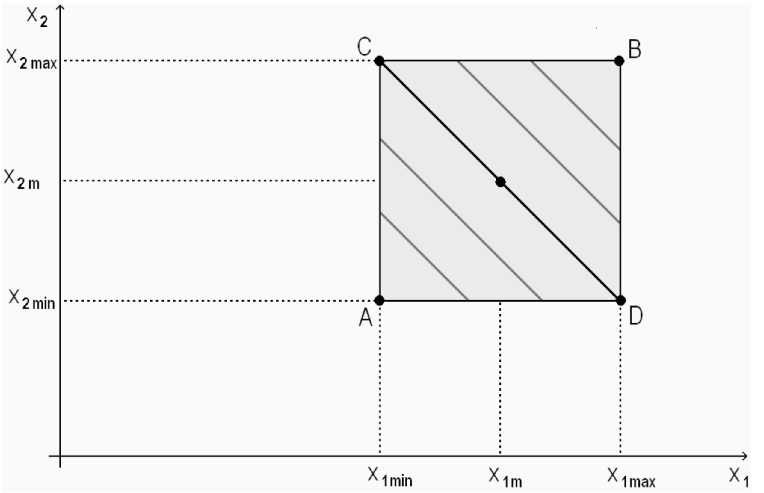
V is called the variance. For a uniform distribution:

$$\sigma=\delta/\sqrt{12}\simeq 0.29\,\delta\ ,$$

and we have $x=x_m\pm\sigma$ with 58% confidence. The standard deviation is an appropriate quantity to characterize the width of a distribution. The range is defined by the extreme values which may be unrepresentative or absurds.

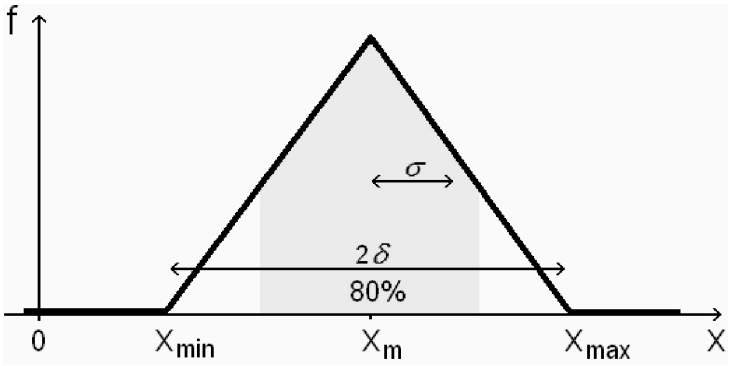
2. LENGTH OF TWO CARDS PUT END TO END

We want to determine the uncertainty on x with $x=x_1+x_2$. If we plot x_2 as a function of x_1 the set of the possible points forms a square domain. The set of points such as x is constant is straight line segment of slope -1 and intercept x : $x_2=-x_1+x$. There is only one case where $x=x_{\min}$ then $\{x_1=x_{1\min}\,;\,x_2=x_{2\min}\}$ at point A in the figure. However on all the segment [CD] we get $x=x_m$. We understand that the different values of x do not have the same probability.



The probability density f of X is computed from those of X_1 and X_2. For a sum of independent random variables the result is given by a convolution [iii] :

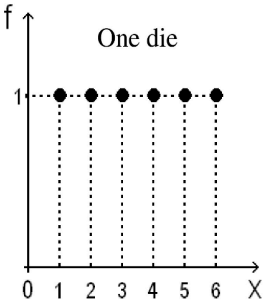
$$f(x)=\int f_1(y)f_2(x-y)\,dy \;\Rightarrow\; \begin{cases} x<x_{min} \Rightarrow f(x)=0 \\ x_{min}<x<x_m \Rightarrow f(x)=\dfrac{(x-x_{min})}{\delta^2} \\ x_m<x<x_{max} \Rightarrow f(x)=\dfrac{(x_{max}-x)}{\delta^2} \\ x>x_{max} \Rightarrow f(x)=0 \end{cases}$$

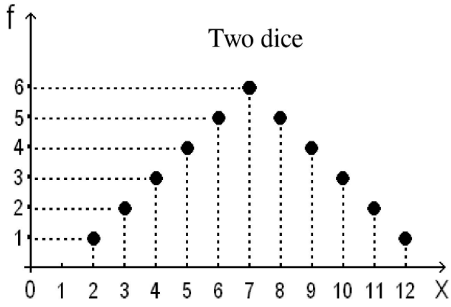


We then have a triangular probability distribution.
We obtain $x = 19.6 \pm 0.1 cm$ with 100% confidence, and $x = 19.6 \pm 0.055 cm$ with 80% confidence.

3. ANALOGY WITH THE THROW OF TWO DICE

For each die the six values are equally likely. Here the law of probability is no longer continuous but discrete. We launch two dice simultaneously, the sum of the values obtained is between two and 12. In this case, there is no equiprobability, a way to get two with a double one, two ways to get three with one and two or two and one ... to get seven we have the maximum of possibilities. We thus find a triangular distribution.





4. LENGTH OF TWO CARDS OF THE SAME DECK PUT END TO END

The cards of a deck are supposed identical then if the length of one of them is overestimated, it will be the same for the second one. In this case the errors are added and can not be compensated. For two different cards, the first measure can be underestimated and the second overestimated, and a compensation can occur. Here it is no longer the case and when $X = X_i + X_i'$ we obtain a uniform distribution of width 2δ. Our random variables are not independent.

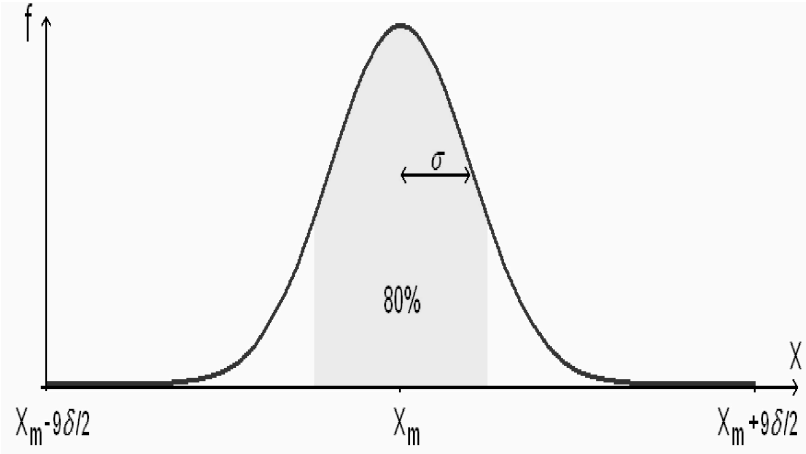
For the deck 1:

$$x_1 = 8.4 \pm 0.04\, cm \Rightarrow x = 2x_1 = 16.8 \pm 0.08\, cm$$

with a confidence of 80%.

5. SUM OF N INDEPENDENT LENGTHS

We have $X = \sum_{i=1}^{N} X_i$. Each length X_i follow a uniform distribution of width δ. For the sum of nine independent random variables after iteration of the calculation we obtain the following curve:



In this case we obtain $x = x_{moy} \pm 0.11\,cm$ with 80% confidence. With 100% confidence: $x = x_{moy} \pm 0.45\,cm$, which leads us to consider domains where the probability of presence of X is really negligible. An uncertainty of 0.45cm seems unnecessary while 99% of the cases were already present with an uncertainty of 0.22cm.

Working with a confidence of 100% it is like considering the range, it is additive for a sum of variables. The range is proportional to N.

	80%	95%	99%
N=1	0.40 δ	0.48 δ	0.50 δ
2	0.55 δ	0.78 δ	0.90 δ
3	0.66 δ	0.97 δ	1.19 δ
4	0.75 δ	1.12 δ	1.41 δ
5	0.84 δ	1.25 δ	1.60 δ
6	0.92 δ	1.38 δ	1.76 δ
7	0.99 δ	1.49 δ	1.91 δ
8	1.06 δ	1.59 δ	2.05 δ
9	1.12 δ	1.69 δ	2.18 δ
10	*1.2* δ	*1.8* δ	*2.3* δ
20	*1.7* δ	*2.5* δ	*3.3* δ
50	*2.6* δ	*4.0* δ	*5.2* δ
100	*3.7* δ	*5.7* δ	*7.4* δ

But this approach does not take into account one thing: the curve narrows around the mean when N increases. There is another additive quantity: the variance. The standard deviation, square root of the variance, is proportional to $\sqrt{N}$ and takes account of error compensations.

We obtain a bell curve. A statistical theorem, called the limit central theorem, indicates that for N large the curve tends to a Gaussian. The range of a Gaussian is infinite for a finite standard deviation.

We summarize the evolution of the uncertainty on the sum of N independent lengths measured with the same resolution δ in a table. In italics, from N=10, these are numerical simulations carried out on a computer by generating random numbers.

The results of the measurements are often given with a confidence of 95%, which corresponds for a Gaussian to an uncertainty of about 2σ .

6. OTHER APPLICATIONS

A runner wishes to measure his travel time. He has a watch with a digital display. The watch shows that he starts at $10\,h\,52\,\mathrm{min}$ and arrives at $11\,h\,11\,\mathrm{min}$. The display is at the minute, so he starts between $10\,h\,52\,\mathrm{min}\,00\,s$ and $10\,h\,52\,\mathrm{min}\,59\,s$. Hence the date of departure is within the interval $t_1 = 10\,h\,52\,\mathrm{min}\,30\,s \pm 30\,s$. The resolution is one minute. The duration of the course is $\Delta t = t_2 - t_1$. The results remain true for a difference. We have N=2 and $\Delta t = 19\mathrm{min} \pm 47\mathrm{s}$ with 95% confidence.

Same procedure if students measure an angular difference with a goniometer. Each measurement is within a minute of arc so the uncertainty of the angular difference is of 47 arc seconds with a confidence of 95%.

Seven persons are in an elevator. Its maximum load is 500kg. Their individual masses are measured with a resolution of one kilogram. The total mass is 499kg. What is the probability of being overloaded?

For N=7 the uncertainty reaches one kilogram with a confidence of 80%. So there is a one out of ten chance for the elevator to be overloaded.

In the laboratory many measuring instruments have digital displays. The resolution is define by the last digit. But the overall uncertainty is much higher. It is necessary to consult the operating instructions of each device.

CONCLUSION

The general approach consists in combining laws of probabilities. The mathematical tool used is a change of variables, then one or more integrations. For the measure with a ruler described in this article, it was a sum of two independent random variables and we obtained a convolution.

If one wants to do a faster calculation, an analysis of variance may be enough. We have a random variable X that depends on N independent random variables X_i : $X = f(X_1, X_2, \ldots, X_i, \ldots, X_N)$. We call σ_i the standard deviation of X_i and σ that of X. For finite σ_i and small variations, we have the propagation of standard deviations formula [iv]:

$$\sigma^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

And, independently of the probability distributions, this relation between the variances remains true. One can thus give its result with an uncertainty for 2σ or 3σ .
Is there a similar formula with the confidences? Yes, but it is approximate, it is the propagation of uncertainties formula:

$$\Delta f^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \Delta x_i^2 \quad ,$$

with $x_i = x_{i\,moy} \pm \Delta x_i$, $f = f_{moy} \pm \Delta f$ and a constant confidence. This formula is very useful and allows a quick and reasonable calculation of the combined uncertainties. Moreover, it is exact if the form of the distributions is the same for X and the X_i . For example, if the distributions X_i are Gaussian, any linear combination is also Gaussian. We thus take account of compensations and avoid using the formula $\Delta f = \sum_{i=1}^{n} |\partial f / \partial x_i| \Delta x_i$ which overestimates the uncertainties, sometimes even with such excess that one loses its physical sense. This last formula does not take into account any compensation, we have the worst situation, statistically unlikely. Here, for example, for N=100, one would have an uncertainty of $50\,\delta$, instead of $5.7\,\delta$ in practice (95% confidence).

In this article we focused on the resolution of an acquisition system that gives a discretization error. But one can also consider systematic errors and random errors. Here the ruler was supposed perfect, that means, accurate and precise.

VI. Mathematical Tools

The mathematical tools used in chapter 1 are studied into the high school scientific section. The partial derivatives used in chapter 2 are taught during the first year of university but we quickly understand the link with the derivatives seen at the high school. It is at the end of the second chapter, with the use of matrices for generalized regression, that we immerse ourselves in university education. My goal is not to review or introduce all these concepts, only a small recap that can be useful to solve the exercises proposed.

A - Derivatives

1- Definition : $f'(x)=\lim_{\epsilon\to 0}\left(\frac{f(x+\epsilon)-f(x)}{\epsilon}\right)$

For example if $f(x)=x^2$:

$(x+\epsilon)^2-x^2=x^2+2x\epsilon+\epsilon^2-x^2\simeq 2x\epsilon$ and $f'(x)=2x$.

On a graph the derivative corresponds to the slope of the curve at each point.

2- Rules for basic and combined functions:

function f	derivative f'	Δf
ax	a (constant)	$\Delta(ax)=\lvert a\rvert\Delta x$
x^α	$\alpha x^{\alpha-1}$	$\Delta(x^\alpha)=\lvert\alpha x^{\alpha-1}\rvert\Delta x$
$\sin(x)$	$\cos(x)$	$\Delta(\sin(x))=\lvert\cos(x)\rvert\Delta x$

$\cos(x)$	$-\sin(x)$	$\Delta(\cos(x))=\lvert\sin(x)\rvert\Delta x$
e^x	e^x	$\Delta(e^x)=e^x\Delta x$
$\ln(x)$	$1/x$	$\Delta(\ln(x))=\Delta x/\lvert x\rvert$
$u+v$	$u'+v'$	(u and v as functions of x)
uv	$u'v+v'u$	(product rule)
$\frac{u}{v}$	$\frac{u'v-v'u}{v^2}$	(quotient rule)
$f(g(x))$	$g'(x)f'(g(x))$	(chain rule)

- $\frac{1}{x}=x^{-1}$ so $(\frac{1}{x})'=(-1)x^{-1-1}=-\frac{1}{x^2}$.
- $\sqrt{x}=x^{\frac{1}{2}}$ then $(\sqrt{x})'=\frac{1}{2}x^{\frac{1}{2}-1}=\frac{1}{2\sqrt{x}}$.
- $(\sin(x^2))'=(x^2)'\cos(x^2)=2x\cos(x^2)$

B - Partial derivatives

A partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant. For example, consider the following function of three variables:

$$f(x,y,z) = x^2 - 2z + xy$$

We can look at the variations of this function with respect to a variable while considering the other constants. We proceed then as for a derivative. So we have:

$$\left(\frac{\partial f}{\partial x}\right)_{y,z}=2x+y \quad , \quad \left(\frac{\partial f}{\partial y}\right)_{x,z}=x \quad \text{and} \quad \left(\frac{\partial f}{\partial z}\right)_{x,y}=-2$$

The first expression is said "partial derivative of f with respect to x" treating y and z like constants. ∂: curly d.

C - Taylor series

With the notion of derivative we have studied the first-order behavior of a function around a point, here we refine to the higher orders.

For every infinitely differentiable function and for $\epsilon \ll 1$ we have the following development in the neighborhood of a point:

$$f(x_0+\epsilon)=f(x_0)+\epsilon f'(x_0)+\frac{\epsilon^2}{2}f''(x_0)+\frac{\epsilon^3}{3!}f^{(3)}(x_0)+...+\frac{\epsilon^n}{n!}f^{(n)}(x_0)+...$$

The more we take high order terms the better is the approximation. For example for $f(x)=\sin(x)$ and $x_0=0$ we find: $\sin(\epsilon)\simeq\epsilon-\frac{\epsilon^3}{3!}$, in the same way $\cos(\epsilon)\simeq 1-\frac{\epsilon^2}{2}$.

Also: $\exp(\epsilon)\simeq 1+\epsilon$ $\quad$ $\ln(1+\epsilon)\simeq\epsilon$ $\quad$ $(1+\epsilon)^\alpha\simeq 1+\alpha\epsilon$

D - Integrals

We show in maths that: $\int_{x=-\infty}^{+\infty} e^{-x^2}dx=\sqrt{\pi}$.

For the standard normal distribution, we can verify that the mean is zero and that the standard deviation is equal to 1:

$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ and $\mu=\int_{-\infty}^{+\infty}x\cdot p(x)dx=0$ because the integral over a symmetric interval of an odd function is zero.

$$\sigma^2=\int_{-\infty}^{+\infty}x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(-x)(-xe^{-\frac{x^2}{2}})dx \quad \text{then}$$

$$\sigma^2\sqrt{2\pi}=[-xe^{-\frac{x^2}{2}}]_{-\infty}^{+\infty}-\int_{-\infty}^{+\infty}(-1)e^{-\frac{x^2}{2}}dx=0+\sqrt{2}\int_{-\infty}^{+\infty}e^{-x^2}dx$$

so the standard deviation equals one.

We have used an integration by parts:

$$\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)dx$$

then we did an integration by substitution: $x' = x/\sqrt{2}$.

We can go further by calculating the skewness $\mu_3 = \int_{-\infty}^{+\infty} x^3 \cdot p(x)dx$, the kurtosis $\mu_4 = \int_{-\infty}^{+\infty} x^4 \cdot p(x)dx$, and, in general, the *moments* of order *n* $\mu_n = \int_{-\infty}^{+\infty} x^n \cdot p(x)\,dx$.

All these moments allow us to characterize a distribution.

For a Gaussian $\mu_3 = 0$ (symmetric). If this coefficient is negative the curve spreads to the left, if it is positive the curve spreads to the right. For a Gaussian $\beta_2 = \mu_4/\sigma^4 = 3$. If this coefficient is less than 3 the curve is more flat than a Gaussian. For a binomial distribution : $\beta_2 = 3 - 8/n$.

Integration by substitution:

Let $u = g^{-1}(x)$ be a new variable with g^{-1} a continuous function strictly monotonic on [a,b] and g the inverse function, then:

$$\int_a^b f(x)dx = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(u))g'(u)du$$

E – Series

Binomial formula: $(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$

Derivatives of geometric series:

$$\text{if} \quad |q|<1 \quad \text{then} \quad \sum_{k=0}^{\infty} q^k=\frac{1}{1-q}$$

we take the derivative with respect to q: $\sum_{k=1}^{\infty} k\,q^{k-1}=\frac{1}{(1-q)^2}$

$$\text{then} \quad \sum_{k=2}^{\infty} k(k-1)q^{k-2}=\frac{2}{(1-q)^3} \quad \ldots$$

And finally we find the negative binomial formula:

$$\sum_{k=r}^{\infty} k(k-1)\ldots(k-r+1)q^{k-r}=\frac{r!}{(1-q)^{r+1}}$$

$$\text{so} \quad \frac{1}{(1-q)^{r+1}}=\sum_{k=r}^{\infty}\binom{k}{r}q^{k-r}$$

A definition of the exponential function: $e^x=\sum_{k=0}^{\infty}\frac{x^k}{k!}$

F - Gamma function $\Gamma(x)=\int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}dt$

This function is an extension of the factorial to real and complex number (except for 0, -1, -2...)[22]. We will use it for half-integer numbers. We demonstrate with an integration by parts: $\Gamma(x+1)=x\Gamma(x)$. $\Gamma(1)=1$ then for n integer $\Gamma(n+1)=n!$. Moreover, $\Gamma(1/2)=\sqrt{\pi}$ allows to calculate the function for the half-integers.

---

22 For example $\pi!\simeq 7.2$.

VII. Answers to Exercises

Chapter I : Random Variable

E1 : Ages *Exercise on page 37*

n=15; mode=18; median=18; mean≃18.333; geometric mean≃18.303; range=4; standard deviation≃1.113; root mean square deviation≃1.075; mean deviation≃0.844.

E2 : Card Game *Exercise on page 37*

Number of possible draws: 32x31x30x29x28 divided by the different ways of arranging 5 cards 5x4x3x2x1=5! permutations then **201 376** hands (32 choose 5).

* Number of possible draws for a four aces hand: only one possiblity for the 4 aces, and 28 possibilities for the fifth card, so **28** possible hands (one hand = 120 possible draws).

Hence the probability p=28/201376=139 chances on a million=1 out of 7192=**0,014%**.

* For a flush: each color has 8 cards. Number of ways to choose 5 cards among 8 = $\frac{8!}{5!(8-5)!} = \binom{8}{5} = C_8^5 = 56$.

