Introduction to probabilities

E. Lantz 2018

Table of contents

Bibliography: Martine Quinio Benamo, "Probabilities and statistics today", Editions L'Harmattan, elementary level. (in French)

Gilbert Saporta, "Probabilities, data analysis and statistics", Editions Technip, advanced level, but still focused on applications (in French or in English).

Chapter 1) Probabilities, conditional probabilities, independence

I) Intuitive concept of probability

A probability is associated with the "chance" that an event will occur.

For example, when rolling a die, you have a 1 in 6 chance of rolling a 5, which is a probability P(5)=1/6.

....or lotto, the probability of getting the grid 3,7, 9,18, 21, 24 is about $1/(14\ 10^6\ )$

Note 1: the probability of obtaining the 1,2,3,4,5,6 grid is identical, which is not intuitive for everyone

Note 2: A grid has six check numbers. There is no notion of order between these numbers. On the other hand, with the drawing, the balls are drawn one by one (without discount), thus in a precise order. We conclude that a winning grid corresponds to several distinct draws (How many?).

In these two examples, a probability P_i is associated with an elementary event i. When rolling a single die, there are 6 elementary events, corresponding to each side, and, for an untapped die Pi =1/6 for i from 1 to 6.

The lotto example is less trivial: one can choose as elementary event a grid (6 numbers, without repetition, between 1 and 49 in any order), or a draw (6 numbers ordered, without repetition, between 1 and 49).

To ensure that a die is not rigged, roll it a large number of times and ensure that the proportion of results is approximately equal to 1/6 for each side. The intuitive idea of probability is thus associated with that of frequency, calculated on a very large number of repetitions of the experiment.

The notion of probability does not always cover the notion of "chance of an outcome in an experiment to be performed". It can translate incomplete knowledge acquired using the results of an experiment already carried out. One example, which we will take up again, is the following (see also TD2):

We know that at a given date, 3% of a population has hepatitis. Screening tests for the disease are available:
If the person is sick, then the test is positive with a 95% probability.
If the person is not sick, then the test is positive with a 10% probability.

A person randomly selected from the population is tested.
What is the probability of a person being sick if their test is positive?

In this example, the test result does not indicate whether the person is ill (or not). On the other hand, we can translate all the knowledge acquired, by the test but also before the test, by a probability for the person tested to be sick.

II) Some definitions

The mathematical theory of probability widely uses the language of sets. This is defined as:

<u>Universe Ω :</u> All elementary events
Example 1: Draw a die Ω={1,2,3,4,5,6}

Example 2: The lotto. There are two possible universe choices: the set of grids, or the set of ordered prints.

This second example shows that, for a given practical problem, the universe is not necessarily imposed. But it's always the first thing to define.

<u>Part of Ω :</u> this is a set A of elementary events included in Ω
Example : A={all values >3 following a roll of the die}={4,5,6}
A is called an event (not elementary if Card(A) $\neq$ 1)

<u>Probability</u>: application of all parts of Ω on[0 1] such that :
-P(Ω)=1
- if A_1,......,A_n are separated parts in Ω (parts with a null intersection between any two parts, also called incompatible events), then

$$P(\cup A_i) = \sum_{i=1}^{n} P(A_i)$$

These properties of probability are called Kolmogorov axiomatics.

Some immediate properties are deduced from this axiomatic:

$P(\emptyset) + P(\Omega) = P(\Omega) \Longrightarrow P(\emptyset) = 0$

More generally, $P(\bar{A}) = 1 - P(A)$, $\bar{A}$ is the complementary part of A (realized if A is not realized).

P(A ∪ B)= P(A)+ P(B)-P(A∩B)

Indeed, three separate parts can be defined: A∩B, C=A- A∩B, D= B- A∩B
We have immediately A ∪ B= C ∪ D ∪ A∩B.
Now, P(C)=P(A) -P(A∩B), P(D)=P(B) -P(A∩B) from where
P(A ∪ B)=P(A)- P(A∩B)+ P(B)- P(A∩B)+ P(A∩B)

III) Combinations of equiprobable elementary events

If all elementary events (elements of Ω) are equiprobable, the probability associated with each of them is obviously 1/ Card(Ω). The probability associated with a part A of Ω is, no less obviously, Card(A) / Card(Ω).
Card(Ω) the number of possible events and Card(A) the number of favorable events.

P(A)=(Number of elements in A)/(Number of elements in Ω)= (Number of favorable cases)/(Number of possible cases)

Example 1. The successive letters of the word CHIENNE are drawn at random. What is the probability of getting the word CHIENNE?

We thus have at the beginning eight letters, including two E which we will number E1 and E2, and two N numbered N1 and N2.

Universe: set of possible ordered prints. There are 8 ways to shoot the first letter, then 7 ways to shoot the second letter, etc., that is 8! possible cases. The inversion of N1 and N2 gives a different draw.

Favorable cases: the word CHIENNE in order. However N1 N2 and N2 N1
are both in favor. Ditto for the E. Giving 4 favorable cases and
P(CHIENNE)=4/8!

Example 2: The lotto. The choice facilitating the calculation is to take as universe the ordered draws of 6 numbers among 49: Card(Ω)=49×48×..×44= 49!/43!

Number of favorable cases: number of ordered draws corresponding to a grid of 6 numbers, or 6!

Let P(A)=1/C_{49}^{6}, with $C_n^k \triangleq \frac{n!}{k!\,(n-k)!}$, C_n^k is the binomial coefficient, and ≜ will be used in this course in the sense of "equal by definition".

IV) Conditional Probabilities

Let's go back to a TD1 exercise: 3 French clubs have qualified for the quarter-finals of the Champions League. What is the probability that two French clubs will meet?

