Statistics
Licence SVP – Paris VI

Leticia F. Cugliandolo
Laboratoire de Physique Théorique et Hautes Energies de Jussieu
Laboratoire de Physique Théorique de l'Ecole Normale Supérieure
Membre de l'Institut Universitaire de France
leticia@lpt.ens.fr

2004

Abstract

This document includes: (1) A detailed schedule of the lectures, TDs, TP and exams. (2) A **draft** of the Lecture notes on Statistics. It presents a summary of the material that will be described during the semester. They are certainly incomplete and may contain errors. Hopefully, we shall improve it with the help of the students. (3) The TDs. (4) The TP. (5) A list of suggested subjects for the project.

1 Schedule Statistiques et Info

Volume horaire : 30h + 30h
Responsable : Leticia Cugliandolo

Equipe :
Philippe Andrey (info) andrey@jouy.inra.fr
Leticia F. Cugliandolo (stats) leticia@lpt.ens.fr
Ramiro Godoy-Diana (info+stats) rgdlod@ladyc.jussieu.fr
Daniela Despan (stats) daniela.despan@chimie.univ-nantes.fr

Semaine	Mardi 8:30	Mardi 10:30	Mercredi (M&AP)	Dates de suivi
27 Sept	CS1		TPI1	
4 Oct	CI1	TDS1		
11 Oct	CS2	TDI1	TPI2	
18 Oct	CS3	CI2		
25 Oct	CI3	TDS2		
1-8 Nov	TDI2	TDS3	TPI3	
15 Nov	CS4		TPI4	
22 Nov	CS5			
29 Nov		TDS4		
6 Dec		TDS5	TPS1	
13 Dec				
3 Janv				
10 Janv				
24 Janv	Exam			

La distribution horaire de Stats est : 10h (cours) + 10h (TDs) + 4h (TP) + 6h (Suivi) par étudiant.

La distribution horaire d'Info est : 6h (cours) + 4h (TDs) + 16h (TP) + 4h (Suivi) par étudiant.

Les projets seront présentés par la moitié des étudiants en forme de rapport écrit (10 pages maxi) et pars l'autre moitié en exposé orale (15'). Une liste des sujets des suivis apparait à la fin de ce document. D'autres sujets sont aussi possibles.

Les étudiants devront preparer et répondre un petit nombre de questions par écrit à rendre avant de commencer la séance du TP de Stats. Les réponses vont faire partie de la note de TP (5/20).

Il y aura un contrôle continu (15') la semaine du 15 Novembre. La note fait parti de la note finale.

Le planning, notes du cours, et ennoncés des TDs, TPs et Projets de Stats (2003, 2004) peuvent être consultés sur la page personelle de leticia en www.lpt.ens.fr

2 Introduction

We are used to reading the word "statistics" in many contexts without knowing exactly what it means. For instance, the newspapers write about the "statistics of the performance of a football team" without ever explaining what they really mean by this expression.

The word statistics has its origin in the Latin "status" as well as the word "state", suggesting that governments used statistical concepts since long ago.

In a few words, statistics is the theory that allows one to make sense out of a list of numbers. In the case of the analysis of the performance of a football team, this list might contain the total number of matches played, the number of matches won, the number of goals scored, and so on since the beginning of the season. And the question one would like to answer is: is the team doing well? The analysis of these numbers with statistical methods does indeed give an answer to this question.

In more technical terms, statistics is the branch of mathematics that deals with the analysis of data. One identifies two sub-branches that we shall study in this course:

1. *Descriptive statistics.* The goal is to obtain useful information from a series of *raw data* that is typically too large to deal with directly. For instance, present experiments in molecular biology present the "difficulty" of yielding too much raw data that need processing before becoming useful. Descriptive statistics is a set of tools, or mathematical manipulations of the raw data, that convert them into a few numbers and plots that are easy to understand.

2. *Inferential statistics.* The goal is to obtain useful information about a very large *population* being able to test only a *sample*, that is to say, a small portion of the total population. The typical example of the application of inferential statistics are election polls. Clearly, one is not able to ask every citizen for which candidate he or she is going to vote. However, one can ask a sample of the population and infer from the result which is going to be the global one. One of the main difficulties in this case is related to the choice of the sample. Clearly, if the sample is taken exclusively from a posh neighbourhood the result will be different from the one obtained from a poor one. Similarly, if only aged people are consulted, the result might be different from the one obtained using a sample of young people.