There are 4 colors : 4x56 =224.

Hence the probability p = 224 / 201 376 = 1 out of 899 = **0,11%**.

E3 : Gravity Field *Exercise on page 37*

a) We have an important dispersion. We could indicate less significant figures.

b) $\bar{g} = 9.32$ m/s^2. σ_g=1.994m/s^2≃2m/s^2

c) We assume that the data follow a Gaussian distribution, then we can use a Student's distribution for the sampling distribution:

$\Delta g = 2.36 \cdot 1.994/\sqrt{8} = 1.66$ m/s^2. Then g=9.32±1.66m/s^2 and the known value is within the confidence interval found:

7.66 m/s^2 < g = 9.81m/s^2 < 10.98 m/s^2 with 95% confidence.

d) We are at the distance σ from the center g≃10m/s^2, then a probability of 68% (prediction interval: Gaussian).

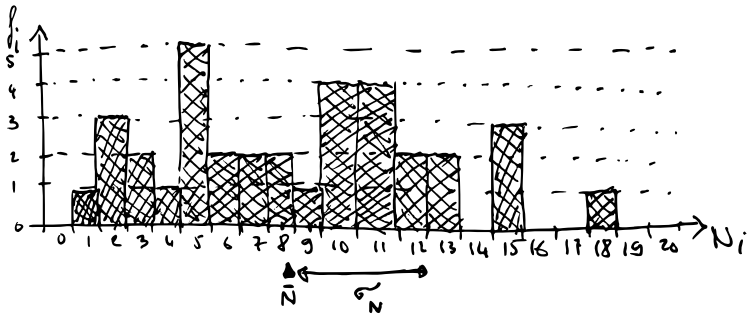
E4 : Elevator *Exercise on page 38* 25 out of 1000.

E5 : Assignment *Exercise on page 38*

a) n=35, $\overline{N}$≃8.29 and σ_N≃4.40 .

b) Grades with their frequencies:

Grades	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Freq.	0	1	3	2	1	5	2	2	2	1	4	4	2	2	0	3	0	0	1	0	0

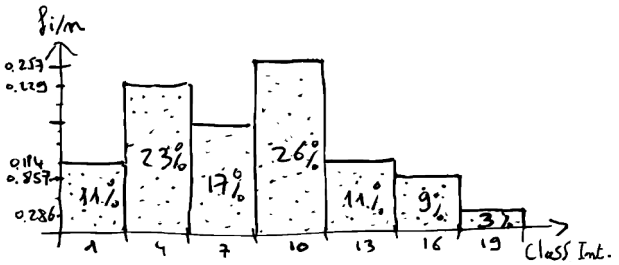


c) Class intervals, frequencies and relative frequencies:

Class I.	[0,1,2]	[3,4,5]	[6,7,8]	[9,10,11]	[12,13,14]	[15,16,17]	[18,19,20]
Freq.	4	8	6	9	4	3	1
F. rel.	11.4%	22.9%	17.1%	25.7%	11.4%	8.57%	2.86%

When we group the data into class intervals there is a loss of informa-

tion. But we gain visibility. We can more easily bring out an overall behavior and categories of students. Relative frequencies directly indicate the percentage of students in each class, for example, 34% of students appear in great difficulty, which is more difficult to interpret on the first diagram.



E6 : Yahtzee *Exercise on page 38*

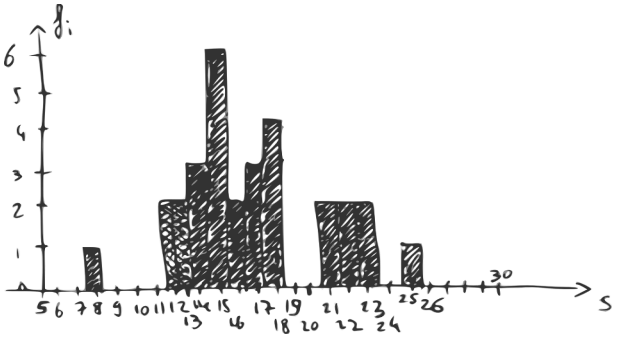
1) 1/1296

2) $(1+5+15+35+70)/6^5=1.65\%$

3) a) n=30, $\overline{S}\simeq 16{,}73$ and $\sigma_S\simeq 3.88$.

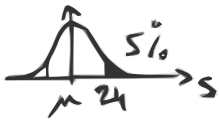
b) For the sum of five dice we get a distribution of the population already close to a Gaussian (see on p18 for four dice) and we use the Student t-value: $\Delta\overline{S}\simeq 1.45$ and $\overline{S}=16.73\pm 1.45$ with a confidence of 95%. Here the population is infinite and known. We find $\mu_{pop}=17.5$. It is consistent. There is no reason to think that the dice are unbalanced.

c)



d) The sum is the result of five independent throws, we can think ac-

cording to the central limit theorem that we approach a Gaussian: $(24-\mu)/\sigma_S \simeq 1.68$. We are close to $t_{\infty,90\%}$, so half 5% on the right tail. So 1 chance out of 20 (on our throws here: 1/30, it's all right).



E7 : Elastic Bands *Exercise on page 39*

H_o : 10 elastic bands out of 1000 are not functional.

Let consider the discrete random variable X_o, its value is 1 if the elastic is not functional and 0 if the elastic works. Probability to have a non functional elastic band: p=0.01

$\mu=0.01x1+0.99x0=0.01$ and $\sigma^2=0.01x0.99^2+0.99x0.01^2$ then $\sigma\simeq 0.0995$

We have $\bar{x}=\mu+t_\infty\sigma/\sqrt{n}$ and $\bar{x}=n_o/n$ then $n_o=n(\mu+t_\infty\sigma/\sqrt{n})$.

*For n=1000: n is large and according to the central limit theorem we have a normal distribution. We look for a value of n_o that has right tail probability 1%.

So $n_o=1000(0.01+2.33\times 0.099/\sqrt{1000})\simeq 17.3$ and we reject the delivery from n_o=18 damaged elastics among the 1000 randomly drawn. Using the binomial law $\underline{n_o=19}$, close to 18 (see below for the explanation).

*For n=200: on the basis of the same reasoning we get n_o=6, indeed $n_o\simeq 200(0.01+2.33\times 0.099/\sqrt{200})\simeq 5.3$. But is the hypothesis of large numbers justified here? *n* must be large, for example greater than 30, but also *np* and *n(1-p)* must also be sufficiently large, for example greater than 5. As here *p* is not around fifty-fifty but on the contrary close to an edge, unlikely or almost certain, we will check the validity by making an exact calculation.

We have a sequence of identical and independent trials (the large number of elastics in the batches delivered make it possible to consider the draws with replacement). Trials with two outcomes, success (S) with the probability *p* or failure (F) with the probability *q=1-p*. We recognize a binomial distribution with parameters *n* and *p* (see on page Erreur : source de la référence non trouvée for more details). We call X the law which corresponds to the number *k* of successes, number of

damaged elastics on *n* tests. We look for the value of $k=n_o$ from which we would reject H_o with less than one chance in a hundred to take a wrong decision: $P(X \geqslant k) \leqslant 1\%$.

$$P(X \geqslant 6) = 1 - P(X \leqslant 5) \simeq 1.6\% \quad \text{and} \quad P(X \geqslant 7) \simeq 0.4\%$$

(on a spreadsheet : LOI.BINOMIALE(k;n;p;1)). So $\underline{n_o=7}$.

*For n=50: with the binomial distribution $\underline{n_o=4}$ (we would have found a different result with the central limit theorem $n_o \simeq 2.1$ and $n_o=3$).

E8 : Testing an insulating panel *Exercise on page 39*

x = 39.28 ± 0.49 mW/m/K with a confidence of 95% and a relative uncertainty of 1.2% (the distribution of the population is assumed to be normal and the sampling distribution of Student). Results in accordance with expectations. According to these results the manufacturer could announce a lower margin. According to a conventional bilateral Student test the value announced by the constructor is well within the confidence interval. The manufacturer indicated a mean but did not give a standard deviation, we estimated it with the sample: $\sigma \simeq s \simeq 0.69$. The probability α of rejecting the null hypothesis when it is true is 23% (with $t = (\bar{x} - \mu)\sqrt{n}/\sigma \simeq 1.29$ and =TDIST(1.29,9,2) in LibreOffice), if the compliance hypothesis were rejected there would be a 23% chance of wrongly rejecting it, which is a far too high risk.

E9 : Coins *Exercise on page 40*

1) $\chi^2 \simeq 2.56 < 3.84$. There is no reason to question the equilibrium of the coins. The difference with respect to the equiprobability is explained by the statistical fluctuations. Or by a conventional bilateral test (prediction interval):

$$\left|(\bar{x} = 42/100) - (\mu = 0.5)\right| \leqslant t_{\infty\,95\%}(\sigma = 0.5)/\sqrt{n = 100}$$

2) $\chi^2 \simeq 0.4 < 3.84$. Likewise. Probability $\alpha \simeq 20.8\%$, we have a one in five chance of rejecting the assumption then what is right, one can not reject it.

3) $\chi^2 \simeq 25.6 >> 3.84$! The coin is certainly not balanced. One chance out of two million to reject the hypothesis when it is right, we reject it!

E10 : Parity *Exercise on page 40*

Expected values	
288.5	288.5
174	174
6	6

$\chi^2 = \frac{(470-288.5)^2}{288.5} + \frac{(107-288.5)^2}{288.5} + \ldots$ and $\chi^2 \simeq 344$ with ddl=2x1=2 the probability to reject the hypothesis when it is true is much less than one chance in 1000 (13.8 <<344 for 0.1%). The parity is not respected.

Understanding these results would require extreme bad faith to argue that there is no reason to believe that parity is not respected by justifying deviations by statistical fluctuations.

E11 : Births *Exercise on page 41*

χ^2 ≃6.16 < 7.8. The hypothesis can not be rejected, we need more data.

χ^2 ≃128 >> 7.8. Hypothesis totally rejected.

E12 : Gaussian distributions in the plane and the space

Exercise on page 42

1D :

1- $\bar{x} = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0$ since the integrated function is defined and odd on a symmetric integration interval (integral property). $\overline{|x|} = \int_{-\infty}^{+\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2\int_{0}^{+\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ because this time the function is even. $\overline{|x|} = \frac{-2}{\sqrt{2\pi}}[e^{-\frac{x^2}{2}}] = \sqrt{\frac{2}{\pi}}$. The distance corresponds to an absolute value, so its mean is non-zero and positive. $\bar{x}$ is the mean of the abscissa, an algebraic quantity.

$\sigma_x = \sigma_{|x|} = 1$, because $x^2 = |x|^2$ and we have a normalized distribution.

2- $P(|x|\leqslant 1)=P(-1\leqslant x\leqslant 1)=\int_{-1}^{-1}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$ we calculate these integrals numerically (with the use of a calculator, on a computer ...) and find the results of the chapter on the Gauss distribution at σ, two σ and three σ.

2D :

1- $p(x,y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}$ *p(x,y)* is a positive function, like we expect for a probability.

On the whole plan we have a probability of 100%:

$$\iint p(x,y)dx\,dy=\int\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\ \int\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=1\times 1=1$$

2- $\iint p(x,y)\,dx\,dy=\int\frac{1}{2\pi}e^{-\frac{\rho^2}{2}}2\pi\rho\,d\rho=1$ $p(\rho)=\rho e^{-\frac{\rho^2}{2}}$

and $\int_0^{+\infty}p(\rho)\ d\rho=-[e^{-\frac{\rho^2}{2}}]=1$.

$$\bar{\rho}=\int_0^{+\infty}\rho\,p(\rho)\ d\rho=\int\rho^2e^{-\frac{\rho^2}{2}}\ d\rho=[\rho^2(-\rho)e^{-\frac{\rho^2}{2}}]-\int 2\rho e^{-\frac{\rho^2}{2}}\ d\rho$$

$\bar{\rho}=0-0+\sqrt{2}\int_0^{+\infty}e^{-t^2}dt=\sqrt{2}\frac{\sqrt{\pi}}{2}=\sqrt{\frac{\pi}{2}}$ after an integration by parts ($u=\rho$) and an integration by substitution ($\rho=\sqrt{2}\ t$ and we recognize a known integral).

$$\sigma_\rho{}^2=\overline{\rho^2}=\int_0^{+\infty}\rho^2p(\rho)\ d\rho=\int\rho^3e^{-\frac{\rho^2}{2}}\ d\rho=[\rho^2(-e^{-\frac{\rho^2}{2}})]+2\int\rho e^{-\frac{\rho^2}{2}}\ d\rho$$

We find an integral computed in 1D: $\sigma_\rho{}^2=0-0+2\times 1$ $\sigma_\rho=\sqrt{2}$

3- $P(\rho\leqslant\sigma_\rho)=\int_0^{\sqrt{2}}\rho e^{-\frac{\rho^2}{2}}\ d\rho=1-\frac{1}{e}$ $P(\rho\leqslant 2\sigma_\rho)=1-\frac{1}{e^4}$

3D :

1- $p(x,y,z)=\frac{1}{(\sqrt{2\pi})^3}e^{-\frac{r^2}{2}}$ $p(r)=\alpha r^2 e^{-\frac{r^2}{2}}$ $\alpha=\frac{4\pi}{(\sqrt{2\pi})^3}$

2- $\int_0^{+\infty} p(r)dr=\alpha\times\sqrt{\frac{\pi}{2}}=1$ (Integral already calculated in 2D)

3- $\bar{r}=\int_0^{+\infty} r\,p(r)dr=\alpha\int_0^{+\infty} r^3 e^{-\frac{r^2}{2}}dr=\alpha\times 2=\frac{4}{\sqrt{2\pi}}$

$\sigma_r^2=\int_0^{+\infty} r^2\,p(r)dr=\alpha\int_0^{+\infty} r^4 e^{-\frac{r^2}{2}}dr=3$ by integration by parts

so $\sigma_r=\sqrt{3}$

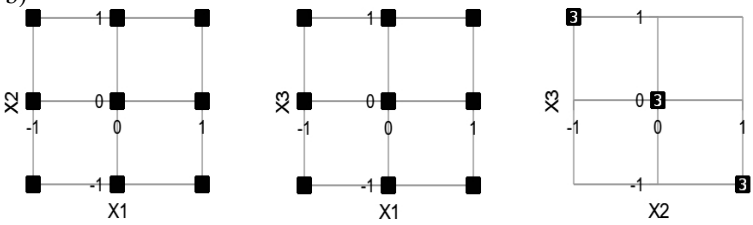
4- $P(r\leqslant\sigma_r)=\int_0^{\sqrt{3}}\rho e^{-\frac{r^2}{2}}dr\simeq 0.608$ and the other two also by numerical calculation.

Chapitre II : Correlation and Independence

E1 : Correlations *Exercise on page 88*

1- a) $\bar{x}_1=(-1-1-1+0+0+0+1+1+1)/9=0$ also $\bar{x}_2=0$ and $\bar{x}_3=0$.
$\sigma_1=\sqrt{\sum_{i=1}^{9}(x_{i1}-\bar{x}_1)^2/(9-1)}=\sqrt{6x1/8}$ so
$\sigma_1=\sigma_2=\sigma_3=\sqrt{3}/2\simeq 0.87$.

b)



c) $r_{12}=0$. $r_{13}=0$. $r_{23}=-1$. X_1 and X_2 are not correlated. The same for X_1 and X_3. X_2 and X_3 are dependents and totally correlated.

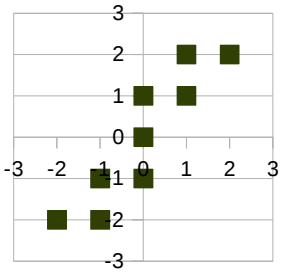
2- a) $\bar{x}_1=0$ and $\bar{x}_2=0$. $\sigma_1\simeq 1.22$ and $\sigma_2\simeq 1.58$.

c) $r_{12}=0.904$. Quantities are positively correlated.

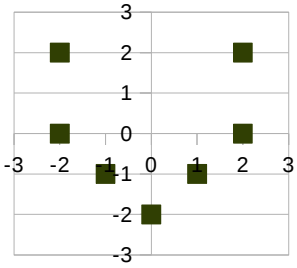
3- a) $\bar{x}_1=0$ and $\bar{x}_2=0$. $\sigma_1\simeq 1.73$ and $\sigma_2\simeq 1.53$.

c) $r_{12}=0$. Quantities completely uncorrelated, do not forget that the correlations sought here are **linear**. There is a correlation in the form of a V.

2- b)



3- c)



E2 : Volumes *Exercise on page 89*

1-

$\overline{V}_1=(100.1+100.0+99.9+100.0)/4$ so $\mathbf{\overline{V}_1 = 100.0\ mL}$.

$$\sigma_1=\sqrt{\frac{0.1^2+0^2+(-0.1)^2+0^2}{4-1}}=\sqrt{\frac{2}{3}}\cdot 0.1\,mL \quad \text{then} \quad \sigma_1\simeq 0.082\,mL$$

According to the central limit theorem and the distribution of the population Gaussian: $\Delta V=t.\sigma/\sqrt{n}=3.18x0.0816/2$

$t_{ddl=3;\ 95\%}=3.18$ so $\mathbf{\Delta\overline{V}_1 \simeq 0.13\ mL}$ and $\Delta\overline{V}_1/\overline{V}1 \simeq 0.13/100$

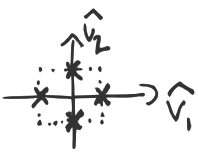
The pipette, withe a confidence of 95%, has a uncertainty of 0.13 mL, so for 100 mL a percent uncertainty of 0.13%.

2-

$\hat{V}^i=V^i-\overline{V}$ and

$$\sum_i[(V_1{}^i-\overline{V_1})(V_2{}^i-\overline{V_2})]=\sum_i[\hat{V_1}{}^i\hat{V_2}{}^i]=0.1\times 0+0\times 0.1+(-0.1)\times 0+0\times 0.1=0$$

by definition $r_{12}=\dfrac{\sum_i[(V_1{}^i-\overline{V_1})(V_2{}^i-\overline{V_2})]}{\sqrt{\sum_i(V_1{}^i-\overline{V_1})^2\sum_i(V_2{}^i-\overline{V_2})^2}}$



so $\boldsymbol{r_{12}=0}$, the quantities are totally uncorrelated and therefore independent.

3-

V={200.1 , 200.1 , 199.9 , 199.9}mL so $\mathbf{\overline{V} = 200\ mL}$.

$$\sigma_V=\sqrt{\frac{0.1^2+0.1^2+(-0.1)^2+(-0.1)^2}{4-1}}=\frac{2}{\sqrt{3}}\cdot 0.1\,mL$$

then $\sigma_V\simeq 0.115\,mL$ and $\mathbf{\Delta\overline{V}\simeq 0.183\ mL}$, $\Delta\overline{V}/\overline{V}\simeq 0.09\%$

4-

$V(V_1,V_2)$ then we have:

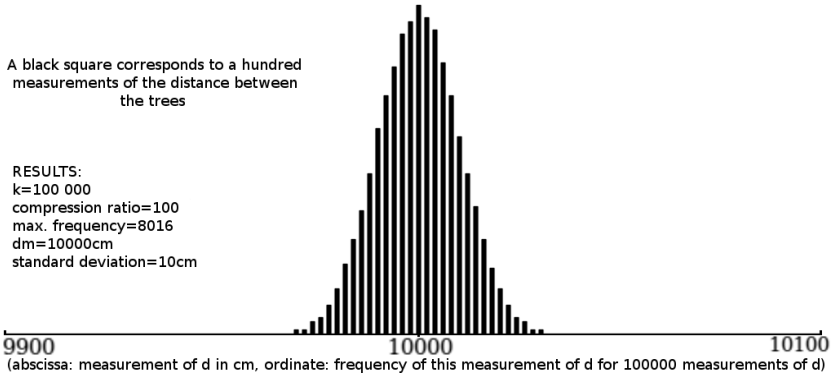
$$\Delta V^2=\left(\frac{\partial V}{\partial V_1}\right)^2_{V_2}\Delta V_1{}^2+\left(\frac{\partial V}{\partial V_2}\right)^2_{V_1}\Delta V_2{}^2=\Delta V_1{}^2+\Delta V_2{}^2$$

and $\Delta V=\sqrt{2}.\Delta V_1\simeq 0.18$ Same result as in question 3-.