To avoid counting possible cases and favorable cases, we can first calculate the probability that one of the French clubs, let's name F1, meets a French club. There are 7 possible opponents including 2 French and this probability is 2/7. We still have to consider the case where F1 meets a foreign club. Then F2 has one chance out of 5 to meet F3, one of the remaining 5 possible opponents. 1/5 is not the probability that F2 meets a French club (this probability is of course 2/7) but the probability that F2 meets a French club knowing that F1 meets a foreign club.
If an element of Ω is formed by a list of four matches (no matter of the order of the matches and the order of opponents in a match), Ω has thus been broken down into two separate parts whose union is equal to Ω (it is the definition of a partition of Ω).
C={elements where F1 meets a French club }. P(C)=2/7
B= { elements whereF1 meets a foreign club}. P(B)=5/7

Let be A the part of Ω where two French clubs meet

P(A|B) reads P(A knowing B) and is defined as (P(A⋂B) in case we take B as universe). Clearly, coming back to the universe Ω=C ⋃ B, we have
P(A⋂B)= P(A|B).P(B) (Formula of compound probabilities).

We will use this equality as definition ; the conditional probability of A knowing B, noted P(A|B), is defined by : P(A|B)≜(P(A⋂B))/(P(B)).

Let's take our example again:
P(A)= P(C)+P(A⋂B)= P(C)+P(A|B).P(B)=2/7+1/5· 5/7=3/7

Bayes's theorem:
In the compound probability formula, A and B can be interchanged, hence:

P(A⋂B)= P(A|B).P(B)= P(B|A).P(A)

that you can write:

P(A|B)= (P(B|A).P(A))/(P(B))

This formula, very simple, is at the origin of a whole section of probability theory and estimation, called Bayesian theory. An overview is given by using the example given in the introduction:
It is known that at a given date, 3% of a population is infected with hepatitis:
If the person is sick, then the test is positive with a 95% probability.
If the person is not sick, then the test is positive with a 10% probability.

A person is randomly tested in the population and the test is positive. How likely is the person tested to be sick?
P(sick|positive)=(P(positive| sick).P(sick))/(P(positive))
=(0.95.0.03)/(0.95.0.03+0.10.0,97)=0.23

This number may seem small. It illustrates the importance of information, called a priori, which gives the probability of being sick before the measurement, here 3%. The test increases this probability in the event of a positive result, but this probability remains very different from what a reasoning based solely on the reliability of the test could give. This information should be estimated with caution. For example, think about what you should think of as P(sick) if you are only testing people going to a health center....

Total Probability Formula

The probability of being positive was calculated by dividing Ω into sick and not sick, i.e. by partitioning Ω. This approach is general

$$P(B) = \sum_i P\ (\mathrm{B}|\mathrm{A}_i)P(\mathrm{A}_i)$$

where the A_i form a partition of Ω : separate parts whose union is equal to Ω

V) Independence

Definition 1: A and B are independent if and only if P(A|B)=P(A)

In other words, restricting the universe to B does not change the probability of A.

An alternative definition immediately derived from the above definition will often be used:

P(A|B)=P(A)⟹P(A and B)= P(A|B). P(B)= P(A) P(B)

Definition 2: A and B are independent if and only if P(A and B)=P(A) P(B)

Independence is a very common assumption, the practical realization of which must be carefully verified. Is it so obvious that getting a coin toss does not affect the next flip?

A more important example, and more difficult, is the following one: one notes, in a manufacture, that all the measurements made at 9 am are close to 1 m and that all those made at 4 pm are close to 1.02 m, for a perfectly known reason: the pieces expand with the temperature, higher in this workshop at 4 pm than at 9 am. Are the measurements of two successive pieces independent?
Although these measurements are very close, compared to the dispersion of all the measurements over the day, the answer can be considered positive: the probability of obtaining a given result on the second piece is not influenced by the result obtained on the first piece, but only by the temperature (or the time of day).
But the conclusion is reversed if we use probabilities to translate an incomplete knowledge: if we measure two successive pieces, without knowing neither the time nor the temperature, the measurement of the first piece gives an indication on the value of the second piece. In this example, the conclusions are opposite if the universe is defined at one time of the day or over the entire day.
We can also use the concept of causality, which we were careful not to use until now, to reformulate the problem. If the probabilities of measuring a given value on two successive pieces are not independent, it is because they are influenced by a common cause: the temperature T at a given time. The Reichenbach's principle stipulates that, if T is the sole cause of the dependence of the lengths of two successive pieces L_N and L_{N+1}, then the conditional probabilities obey the definition 2 of independence: $P(L_N \text{ and } L_{N+1}|T)= P(L_N \ |T) \ . \ P(L_{N+1}|T)$. Using causality in this way is obviously equivalent to restricting the universe to pieces measured at a given time (at a given temperature).

Chapter 2) Discrete random variables

I) Introduction and definition

Let us take again the example of the dice. The fact that each side has a numerical value has no importance in the reasoning followed so far. You could have had a drawing on each side, for example an animal, and got P(dog)=P(cat)=1/6. Similarly, you can roll two dice and get P(dog, cat), rather than P(1,5).
On the other hand, the numerical values of the faces are essential if we are interested in the probability of the sum of two dice. An S result is obtained between 2 and 12, with, for example, P(S=2)= 1/36, but P(S=6)=5/36. Unlike most cases considered so far, the different values of S are not equiprobable. The calculation is done by determining the number of elementary events associated with a value of S. These elementary events form a part A of 5 elements in the universe Ω of the ordered results of the roll of two dice (36 elements). This example suggests the following definition of a discrete random variable:

Definition 1 (mathematical) of a discrete random variable (r.v.):

A discrete r.v. is an application that associates a part A of Ω, and the corresponding probability, with a number of R (often, but not always, of N for discrete r.v.).

With this definition, a random variable is neither variable nor random. If we take again our example, the result of a draw of two dice will however give for S a number between 2 and 12, therefore variable, and random since the value obtained is due to chance. Hence a second definition:

Definition 2 (experimental) of a discrete random variable (r.v.):

(The result of) a r.v. is the numerical result of a random experiment.

In the example of the sum of two dice, the possible results were in finite number (the numbers from 2 to 12). We will see cases where a non-zero probability is associated with all values of N. But, of course, if the r.v. X takes its values from N, we have $(\sum_{n=0}^{\infty} P(X = n)\ )$ =1.

II Basic examples

Bernouilli's Law:

A Bernouilli's r.v. can only take two values, 1 and 0, with a respective probability p and (1-p): P(X=1)=p, P(X=0)=1-p.