In the rest of this course we shall study these two branches of Statistics. We shall start by studying *Descriptive statistics* and a set of definitions that will be useful henceforth. Next, we shall review concepts of *Probability Theory* that allow one to identify generic behaviour of *random events* (and will be of use in other courses too, such as, e.g. Thermodynamics). Finally, we shall enter the field of *Inferential Statistics.* We shall also spend some time explaining several statistical tools for *experimental design and data analysis.* In the Appendices we shall present the proofs of some simple equations that appear in the main text.

The mathematical theory of Statistics developed mostly in England, during the beginning of the XXth century. It was also the time when it became clear that a purely deterministic description of physics was not feasible (with the development of Statistical Mechanics and Quantum Mechanics). [Interestingly enough, many developments in the Theory of Statistics were the consequence of reseaerch in Agriculture (the analysis of the effect of fertilizers!) by Sir R. Fisher.]

3 Basic notions in descriptive statistics

We want to draw some conclusions out of a (long) list of numbers, $x_1, \ldots, x_n$, that represent the outcome of some measurement (e.g. the number of goals scored by a football team in each game during the season). With this aim, let us define a number of quantities, that are simple operations on the list of numbers and will tell us global features of the raw data.

Average. It represents the mean value of the list:

$$\mu_x \equiv \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

The word average originates in the French *avérie* (itself from the Italian *averia* that originates in the Arabic) and the reason is that shippers used to contribute to their guild a certain amount of money to cover for eventual wreckages.

Median. It is the "middle" value in the list. To define it one first orders the list in such a way that $y_1 \leq y_2 \leq \ldots y_n$ where y_i for each i is equal to an element in the original list $x_1, x_2, \ldots, x_n$. The median is

$$x_{median} \equiv \begin{cases} y_{\frac{n+1}{2}} & \text{if} \quad n \text{ is odd} , \\ \frac{1}{2}\left(y_{\frac{n}{2}} + y_{\frac{n}{2}+1}\right) & \text{if} \quad n \text{ is even} , \end{cases} \tag{2}$$

The word median originates in the Latin *medius*.

It is important to note that, in general, $\mu_x \neq x_{median}$. If the list of numbers is such that there are some very large (small) values of x_i one has $\mu_x > x_{median}$ ($\mu_x < x_{median}$). Let us give an example. The list x_i represents the times needed for 5 mice to solve a maze and the outcomes are $x_1 = 9'$, $x_2 = 10'$, $x_3 = 12'$, $x_4 = 9.5'$, $x_5 = 24$h. In this case, the average is not very representative of most mice since it will yield a result of the order of a few hours, while the median is of the order of $10'$. This tells us that the median is useful when one wants to get rid of *rare events*, like the silly mouse.

Mode. It is the most frequent value in the list. It is not necessarily unique, they can be several modes (see, e.g., the case in Fig. **??**).

Range. It is the interval of variation of the data:

$$r \equiv x_{max} - x_{min} \tag{3}$$

with $x_{max} \equiv \max\{x_1, \ldots, x_n\}$ and $x_{min} \equiv \min\{x_1, \ldots, x_n\}$.

The range is obviously very sensitive to extreme values.

Variance. It measures the spreading of the data, i.e. the variance is small if all data are concentrated around the mean and it is large otherwise. It is defined as

$$\sigma_x^2 \equiv \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)^2 = \langle\, x^2 \,\rangle - \mu_x^2 \tag{4}$$

(see the Appendix A for a proof of the second identity). Hereafter $\langle \cdots \rangle$ indicates the sum $n^{-1} \sum_{i=1}^{n} \cdots$. Note that $(x_i - \mu_x)^2$ is a measure of the individual distance from the average and that from its very definition σ_x^2 is positive.

Standard deviation. It is just given by

$$\sigma_x \equiv \sqrt{\sigma_x^2} \; . \tag{5}$$

Relative variability. It is a comparison between the standard deviation and the average:

$$rv_x \equiv \frac{\sigma_x}{\mu_x} \; . \tag{6}$$

Note that this ratio is an adimensional quantity and as such it is relevant to comparing the spread to the average of the data.