E3 : Trees *Exercise on page 90*

We have $d=\sum x_i$ (also written $d=x_1+x_2+...+x_i+...+x_n$), where the x_i are the length measures for each displacement of the stick. Δx_i is the uncertainty on each measurement x_i. Here $\Delta x_i=1cm$ and $x_i=100\pm1cm$. What is then the uncertainty Δd ?

One would think that $\Delta d=\sum \Delta x_i$, so $\Delta d=100cm$ and 1m of uncertainty per 100m measured from one tree to another. But there is a problem, this is absolutely not what the simulation indicates below!



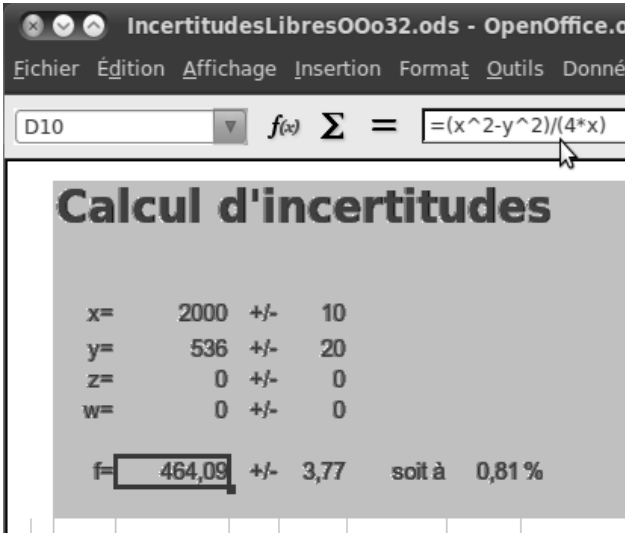
Rather, it indicates $\Delta d=0.1m$. Indeed a simple calculation of probabilities shows that it is extremely unlikely to obtain for d, $99m$ or $101m$, whereas one has a chance in two, for xi, to have $0.99m$ or $1.01m$. Now we want to have Δd with equal probability of Δxi.

We get $\partial d/\partial x_i=1$ and $\Delta x_1=\Delta x$, whatever $i=1...n$, so $\Delta d=\sqrt{n}.\Delta x$ and $\Delta d=0.1m$, which corresponds well to the expected result!

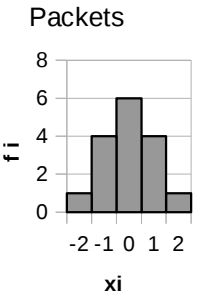
E4 : The two-position method *Exercise on page 90*

$(\partial f'/\partial D)_d=(D^2+d^2)/4D^2$ and $(\partial f'\partial/d)_D=-d/2D$, then
$\Delta f'=\sqrt{[((D^2+d^2)/4D^2)^2(\Delta D)^2+(d/2D)^2(\Delta d)^2]}$
so **f' = 464.1±3,8 mm**
and $\Delta f'/f'=0.8\%$.

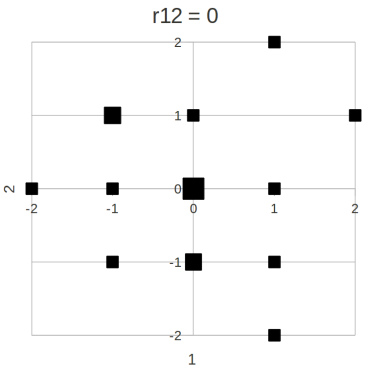
Either by a numerical calculation with a spreadsheet and macros (no need then to perform calculations of partial derivatives):

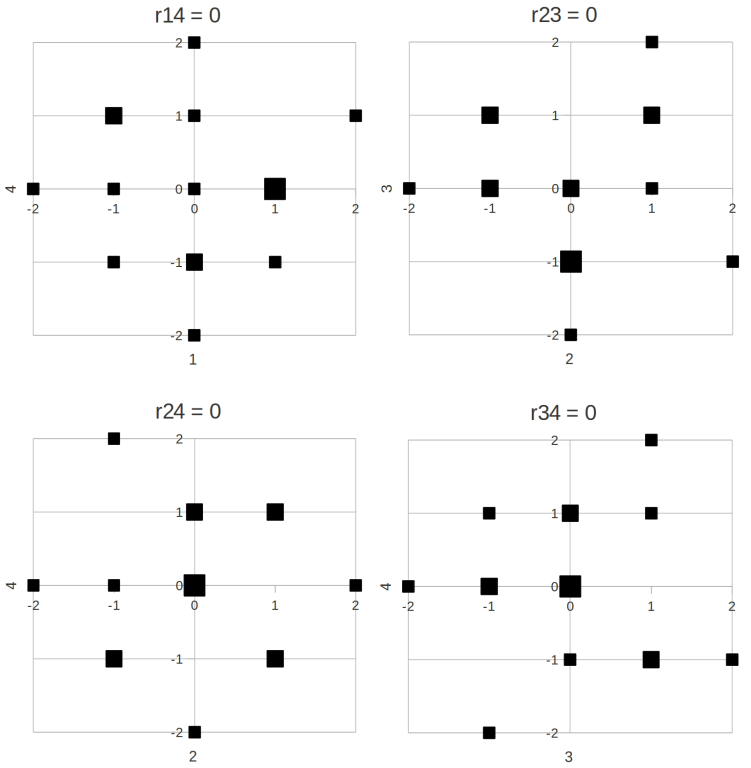


The spreadsheet uses here 2 of the 4 perfectly not correlated globally Gaussian packets available:



1	2	3	4	x	y	f
-2	0	-2	0	1973.71	536	457.04
-1	-1	0	0	1986.86	509.71	464.02
-1	0	-1	1	1986.86	536	460.56
-1	1	1	1	1986.86	562.29	456.93
-1	1	2	-1	1986.86	562.29	456.93
0	0	0	0	2000	536	464.09
0	0	-1	-2	2000	536	464.09
0	-1	1	-1	2000	509.71	467.52
0	1	1	-1	2000	562.29	460.48
0	-1	1	2	2000	509.71	467.52
0	0	0	1	2000	536	464.09
1	-2	0	0	2013.14	483.43	474.26
1	0	-1	0	2013.14	536	467.61
1	-1	0	-1	2013.14	509.71	471.02
1	2	-1	0	2013.14	588.57	460.27
2	1	0	1	2026.29	562.29	467.56





E5 : Refractive Index *Exercise on page 90*

According the Snell's law for the refraction: $n_2 = n_1 \frac{\sin(i_1)}{\sin(i_2)}$

so $\mathbf{n_2 \simeq 1.462}$

$n_2(i_1,i_2)$ and

$$\Delta n_2^2 = \left(\frac{\partial n_2}{\partial i_1}\right)_{i_2}^2 \Delta i_1^2 + \left(\frac{\partial n_2}{\partial i_2}\right)_{i_1}^2 \Delta i_2^2$$

(The variation with respect to the index n_1 is not included because n_1 is assumed to be known with a great accuracy)

So $\left(\frac{\partial n_2}{\partial i_1}\right)_{i_2} = \frac{n_1}{\sin(i_2)}\cos(i_1)$ and $\left(\frac{\partial n_2}{\partial i_2}\right)_{i_1} = n_1 \sin(i_1)\frac{-\cos(i_2)}{\sin^2(i_2)}$.

hence $\Delta n_2 = \left[\left(\frac{\cos(30°)}{\sin(20°)}\frac{\pi}{180}\right)^2 + \left(\frac{\sin(30°)\cos(20°)}{\sin^2(20°)}\frac{2\pi}{180}\right)^2\right]^{1/2}$

(the angles are placed in their natural unit, the radians, dimensionless quantity: π rad = 180°)

Finally $\Delta n_2 = 0.1470$ with 95% confidence
then ***n_2=1.46±0.15*** and $\Delta n_2 / n_2$=10%.

With the spreadsheet: n_2=*1.46±0.16* and $\Delta n_2/n_2$=11%.The numerical method is approximate but we verify that we have not made a calculation error.

Moreover, with a goniometer we can make measurements of angles much more accurate to a few minutes of arc (one arc minute = 1' = 1°/60).

E6 : Cauchy's equation *Exercise on page 90*

1 -

$$\Delta n^2 = \left(\frac{\partial n}{\partial D_m}\right)_A^2 \Delta D_m{}^2 + \left(\frac{\partial n}{\partial A}\right)_{D_m}^2 \Delta A^2$$

with $\left(\frac{\partial n}{\partial D_m}\right)_A = \frac{\cos[(A+D_m)/2]}{2\sin(A/2)}$ and

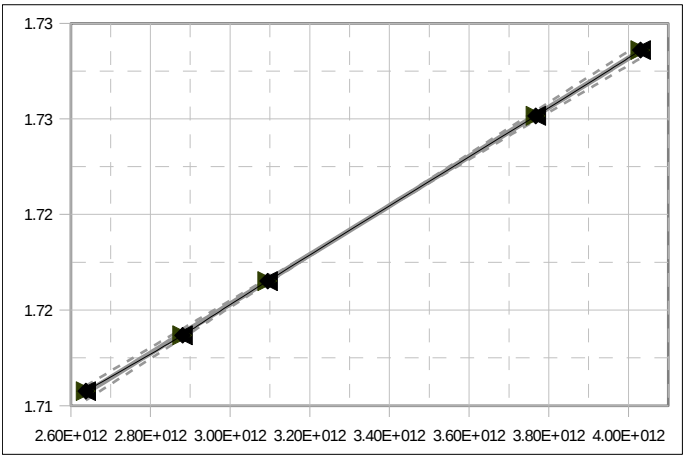
$$\left(\frac{\partial n}{\partial A}\right)_{D_m} = \frac{\cos[(A+D_m)/2]\sin(A/2) - \sin[(A+D_m)/2]\cos(A/2)}{2\sin^2(A/2)}$$

with $A = 60° \simeq D_m$: $\left(\frac{\partial n}{\partial D_m}\right)_A \simeq \frac{1}{2}$ and $\left(\frac{\partial n}{\partial A}\right)_{D_m} \simeq -1$

then $\underline{\Delta n \simeq 0.00065}$ and $\underline{\Delta n/n \simeq 0.04\%}$
(we find the same result with a numerical calculation)

2-
With error bars, $\Delta n = 0.00065$ and $\Delta(1/\lambda^2) = 2\Delta\lambda/\lambda^3$ with $\Delta\lambda = 0.05$nm, we find **$A = 1.67877 \pm 0.00182$**; $\Delta A/A = 0.109\%$;
$B = (1.285 \pm 0.055) \times 10^{-14}\ m^2$ and $\Delta B/B = 4.26\%$

n as a function of $1/\lambda^2$ ($1/m^2$) :



A simple regression give **r=0.99995 ...**

(If we miss the error bars $A = 1.67877 \pm 0.00075$; $\Delta A/A = 0.044\%$; $B = (1.285 \pm 0.022) \times 10^{-14}\ m^2$ et $\Delta B/B = 1.75\%$)

3- With a simple regression:

- r=0.99947... for *n(1/λ)*
- r=0.99995... for *n($1/\lambda^2$)*
- r=0.99982... for *n($1/\lambda^3$)*

An expression in $1/\lambda^2$ is therefore better,
But what value of α allows to optimize the regression?

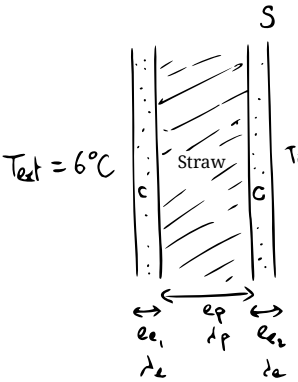
To determine this value we plot $y=\ln(n-A)=\alpha x+\ln(B)$ with $x=\ln(\lambda)$. α and ln(B) will be obtained with their uncertainties and A will be chosen to maximize *r*.

The results are as follows $r_{max}\simeq$-0.99997024 pour A≃1.6843744 and we find: α= - 2.306 ± 0.033 . Then with the error bars: **α= - 2.3 ± 0.1** . Conclusion α= - 2 is not in the interval, the model with $1/\lambda^2$ is not validated.

That said, it works properly. The difference can be explained simply because the theory of light-matter interaction used to find the Cauchy's equation give a more complicated function. Here we consider a Taylor expansion of this function and actually there are more terms: $n=A+\frac{B}{\lambda^2}+\frac{C}{\lambda^4}$. A more advanced study could verify this by including the new $1/\lambda^4$ term.

We should also check our experimental uncertainties, it is possible that we have neglected or underestimated sources of uncertainty (for example by tripling the angular uncertainties up to 6': α= - 2.3 ± 0.3 ...).

E7 : Wall *Exercise on page 92*



1-

$$R_p=\frac{e_p}{\lambda_p S}=\frac{0.4\,mK\,m}{0.045\,W\times 72\,m^2}$$

so $R_p=0.123\,K/W$

$$\left(\frac{\Delta R_p}{R_p}\right)^2=\left(\frac{\Delta e_p}{e_p}\right)^2+\left(\frac{\Delta \lambda_p}{\lambda_p}\right)^2$$

and $\Delta R_p/R_p=0.103$

$$R_p = 0.123 \pm 0.013\, K/W$$

2- $R_e = 0.0069 \pm 0.0010\, K/W$

3- The resistances are in series: $R=R_p+R_e$
et $\Delta R = \sqrt{(\Delta R_p)^2 + (\Delta R_e)^2}$ and $R = 0.130 \pm 0.015\, K/W$

4- $\Phi = \dfrac{\Delta T}{R} = \dfrac{12}{0.130} = 92.3\, W$

$(\Delta\Phi/\Phi)^2 = (\Delta(\Delta T)/\Delta T)^2 + (\Delta R/R)^2 \qquad \Delta T = T_{intérieur} - T_{extérieur}$

then $\Delta(\Delta T) = \sqrt{2} \times 0.5\,°C$, $\Delta\Phi/\Phi$ = 12.9% et
$80\, W \leqslant \Phi \leqslant 104.3\, W$ hence a minimum heating power:

P_{min} = 105 W

E8 : Insulation et Inertia *Exercise on page 93*

1- In parallel: $\dfrac{1}{R_{éq}} = \sum_i \left(\dfrac{1}{R_i}\right) = \sum_i \left(\dfrac{\lambda_i S_i}{e_i}\right)$

or wit the thermal conductances $G=1/R$: $G_{éq} = \sum_i G_i$

then in the present case:

$$\frac{1}{R} = \frac{S_s}{(e_s/\lambda_s)} + \frac{S_t}{(e_t/\lambda_t)} + \frac{S_m}{(e_m/\lambda_m)} + \frac{S_h}{(e_h/\lambda_h)} = \frac{36}{4} + \frac{54}{8} + \frac{82}{4} + \frac{8}{1}$$

concerning the uncertainties:

$\Delta G = \sqrt{\sum_i \Delta G_i^2}$, $\Delta G_i = S_i \dfrac{\Delta(e_i/\lambda_i)}{(e_i/\lambda_i)^2}$, $\Delta G \simeq 2.471\, W/K$

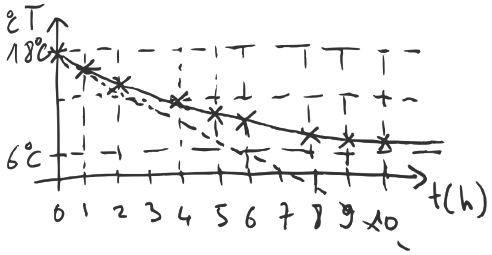
and $\Delta R = R^2 \Delta G \simeq 1.26\, mK/W$

then $R = 22.6 \pm 1.3\, mK/W$ and $\Delta R/R$ = 5.6 %

2- $\Phi = \frac{\Delta T}{R} = \frac{12}{22.6 \times 10^{-3}} \simeq 531\,W$ ± 8.1% (As in the exercise *Wall*)
and a minimum heating power:

P_{min} = 575 W

3-

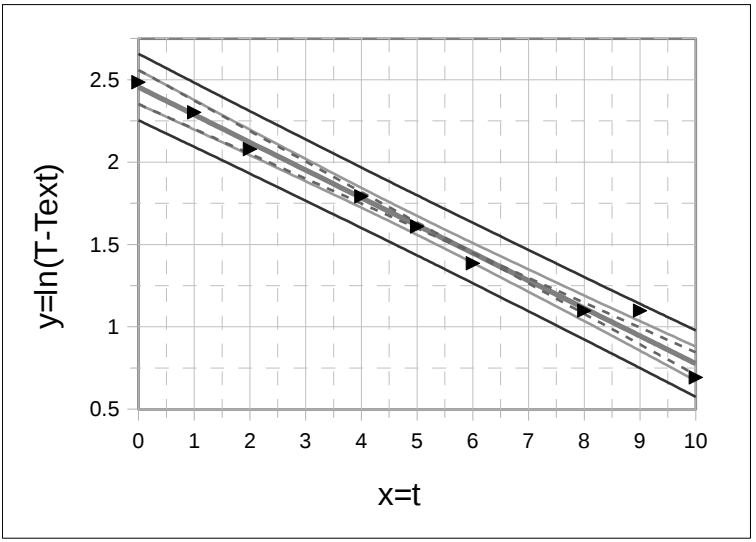


We necessarily have an horizontal asymptote $T=T_{ext}$: the temperature can not drop spontaneously below T_{ext}.

If the curve was linear the temperature would tend towards -∞ with time.

$T-b=a\,e^{-t/\tau}$ and $\ln(T-b)=\ln a-\frac{1}{\tau}t$,

we find ***b=T_{ext}= 5.7°C*** (maximum correlation coefficient *r*),
Then we do a regression on a spreadsheet with $y=\ln(T-b)$ and $x=t$ (on page 2 in the file IncertitudesLibres):



slope = $-1/\tau = -0.156 \pm 0.0157\ h^{-1}$

y-intercept = $\ln a = \ln (T_{int}-T_{ext}) = 2.474 \pm 0.0949$

$r_{max} \simeq -0.9937170474$ (the best alignment of the points)

we have: $\frac{\Delta \tau}{\tau} = \frac{\Delta (1/\tau)}{(1/\tau)}$ and $\Delta x = x \,.\, \Delta(\ln x)$

Then : $\boldsymbol{\tau = 6\ h\ 25\ min}$ **within** ***10.1%*** and $\boldsymbol{T_{int} = 17.5 \pm 1.2\ ^{\circ}C}$.

4-

a) energy outflux during dt = thermal energy lost during dt

so $\Phi\, dt = \frac{\Delta T}{R}\, dt = \frac{T(t) - T_{ext}}{R} dt = C\, dT$

and $\frac{dT}{dt} + \frac{1}{\tau} T = \frac{1}{\tau} T_{ext}$ with $\tau = RC$

The general solution has been given in question 3-.

b) $C = \tau / R$ and $(\Delta C / C)^2 = (\Delta \tau / \tau)^2 + (\Delta R / R)^2$

and $\boldsymbol{C = 1034 \pm 119\ kJ/K}$

E9 : Yield *Exercise on page 94*

1-
On a spreadsheet (regression line on the graph):

x	y	(xi-xm) *(yi-ym)	(xi-xm)²	yim	(yi-yim)²	xi^2	(xi-xm)²	(yi-ym)²	(xi-xm) *(yi-ym)
100	**41**	5271.43	90000	40.79	0.05	10000	90000	308.76	5271.4
200	**44**	2914.29	40000	46.71	7.37	40000	40000	212.33	2914.3
300	**53**	557.14	10000	52.64	0.13	90000	10000	31.04	557.14
400	**63**	0	0	58.57	19.61	160000	0	19.61	0
500	**66**	742.86	10000	64.5	2.25	250000	10000	55.18	742.86
600	**65**	1285.71	40000	70.43	29.47	360000	40000	41.33	1285.7
700	**78**	5828.57	90000	76.36	2.7	490000	90000	377.47	5828.6

xm =	ym =	Cov(x,y) =	Var(x) =	sy/x =	sx2 =	sx =	sy =	s2xy =
400	58.57	16600	280000	3.51	1400000	216.02	13.2	2766.7

da = 0,017	**da/|a| = 29%**
db = 8	**db/|b| = 22%**

dy = 9.021

r = 0.9701

r: linear and algebraic correlation coefficient.

a =	**0.059**	**+ ou -**	**0.017**
b =	**35**	**+ ou -**	**8**
0.042	< a <		0.076
27	< b <		42

sa = 0.00663

sb = 2.96579

Critical linear correlation coefficient:

rc = 0.669

Linear correlation test:

if |r|<rc then quantities not correlated

if |r|>rc then quantities correlated

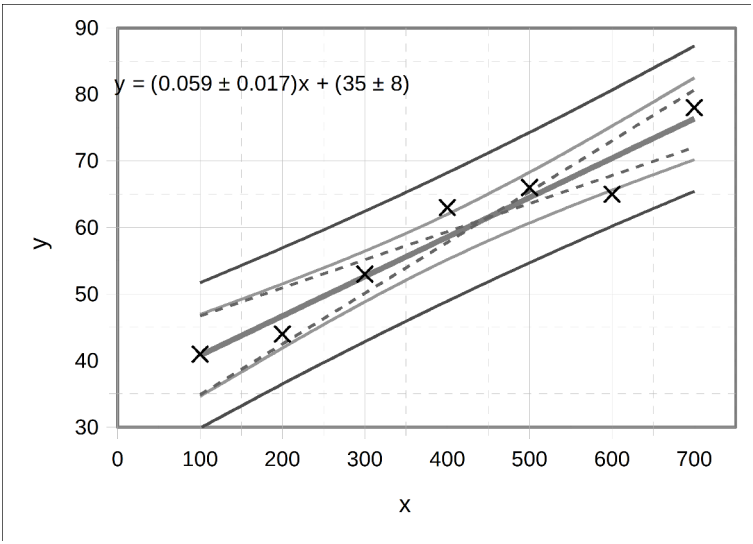
Here the quantities are correlated.