For example, we can model the heads or tails by associating X=1 to heads and X=0 to tails, with p=1/2. A second example can associate X=1 with a die roll giving a 6, with a probability P=1/6, and X=0 for any other value, with the probability 5/6.

Binomial Law :

It is the sum of N independent Bernouilli r.v.: $Y = \sum_{i=1}^{N} X_i$

Example : Y is the number of tails when drawing a series of 100 tails or heads. It has been shown in exercise that $P(Y = k) = C_N^k\, p^k\, (1 - p)^{N-k}$

Poisson Law

Let be a time interval Δt. We consider a random process generating discrete events **independent of each other** with a probability of generating an event constant over time. Let us consider in Δt a time interval dt small enough that:
P(1 event during dt)=λ dt/ Δ t<<<1. λ is a parameter without unit, which we will see later that it is the expectation number of events during Δt. Then, we can neglect the probability that two or more events occur during dt and the law Y followed by the number of events k during Δt is written as the sum of a large number n=Δt/dt of Bernouilli's laws of probability p=λ dt/Δt. Although n⟶∞ and n⟶0, np= λ is a finite number not zero.

This is a binomial law, hence $P(Y = k) = C_n^k\, p^k\, (1 - p)^{n-k}$

but
$p \to 0 \Longrightarrow (1 - p) \simeq exp(-p) \Longrightarrow (1 - p)^{n-k} = exp(-np)exp(kp) \simeq exp(-np)$

Indeed k is a finite integer and $kp \to 0$, unlike np= λ. Moreover:

$$C_n^k\, p^k = \frac{n(n-1)\dots(n-k+1)}{k!}\, p^k = \frac{(np)^k(1-1/n)\dots(1-(k-1)/n)}{k!} \cong \frac{\lambda^k}{k!}$$

Hence the final result:

$$P(Y=k) = \frac{exp(-\lambda)\lambda^k}{k!}$$

III Moments of a discrete r.v.

Expectation or true mean:

The expectation E(X) of the r.v. X, which can take the values x_i, is defined as:

$$E(X) \triangleq < X > \triangleq \sum_i P(X = x_i)\ \ x_i$$

Variance:

The variance V(X) is defined by

$$V(X) \triangleq E((X - E(X))^2) = \sum_i P(X = x_i)\ \ (x_i - E(X))\ ^2$$

The last equality is not obvious. Indeed, one uses $P(X = x_i)$, whereas the definition leads to use $P(Y = (x_i{-}E(X))^2\ )$. But, we find in the sum on i the different probabilities corresponding to a value of Y.

In developing this last equality, we show (to lighten writing, we note, if there is no ambiguity, $P(x_i)$ for $P(X = x_i)$) :

$$V(X) = \sum_i P(x_i)\, x_i\ ^2 - 2\, E(X)\left(\sum_i P(x_i) x_i\right) + E(X)^2 \sum_i P(x_i) = E(X^2) - E(X)^2$$

Standard deviation:

The standard deviation σX is defined as: $\sigma_X = \sqrt{V(X)}$. Very often the variance will be noted using the standard deviation: $V(X) = \sigma_X^2$.

Centered moments centered of order n:

They are defined by: $\mu_n \triangleq E((X - E(X))^n)$

So $\sigma_X^2 = \mu_2$.

IV) Joint and marginal laws, sum of r.v.., product, (in)dependence

IV.1) Expectation of a sum

Let us return to the example of the sum of two dice and look for E(S)=E(N1+N2), where N1 denotes the r.v. which can take the values 1 to 6, with P(n)=1/6, n∈ {1,...,6}.
Going back to the definition, we have:

$< S > = \sum_{s_i=2}^{12} P(s_i) \ \ s_i$

But P(s_i) is a relatively complicated law, whereas <N1>, or <N2>, is calculated immediately because P(n) has only one possible value: <N1>=<N2>=3.5. We will show that <S> is simply given by the sum of <N1> and <N2>. We need to use the notion of joint probability:

Definition :

Let be two r.v. X and Y, taking the values x_i and y_i respectively. The joint probability P_{ij} is defined by :

P_{ij} = P(X= x_i and Y=y_j)

Reminder: in general, $P_{ij} \neq$ P(X= x_i). P(Y=y_j), unless X and Y are independent.

Theorem : E(X+Y)=E(X)+E(Y)

This theorem is valid whether X and Y are independent or not.

Demonstration :

$$E(X+Y) = \sum_i \sum_j P_{ij}. \ (x_i + y_j) = \sum_i x_i \sum_j P_{ij} + \sum_j y_j \sum_i P_{ij}$$

But (total probability formula):

$$P(X = x_i) = \sum_j P\,(X = x_i \ \text{ et } Y = y_j) = \sum_j P_{ij}$$

In this context, P(X=x_i) is called the marginal law of X.
Hence:

$$E(X+Y) = \sum_i x_i \ P(X = x_i) + \sum_j y_j \, P(Y = y_j) = E(X) + E(Y)$$

IV.2) Expectation of a product

$$E(XY) = \sum_i \sum_j P_{ij}. \ (x_i y_j)$$

If X and Y are independent, P_{ij}= P(X= x_i). P(Y=y_j)

Then, only in this case:

$$E(XY) = \sum_i \sum_j P(x_i)x_i\, P(y_j)y_j = \sum_i P(x_i)x_i \sum_j P(y_j)y_j = E(X)E(Y)$$

IV.3) Variance of a sum .

V(X+Y)=E((X+Y)2)-(E(X)+E(Y))2=E(X2)+E(Y2)+2E(XY)-E(X)2-2E(X)E(Y)=V(X)+V(Y)+2(E(XY)-E(X)E(Y))= V(X)+V(Y)+2 COV(X,Y)

where the last equality defines the covariance between X and Y :

COV(X,Y)≜ E(XY)-E(X)E(Y) =E[(X-E(X))(Y-E(Y))]

If X and Y are independent, COV(X,Y)=0. The reverse is not always true.
Two cases are particularly remarkable:

1) X and Y independent : V(X+Y)=V(X)+V(Y)

2) Y=X: V(2X) = 4 V(X)

More generally, let be c a real constant, V(cX)=c^2 V(X)

This last formula is better understood at the level of standard deviations:
$\sigma_{c\text{⍰}}$= c $\sigma_{\text{⍰}}$. The standard deviation represents the dispersion of the variable and c represents a change in scale.