R= 0.941

R: coefficient of determination (R=r^2).

94.1% of the variation of y is explained with that one of x

Yield as a function of the amount of fertilizer:



2- The yield would then be of 67 Q/ha ± 4 Q/ha with a confidence of 95 pour cent.

For xo= 550

Estimate of the mean of yo:

63	< yo <	72
yom=		67
dyom=		4
dyom/|yom|=		6%

3- The yield would then be of 35 Q/ha ± 8 Q/ha with a confidence of 95 pour cent (y-intercept).

For xo= 0

Estimate of the mean of yo:

27	< yo <	42
yom=		35
dyom=		8
dyom/|yom|=		22%

4- The probability is of 95% (2σ prediction interval).

For xo= 250

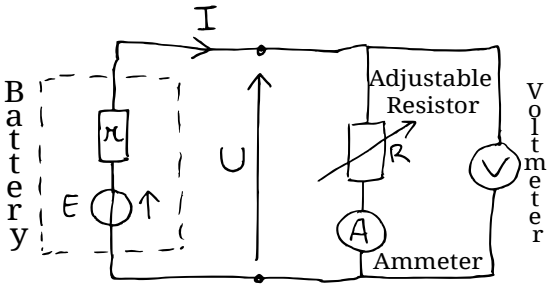
Prediction for yo:

40	< yo <	60
yom=		50
dyom=		10
dyom/|yom|=		20%

E10: Study of a battery *Exercise on page 95*

To characterize an electrical component we can plot U(I) with the following circuit:

In this case we study a battery.
By varying R the circuit works on different points of the voltage-current characteristics of our battery.



There are two possible circuits depending on the positions of the ammeter and the voltmeter ($R_A \to 0$ and $R_V \to \infty$).

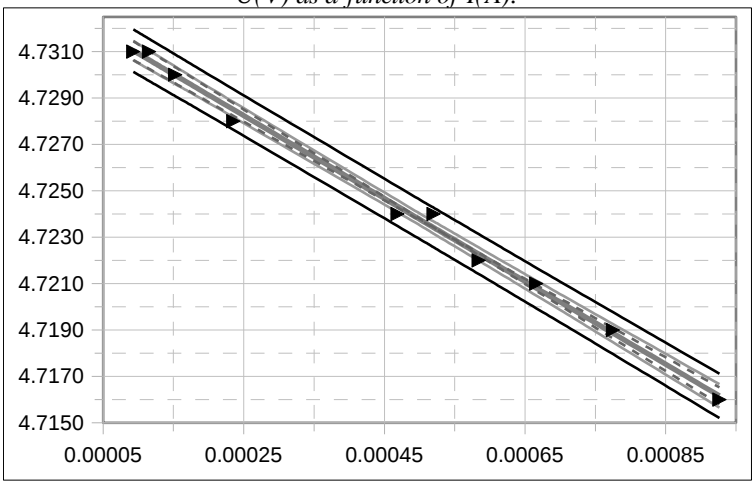
1- The spreadsheet give:

a =	-17.85	±	0.92
b =	4.7327	±	0.00049

r = -0.9980 Confidence: 95 %

then:
E=4.7327±0.0005 V
accuracy: 0.010 %
r=17.85±0.92 Ω
accuracy: 5.16 %

U(V) as a function of I(A):



2-

I(A)	dI(A)	U(V)	dU(V)	poids wi	wi/S
0.000093	2.16E-07	4.731	0.00537	3.47E+04	*10%*
0.000115	2.61E-07	4.731	0.00537	3.47E+04	*10%*
0.000153	3.35E-07	4.730	0.00537	3.47E+04	*10%*
0.000235	7.70E-07	4.728	0.00536	3.48E+04	*10%*
0.000469	1.24E-06	4.724	0.00536	3.48E+04	*10%*
0.000520	1.34E-06	4.724	0.00536	3.48E+04	*10%*
0.000584	1.47E-06	4.722	0.00536	3.48E+04	*10%*
0.000666	1.63E-06	4.721	0.00536	3.48E+04	*10%*
0.000775	1.85E-06	4.719	0.00536	3.48E+04	*10%*
0.000926	2.15E-06	4.716	0.00536	3.48E+04	*10%*

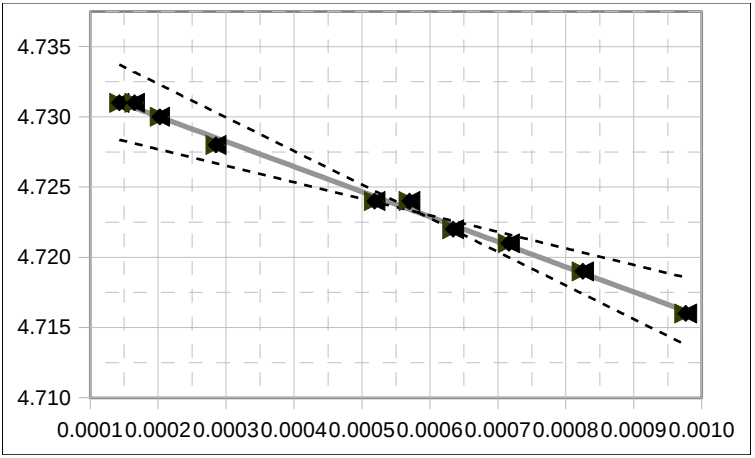
S= 3.48E+05

da = 6.09 da/|a| = 34.1% sa = 6.09
db = 0.00324 db/|b| = 0.0685% sb = 0.00324

a =	**-17.85**	**±**	**6.1**
b =	**4.7327**	**±**	**0.0032**

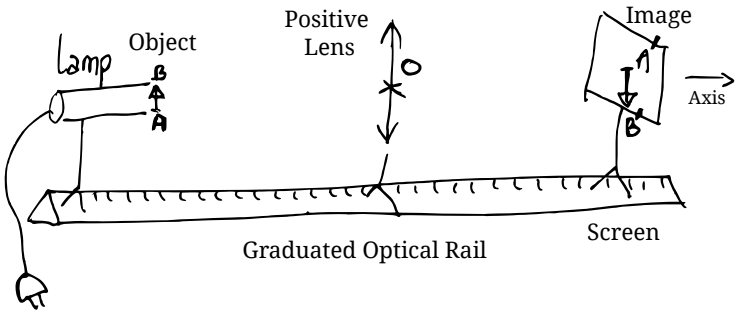
then: **E = 4.7327 ± 0.0032 V** , accuracy: 0.07 %
and **r = 17.85 ± 6.1 Ω** , accuracy: 34 %

U(V) as a function of I(A):



The results in question 2- are the good ones, because those in 1- do not take into account important sources of uncertainty.

E11: Thin lens formula *Exercise on page 95*



In the context of the small angle approximation we obtain the thin lens formula:

$$\frac{1}{\overline{OA'}}-\frac{1}{\overline{OA}}=\frac{1}{f'} \qquad , \ f' : \text{image focal length.}$$

Here the object and the image are real and according to the sign convention chosen $\overline{OA}<0$ and $\overline{OA'}>0$. To have positive quantities we set: $\overline{OA}=-OA$ and $\overline{OA}=OA'$.

The fastest method for measuring the focal length of a converging lens is the autocollimation method. Then we have the Bessel method, more precise but longer to implement. The advantage with the other method of this exercise, certainly long, is that we do not work with a single measure but with *n* measures. The larger the number is, the more accurate is the result.

ΔOA contains geometric uncertainties (length measurements) and modeling uncertainties (lens thick, small angles approximate). ΔOA' contains in addition the optical uncertainties (focusing latitudes):

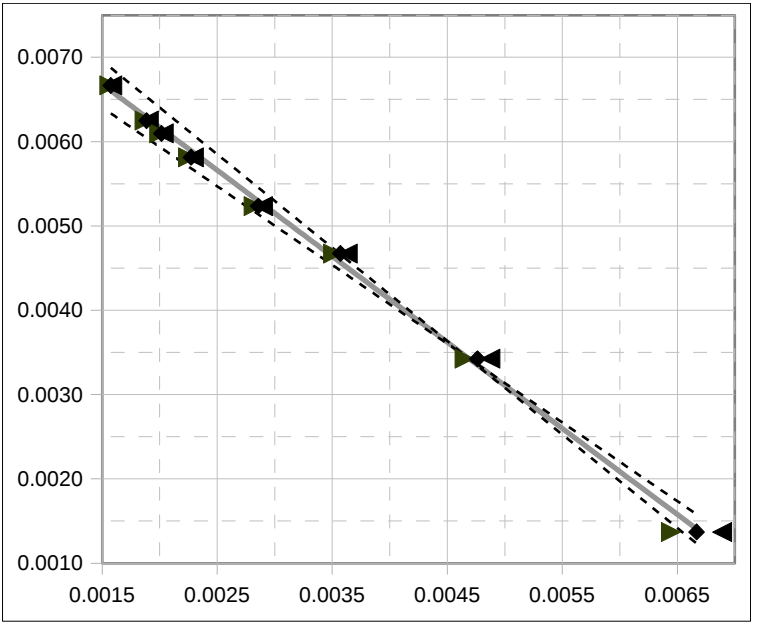
$$\Delta OA'=\sqrt{(\Delta OA'_{geo}{}^2+\Delta OA'_{mod}{}^2+\Delta OA'_{opt}{}^2)}$$

The relationship between OA and OA' is hyperbolic, then we consider x=1/OA and y=1/OA'. We have then a linear relationship $y=ax+b$ where the slope a as to be equal to -1 and the y-intercept b to 1/f'. $\Delta x=\Delta(1/OA)=\Delta OA/OA^2$, as well for Δy.

With a spreadsheet we obtain:
(here the file www.incertitudes.fr/livre/TPlentille.ods)
a= -1.02±0.09 and b=0.0082±0.0004 so **f' = 122 ± 6 mm** (consistent with the manufacturer statement f'=125 mm).

a= -1 is correctly in the interval, we also verify the validity of the linear relationship between 1/OA and 1/OA'. To further clarify the validity of the thin lens formula, we should also show that this linear relation is the one which gives the best correlation.

y as a function of x:



n = 8		a estim.= -1.0210		a=
x	dx	y	dy	poids wi
0.00157	0.0000124	0.00667	0.000667	2249189
0.00189	0.0000178	0.00625	0.000664	2265985
0.00202	0.0000203	0.00610	0.000558	3210643
0.00227	0.0000258	0.00581	0.000608	2696211
0.00286	0.0000408	0.00524	0.000548	3308045
0.00357	0.0000638	0.00467	0.000546	3308566
0.00476	0.0001134	0.00342	0.000328	8248034
0.00667	0.0002222	0.00137	0.000191	11349080
			S=	36635752

da = 0.0871	da/|a| = 8.5%
db = 0.000408	db/|b| = 5.0%

a =	-1.021	+ ou -	0.087
b =	0.00821	+ ou -	0.00041
	-1.11	< a <	-0.93

f'=1/b=	122
df'=d(1/b)=	6
df'/f'=d(1/b)/|1/b|=	5%

Theory

E12 :

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-2\sum_{i=1}^{n}x_i\bar{x}+\sum_{i=1}^{n}\bar{x}^2=n\overline{x^2}-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2$$

$$n\overline{x^2}-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2=n\overline{x^2}-2\bar{x}\,n\bar{x}+n\bar{x}^2=n(\overline{x^2}-\bar{x}^2)$$

E13 :

$$s_b=s_r\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}=s_r\sqrt{\frac{\sum(x_i-\bar{x})^2+n\bar{x}^2}{n\sum(x_i-\bar{x})^2}}=s_r\sqrt{\frac{n(\overline{x^2}-\bar{x}^2)+n\bar{x}^2}{n\sum(x_i-\bar{x})^2}}$$

then $s_b = s_r \sqrt{\frac{n\overline{x^2}}{n\sum (x_i - \bar{x})^2}} = s_r \sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}$

E14 :

$s_a = \frac{s_r}{\sqrt{\sum_i (x_i - \bar{x})^2}}$ and $s_b = s_r \sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}$ so $s_b = s_a \sqrt{\frac{\sum x_i^2}{n}}$

Which brings us immediately to the desired result.

E15 : Asymptotes *Exercise on page 97*

For the extreme line of greater slope:

$$\Delta y_{o+} = y_{o+} - y_o = (a_+ x_o + b_-) - (a x_o + b) = \Delta a \,.\, x_o - \Delta b$$

$$\lim_{x_o \to \infty} \Delta y_{o+} = \lim_{x_o \to \infty} (\Delta a \,.\, x_o - \Delta b) = \Delta a \,.\, x_o$$

For the confidence curve:

$$\Delta y_{conf} = t_{n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = t_{n-2} s_r \sqrt{\frac{1}{n x_o^2} + \frac{\left(1 - \frac{\bar{x}}{x_o}\right)^2}{\sum (x_i - \bar{x})^2}} x_o$$

so $\lim_{x_o \to \infty} \Delta y_{conf} = t_{n-2} s_r \sqrt{0 + \frac{1}{\sum (x_i - \bar{x})^2}} x_o = \Delta a \,.\, x_o$

For the prediction curve we arrive at the same conclusion. As we experimentally guessed all the asymptotes meet when we tend to infinity.

E16 : Confidence Interval and Prediction Interval for Linear Regression With error bars *Exercise on page 97*

1- $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ and $\overline{x^2} = \frac{\sum w_i x_i^2}{\sum w_i}$

.

$$\Delta = \sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2 = \left(\sum w_i\right)^2 (\overline{x^2} - \bar{x}^2)$$

2- Simple regression: $\Delta y_o = t_{n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

$$s_a = \frac{s_r}{\sqrt{\sum_i (x_i - \bar{x})^2}} \quad \text{and} \quad \Delta a = t_{n-2}\, s_a$$

$$\text{then} \quad t_{n-2}\, s_r = \Delta a \sqrt{\sum (x_i - \bar{x})^2}$$

so $$\Delta y_o = \Delta a \sqrt{n(\overline{x^2} - \bar{x}^2)} \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{n(\overline{x^2} - \bar{x}^2)}}$$

Error bars regression: $\Delta a = \sqrt{\frac{\sum w_i}{\Delta}}$

and $\Delta = \left(\sum w_i\right)^2 (\overline{x^2} - \bar{x}^2)$ hence by analogy we have:

$$\Delta y_o = \sqrt{\frac{\sum w_i}{\Delta}} \sqrt{n} \sqrt{\frac{\Delta}{\left(\sum w_i\right)^2}} \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{n(\overline{x^2} - \bar{x}^2)}}$$

so $$\Delta y_o = \frac{1}{\sqrt{n \sum w_i}} \sqrt{1 + \frac{(x_o - \bar{x})^2}{\overline{x^2} - \bar{x}^2}}$$ for the confidence,

and $$\Delta y_o = \frac{1}{\sqrt{n \sum w_i}} \sqrt{1 + n + \frac{(x_o - \bar{x})^2}{\overline{x^2} - \bar{x}^2}}$$ for the prediction,

3- By the same calculation as in simple regression $\lim_{x_o \to \infty} \Delta y_{o+} = \Delta a \,.\, x_o$

$$\Delta y_{conf} = \frac{1}{\sqrt{n \sum w_i}} \sqrt{1 + \frac{(x_o - \bar{x})^2}{\overline{x^2} - \bar{x}^2}} = \frac{1}{\sqrt{n \sum w_i}} \sqrt{\frac{1}{x_o^2} + \frac{\left(1 - \frac{\bar{x}}{x_o}\right)^2}{\overline{x^2} - \bar{x}^2}}\, x_o$$

$$\lim_{x_o \to \infty} \Delta y_{conf} = \frac{1}{\sqrt{n \sum w_i}} \sqrt{\frac{1}{\overline{x^2} - \bar{x}^2}}\, x_o = \frac{1}{\sqrt{n \sum w_i}} \sqrt{\frac{\left(\sum w_i\right)^2}{\Delta}}\, x_o$$

$$\lim_{x_o \to \infty} \Delta y_{conf} = \frac{1}{\sqrt{n}} \sqrt{\frac{\sum w_i}{\Delta}}\, x_o = \frac{1}{\sqrt{n}} \Delta a \,.\, x_o$$

By analogy we have to take **n=1** into the error bars regression formu-

las. In fact we had to expect a factor of difference, indeed the w_i weights are defined by a factor. For example, we could have taken $w_i' = \frac{1/a^2}{(\Delta y_i/a)^2 + (\Delta x_i)^2}$. Also in simple regression the number of points is clearly defined, while with the error bars some points count more than others, n=1 means that all the points correspond to 100% of the data. To find the formulas of the statement we replace n by 1.

E17 Others expressions *Exercise on page 98*

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad \overline{x^2} = \frac{\sum w_i x_i^2}{\sum w_i} \quad \overline{xy} = \frac{\sum w_i x_i y_i}{\sum w_i} \quad \frac{\Delta}{\left(\sum w_i\right)^2} = \overline{x^2} - \bar{x}^2$$

$$a = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\Delta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \frac{\sum w_i y_i \sum w_i x_i^2 - \sum w_i x_i \sum w_i x_i y_i}{\Delta} = \frac{\bar{y}\,\overline{x^2} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$$

$$\Delta a = \sqrt{\frac{\sum w_i}{\Delta}} = \frac{1}{\sqrt{\sum w_i}} \frac{1}{\sqrt{\overline{x^2} - \bar{x}^2}}$$

$$\Delta b = \sqrt{\frac{\sum w_i x_i^2}{\Delta}} = \frac{1}{\sqrt{\sum w_i}} \sqrt{\frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}}$$

Simple regression: $a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$ (same formulas with the bars)

and $\bar{y} = a\bar{x} + b$ imply $b = \frac{\bar{y}\,\overline{x^2} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$ (same formula)

$$\Delta a = t_{n-2} s_r \frac{1}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{t_{n-2} s_r}{\sqrt{n}} \frac{1}{\sqrt{\overline{x^2} - \bar{x}^2}} \text{ et}$$

$$\Delta b = t_{n-2} s_r \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}} = \frac{t_{n-2} s_r}{\sqrt{n}} \sqrt{\frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}}$$

This formulas are similar. Like $n \sim 1$ we have $\frac{1}{\sqrt{\sum w_i}} \sim t_{n-2} s_r$ so $\frac{1}{\sqrt{\sum w_i}}$ for the error bars regression, it is equivalent to the residuals uncertainty in simple regression.