V) Esperance (or mean) and variance of Bernouilli, binomial and Poisson laws

Bernouilli's Law:

A Bernouilli r.v. X of parameter p has a mean:

<X>=p.1 +(1-p).0= p.

and a variance :

$$V(X) \;=\; p \;\; (1-p)^2 + (1-p)(0-p)^2 = p - p^2$$

This can also be established by noting that X^2=X since X= 0 or 1. Hence :

$$V(X) = E(X^2) - E(X)^2 = p - p^2$$

Binomial law:

It is the sum of n independent Bernouilli's r.v of parameter p. Hence, using the formulas on the variance and mean of a sum:

<X>=n p

V(X)=n (p-p^2)

N.B: variances can only be added because the laws are independent.

Poisson law:

It is a binomial law with n⟶ ∞, $p \to 0$, np= λ finite, hence

$$<X>=n\ p=\lambda$$

$$V(X) = n\ (p - p^2) = \lambda$$

We can also reason in the following way:

Mean

on dt : E (X)=P(1).1= λ dt/Δt

Indeed, P(2 or more) is negligible

on Δt, we add n=Δt/dt laws of means λ dt/Δt, which gives a sum of means equal to λ. λ is therefore the mean of the Poisson's law, i.e. the mean number of events over the interval Δt.

Variance

on dt : $E(X^2)$-$E(X)^2$=E(X)- $E(X)^2$= $\lambda\frac{dt}{\Delta t} - \left(\lambda\frac{dt}{\Delta t}\right)^2 \simeq \lambda\frac{dt}{\Delta t}$

We used the fact that X can only be 0 or 1, hence X^2=X, and the fact that λ dt/Δt<<<1, which allows $\left(\lambda\frac{dt}{\Delta t}\right)^2$ to be neglected

on Δt : the n Bernouilli's laws are independent (the independence of events is the fundamental hypothesis of Poisson's law), from which we can add variances as well as means, which gives V(X)=λ.

A Poisson r.v. has a variance equal to its mean

N.B This result only makes sense because the possible values are numbers without units. In the next chapter, devoted to continuous r.v., the accessible values will most often be physical quantities with units, and the variance will be homogeneous to the square of the mean.

VI Weak law of large numbers

Position of the problem

Can we give a precise content to the intuitive idea: the frequency of realization of a given event is close to its probability if we repeat the experiment a large number of times? For example, if you flip a coin 10 times, the probability of finding 4 to 6 heads, or a "frequency" between 0.4 and 0.6, is 65.6%. If 100 flips are made, the probability of finding 40 to 60 heads is 96.5%: the experimental frequency approaches the theoretical probability of 0.5. It may be noted, however, that while frequency fluctuations decrease as the number of flips increases, fluctuations in the number of flips increase. For example, for 10 flips P(5 heads)=24.6%, while for 100 flips, P(50 heads)=8.0%.
Note also that the frequency of the heads can be calculated using the arithmetic average X of the N flips (X=1 for heads, 0 for tails) : $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Warning: do not confuse the arithmetic average $\bar{X}$ with the true mean, or expectation, E(X)=<X>. For a coin toss, the true mean is a parameter of the r.v., with value of exactly 1/2.

The arithmetic average is, however, a new r.v., for example with values in {0 0,1......,1 } for N=10. Let's calculate its mean and its variance:

E($\bar{X}$)=E(X) (Obvious demonstration, using the expectation value of a sum)

$$V(\bar{X}) = \frac{1}{N^2}\sum_{i=1}^{N} V(X) = \frac{V(X)}{N}$$

where we used the fact that the flips are independent.

We will now prove an inequality that links the variance of a r.v. to the probability of the difference between an achievement of this r.v. and its mean

Inequality of Bienaymé-Tchebytchev

Let X be a r.v. of expectation E(X) and variance V(x) and let be $x_i$the result of a realization of X. We have:

For all real e>0, P($|x_i - E(X)| > e) \leq \frac{V(X)}{e^2}$

For example, P($|x_i - E(X)| > 2\sigma_X) \leq 1/4$

For many laws, much lower limits can be proven.

Demonstration:

Let be a real e >0. By definition, $V(X) = \sum_i P(X = x_i) \ (x_i - E(X))^{2}$

Among the possible $x_i$values for X are those for which $|x_i - E(X)| > e$ is verified. Let A be this set.

$$V(X) \geq \sum_{x_i \in A} P(x_i)\,(x_i - E(X))^{2} \geq e^2 \left(\sum_{x_i \in A} P(x_i) \right)$$

However, $\sum_{x_i \in A} P(x_i) = P(|x_i - E(X)| > e)$

Hence $V(X) \geq e^2[P(|x_i - E(X)| > e)]$

Combining this inequality with the fact that the arithmetic mean has a variance proportional to the inverse of the number of experiments results in the weak law of large numbers:

<u>Theorem (weak law of large numbers)</u>: Let be N independent r.v. with the same law of probability P(X) and of expectation E(X); then:

whatever e>0, P$(|\bar{X} - E(X)| > e) \to 0 \ \ for\ N \to \infty$

Indeed $\bar{X}$ has for mean E(X), for variance V(X)/N and :

P$(|\bar{X} - E(X)| > e) \leq \frac{V(X)}{Ne^2}$

Let's take the example of 100 flips:

P$(|\bar{X} - 0{,}5)| > 0{,}1) \leq \frac{V(X)}{100\ .0{,}1^2} = 0{,}25$

We saw that P$(|\bar{X} - 0{,}5)| > 0{,}1) = 0{,}035$. This result is well below the maximum indicated by the inequality of Bienaymé-Tchebytchev.

Despite its lack of "performance", this inequality nevertheless has the great interest of proving that frequency tends towards probability for large N.

Chapter 3) Continuous random variables and normal distribution

I) Continuous random variables

Let us repeat the measurement of a quantity, for example the period of a pendulum. We'll get a measurement chart.