E18 : Least Squares Method *Exercise on page 99*

❶ Simple regression: $\sum d^2 = \sum_i (y_i - a x_i - b)^2$

$$\frac{\partial (\sum d^2)}{\partial b} = \sum_i 2 \frac{\partial (y_i - a x_i - b)}{\partial b} (y_i - a x_i - b) = -2 \sum_i (y_i - a x_i - b)$$

then $\sum (y_i - a x_i - b) = 0$, $\sum y_i - a \sum x_i - nb = 0$

and finally by dividing by n: $\underline{\bar{y} = a \bar{x} + b}$

$$\frac{\partial (\sum d^2)}{\partial a} = \sum_i 2 \frac{\partial (y_i - a x_i - b)}{\partial a} (y_i - a x_i - b) = -2 \sum_i x_i (y_i - a x_i - b)$$

then $\sum (y_i - a x_i - b) x_i = 0$, $\sum y_i x_i - a \sum x_i^2 - b \sum x_i = 0$

and we divide by n: $\overline{x y} - a \overline{x^2} - b \bar{x} = 0$

so $\overline{x y} - a \overline{x^2} - (\bar{y} - a \bar{x}) \bar{x} = 0$ and $\underline{a = \frac{\overline{x y} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2}}$

❷ Error bars regression: $S^2 = \sum_i w_i (y_i - a x_i - b)^2$

$\frac{\partial (\sum S^2)}{\partial b} = -2 \sum w_i (y_i - a x_i - b)$ because w_i does not depend on b.

And we obtain also $\underline{\bar{y} = a \bar{x} + b}$.

$$\frac{\partial w_i}{\partial a} = \frac{\partial}{\partial a} \left(\frac{1}{\Delta y_i^2 + a^2 \Delta x_i^2} \right) = -\frac{2 a \Delta x_i^2}{(\Delta y_i^2 + a^2 \Delta x_i^2)^2} = -2 a \Delta x_i^2 w_i^2$$

$$\frac{\partial (\sum S^2)}{\partial a} = -2a \sum_i \Delta x_i^2 w_i^2 (y_i - a x_i - b)^2 + \sum_i 2 w_i (-x_i)(y_i - a x_i - b)$$

but Δx_i is small compared to x_i and y_i, so we can neglect the first term and to keep only the second one, and thus find the same result as before in simple regression.

Method 1:

1- $$a=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\overline{x^2}-\bar{x}^2}=\frac{\sum(x_i y_i-\bar{x}\,y_i)}{n(\overline{x^2}-\bar{x}^2)}=\ldots$$

2- $$V(a)=\sum p_i^2 V(y_i)=\frac{\sum(x_i-\bar{x})^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\sigma_r^2=\ldots$$

Because the variables y_i are independent and are assumed to have the same variance.

3- $$b=\frac{\bar{y}\,\overline{x^2}-\bar{x}\,\overline{xy}}{\overline{x^2}-\bar{x}^2}=\frac{\sum(y_i\overline{x^2}-\bar{x}\,x_i y_i)}{n(\overline{x^2}-\bar{x}^2)}=\ldots$$

$$V(b)=\frac{\sum(\overline{x^2}-\bar{x}\,x_i)^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\sigma_r^2=\frac{n\left(\overline{x^2}\right)^2-2\overline{x^2}\,\bar{x}^2+\bar{x}^2\overline{x^2}}{\left(\sum(x_i-\bar{x})^2\right)^2}\sigma_r^2=\ldots$$

Method 2:

$$a=\frac{\sum x_i y_i-\bar{x}\sum y_i}{n(\overline{x^2}-\bar{x}^2)} \quad \text{or}$$

$$\frac{\partial\sum y_i}{\partial y_j}=\frac{\partial(y_1+\ldots+y_j+\ldots+y_n)}{\partial y_j}=0+\ldots+1+\ldots+0=1$$

and $$\frac{\partial\sum x_i y_i}{\partial y_j}=\frac{\partial(x_1y_1+\ldots+x_j y_j+\ldots+x_n y_n)}{\partial y_j}=0+\ldots+x_j+\ldots+0=x_j$$

then $$\frac{\partial a}{\partial y_j}=\frac{x_j-\bar{x}}{n(\overline{x^2}-\bar{x}^2)} \quad \text{and} \quad s_a{}^2=\sum\left(\frac{x_j-\bar{x}}{n(\overline{x^2}-\bar{x}^2)}\right)^2 s_{y_i}{}^2=\ldots$$

Analogous procedure for *b*.

Matrix approach: the system of equation p201 is equivalent to

$$\begin{cases} \bar{y}-b-a\bar{x}=0 \\ \overline{xy}-b\bar{x}-a\overline{x^2}=0 \end{cases}, \quad H=\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}, \quad A=\begin{pmatrix} b \\ a \end{pmatrix} \text{ and } B=\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}.$$

$$HA=B,\ A=H^{-1}B,\ H^{-1}=\frac{1}{\overline{x^2}-\bar{x}^2}\begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \text{ and } \sigma_b{}^2=(H^{-1})_{11}\,\sigma_r{}^2$$

We have a matrix of non-zero determinant, and we have applied the inversion formula for a 2×2 matrix (we verify that $HH^{-1}=H^{-1}H=I$). We then quickly have the formulas sought.

E19 Expectation of a *Exercise on page 99*

$$E(a)=E\left(\sum_i p_i y_i\right)=\sum_i p_i E(y_i)=\sum_i p_i(\alpha x_i+\beta)$$

$$E(a)=\alpha\sum_i p_i x_i+\beta\sum_i p_i \qquad \sum p_i=\frac{1}{\sum(x_i-\bar{x})^2}\sum(x_i-\bar{x})=0$$

$$\sum p_i x_i=\sum\frac{x_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}=\frac{1}{\sum(x_i-\bar{x})^2}\sum((x_i-\bar{x})^2+\bar{x}(x_i-\bar{x}))=1+0$$

then $E(a)=\alpha$

E20 Standard Deviations proportional to y *Exercise on page 100*

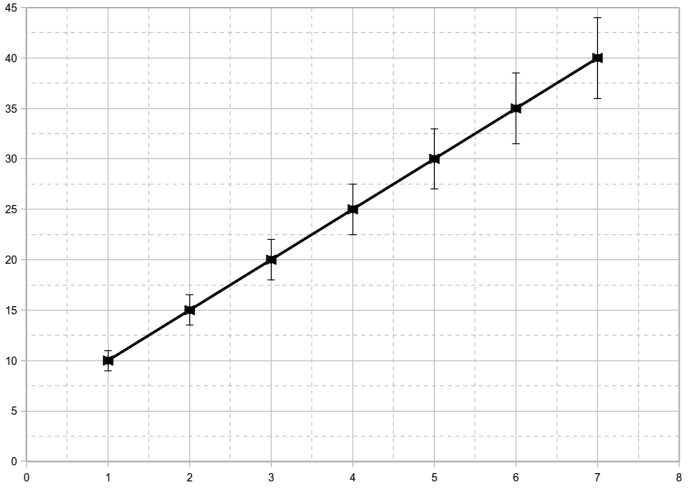
1- $$a=\frac{\sum w_i\sum w_i x_i y_i-\sum w_i x_i\sum w_i y_i}{\sum w_i\sum w_i x_i^2-\left(\sum w_i x_i\right)^2} \quad \text{with} \quad w_i=\frac{1}{k^2 y_i^2}.$$

2- $$a=\frac{u}{v} \qquad \frac{\partial a}{\partial y_j}=\frac{u'v-v'u}{v^2}$$

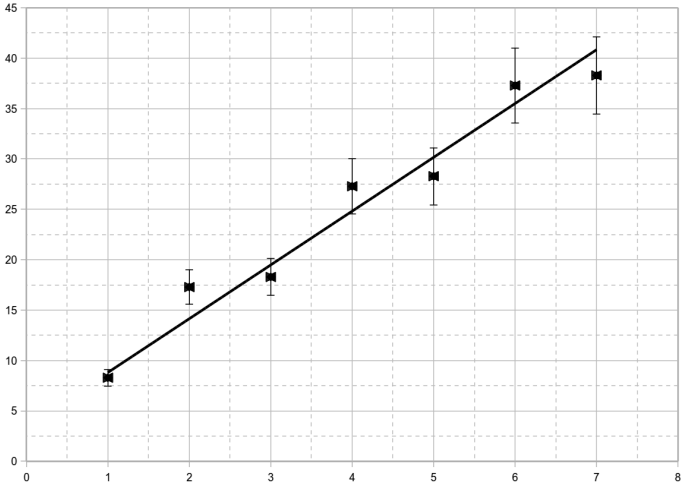
$$u'=\frac{\partial u}{\partial y_j}=-\frac{2}{y_j^3}\sum\frac{x}{y}+\sum\frac{1}{y^2}\frac{-x_j}{y_j^2}-\frac{-2x_j}{y_j^3}\sum\frac{1}{y}-\sum\frac{x}{y^2}\frac{-1}{y_j^2}$$

$$v' = \frac{\partial v}{\partial y_j} = \frac{-2}{y_j^3} \sum \frac{x^2}{y^2} + \sum \frac{1}{y^2} \frac{-2x_j^2}{y_j^3} - 2 \sum \frac{x}{y^2} \frac{-2x_j}{y_j^3}$$

We thus have the calculation of the partial derivative, we add them to the square according to all the values of j on a spreadsheet in order to obtain s_a:



y=5x+5



y=5.34x+3.46

Data set 1 for $k=0.1$: $s_a \simeq 0.422$

Data set 2 for $k=0.1$: $s_a \simeq 0.427$

We find exactly the same results with small variations method:

	j=1	*j=2*	*j=3*	*j=4*	*j=5*	*j=6*	*j=7*
$\frac{\partial a}{\partial y_j}$	-0.2237	-0.0203	0.0331	0.0496	0.0542	0.0543	0.0528

and $s_a \simeq 0.422$

	j=1	*j=2*	*j=3*	*j=4*	*j=5*	*j=6*	*j=7*
$\frac{\partial a}{\partial y_j}$	-0.2232	-0.0016	0.0514	0.0343	0.0667	0.0410	0.0611

and $s_a \simeq 0.427$

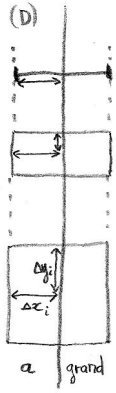
We had with the classic method p67:

Data set 1 for $k=0.1$: $s_a \simeq 0.422$

Data set 2 for $k=0.1$: $s_a \simeq 0.399$

Again the method with the propagation formula and more realistic than the classical method.

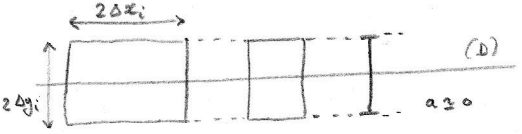
E21 Interpretation of w_i *Exercise on page 100*



1- We want to find the straight line which passes at best through all the rectangles of uncertainties, and if possible closer to their center. If the slope of the line is small, the line is close to the horizontal, $a \simeq 0$ and only Δy_i will act to adjust the line and $w_i \simeq \frac{1}{(\Delta y_i)^2}$. For example, a rectangle such as $\Delta x_i = \Delta y_i$ will determine as much the line as another one for which $\Delta x_i = 0$. On the contrary, if

the line is vertical, $a \simeq \infty$ and only Δx_i will act to adjust the line and $w_i \simeq \frac{1}{(a\Delta x_i)^2}$.

In the intermediate case where $a=1$, Δy_i and Δx_i have the same importance, which is satisfied by the formula $w_i = \frac{1}{(\Delta y_i)^2 + (\Delta x_i)^2}$.



E22 Decomposition into Gaussians *Exercise on page 101*

We consider a decomposition of the form:

$$f(x) = a_1 e^{-\frac{1}{2}\left(\frac{x-a_2}{a_3}\right)^2} + a_4 e^{-\frac{1}{2}\left(\frac{x-a_5}{a_6}\right)^2}$$

We have six parameters:

$$\frac{\partial f}{\partial a_1} = e^{-\frac{1}{2}\left(\frac{x-a_2}{a_3}\right)^2} \quad , \quad \frac{\partial f}{\partial a_2} = \frac{a_1}{a_3^2}(x-a_2) e^{-\frac{1}{2}\left(\frac{x-a_2}{a_3}\right)^2} \quad , \text{ etc.}$$

We start with the following estimated parameters:

a_1	a_2	a_3	a_4	a_5	a_6
5.8	59.99	0.012	10	60.03	0.012

If you start with any parameters, the iteration may be divergent. You first see if the set of estimated parameters give a graphically correct result and, above all, you try manually, groping, to minimize S^2.

We obtain the following expressions for the first iteration:

$$H=\begin{vmatrix} 2.13 & 2.36 & 508 & 0.132 & -184 & 361 \\ 2.36 & 245572 & 7176 & 107 & -121329 & 202215 \\ 508 & 7176 & 354716 & 210 & -202215 & 303007 \\ 0.132 & 107 & 210 & 2.13 & -4.06 & 876 \\ -184 & -121329 & -202215 & -4.06 & 730001 & -21333 \\ 361 & 202215 & 303007 & 876 & -21333 & 1054447 \end{vmatrix}$$

$$B=\begin{vmatrix} 0.933 \\ 397 \\ 959 \\ 0.029 \\ -2767 \\ 520 \end{vmatrix} \quad \text{et } A=H^{-1}B=\begin{vmatrix} -0.0128 \\ -0.0009 \\ -0.0001 \\ -0.2927 \\ -0.0039 \\ 0.0008 \end{vmatrix}$$

We got the variations of the parameters then we obtain a new set of parameters for the second iteration. We iterate as much as necessary for convergence and stability of the quantities:

	a_1	a_2	a_3	a_4	a_5	a_6	S^2
It. 0	5.8	59.99	0.012	10	60.03	0.012	13.2711
It. 1	5.787	59.9891	0.01194	9.707	60.0261	0.01285	2.9499
It. 2	5.784	59.9892	0.01210	10.283	60.0261	0.01197	2.3998
It. 3	5.764	59.9890	0.01191	10.271	60.0260	0.01218	2.3848
It. 4	5.762	59.9891	0.01200	10.279	60.0260	0.01212	2.3839
It. 5	5.762	59.9890	0.01197	10.276	60.0260	0.01214	2.3838
It. 6	5.762	59.9890	0.01198	10.277	60.0260	0.01214	2.3838

Hence the following final parameters and inverse matrix:

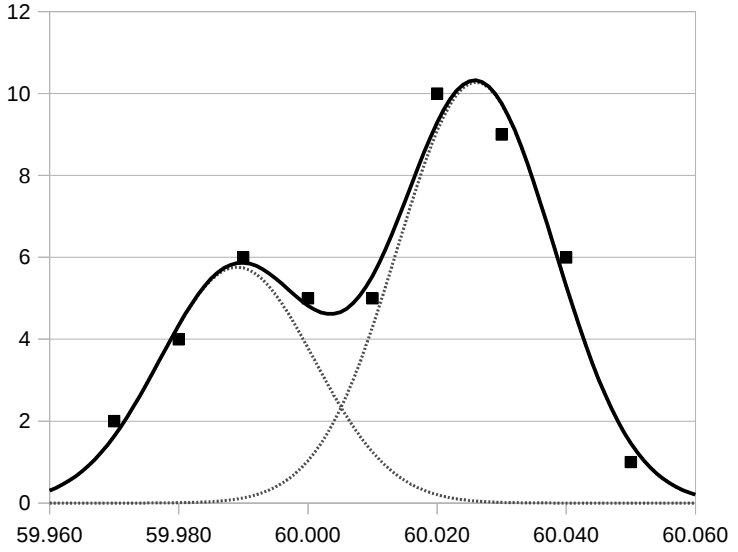
a_1	a_2	a_3	a_4	a_5	a_6
5.76	59.989	0.0120	10.28	60.026	0.0121

$$H^{-1}=\begin{vmatrix} 0.11299386 & 0.00017946 & -0.00014445 & 0.72254732 & 0.00002639 & -0.00062087 \\ -0.00011583 & 0.00000361 & 0.00000392 & 0.00002639 & 0.00000316 & -0.00000182 \\ -0.00006533 & -0.00000371 & -0.00000353 & -0.00062087 & -0.00000182 & 0.00000317 \end{vmatrix}$$

We can thus determine the standard deviations on the parameters:

$$s_r=\sqrt{\frac{\sum (y_i-f_i)^2}{n-6}}\simeq 0.891 \quad \text{et} \quad s_{a_k}{}^2=(H^{-1})_{kk}s_r{}^2 \quad \text{d'où}$$

σ_{a1}	σ_{a2}	σ_{a3}	σ_{a4}	σ_{a5}	σ_{a6}
0.77	0.003	0.003	0.76	0.002	0.0016



A nail corresponds to an area of 0.01x1, so for a unit area: u.a. = 100 nails.

$$A=\int_{-\infty}^{+\infty}\alpha\, e^{-\frac{1}{2}\left(\frac{x-\beta}{\delta}\right)^2}dx=\sqrt{2\pi}\,\alpha\,\delta$$

with an integration by substitution: $x'=\frac{x-\beta}{\delta}$

For the left peak: $N_1 = \sqrt{2\pi}\, a_1\, a_3 = 17.33 \pm 4.91$

For the right peak: $N_2 = \sqrt{2\pi}\, a_4\, a_6 = 31.18 \pm 4.72$

In total: $N = N_1 + N_2 = 48.5 \pm 6.8$ (consistent with 48 nails)

Chapter III : Probability Distributions

E1: Binomial Distribution *Exercise on page 122*

Positive probabilities. Moreover, according to the binomial formula:

$$\sum_{k=0}^{n} P(X=k)=\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}=(p+q)^n=1$$

Probabilities between 0 and 1.

Expectation: $E(X)=\sum_{k=0}^{n} k\,P(X=k)=\sum_{k=1}^{n} k\,\frac{n!}{(n-k)!\,k!}\,p^k q^{n-k}$

then $E(X)=\sum_{k=1}^{n}\frac{n(n-1)!}{(n-1-k+1)!(k-1)!}\,p\,p^{k-1}q^{n-1-k+1}$

We extract from the sum the quantities independent of k:

$$E(X)=np\sum_{k=1}^{n}\frac{(n-1)!}{(n-1-k+1)!(k-1)!}\,p^{k-1}q^{n-1-k+1}$$

We perform the index substitution *j=k-1*:

$$E(X)=np\sum_{j=0}^{j=n-1}\frac{(n-1)!}{(n-1-j)!\,j!}\,p^j q^{n-1-j}=np(p+q)^{n-1}=np$$

Variance:

$$E(X^2)=\sum_{k=0}^{n} k^2\,P(X=k)=\sum_{k=1}^{n}[k(k-1)+k]\,\frac{n!}{(n-k)!\,k!}\,p^k q^{n-k}$$

then $E(X^2)=\sum_{k=2}^{n}\frac{n!}{(n-k)!(k-2)!}\,p^k q^{n-k}+E(X)$ and

$$E(X^2)=\sum_{k=2}^{n}\frac{n(n-1)(n-2)!}{(n-2-k+2)!(k-2)!}\,p^2\,p^{k-2}q^{n-2-k+2}+E(X)$$

$$E(X^2)=n(n-1)p^2\times 1+E(X)=n(n-1)p^2+n\,p=n^2p^2-np^2+np$$

and $V(X)=E(X^2)-E(X)^2=np(1-p)=npq$

E2: Sum of binomial distributions *Exercise on page 122*

$$P(Z=k)=\sum_i P([X=i]\cap[Y=k-i])=\sum_i P(X=i)P_{X=i}(Y=k-i)$$

If the random variables are independent:

$$P(Z=k)=\sum_i P(X=i)P(Y=k-i)$$

for the binomial distributions $\mathcal{B}_X(n_1,p)$ and $\mathcal{B}_Y(n_2,p)$:

$$P(Z=k)=\sum_{i=0}^{n_1+n_2}\binom{n_1}{i}p^i q^{n_1-i}\binom{n}{k-i}p^{k-i}q^{n_2-k+i}$$

then $$P(Z=k)=\left[\sum_{i=0}^{n_1+n_2}\binom{n_1}{i}\binom{n}{k-i}\right]p^k q^{n_1+n_2-k}$$

And finally using the Vandermonde identity for the binomial coefficients we get:

$$P(Z=k)=\binom{n_1+n_2}{k}p^k q^{n_1+n_2-k}$$ so a $\mathcal{B}_Z(n_1+n_2,p)$.

E3: Geometric distribution *Exercise on page 122*

Positive probabilities.

$$\sum_{k=1}^{\infty}P(X=k)=\sum_1^{\infty}q^{k-1}p=p\lim_{k\to\infty}\frac{1-q^k}{1-q}=1$$

Probabilities between 0 and 1.

Expectation: $$\sum_{k=1}^{\infty}kP(X=k)=p\sum_1^{\infty}k\,q^{k-1}=p\frac{1}{(1-q)^2}=\frac{1}{p}$$

Variance: $$E(X^2)=\sum_{k=1}^{\infty}k^2P(X=k)=p\sum_1^{\infty}[k(k-1)+k]q^{k-1}$$

and $E(X^2)=pq\sum_2^\infty k(k-1)q^{k-2}+E(X)=pq\frac{2}{(1-q)^3}+\frac{1}{p}$

then $V(X)=\frac{2q}{p^2}+\frac{1}{p}-\frac{1}{p^2}=\frac{2q+p-1}{p^2}=\frac{q}{p^2}$

Sum of independent geometric random variables of parameter *p*:

$$P(Z=k)=\sum_i P(X=i)P(Y=k-i)=\sum_{i=1}^{k-1} q^{i-1}pq^{k-i-1}p$$

so $P(Z=k)=(k-1)q^{k-2}p^2$

Probability of having a second success in rank *k*. The first success has *k-1* possibilities from position *1* to *k-1*.

E4: Firsts successes *Exercise on page 122*

1- $P(X=5)=(1/2)^4.1/2 = 1/32 \simeq 3\%$. $P_{X>3}(X=8)=P(X=8-3)=P(X=5)$.

2- $E(X)=1/p=6$, on average after six throws.

$$P(X\leqslant 6)=\sum_{k=1}^{6} q^{k-1}p=p\sum_{i=0}^{5} q^i=p\frac{1-q^6}{1-q}=1-q^6\simeq 66\,\%$$

E5: Poisson distribution *Exercise on page 123*

Positive probabilities.

$$\sum_{k=0}^{\infty} P(X=k)=\sum_0^\infty \frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_0^\infty \frac{\lambda^k}{k!}=e^{-\lambda}e^{\lambda}=1$$

(definition of the exponential function in terms of series)

Probabilities between 0 and 1.

Expectation:

$$\sum_{k=0}^{\infty} k\,P(X=k)=\sum_{1}^{\infty}\frac{\lambda^k}{(k-1)!}e^{-\lambda}=\sum_{j=0}^{\infty}\frac{\lambda^{j+1}}{j!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}=\lambda$$

Variance: $E(X^2)=\sum_{k=0}^{\infty}k^2P(X=k)=\sum_{1}^{\infty}[k(k-1)+k]\frac{\lambda^k}{k!}e^{-\lambda}$

$$E(X^2)=\sum_{2}^{\infty}\frac{\lambda^k}{(k-2)!}e^{-\lambda}+E(X)=e^{-\lambda}\sum_{2}^{\infty}\frac{\lambda^k}{(k-2)!}+\lambda=\lambda^2+\lambda$$

then $V(X)=\lambda^2+\lambda-\lambda^2=\lambda$

Sum of independent Poisson distributions with parameters λ_1 and λ_2:

$$P(Z=k)=\sum_i P(X=i)P(Y=k-i)=\sum_{i=0}^{k}\frac{\lambda_1^{\,i}}{i!}e^{-\lambda_1}\frac{\lambda_2^{\,k-i}}{(k-i)!}e^{-\lambda_2}$$

and $P(Z=k)=\sum_{i=0}^{k}\frac{\lambda^k}{i!(k-i)!}e^{-\lambda_1-\lambda_2}=\left[\sum_{i=0}^{k}\binom{k}{i}\lambda_1^{\,i}\lambda_2^{\,k-i}\right]\frac{e^{-\lambda_1-\lambda_2}}{k!}$

With the binomial formula: $P(Z=k)=\frac{(\lambda_1+\lambda_2)^k}{k!}e^{-(\lambda_1+\lambda_2)}$

The expression obtained corresponds to a
Poisson distribution with parameter $\lambda_1+\lambda_2$.

E6: Uniform distribution *Exercise on page 123*

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)dx=\int_a^b\frac{x^2}{b-a}dx=\frac{1}{3}\frac{b^3-a^3}{b-a}=\frac{1}{3}\frac{(b-a)(b^2+ab+a^2)}{b-a}$$

then $V(X)=E(X^2)-E(X)^2=\frac{1}{3}(b^2+ab+a^2)-\frac{(a+b)^2}{4}$

and $V(X)=\frac{(b-a)^2}{12}$

E7: Exponential distribution *Exercise on page 123*

Positive density of probability.

$$\int_{-\infty}^{+\infty} f(t)\,dt = \int_{0}^{+\infty} \lambda e^{-\lambda t}\,dt \quad \text{(improper integral)}$$

$$I_A = \int_0^A \lambda e^{-\lambda t}\,dt = \left[-e^{-\lambda t}\right]_0^A = 1 - e^{-\lambda A} \quad \text{then} \quad \lim_{A \to +\infty} I_A = 1$$

Expectation: $E(T) = \int_{-\infty}^{+\infty} t f(t)\,dt = \int_0^{+\infty} \lambda t e^{-\lambda t}\,dt$

$$J_A = \int_0^A \lambda t e^{-\lambda t}\,dt = \left[-t e^{-\lambda t}\right] + \int e^{-\lambda t}\,dt = -A e^{-\lambda A} - \frac{1}{\lambda} e^{-\lambda A} + \frac{1}{\lambda}$$

(integration by parts)

and $E(T) = \lim_{A \to +\infty} J_A = \frac{1}{\lambda}$

Variance: $E(T^2) = \int_{-\infty}^{+\infty} t^2 f(t)\,dt = \int_0^{+\infty} \lambda t^2 e^{-\lambda t}\,dt$

After two successive integrations by parts:

$$K_A = \int_0^A \lambda t^2 e^{-\lambda t}\,dt = \left[-t^2 e^{-\lambda t}\right] + 2\int t e^{-\lambda t}\,dt$$

then $E(T^2) = \lim_{A \to +\infty} K_A = \frac{2}{\lambda^2}$ et $V(T) = \frac{1}{\lambda^2}$

Sum of independent exponential distributions of the same parameter:

$$f_Z(x) = \int_{-\infty}^{+\infty} f_X(y) f_Y(x-y)\,dy$$

- if $x<0$ then $f_Z(x) = 0$
- if $x>0$ then $f_Z(x) = \int_0^x \lambda e^{-\lambda y} \lambda e^{-\lambda(x-y)}\,dy$

and $f_Z(x)=\lambda^2 e^{-\lambda x}\int_0^x dy=\lambda^2 x\, e^{-\lambda x}$, It is not an exponential distribution, it is a Gamma distribution with parameters 2 and 1/λ.

E8: Sum of Gaussians *Exercise on page 123*

1- $f_Z(x)=\int_{-\infty}^{+\infty} f_X(y)f_Y(x-y)\,dy=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{\frac{-y^2}{2}} e^{\frac{-(x-y)^2}{2}} dy$

$$f_Z(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}[y^2+(x-y)^2]}dy=\frac{1}{2\pi}e^{-\frac{x^2}{4}}\int_{-\infty}^{+\infty} e^{-\left(y-\frac{x}{2}\right)^2}dy$$

then $f_Z(x)=\frac{1}{2\sqrt{\pi}}e^{-\frac{x^2}{4}}=\frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{1}{2}\frac{x^2}{\sqrt{2}^2}}$ we recognize a normal distribution. Distribution centered and as expected with a standard deviation $\sqrt{2}$, indeed $\sigma_Z^2=\sigma_X^2+\sigma_Y^2=1+1=2$.

2- $f_Z(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{y-\mu_1}{\sigma_1}\right)^2} e^{-\frac{1}{2}\left(\frac{x-y-\mu_2}{\sigma_2}\right)^2} dy$ We then have a somewhat long calculation: $f_Z(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-\frac{\sigma_2^2(y-\mu_1)^2+\sigma_1^2(x-y-\mu_2)^2}{2\sigma_1^2\sigma_2^2}} dy=...$ which, once conducted, does indeed provide a normal distribution with mean $\mu_Z=\mu_1+\mu_2$ and variance $\sigma_Z^2=\sigma_1^2+\sigma_2^2$.

E9: First Students *Exercise on page 123*

$$f_1(x)=\frac{1}{\sqrt{\pi}}\frac{\Gamma(1)}{\Gamma\left(\frac{1}{2}\right)(1+x^2)} \quad \text{then} \quad f_1(x)=\frac{1}{\pi}\frac{1}{1+x^2}$$

also $f_2(x)=\frac{1}{2\sqrt{2}}\frac{1}{\left(1+\frac{x^2}{2}\right)^{\frac{3}{2}}}$, $f_3(x)=\frac{2}{\sqrt{3}\pi}\frac{1}{\left(1+\frac{x^2}{3}\right)^2}$

and $f_4(x)=\frac{1}{\sqrt{4\pi}}\frac{\Gamma\left(\frac{5}{2}\right)}{\Gamma(2)\left(1+\frac{x^2}{4}\right)^{\frac{5}{2}}}=\frac{1}{2\sqrt{\pi}}\frac{\frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right)}{1!\left(1+\frac{x^2}{4}\right)^{\frac{5}{2}}}$

so $f_4(x)=\frac{3}{8}\frac{1}{\left(1+\frac{x^2}{4}\right)^{\frac{5}{2}}}$

E10: Student's t-distribution *Exercise on page 124*

Positive density of probability.

We want to show that $k\geqslant 1$, $\int_{-\infty}^{+\infty} f_k(x)dx=1$.

We assume that the integrals are convergent.

$dx=\sqrt{k}\,du$ then with the integration by substitution and the symmetry of the function:

$$I_k=\int_{-\infty}^{+\infty}\left(1+u^2\right)^{-\frac{k+1}{2}}\sqrt{k}\,du=2\sqrt{k}\int_0^{+\infty}\left(1+u^2\right)^{-\frac{k+1}{2}}du$$

Let J_k be the integral which verify: $I_k=2\sqrt{k}J_k$.

We first consider the case where k is even: $k=2n$

$$\text{so}\quad J_n=\int_0^{+\infty}\left(1+u^2\right)^{-n-\frac{1}{2}}du=\int_0^{+\infty}\frac{1}{\sqrt{1+u^2}}\frac{1}{\left(1+u^2\right)^n}du$$

$$\text{but}\quad J_n=\int_0^{+\infty}\frac{1}{\sqrt{1+u^2}}\frac{1+u^2-u^2}{\left(1+u^2\right)^n}du\quad\text{then}$$

$$J_n=J_{n-1}-\int_0^{+\infty}\frac{1}{\sqrt{1+u^2}}\frac{u^2}{\left(1+u^2\right)^n}du\text{ , let }K_n=\int_0^{+\infty}\frac{1}{\sqrt{1+u^2}}\frac{u^2}{\left(1+u^2\right)^n}du$$

Integration by parts:

$$K_n = 0 + \frac{1}{2\left(n-\frac{1}{2}\right)} J_{n-1} = \frac{J_{n-1}}{2n-1} \quad \text{and} \quad J_n = \frac{2n-2}{2n-1} J_{n-1}$$

$$\text{then} \quad J_n = \frac{2n-2}{2n-1}\frac{2n-4}{2n-3} J_{n-2} = \ldots = \frac{2n-2}{2n-1}\frac{2n-4}{2n-3}\ldots\frac{2}{3} J_1$$

For J_1 : $J_1 = \int_0^{+\infty} \frac{1}{\left(1+u^2\right)^{3/2}} du$ we do an integration by substitution

$u = sh\,x$ [23] so $du = ch\,x\,dx$ give $J_1 = \int_0^{+\infty} \frac{1}{ch^2 x} dx$

$$\text{and} \quad J_1 = \left[th\,x\right]_0^{+\infty} = 1$$

For J_n : $J_n = \frac{n-1}{n-1/2}\frac{n-2}{n-3/2}\ldots\frac{1}{3/2} = \frac{\sqrt{\pi}\,\Gamma(n)}{2\,\Gamma(n+1/2)}$

$$\text{then} \quad J_k = \frac{\sqrt{\pi}\,\Gamma\left(\frac{k}{2}\right)}{2\,\Gamma\left(\frac{k+1}{2}\right)}$$

If we consider the case where k is odd we obtain the same expression.

Eventually:

$$\int_{-\infty}^{+\infty} f_k(x)\,dx = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} I_k = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} 2\sqrt{k} \frac{\sqrt{\pi}\,\Gamma\left(\frac{k}{2}\right)}{2\,\Gamma\left(\frac{k+1}{2}\right)} = 1$$

---

23 $sh\,x$: hyperbolic sine with $sh\,x = \frac{e^x - e^{-x}}{2}$, $ch\,x$: hyperbolic cosine with $ch\,x = \frac{e^x + e^{-x}}{2}$ and $th\,x = \frac{sh\,x}{ch\,x}$.

E11: Variance of Student's t-distribution *Exercise on page 124*

$$V(X)=E(X^2)-E(X)^2=E(X^2)=\int_{-\infty}^{+\infty} x^2 f_k(x)\,dx$$

This integral is convergent if $k\geqslant 3$.

$$V(X)=\frac{1}{\sqrt{k\pi}}\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\int_{-\infty}^{+\infty}\frac{x^2}{\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}}dx=\frac{2}{\sqrt{k\pi}}\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\int_{0}^{+\infty}\frac{ku^2\sqrt{k}}{\left(1+u^2\right)^{\frac{k+1}{2}}}du$$

First if k is even, with the results of the previous exercise, $k=2n$:

$$V(X)=\frac{2k}{\sqrt{\pi}}\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\int_{0}^{+\infty}\frac{u^2}{\left(1+u^2\right)^{\frac{k+1}{2}}}du=\frac{4n}{\sqrt{\pi}}\frac{\Gamma\left(n+\frac{1}{2}\right)}{\Gamma(n)}K_n$$

We had found: $K_n=\frac{J_{n-1}}{2n-1}=\frac{\sqrt{\pi}\Gamma(n-1)}{2(2n-1)\Gamma(n-1/2)}$

$$K_n=\frac{\sqrt{\pi}\Gamma(n)(n-1/2)}{2(2n-1)(n-1)\Gamma(n+1/2)}=\frac{\sqrt{\pi}\Gamma(n)}{4(n-1)\Gamma(n+1/2)}$$

then: $V(X)=\frac{n}{n-1}=\frac{k/2}{k/2-1}=\frac{k}{k-2}$

For k odd we obtain the same expression.

E12: Sum of Student's t-distributions *Exercise on page 124*

• For $k=1$:

$$f_Z(x)=\int_{-\infty}^{+\infty} f_1(y)f_1(x-y)\,dy=\frac{1}{\pi^2}\int_{-\infty}^{+\infty}\frac{1}{1+y^2}\frac{1}{1+(x-y)^2}\,dy$$

Then using a software of symbolic calculation we find:

$$f_Z(x)=\frac{1}{2\pi}\frac{1}{1+x^2/4}$$

In this case we find the general form of a Student for which $k=1$.

• For $k=3$:

$$f_{f_3*f_3}(x)=\int_{-\infty}^{+\infty} f_3(y)\,f_3(x-y)\,dy=\frac{5\sqrt{3}}{12\pi}\,\frac{1+x^2/60}{\left(1+x^2/12\right)^3}$$

We do not recognize the general form of a Student.

• For $k=5$:

$$f_{f_5*f_5}(x) \;=\; \frac{400\sqrt{5}}{3\pi}\;\;\frac{x^4+120\,x^2+8400}{\left(x^2+20\right)^5}$$

We do not recognize the general form of a Student. The Kurtosis coefficient β_2 is 6 whereas the degree of the polynomial in the denominator minus the degree of the polynomial in the numerator is 6. This data does not correspond to a Student. A table presented below summarizes the different properties of the first Students.

• For $k=7$:

$$f_{f_7*f_7}(x) \;=\; \frac{10976\sqrt{7}}{25\pi}\;\;\frac{5\,x^6+1092\,x^4+112112\,x^2+9417408}{\left(x^2+28\right)^7}$$

We do not recognize the general form of a Student. We can again study the form factors: the kurtosis β_2 is 4, β_2 is 50 whereas the degree of the polynomial in the denominator minus that of the numerator is 8. This data does not correspond to a Student.

k	d°	V	β_2	β_4	$f_k(x)$	$f_k(0)$
1	2	-	-	-	$\frac{1}{\pi}\frac{1}{1+x^2}$	≃0.318
2	3	-	-	-	$\frac{1}{2\sqrt{2}}1/\left(1+\frac{x^2}{2}\right)^{\frac{3}{2}}$	≃0.354
3	4	3	-	-	$\frac{2}{\sqrt{3}\pi}1/\left(1+\frac{x^2}{3}\right)^{2}$	≃0.366
4	5	2	-	-	$\frac{3}{8}1/\left(1+\frac{x^2}{4}\right)^{\frac{5}{2}}$	≃0.375
5	6	5/3 ≃1.67	9	-	$\frac{8}{3\sqrt{5}\pi}1/\left(1+\frac{x^2}{5}\right)^{3}$	≃0.380
6	7	3/2 =1.5	6	-	$\frac{15}{16\sqrt{6}}1/\left(1+\frac{x^2}{6}\right)^{\frac{7}{2}}$	≃0.383
7	8	7/5 =1.4	5	125	$\frac{16}{5\sqrt{7}\pi}1/\left(1+\frac{x^2}{7}\right)^{4}$	≃0.385
8	9	4/3 ≃1.33	9/2 =4.5	67.5	$\frac{35}{64\sqrt{2}}1/\left(1+\frac{x^2}{8}\right)^{\frac{9}{2}}$	≃0.387
9	10	9/7 ≃1.29	21/5 =4.2	49	$\frac{128}{105\pi}1/\left(1+\frac{x^2}{9}\right)^{5}$	...
10	11	5/4 =1.25	4	...	...	
11	12	11/9 ≃1.22	27/7 ≃3.9			
12	13	6/5 =1.2	15/4 =3.75			
...						
∞		1	3	15	$1/\sqrt{2\pi}\ \ e^{-x^2/2}$	≃0.399

E13: Chi-squared distribution *Exercise on page 124*

$$f_1(x)=\frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}} \qquad f_2(x)=\frac{1}{2}e^{-\frac{x}{2}}$$

$$f_3(x)=\sqrt{\frac{x}{2\pi}}e^{-\frac{x}{2}} \qquad f_{10}(x)=\frac{x^4}{768}e^{-\frac{x}{2}}$$

E14: Product of distributions *Exercise on page 124*

1. W = ln Z = ln X +ln Y = U + V. First we are looking for the probability density of U and V, after we use a convolution to determine the distribution of the sum, and eventually we find the $Z = e^W$ density.

2. a. U and V densities: $f_{U,V}(x)=\begin{cases} 0 & x\leqslant 0 \\ e^x & 0<x<\ln 2 \\ 0 & x\geqslant \ln 2 \end{cases}$

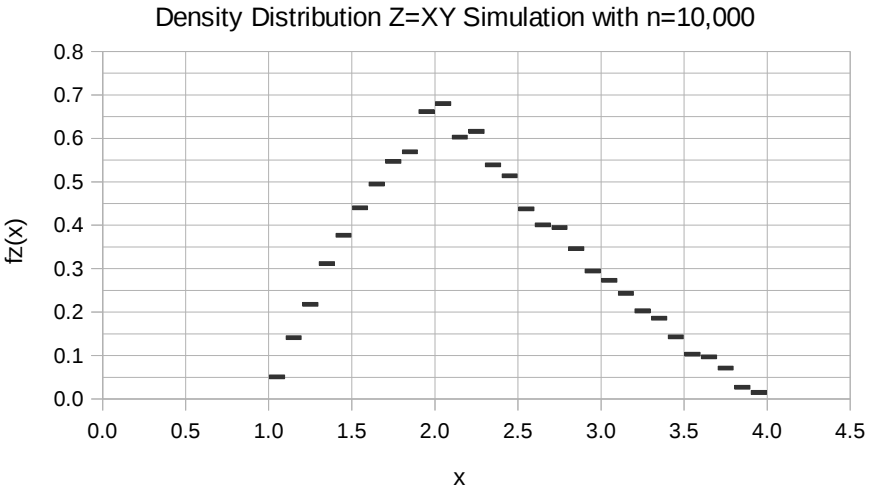
W density: $f_W(x)=\int_{-\infty}^{+\infty} f_U(y)f_V(x-y)dy$

After calculation: $f_W(x)=\begin{cases} 0 & x\leqslant 0 \\ xe^x & 0<x\leq \ln 2 \\ (2\ln 2-x)e^x & \ln 2<x<2\ln 2 \\ 0 & x\geqslant 2\ln 2 \end{cases}$

Z density: $f_Z(x)=\frac{1}{x}f_W(\ln x)$ if x>0 and zero otherwise.