Example : T(s) : 15.21 15.23 15.29 15.14

and we will often try to model the random part of the measurement by a continuous r.v. , with a non-zero probability over an interval. The cumulative density function F(x) formalizes this intuitive idea:

Definition of the cumulative density function F(x):

$$\mathrm{F(x)} \triangleq \mathrm{P(X < x)}, \qquad \mathrm{F}(-\infty) = 0, \qquad \mathrm{F}(\infty) = 1$$

Although more important for a continuous r.v. , this definition is also valid for a. discrete r.v.

We deduce the probability for a continuous r.v to have a result over an interval:

$$\mathrm{P}(x_1 < \mathrm{X} < x_2) = \ \mathrm{F}(x_2) - \ \mathrm{F}(x_1)$$

But, unlike a discrete r.v., we can't define a probability at a point. Indeed, the probability is spread over an interval (more or less long) and we feel that the probability at a point is zero because a point can be seen as an infinitely small interval.
On the other hand, the derivative can easily be defined at a point of the distribution function. Hence:

Definition :

The probability density function (p.d.f.) p(x) of the continuous r.v. X is defined as:

$$p(x) \triangleq \lim_{\varepsilon \longrightarrow 0} \frac{P\left(x-\frac{\varepsilon}{2}<X<x+\frac{\varepsilon}{2}\right)}{\varepsilon} = \frac{dF(x)}{dx}$$

N.B.1: this notation, introduced here for its simplicity, is ambiguous if several r.v. are considered. For example, if X and Y are two continuous r.v., the two distinct values $p_X(x_0)\ and\ p_Y(x_0)$ should be distinguished at point x_0. This more precise notation will be used if necessary.

N.B.2: Contrary to a probability, the p.d.f generally has a dimension, inverse to that of the r.v. X. In the example that opens this chapter, T is in seconds and its probability density in s^{-1}.

Of course, if you define the r.v. by its probability density, you find the cumulative density function as :

$$F(x) = \int_{-\infty}^{x} p(x'\ )\, dx'$$

Hence, immediately:

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

It is very easy to generalize to continuous variables the results defined or demonstrated on the discrete r.v. by replacing probabilities by probability densities and sums by integrals :

Expectation or true mean:

E(X)=<X>= $\int_{-\infty}^{\infty} x\, p(x)\, dx$

The results on the expectation of a sum and the expectation of a product are identical to the discrete case.

Variance:

V(X)=$\sigma_X^2 = \mathrm{E}((\mathrm{X}-<\mathrm{X}>)^2) = \int_{-\infty}^{\infty} (x-<X>)^2\, p_X(x)\, dx$

The last equality is not obvious: according to the definition of expectation, we should use, with Y=$(\mathrm{X}-<\mathrm{X}>)^2$, the probability elements $p_Y(y)dy$. We will admit that using $p_X(x)dx$ is correct. One can understand this result by noting that the probability for Y to be equal to a value within a width segment dy around y is also the probability for X-<X> to be equal to a value within one of the segments of width dx around $\pm\sqrt{}$y.

Just like the discrete r.v., we show that $\sigma_X^2 = E(X^2) - E(X)^2$.

...and we define the centered moments of order n: $\mu_n = \mathrm{E}((\mathrm{X}-<\mathrm{X}>)^n)$, i.e. V(X)= μ_2

Covariance and joint laws:

The definition of covariance is identical to that of discrete variables:

Cov(X,Y)=E(XY)-E(X)E(Y)=$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, xy\, p_{XY}(x,y)\ -$ E(X)E(Y)

where $p_{XY}$is a joint probability density : $p_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$

where F(x,y)=P(X<x and Y<y).

As for the discrete variables, we can calculate from the joint probability density the marginal laws $p_X(x)$ and $p_Y(y)$ with

$$p_X(x) = \int_{-\infty}^{\infty} dy\ p_{XY}(x,y), \quad p_Y(y) = \int_{-\infty}^{\infty} dx\ p_{XY}(x,y)$$

As for the discrete r.v., we have:

V(X+Y)= V(X)+V(Y)+2 COV (X,Y)

The demonstration is identical to the discrete case.

Probability density of a sum of independent r.v.:

Let Z =X+Y be the sum of two continuous independent r.v.. We propose to find a relationship linking p_Z to p_X and p_Y.

Let's set the value of X to x_0. We have $p_z(z|x_0) = p_Y(y = z - x_0)$

Leading, by considering all possible values of x and using independence, i.e. p(X=x et Y=y)=p_X(x). p_Y(y), to:

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\,dx \triangleq p_X(z) * p_Y(z)$$

where * means "convolution product". You have seen or will see that the convolution product of two functions f and g at point x is a function s(x) defined by :

$$f(x) * g(x) \triangleq s(x) = \int_{-\infty}^{\infty} f(t)g(x-t)\,dt$$

II) Normal (or Gaussian) distribution

This law, which is extremely common in practice, has for p.d.f.:

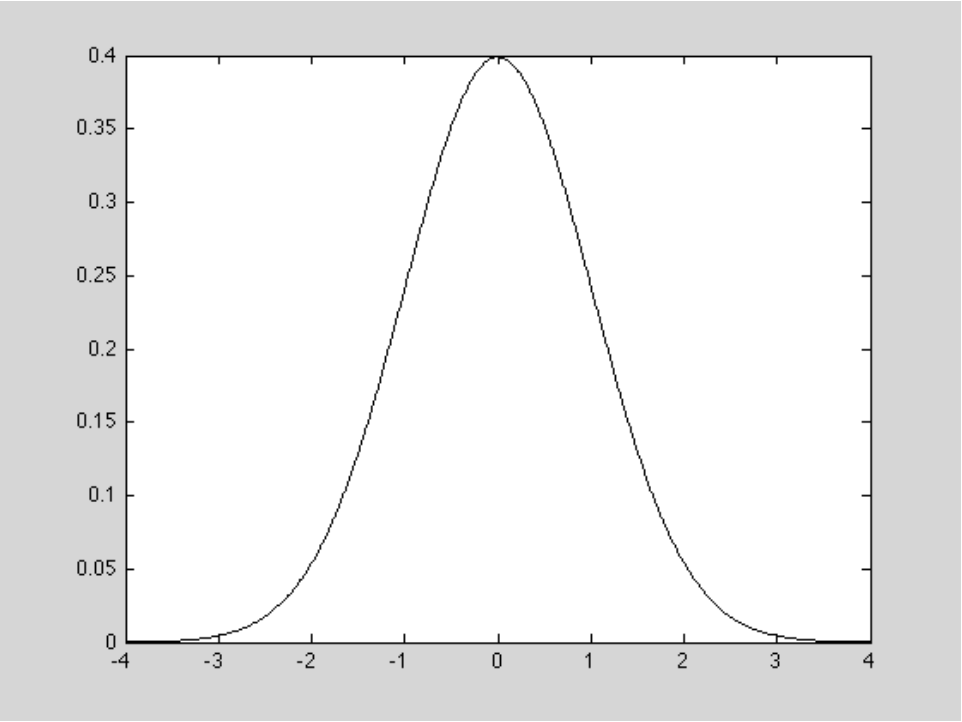
$p(x) = LG(m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\frac{(x-m)^2}{2\sigma^2}$, σ positive and m real