After calculation: $f_Z(x)=\begin{cases} 0 & x\leqslant 1 \\ \ln x & 1<x\leq 2 \\ 2\ln 2-\ln x & 2<x<4 \\ 0 & x\geqslant 4 \end{cases}$

2. b. File for the simulation on a spreadsheet:
www.incertitudes.fr/livre/ProduitXY.ods



E15: Sum of exponentials *Exercise on page 125*

1. Exponential distribution: $f_X(x)=\begin{cases} 0 & if \quad x\le 0 \\ \lambda e^{-\lambda x} & if \quad x>0 \end{cases}$

Sum of exponential distributions: $f_{S_2}(x)=\int_{-\infty}^{+\infty} f_{X_1}(y) f_{X_2}(x-y)\,dy$

Non-zero integral if y>0 and x-y>0 and then 0<y<x:

$$f_{S_2}(x)=\lambda^2 e^{-\lambda x}\int_0^x dy \quad \text{and} \quad f_{S_2}(x)=\begin{cases} 0 & if \quad x\le 0 \\ \lambda^2 x e^{-\lambda x} & if \quad x>0 \end{cases}$$

2. If we continue we obtain the law of $S_3=S_2+X_3$ using the same method: $f_{S_3}(x)=\int_{-\infty}^{+\infty} f_{S_2}(y) f_{X_3}(x-y)\,dy$, nonzero if 0<y<x.

$$f_{S_3}(x)=\lambda^3 e^{-\lambda x}\int_0^x y\,dy \quad \text{and} \quad f_{S_3}(x)=\begin{cases} 0 & if \quad x\le 0 \\ \lambda^3\dfrac{x^2}{2}\mathrm{e}^{-\lambda x} & if \quad x>0 \end{cases}$$

We thus prove by induction the law of S_n:

$$f_{S_n}(x) = \begin{cases} 0 & if \quad x\le 0 \\ \lambda^n\dfrac{x^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda x} & if \quad x>0 \end{cases}$$

3. The law of M_n is an affine function of S_n: $f_{M_n}(x)=n f_{S_n}(nx)$

$$\text{then:} \quad f_{M_n}(x) = \begin{cases} 0 & if \quad x\le 0 \\ n^n\lambda^n\dfrac{x^{n-1}}{(n-1)!}\mathrm{e}^{-n\lambda x} & if \quad x>0 \end{cases}$$

E16: Inverse distribution *Exercise on page 125*

1. $F_Y(y)=P(Y\le y)=P(X\ge 1/y)=1-P(X<1/y)$

so $F_Y(y)=1-F_X(1/y)$ and $f_Y(y)=\dfrac{1}{y^2}f_X\left(\dfrac{1}{y}\right)$.

2. a.

$$f_{T_n}(x) = \begin{cases} 0 & if \quad x\le 0 \\ \dfrac{n^n\lambda^n}{(n-1)!x^{n+1}}\mathrm{e}^{-\frac{n\lambda}{x}} & if \quad x>0 \end{cases}$$

b. $f_Y(y)=\dfrac{1}{y^2}\dfrac{1}{\pi}\dfrac{1}{1+\frac{1}{y^2}}=\dfrac{1}{\pi}\dfrac{1}{1+y^2}$

The inverse of a Cauchy distribution is also a Cauchy distribution.

Chapter IV : Estimators

E1: Estimators of the mean *Exercise on page 148*

$E(A_3)=(E(X_1)+E(X_2)+E(X_3))/3$ (expectation linearity).
Then: $E(A_3)=3xm/3$ and $E(A_3)=m$. Also: $E(B_3)=m$.

On the other hand: $E(C_3)=4/3$ m.

Conclusion: A_3 and B_3 are estimators of m. C_3 is not an estimator of m. A_3 and B_3 are unbiased.

$V(A_3)=1/3^2\ (V(X_1)+V(X_2)+V(X_3))$ (independent variables).
Then: $r_{A_3}(m)=V(A_3)=\sigma^2/3$.
$V(B_3)=1/6^2\ V(X_1)+2^2/6^2\ V(X_2)+3^2/6^2\ V(X_3)$
and $r_{B_3}(m)=V(B_3)=7\sigma^2/18$.

Conclusion: $r_{A_3}(m)<r_{B_3}(m)$ and A_3 is a better estimator of the mean than B_3.

E2: Homokinetic Beam *Exercise on page 148*

$\sum_{i=1}^{100} v_i=121.3\times10^3$ and $\sum_{i=1}^{100} v_i^{\,2}=152.3\times10^6$

From the course we have the following unbiased estimator of the mean v_m: $T_n=\overline{V_n}=\frac{1}{n}\sum_{i=1}^{n} V_i$.

So: $v_m=\frac{1}{100}\sum_{i=1}^{100} v_i$ and $v_m=1213$ m/s.

From the course we have the following unbiased estimator of the variance $\sigma_V^{\,2}$:

$$R_n^{\,2}=\frac{\sum_{i=1}^{n}(V_i-\bar{V}_n)^2}{n-1}=\frac{n}{n-1}\frac{\sum_{i=1}^{n}(V_i-\bar{V}_n)^2}{n}=\frac{n}{n-1}\left(\frac{\sum_{i=1}^{n}V_i^{\,2}}{n}-\bar{V}_n^{\,2}\right) .$$

Then: $\sigma_V^{\,2}=\frac{100}{99}\left[\frac{1}{100}\sum_{i=1}^{100} v_i^2-\left(\frac{1}{100}\sum_{i=1}^{100} v_i\right)^2\right]$ and $\sigma_V^{\,2}=52.15\times10^3$.

Now in accordance with the central limit theorem we have shown that $V(T_n)=\frac{\sigma_V{}^2}{n}$. The sample size is sufficient and the distribution of the estimator has a Gaussian profile:

$$v=v_m\pm t\,\sigma_{T_n}=v_m\pm t_{\infty,95\%}\sigma_V/\sqrt{n}=1213\pm 1.96\times 228/10$$

and $v=1213\pm 45$ m/s with 95% confidence.

E3: Two estimators *Exercise on page 149*

1. Whatever θ $\Sigma p_i=1$, besides $0<p_i<1$ that's why $\theta \in]0,1/4[$.

2. $E(X)=0x3\theta+1x\theta+2x(1-4\theta)$ so $E(X)=2-7\theta$. $E(X^2)=4-15\theta$.
$V(X)=4-15\theta-(2-7\theta)^2$ and $V(X)=\theta(13-49\theta)$.

3. $E(T_n)=aE(\overline{X}_n)+b$ (expectation linearity).
and $E(\overline{X}_n)=1/n.\Sigma E(X_i)=E(X)$ then $E(T_n)=a(2-7\theta)+b=\theta$ (not biased)
and a=-1/7 and b=2/7. So $T_n=(2-\overline{X}_n)/7$.
$V(T_n)=1/49.V(\overline{X}_n)$ but $V(\overline{X}_n)=V(X)/n$ then $V(T_n)=\theta(13-49\theta)/49n$

4. $E(Y_i)=\theta$ then $E(Z_n)=n\theta$ and $E(U_n)=\theta$, estimator unbiased.

Values of Y	0	1
Probabilities	1-θ	θ

$V(U_n)=1/n^2.V(Z_n)$ but $V(Z_n)=nV(Y)$ and $V(Y)=\theta-\theta^2$, $V(U_n)=\theta(1-\theta)/n$.

5. $\overline{X}_{100}=(0x31+1x12+2x57)/100=1.26$ $T_{100}=(2-1.26)/7$ and $\underline{\theta_T\simeq 0.106}$
$r_T(\theta)=V(T_{100})\simeq 0.106(13-49\times 0.106)/4900$ and $r_T(\theta)=V(T_{100})\simeq 1.7\times 10^{-4}$.
$U_{100}=12/100$ then $\underline{\theta_U=0.12}$. $V(U_{100})=0.12\times 0.88/100$ and
$r_U(\theta)=V(U_{100})\simeq 10.6\times 10^{-4}$. $r_T(\theta) < r_U(\theta)$: We prefer T_{100}.
And T_n in general because $r_T - r_U = -36\theta/49n < 0$ whatever θ and n.

E4 : Ballot boxes *Exercise on page 150*

$E(P_i)=p$ wit i= 1 or 2 (expectation of a binomial distribution).
$E(T)=(E(P_1)+E(P_2))/2=2p/2$ and $E(T)=p$.
$E(U)=xp+(1-x)p$ and $E(U)=p$.
We have two estimators unbiased of p.
$V(P_i)=n_ipq$ with q=1-p (variance of a binomial distribution).
$V(T)=1/4(V(P_1)+V(P_2))$ because P_1 and P_2 are independent variables.

Then $V(T)=(n_1+n_2)pq/4$. Also $V(U)= (x^2n_1 + (1-x)^2n_2)pq$.
$dV(U)/dx= (2xn_1 -2(1-x)n_2)pq=0$ and $x_{opt}=n_2/(n_1+n_2)$.
$V(U)=n_1n_2pq/(n_1+n_2)$ and $V(U)-V(T)= - (n_1-n_2)^2/4(n_1+n_2)$.
U has a mean square error smaller than V, then U is a better estimator than V for x_{opt} (U=V if $n_1=n_2$).

E5: Continuous variable *Exercise on page 150*

1. We have a positive density.

Let calculate the integral $\int_{-\infty}^{\infty} f(x)\,dx$ and show the value is 1:

$$I_A=\int_1^A \frac{a}{x^{a+1}}dx=a\left[-x^{-a}/a\right]_1^A=1-1/A^a \text{ and } \lim_{A\to+\infty} I_A=1 .$$

2. $J_A=\int_1^A \frac{ax}{x^{a+1}}dx=a\left[x^{-a+1}/(-a+1)\right]_1^A$ and $E(X)=\frac{a}{a-1}$

3. $$L(x_i;a)=\prod_{i=1}^n f(x_i;a)=\prod_{i=1}^n \frac{a}{x_i^{a+1}}=\frac{a^n}{\left(\prod x_i\right)^{a+1}}$$

then $\ln L=n\ln a-(a+1)\sum_{i=1}^n \ln x_i$ and $\frac{d\ln L}{da}=\frac{n}{a}-\sum_{i=1}^n \ln x_i$

and $\frac{d\ln L}{da}=0$ give $a=\frac{n}{\sum \ln x_i}=\frac{n}{\ln(\prod x_i)}$.

Method of Moments: $\bar{x}=\frac{a}{a-1}$ and $a=\frac{\bar{x}}{\bar{x}-1}$.

4.

• Maximum Likelihood: $\hat{a}\simeq\frac{8}{\ln(14.18)}$ and $\hat{a}\simeq 3.02$.

• Method of Moments: $\bar{x}\simeq 1.52$ and $\hat{a}\simeq 2.94$.

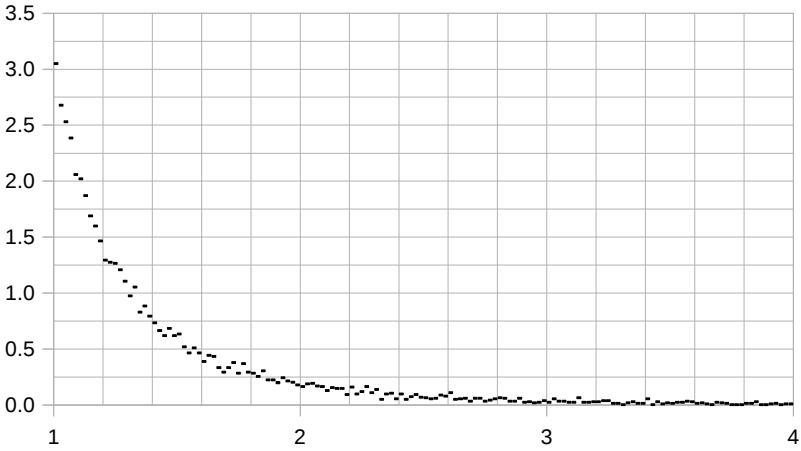
5. $K_A = \int_1^A \frac{a}{x^{a-1}} dx = \frac{a}{2-a}\left[1/x^{a-2}\right]_1^A$ this integral tends to a finite limit only if a> 2. Otherwise the integral is divergent and the variance is not defined.

If a>2 $E(X^2) = \frac{a}{a-2}$ and $V(X) = \frac{a}{(a-2)(a-1)^2}$.

6. We want to simulate X with a uniform distribution U(0,1). We use the inversion method. The cumulative distribution function give:

$F(x) = 1 - \frac{1}{x^a} = y$ and $x = \frac{1}{(1-y)^{1/a}}$ so $X = \frac{1}{U^{1/a}}$.

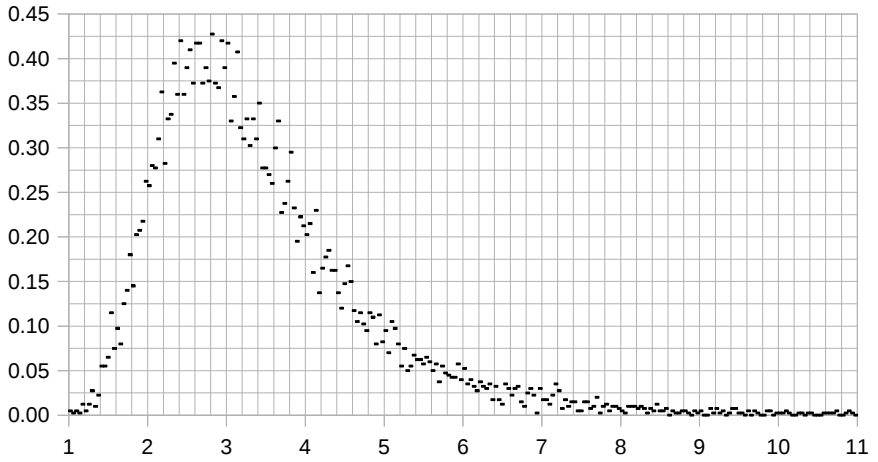
We obtain for our sample of size n=8 and N=10,000 trials the following distribution:



Simulation of $f_X(x)$ for a=3.02

We find by the simulation: E(X)≃1.49 and σ_X≃0.85, values that correspond to the theoretical values.

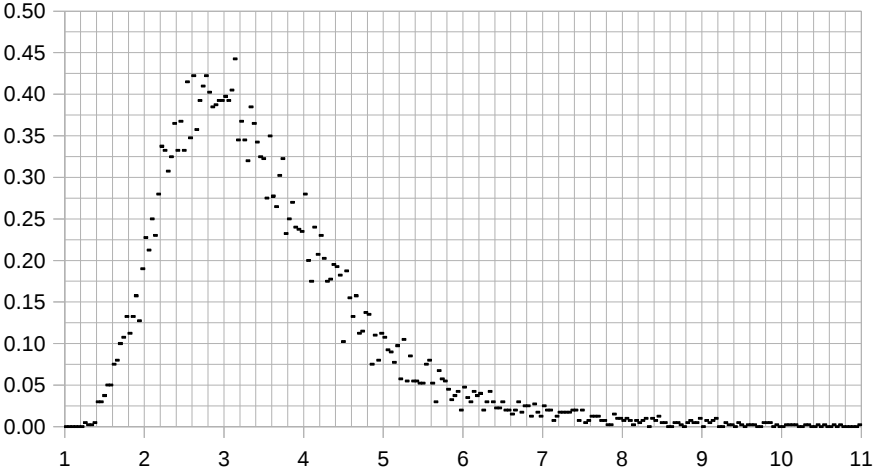
For the maximum likelihood estimator we have the following distribution for N=10,000:



Simulation with the maximum likelihood estimator for a=3.02

We find by the simulation a mean 3.44 and a standard deviation 1.4 .

For the method of moments estimator we have the following distribution for N=10,000:



Simulation with the method of moments estimator for a=2.94

We find by the simulation a mean 3.54 and a standard deviation 1.4 .

In both cases we have significant biases. The mean, the expectation of the estimator, is well above the point estimate of *a*. The variance is also important, for example at 90% with the method of the maximum likelihood, the interval of *a* is between 1.8 and 6.1 approximately.
At this stage of the study we are not assured of an asymptotically zero bias, nor of a convergence. Simulations should be carried out for different sample sizes in order to conjecture the evolution as a function of *n* of bias and mean square error.
We can also undertake an analytical calculation of the distribution of estimators by product of convolutions and composition of functions.

Simulations on the LibreOffice spreadsheet:
www.incertitudes.fr/livre/Simul5.ods

E6: Linear Density *Exercise on page 151*

1. $\int_0^1 (a\,x+b)\,dx=\frac{a}{2}+b=1$ and $b=1-\frac{a}{2}$

Moreover f(0) and f(1) must be positive hence: $-2 \le a \le 2$.

2. $E(X)=\int_0^1 (a\,x^2+b\,x)\,dx=\frac{a}{3}+\frac{b}{2}$ and $E(X)=\frac{1}{2}+\frac{a}{12}$

$E(X^2)=\frac{a}{4}+\frac{b}{3}=\frac{1}{3}+\frac{a}{12}$ and $V(X)=\frac{1}{3}+\frac{a}{12}-\left(\frac{1}{2}+\frac{a}{12}\right)^2$

so $V(X)=\frac{1}{12}-\frac{a^2}{144}=\frac{12-a^2}{144}$

3. $\overline{X_n}=\frac{1}{n}\sum_{i=1}^{n} X_i=\frac{1}{2}+\frac{\hat{a}}{12}$ and $T_n=\hat{a}=12\left(\overline{X_n}-\frac{1}{2}\right)$

$E(T_n)=a$, estimator unbiased.

$V(T_n)=144\times\frac{1}{n^2}\times n\,V(X)$ and $V(T_n)=\frac{12-a^2}{n}$

4. With the estimator of question 3.: $a_m \simeq 0.59$.

We have a linear relation between the estimator of *a* and that one of the mean, thus, as the mean, the estimator of *a* obeys the central limit theorem and in the case of the large numbers: $a=a_m \pm t.s/\sqrt{n}$. Here, *n=9*, the sample is too small for this method. Nor can we use Student's distribution because the distribution of the population is linear and not Gaussian. We could do then an integral calculation or a numerical simulation.

E7: Estimators for the exponential law *Exercise on page 151*

1. $E(T_n)=\frac{n^n\lambda^n}{(n-1)!}\int_0^{+\infty}\frac{e^{-\frac{n\lambda}{x}}}{x^n}dx$, let $u=\frac{n\lambda}{x}$ and

$E(T_n)=\frac{n\lambda}{(n-1)!}\int_0^{+\infty}u^{n-2}e^{-u}dx=\frac{n\lambda}{(n-1)!}I_n$. With an integration by parts: $I_n=(n-2)I_{n-1}$ then $I_n=(n-2)!$ because $I_3=1$.

So: $E(T_n)=\frac{n}{n-1}\lambda$ and $b_{T_n}(\lambda)=\frac{\lambda}{n-1}$.

$E(T_n^{\,2})=\frac{n^2\lambda^2}{(n-1)!}I_{n-1}=\frac{n^2\lambda^2}{(n-1)(n-2)}$, $V(T_n)=\frac{n^2\lambda^2}{(n-2)(n-1)^2}$

and $r_{T_n}(\lambda)=\lambda^2\frac{(n+2)}{(n-2)(n-1)}$.

2. $E(W_n)=\lambda$ and $V(W_n)=r_{W_n}(\lambda)=\frac{\lambda^2}{n-2}$ (using the properties of expectation and variance).

3. We prefer W_n to estimate λ because unlike T_n its bias is zero and moreover its mean square error is lower.

E8 : Decays *Exercise on page 152*

1. Let us make a false reasoning:
Average of the ten measured lifetimes: m≃2.60 μs.
Sample standard deviation: s≃2.24 μs.
Student's t-value: t≃1.83 (9 degrees of freedom and 90% confidence).
Central limit theorem: m=2.6 ± 1.3 μs with 90% confidence.
This result is a good estimate for the mean (zero bias, controlled risk-mean square error).
For λ, the inverse of the mean, we now use the uncertainty propagation formula: $\lambda=1/m$ then $\Delta\lambda=\Delta m/m^2$

So: $\lambda = 0.39 \pm 0.19$ /μs with a confidence of 90%.

Comments :
1. Here the central limit theorem can not be applied because we clearly have a small sample (n=10). Moreover, although Student's distribution can be used for small numbers, it is used only when the population is normal, which is not the case here because it is an exponential distribution. If we had taken t_∞ it was also false because *n* small. One can imagine that to enlarge with the Student would make it possible to counterbalance the small number, but it is a rush job and it is false!
2. The uncertainty propagation formula is an approximation, it is exact if all distributions of probabilities have the same form, here it is not the case. However, the approximation is correct and the calculation is fast. But nothing assures us that we are with a confidence of 90%, maybe a little less or a little more. We do not control our estimate here.

2. By applying a linear function and according to the exercise on page 151:

$$f_{W_n}(x)=\frac{n}{n-1}f_{T_n}\left(\frac{nx}{n-1}\right)=\frac{(n-1)^{n-1}\lambda^n}{(n-2)!\,x^{n+1}}e^{-\frac{(n-1)\lambda}{x}}$$

We then integrate on the first 5 percentiles and then the first 95 percentiles to obtain the limits of the confidence interval:

$$\int_0^{\lambda_{min}} f_{W_{10}}(x)\,dx=0.05 \quad \text{and} \quad \int_0^{\lambda_{max}} f_{W_{10}}(x)\,dx=0.95$$

By numerical calculation of the integrals: $\lambda_{min}\simeq 0.221$ and $\lambda_{max}\simeq 0.639$.

And finally: $\lambda = 0{,}385\,^{+0{,}254}_{-0{,}164}$ /μs with 90% confidence.

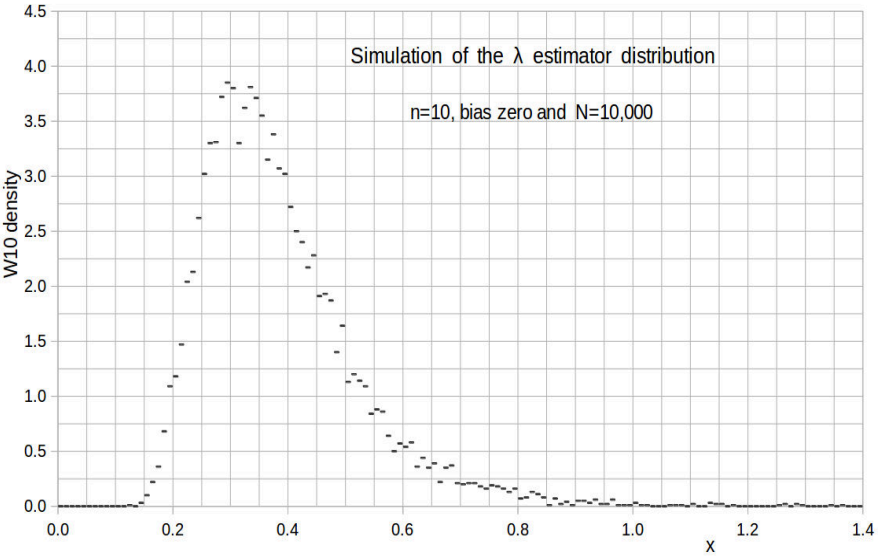
3. We generate 10 identical and independent exponential distributions X_i from ten independent uniform continuous distributions $U_i(0,1)$:

$$X_i = - \ln(U_i) / \lambda$$

We take λ as the point estimate obtained with the sample.
Here n=10, then: $W_{10}=9/(X_1+X_2+...+X_{10})$. The computer performs this calculation N=10,000 and the 10000 results for W_{10} are sorted by class and placed on a graph to create the W_{10} sampling distribution. With the discretized cumulative distribution function we evaluate the positions of the 5th and 95th centiles: $x_{5\%}\simeq 0.22$ and $x_{95\%}\simeq 0.64$.

Then: $\lambda = 0.39 \,^{+0.25}_{-0.17}$ /µs with a confidence of 90%.

This result fully corroborates the exact calculation.



File created with the LibreOffice spreadsheet:

www.incertitudes.fr/livre/Expos.ods

VIII. Bibliography / Sources / Softwares / Illustrations

Books

[vi] WONNACOTT. *Introductory Statistics for Business and Economics*, 1972. 919 p.

SHELDON M. ROSS. *A first course in probability*.1976. 585 p.

Livres

[iv] PROTASSOV Konstantin. *Probabilités et incertitudes dans l'analyse des données expérimentales*. Presses Universitaires de Grenoble, 1999. 128 p.

[x] JOURNEAUX Roger. *Traitement des mesures*. Ellipses 2009 377p

[iii] SAPORTA Gilbert. *Probabilités, analyse des données et statistique*. Technip, 2006. 622 p.

RONDY SYLVAIN. *Maths - Économique et Commerciale option Scientifique - 2e année*. Collection Prépas Sciences, ellipse, 2014. 810 p.

Web

Place http:// in front of the site name. Most files are copied to the folder <www.incertitudes.fr/livre/>.

ROUAUD Mathieu. *Calculs d'incertitudes*. <www.incertitudes.fr/Incertitudes.html>

Online integral calculator: <www.integral-calculator.com>

[v] *Nombres, mesures et incertitudes*. mai 2010. <www.educnet.education.fr/rnchimie/recom/mesures_incertitudes.pdf>

[vii] *Évaluation des données de mesure — Guide pour l'expression de l'incertitude de mesure.* <www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_F.pdf>, 2008. 132p.

Breuil P, Di Benedetto D. *Incertitude et étalonnage.*2000. 16p.

[ix] *Gauss–Newton algorithm* - Wikipédia (26 February 2017)

Articles

[i] *Regression line analysis* K.S.KRANE et L.SCHECTER American J. of Physics **50**,82 (1982).

[ii] Rouaud Mathieu. *Mesure avec une règle.* Bulletin de l'Union des Physiciens n°913, avril 2009.

Softwares

All the software used are free and open source.

- Word processing and spreadsheet: *LibreOffice*
- Graphics: *Gimp* (points), *Inkskape* (vectors) and *Blender* (3D).
- Computation: *Scilab* (numerical), *XCAS* (symbolic) et *PHP* programs (on server).
- Plotters: *KmPlot, TeXgraph.*
- Operating System: *Ubuntu*.

Illustrations

3D representations in front of the table of contents and third thumbnail of the back cover.

A random walker moves in a plane. He throws two coins and looks at the results: two tails, one tail and one head, one head and one tail or two heads. The first coin tells him if he has to make a first move towards the East or the West, the second if he has to make a second step towards the South or the North. Thus at each time interval Δt he moves in the plane of two steps. During Δt the distance traveled by the walker is $\Delta d = 2\,p$ (p length of one step).

At what distance from the starting point is the walker at the instant t?
(after n time intervals: $t = n\,\Delta t$)

For $n=1$, we draw the following tables:

0.5	1	1
-0.5	1	1
	-0.5	0.5

25%	25%
25%	25%

The center of the table is its starting point. On the abscissa x (direction East-West) the displacement is more or less one step ($x=\pm p$, $p= \Delta d /2$ and we have fixed $\Delta d=1$). Similarly on the y-axis: $y=\pm p$. In each square is indicated the number of possibilities to meet at this place. The second table indicates the probabilities (1/4, 4=2x2).

In the four possibilities he is at a distance $\sqrt{2}/2$ from the point of origin. And, in terms of standard deviation, the characteristic distance is $s=\sqrt{(4\times 1/\sqrt{2})^2/(4-1)}\simeq 0.816$.

For $n=2$, we draw the following tables:

1	1	2	1
0	2	4	2
-1	1	2	1
	-1	0	1

6%	13%	6%
13%	25%	13%
6%	13%	6%

For example to be at (x=0 ; y=-1), there are two possible paths: (PF,PP) and (FP,PP). Standard deviation $s_2 \simeq 1.033$.
The walker has a one in four chance to be back to the starting point.

For *n=3*, we draw the following tables:

1.5	1	3	3	1
0.5	3	9	9	3
-0.5	3	9	9	3
-1.5	1	3	3	1
	-1.5	-0.5	0.5	1.5

2%	5%	5%	2%
5%	14%	14%	5%
5%	14%	14%	5%
2%	5%	5%	2%

$s_3 \simeq 1.234$

For *n=4*, we draw the following tables:

2	1	4	6	4	1
1	4	16	24	16	4
0	6	24	36	24	6
-1	4	16	24	16	4
-2	1	4	6	4	1
	-2	-1	0	1	2

0.4%	1.6%	2.3%	1.6%	0.4%
1.6%	6.3%	9.4%	6.3%	1.6%
2.3%	9.4%	14%	9.4%	2.3%
1.6%	6.3%	9.4%	6.3%	1.6%
0.4%	1.6%	2.3%	1.6%	0.4%

For *n=5*, we draw the following tables:

2.5	1	5	10	10	5	1
1.5	5	25	50	50	25	5
0.5	10	50	**100**	**100**	50	10
-0.5	10	50	**100**	**100**	50	10
-1.5	5	25	50	50	25	5
-2.5	1	5	10	10	5	1
	-2.5	-1.5	-0.5	0.5	1.5	2.5

0.1%	0.5%	1.0%	1.0%	0.5%	0.1%
0.5%	2.4%	4.9%	4.9%	2.4%	0.5%
1.0%	4.9%	9.8%	9.8%	4.9%	1.0%
1.0%	4.9%	9.8%	9.8%	4.9%	1.0%
0.5%	2.4%	4.9%	4.9%	2.4%	0.5%
0.1%	0.5%	1.0%	1.0%	0.5%	0.1%

For $n=6$, we draw the following tables:

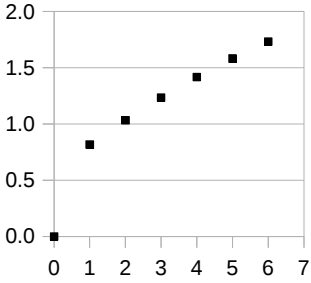
3	1	6	15	20	15	6	1
2	6	36	90	**120**	90	36	6
1	15	90	**225**	**300**	**225**	90	15
0	20	**120**	**300**	**400**	**300**	**120**	20
-1	15	90	**225**	**300**	**225**	90	15
-2	6	36	90	**120**	90	36	6
-3	1	6	15	20	15	6	1
	-3	-2	-1	0	1	2	3

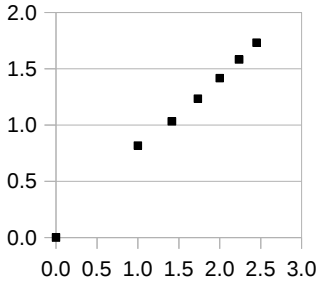
0.0%	**0.1%**	**0.4%**	**0.5%**	**0.4%**	**0.1%**	**0.0%**
0.1%	**0.9%**	**2.2%**	**2.9%**	**2.2%**	**0.9%**	**0.1%**
0.4%	**2.2%**	**5.5%**	**7.3%**	**5.5%**	**2.2%**	**0.4%**
0.5%	**2.9%**	**7.3%**	**9.8%**	**7.3%**	**2.9%**	**0.5%**
0.4%	**2.2%**	**5.5%**	**7.3%**	**5.5%**	**2.2%**	**0.4%**
0.1%	**0.9%**	**2.2%**	**2.9%**	**2.2%**	**0.9%**	**0.1%**
0.0%	**0.1%**	**0.4%**	**0.5%**	**0.4%**	**0.1%**	**0.0%**

Hence the evolution of the quadratic mean distance from the starting point as a function of time:

t (Δt)	0	1	2	3	4	5	6
√t	0.00	1.00	1.41	1.73	2.00	2.24	2.45
s (2p)	**0.000**	**0.816**	**1.033**	**1.234**	**1.417**	**1.582**	**1.732**

We plot the curves s as a function of t and s as a function of √t:





We find a much better correlation in $\sqrt{t}$. Indeed we saw directly that the distance at the origin did not evolve proportionally to time, for $n=2$ we are about one unit of the origin, so we should be towards 3 for $n=6$.
This variation in $\sqrt{t}$ is characteristic of diffusion phenomena and here finds its analogy with the compensation of errors in $\sqrt{n}$.

IX. TABLES / Index

A. Standard normal distribution

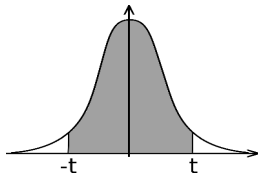
$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

$$F(z)=\int_{-\infty}^{z} f(z)dz=P(Z\leqslant z) \qquad P(Z>z)=1-P(Z\leqslant z)$$

$$P(Z\leqslant -z)=P(Z>z) \quad \text{Example :} \quad P(Z\leqslant 1.67)\simeq 0.95254$$

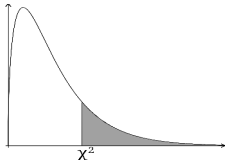
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

B. Student's t-values



Student value t		Confidence (%)								
		50	**80**	**90**	**95**	**98**	**99**	**99.5**	**99.8**	**99.9**
Degrees of freedom (sample size minus the number of parameters)	**1**	1.00	3.08	6.31	12.7	31.8	63.7	127	318	637
	2	0.82	1.89	2.92	4.30	6.96	9.92	14.1	22.3	31.6
	3	0.76	1.64	2.35	3.18	4.54	5.84	7.45	10.2	12.9
	4	0.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61
	5	0.73	1.48	2.02	2.57	3.36	4.03	4.77	5.89	6.87
	6	0.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96
	7	0.71	1.41	1.89	2.36	3.00	3.50	4.03	4.79	5.41
	8	0.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04
	9	0.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78
	10	0.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59
	11	0.70	1.36	1.80	2.20	2.72	3.11	3.50	4.02	4.44
	12	0.70	1.36	1.78	2.18	2.68	3.05	3.43	3.93	4.32
	13	0.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22
	14	0.69	1.35	1.76	2.14	2.62	2.98	3.33	3.79	4.14
	15	0.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07
	16	0.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.01
	17	0.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97
	18	0.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92
	19	0.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88
	20	0.69	1.33	1.72	2.09	2.53	2.85	3.15	3.55	3.85
	22	0.69	1.32	1.72	2.07	2.51	2.82	3.12	3.50	3.79
	24	0.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.75
	26	0.68	1.31	1.71	2.06	2.48	2.78	3.07	3.43	3.71
	28	0.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67
	30	0.68	1.31	1.70	2.04	2.46	2.75	3.03	3.39	3.65
	40	0.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55
	50	0.68	1.30	1.68	2.01	2.40	2.68	2.94	3.26	3.50
	60	0.68	1.30	1.67	2.00	2.39	2.66	2.91	3.23	3.46
	70	0.68	1.29	1.67	1.99	2.38	2.65	2.90	3.21	3.44
	80	0.68	1.29	1.66	1.99	2.37	2.64	2.89	3.20	3.42
	90	0.68	1.29	1.66	1.99	2.37	2.63	2.88	3.18	3.40
	100	0.68	1.29	1.66	1.98	2.36	2.63	2.87	3.17	3.39
	200	0.68	1.29	1.65	1.97	2.35	2.60	2.84	3.13	3.34
	300	0.68	1.28	1.65	1.97	2.34	2.59	2.83	3.12	3.32
	500	0.67	1.28	1.65	1.96	2.33	2.59	2.82	3.11	3.31
	1000	0.67	1.28	1.65	1.96	2.33	2.58	2.81	3.10	3.30
	∞	**0.67**	**1.28**	**1.64**	**1.96**	**2.33**	**2.58**	**2.81**	**3.09**	**3.29**

C. Chi-square values



χ^2	Probability α to reject the hypothesis while it is true (%)										
number of degrees of freedom	**99**	**90**	**75**	**50**	**30**	**20**	**10**	**5**	**2**	**1**	**0,1**
1	0.000157	0.0158	0.102	0.455	1.07	1.64	2.71	3.84	5.41	6.63	10.8
2	0.0201	0.211	0.575	1.386	2.41	3.22	4.61	5.99	7.82	9.21	13.8
3	0.115	0.584	1.21	2.366	3.66	4.64	6.25	7.81	9.84	11.3	16.3
4	0.297	1.06	1.92	3.357	4.88	5.99	7.78	9.49	11.7	13.3	18.5
5	0.554	1.61	2.67	4.351	6.06	7.29	9.24	11.1	13.4	15.1	20.5
6	0.872	2.20	3.45	5.348	7.23	8.56	10.6	12.6	15.0	16.8	22.5
7	1.24	2.83	4.25	6.346	8.38	9.80	12.0	14.1	16.6	18.5	24.3
8	1.65	3.49	5.07	7.344	9.52	11.0	13.4	15.5	18.2	20.1	26.1
9	2.09	4.17	5.90	8.343	10.7	12.2	14.7	16.9	19.7	21.7	27.9
10	2.56	4.87	6.74	9.342	11.8	13.4	16.0	18.3	21.2	23.2	29.6
11	3.05	5.58	7.58	10.34	12.9	14.6	17.3	19.7	22.6	24.7	31.3
12	3.57	6.30	8.44	11.34	14.0	15.8	18.5	21.0	24.1	26.2	32.9
13	4.11	7.04	9.30	12.34	15.1	17.0	19.8	22.4	25.5	27.7	34.5
14	4.66	7.79	10.2	13.34	16.2	18.2	21.1	23.7	26.9	29.1	36.1
15	5.23	8.55	11.0	14.34	17.3	19.3	22.3	25.0	28.3	30.6	37.7
16	5.81	9.31	11.9	15.34	18.4	20.5	23.5	26.3	29.6	32.0	39.3
17	6.41	10.1	12.8	16.34	19.5	21.6	24.8	27.6	31.0	33.4	40.8
18	7.01	10.9	13.7	17.34	20.6	22.8	26.0	28.9	32.3	34.8	42.3
19	7.63	11.7	14.6	18.34	21.7	23.9	27.2	30.1	33.7	36.2	43.8
20	8.26	12.4	15.5	19.34	22.8	25.0	28.4	31.4	35.0	37.6	45.3
21	8.90	13.2	16.3	20.34	23.9	26.2	29.6	32.7	36.3	38.9	46.8
22	9.54	14.0	17.2	21.34	24.9	27.3	30.8	33.9	37.7	40.3	48.3
23	10.2	14.8	18.1	22.34	26.0	28.4	32.0	35.2	39.0	41.6	49.7
24	10.9	15.7	19.0	23.34	27.1	29.6	33.2	36.4	40.3	43.0	51.2
25	11.5	16.5	19.9	24.34	28.2	30.7	34.4	37.7	41.6	44.3	52.6
26	12.2	17.3	20.8	25.34	29.2	31.8	35.6	38.9	42.9	45.6	54.1
27	12.9	18.1	21.7	26.34	30.3	32.9	36.7	40.1	44.1	47.0	55.5
28	13.6	18.9	22.7	27.34	31.4	34.0	37.9	41.3	45.4	48.3	56.9
29	14.3	19.8	23.6	28.34	32.5	35.1	39.1	42.6	46.7	49.6	58.3
30	15.0	20.6	24.5	29.34	33.5	36.3	40.3	43.8	48.0	50.9	59.7
31	15.7	21.4	25.4	30.34	34.6	37.4	41.4	45.0	49.2	52.2	61.1
32	16.4	22.3	26.3	31.34	35.7	38.5	42.6	46.2	50.5	53.5	62.5
33	17.1	23.1	27.2	32.34	36.7	39.6	43.7	47.4	51.7	54.8	63.9
34	17.8	24.0	28.1	33.34	37.8	40.7	44.9	48.6	53.0	56.1	65.2
35	18.5	24.8	29.1	34.34	38.9	41.8	46.1	49.8	54.2	57.3	66.6
36	19.2	25.6	30.0	35.34	39.9	42.9	47.2	51.0	55.5	58.6	68.0
37	20.0	26.5	30.9	36.34	41.0	44.0	48.4	52.2	56.7	59.9	69.3
38	20.7	27.3	31.8	37.34	42.0	45.1	49.5	53.4	58.0	61.2	70.7
39	21.4	28.2	32.7	38.34	43.1	46.2	50.7	54.6	59.2	62.4	72.1
40	22.2	29.1	33.7	39.34	44.2	47.3	51.8	55.8	60.4	63.7	73.4

Index

This book is made for anyone interested in experimental sciences and mathematics. Statistics and surveys are very common in our society, and the application area is supposed to be as large as possible. We are willing to go beyond the theory so that the invistigator may find necessary tools to solve simple and rigourous uncertainties quantification. Indeed, science tries to link natural phenomena to a mathematical logic. Overall coherence and truth need to be practised by a critical mind that lays on measures accompanied with their uncertainties.