It is easy to show: E(X)=m, V(X)= σ^2, $\mu_{n>2}$=0

Some values of the cumulative density function are remarkable:

F(m-1,96σ)=0,025, F(m+1,96σ)=0,975

The normal distribution LG(0.1) of mean 0 and standard deviation 1, shown below, is called the reduced centered normal distribution.



Sum of two normal distributions

Let be X~LG(m,σ) and Y~LG(m',σ') two r.v. following independent (here ~ means "follows") normal laws, then one shows:

$$X + Y \sim LG(m + m' \ , \sqrt{\sigma^2 + \sigma'^{\ 2}})$$

Moments of a normal law

It is possible to show that all the moments $\mu_n$are null if n>2.

Independence of two normal laws

Two normal r.v. are independent if and only if their covariance is zero. Reminder: for any law, zero covariance is a necessary condition for independence, but is in general not sufficient.

III Central-limit theorem or strong law of large numbers

Let X_1,..., X_i, ..., X_n n independent random variables, of respective expectation m_i and with the same variance σ^2, but not necessarily with the same probability distribution. We have:

$$\begin{matrix} lim \\ n \longrightarrow \infty \end{matrix} \left(\frac{\sum_{i=1}^{n}(X_i - m_i)}{\sqrt{n\sigma^2}} \right) \sim LG(0,1)$$

In fact, the "same variance" condition is not exactly necessary. It is sufficient that the variances have the same order of magnitude.

This theorem, which we will admit, has multiple applications. Let us quote some of them:

- a Poisson's distribution is defined as the sum of a large number of Bernouilli variables. If the mean of Poisson's distribution is greater than 20, one can consider that the Poisson's distribution tends towards a normal distribution of variance equal to the mean.

- consider an election survey conducted on 1000 people. If each person has the probability p to vote for a candidate and if p is not too small, the proportion of respondents voting for that candidate is a r.v. of mean 1000 p and variance 1000 p (1-p).

- in a measurement process of good quality, said under control, all important causes of error have been eliminated. The residual uncertainty is due to a large number of independent causes, of various origins and of comparable weight. The measurement error is then expected to be a. Gaussian r.v..

Chapter 4) Notions of Estimation, χ2 and Student laws, confidence intervals

1) Estimators: definitions and general properties

Throughout this chapter we will consider a simple and repeated measurement process. The value to be measured is a quantity θ, a single scalar in this introduction to estimation. In the measurement modeling, it is therefore a number, having a determined value. Despite repeated measurements, the exact value of this quantity will remain unknown for the experimenter, but the aim of the estimation is, on the one hand, to give the closest possible value, and on the other hand to specify as much as possible the uncertainty range within we are (almost) sure to find θ. Suppose we do N measurements d_n of θ, which can be written as:

$$d_n = \theta + \ eral_n, n = 1, \dots, N$$

The random error $eral_n$ obeys a Gaussian distribution with zero mean, which means on the one hand that the measurement process is under control (see above), on the other hand that the measurement process is without systematic error.

Indeed, the systematic error is by definition the part of the error that is found in all measurements, therefore the mean of the error. The measurements d_n have therefore a mean θ and we will assume that they all have the same (often unknown) variance σ^2 . We will further assume that all measurements are independent. This is an important assumption, not always verified in practice.

An estimator of θ, noted σ^2 is constructed from the measurements, and possibly from the information known before the measurements, called a priori:

$\hat{\theta} = f(d_1, \ldots., d_N, , info\ a\ priori)$.

<u>Example, the arithmetic average</u>:

$$\hat{\theta} = \bar{d} = \frac{1}{N}\sum_{n=1}^{N} d_n$$

This example shows us that an estimator is a random variable, just like the measures from which it is derived. However, Chapter 2 has shown that the variance of the arithmetic mean is σ^2/N. $\bar{d}$ is thus of expectation θ, just like the measurements d$_n$, but fluctuates less around θ, since with a standard deviation divided by √ N .

<u>Convergent (asymptotic unbiased) estimator</u>:

An estimator is said to be asymptotic unbiased if:

$$lim_{N\to\infty}\hat{\theta} = \theta$$

<u>Unbiased estimator:</u>

An estimator is unbiased if:

$$E(\hat{\theta}) = \theta$$

Any reasonable estimator is asymptotic unbiased: if we have an infinite number of measures, we completely know the law of probability and therefore the true value. For example, it has been shown that the arithmetic mean tends towards the true mean for a very large number of measures (weak law of large numbers).
On the other hand, there are good estimators that are biased. Indeed, an unbiased estimator has a variance greater than or equal to a limit σ_0^2, known as the Cramer-Rao limit. This limit is given by a somewhat barbaric formula:

$$\sigma_0^2 = \frac{1}{E\left\{\left[\frac{\partial}{\partial\theta} ln(p(\hat{\theta} and\ \theta))\right]^2\right\}}$$

An unbiased estimator of variance $\sigma_0^2$is said to be efficient or minimum variance unbiased and is, of course, the best unbiased estimator. On the other hand, there are sometimes biased estimators, which have therefore a mean different from the true value (this difference is called bias), whose variance is much lower than the Cramer-Rao limit. They may then be "better" than the efficient estimator, where "better" is defined in a sense that will not be specified in this short intro.

2) Estimators of the mean

Two are in common use:

- the arithmetic mean $\bar{d} = \frac{1}{N}\sum_{n=1}^{N} d_n$

One immediately demonstrates, if the error is Gaussian, that $\bar{d}$ follows a Gaussian law, of variance σ^2/N and mean θ.

- the median. The measurements are ordered from the smallest (d_1) to the largest (d_N). The median is then defined as $d_{(N+1)/2}$ if N is odd, $(d_{N/2} + d_{(N+1)/2})/2$ if N is even.

The median is much less sensitive than the mean to outliers (N.B.: if the series of measurements includes outliers, the error is no longer a Gaussian r.v., unlike in the remaining of the chapter). We can compare the two estimators on an example: measurements of the period of a pendulum made on the chronometer by first year students:

T (seconds) : 10.62 10.38 10.34 10.35 10.40 10.36

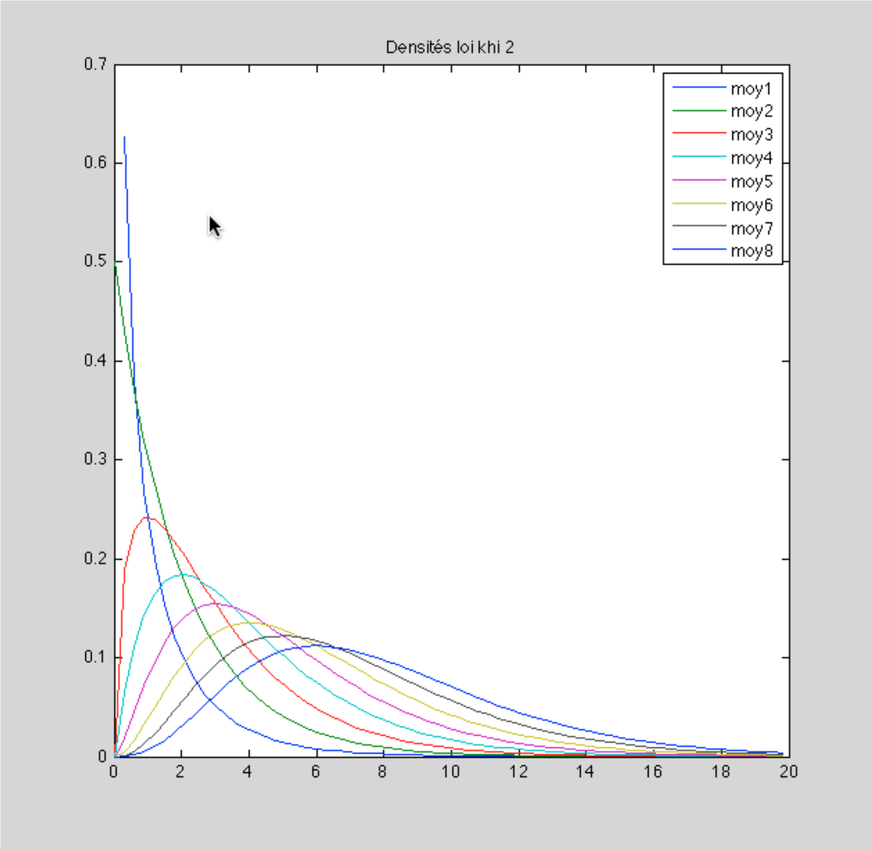
A graphical representation of the data is very useful to conclude.....

3) χ2 distribution: definition

This paragraph provides the tools necessary for the study of variance estimators, the subject of the next paragraph.

Let be X_1,......,X_N N Gaussian r.v. , centered, reduced and independent. The χ2 distribution with N degrees of freedom is, by definition, the law followed by the Y r.v.:

$$Y = X_1^2 + \ldots\ldots + X_N^2$$



One shows that the mean of such a distribution is N and its variance 2N. The figure above shows the distributions for N from 1 to 8. The Gaussian approximation is excellent for N ≥20.

4) Estimators of variance

A) Known mean:

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(d_i - \theta)^2$$

B) Estimated mean:

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2$$

If the multiplicative coefficient were 1/N and not 1/(N-1), $\widehat{\sigma^2}$ would be the arithmetic mean of $\left(d_i - \bar{d}\right)^2$, just as the true variance is the (true) mean of $(d_i - E(d))^2$. We understand the need to use 1/(N-1) when thinking about the case where we have only one measurement. So $\sigma^2$is indeterminate, which seems correct since we have no idea of the dispersion of the measurements, represented by the variance. Using 1/N would give zero variance, which is clearly incorrect. In fact, the $d_i - \bar{d}$ are not independent, unlike the d_i . For example, for N=2 measurements, $d_1 - \bar{d}$ =-$\left(d_2 - \bar{d}\right)$

Theorem: $\frac{\sum_{i=1}^{N}(d_i-\bar{d})^2}{\sigma^2}$follows a χ² distribution with N-1 degrees of freedom, of mean N-1.
$\widehat{\sigma^2}$ thus has a mean σ^2, and one shows that it is the efficient estimator of σ^2.

The demonstration of this theorem for N=3 can be done by noting that:

$$\sum_{i=1}^{3}\left(d_i-\bar{d}\right)^2 = P_1^2 + P_2^2, \text{with:}$$
$$P_1 = (d_1 - d_2)/\sqrt{2}, P_2 = (2\, d_3 - d_1 - d_2)/\sqrt{6}$$

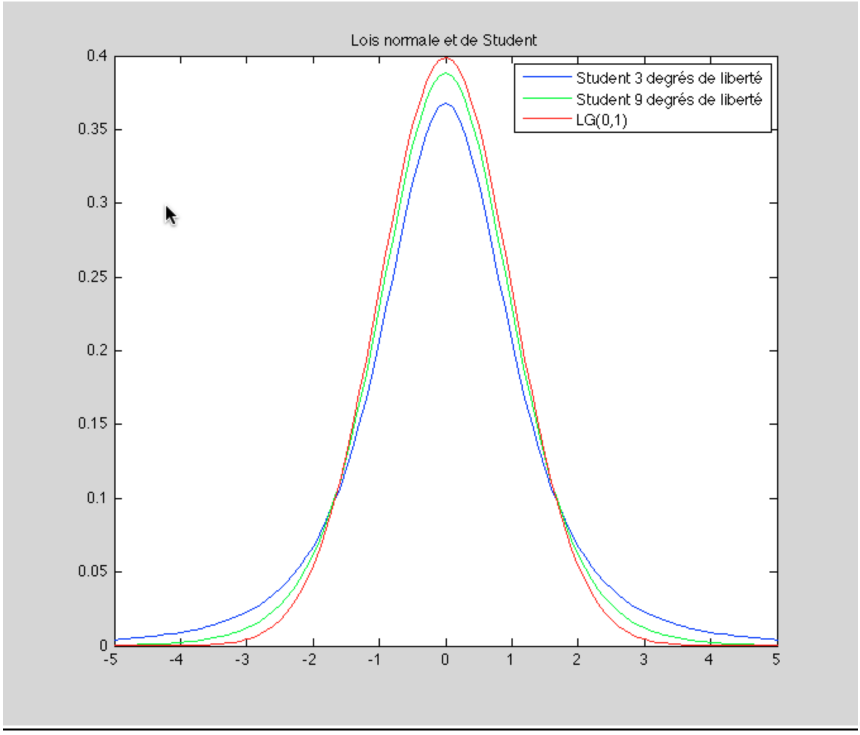
We can easily verify that P1 and P2 are Gaussian (because sum of Gaussian), of variance σ^2, and of zero covariance, which is equivalent to independence for Gaussian distributions.
It will be admitted that this demonstration can be generalized to all N ≥2 .

$\widehat{\sigma^2}$ has a mean $\sigma^2$and is therefore unbiased. It is the efficient variance estimator. On the other hand, the estimator of the standard deviation $\hat{\sigma} = \sqrt{\widehat{\sigma^2}}$ is biased, but of course asymptotic unbiased.

5) Student's Law

To estimate the true quantity θ, two estimators are available:
- the average $\bar{d}$, which follows a Gauss distribution, of unknown mean θ and standard deviation σ/√ N.
- the estimated standard deviation on measurements $\sqrt{\widehat{\sigma^2}}$, i.e. an estimated standard deviation on the arithmetic mean $\sqrt{\frac{\widehat{\sigma^2}}{N}}$where (N-1) $\widehat{\sigma^2}/\sigma^2$follows a χ² distribution with to N-1 degrees of freedom.

Their quotient $\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$follows a law, called Student's law, with N-1 degrees of freedom, whose figure on the following page gives the graph for two values of N-1:



[5]

We see that, for N-1=3, the probability that the r.v. is greater than 2 in absolute value is clearly higher than the ≈5% obtained for a Gaussian : for a weak N, a Student's law can be seen as a Gaussian whose standard deviation is poorly known, which makes more probable values, expressed in estimated standard deviations, far from the mean.

6) Confidence intervals

In the previous paragraph, the sentence "$\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$ follows a Student distribution" refers to the modeling of measurements. In this model world, $\bar{d}$ and $\sqrt{\widehat{\sigma^2}}$)are random variables and θ a precise value, even if unknown. This situation does not reflect the reality of the experimenter: for him, the measurements are data with perfectly determined values, having allowed him to construct mean and variance estimators, which have also determined values. On the other hand, he can only hope to determine a probability distribution on θ, whose exact value will remain unknown.

It is tempting, to build this law of probability, to simply consider that this is now θ the random variable in $\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$. The situation is analogous to the example given in Chapter 1 to illustrate Bayes' theorem. The true value here is θ "being sick or not", and the estimator of the mean has replaced the test result....but it is just as dangerous, in principle, to confuse $\mathrm{p}(\bar{d}|\theta)$ and $\mathrm{p}(\theta|\bar{d})$. In order to link these two expressions, let us write Bayes's theorem for these probability densities:

$$\mathrm{p}(\theta|\bar{d}) \propto \mathrm{p}(\bar{d}|\theta)\, p_{prior}(\theta)$$

Compared to chapter 1, equality has been replaced by an operator $\propto$ which means "proportional to" and the denominator, which should be p$(\bar{d})$, has been "forgotten". Indeed, from the point of view of the experimenter, d, and therefore p$(\bar{d})$, are constants, which are taken into account in the form of a proportionality coefficient. If necessary, this proportionality coefficient shall be determined bearing in mind that, for any r.v. X,

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

We therefore have the equality p$(\bar{d}|\theta)$ = p$(\theta|\bar{d})$ only if $p_{prior}(\theta)$= Cste. This is a reasonable assumption for a controlled measurement process, where the measurement error is low and Gaussian. The uncertainty range is then low enough to consider that, within this range, the probability density of θ before measurements is a constant. Of course, if we have an explicit expression of $p_{prior}(\theta)$, we must renounce this assumption and calculate p$(\theta|\bar{d})$ with $p_{prior}(\theta)$

If $p_{prior}(\theta)$= Cste, $\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$follows a Student law with N-1 degrees of freedom, where θ is the random variable. The range around the arithmetic mean where θ has a 95% chance of being found can then be determined: this range, called the confidence interval, is given by $\bar{d} \mp \alpha\frac{\hat{\sigma}}{\sqrt{N}}$, where α depends on N :

N:	3	5	10	20	40
α :	4.3	2.8	2.3	2.1	2.0

Thus, from 40 measurements, Student's law merges with a Gaussian law. For a very small number of measures, however, there is a greater chance of underestimating the standard deviation, which gives a greater chance that the true value deviates from the arithmetic mean of more than two estimated standard deviations.

Warning: Using 2 or α has hardly any consequences as soon as you make at least ten measurements. However, it should not be forgotten that the estimated standard deviation of θ from the arithmetic mean is not the estimated standard deviation of the measures $\hat{\sigma}$, but $\frac{\hat{\sigma}}{\sqrt{N}}$! That's the point of repeating the measurements! ...and we will not forget either that this division by $\sqrt{N}$ is intimately linked to the assumption of independence of the measurements. If this assumption is not fully verified, the size of the confidence interval may be underestimated.