The variance is well-suited to distinguish rare events. For instance, a professor will try to prepare an exam such that its evaluation yield rather spread notes (certainly bounded between 0 and 20 in France) so as to distinguish bright students. Instead, a constructor of explosives will prefer to have a small variance in the delay time for explosion to avoid accidents.

The standard deviation is measured in the same units as the average and it is hence directly comparable to it. This is why one uses the adimensional quantity rv_x to give a concrete idea on the spreading of data.

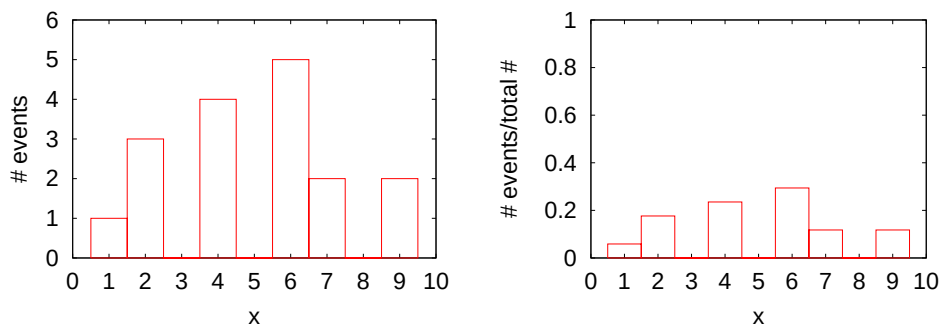

Figure 1: Bar diagram (left) and frequency histogram (right) for the set of data in the Table below.

It is often useful to characterise the ensemble of data with the averages of more complicated functions. One can then define the average of a generic function $f(x_i)$:

Generic average and momenta

$$\langle\, f(x) \,\rangle = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \; . \tag{7}$$

The angular brackets usually denote generic averages of this type. A particular case is given by $f(x) = x^k$:

$$\mu_x^{(k)} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^k \; . \tag{8}$$

Note that setting $k = 1$ one recovers the definition of the average and with $k = 2$ one obtains one of the two terms appearing in the second expression in (4) for the variance. The knowledge of all moments allows one to reconstruct the form of the frequency diagram. The knowledge of some of them gives partial knowledege about the form of this diagram.

Bar diagrams. These are graphs in which one writes, on the x-axis, all the values of x taken on the list, and on the y axis the number of times each value appears on the list. See Fig. 1-left.

A very important parameter[1] used to construct these graphs is the *width of the bins*. Indeed, one needs to choose a distance between points on the x-axis that will be considered as being different. Thus, data spanning an interval $x_{max} - x_{min}$ are classified in n_{bin} groups, each of length $\Delta x = (x_{max} - x_{min})/n_{bin}$. Clearly, the size of Δx should not be comparable to $x_{max} - x_{min}$ (to avoid putting all data in the same bin), and it should not be very small either (to avoid having at most a single event in each bin).

The location and width of the bar in the diagram is arbitrary. One can choose to center the bar at the center of the bin, at the left-end of the bin, at the right-end of the bin... one can also choose its width. The most natural choices, and the ones we use here, are to center the bar on the center of the bin, and to use a width such that neighbouring bars touch.

In many cases, the width of the bar diagram at about half height is given by twice the standard deviation.

Frequency histograms. In this case, the y-axis is normalized by the total number of data n. Thus, the y-axis is bounded between 0 and 1. See Fig. 1-right. The *form* of the diagram is not changed by this transformation.

From the bar and frequency diagrams one can check whether the data are or are not evenly distributed about the average. In *symmetric* cases they are, in asymmetric cases instead (*skewed distribution*) the data pile up on one side of the average.

In Fig. 1 we plot the data in the following Table

6	2	4	2	9	6	4
2	4	6	1	7	7	6
6	9	4				

in the form of a bar diagram (left) and a frequency histogram (right). The construction rule is very simple. First we chose the bin-width to be one, the most natural choice. Next, we count how many times each number appears on the table and draw a bar of this height on the diagram. For example, 1 appears only once and hence its associated bar has height equal to one. Instead, 6 appears five times and its bar has then height equal to five. Note that the sum of the heights of the bars in the bar diagram should be equal to the number of data on the table, while the sum of the heights in the frequency plot equals one.

Correlation. The correlation quantifies the similarity between two (or more) data sets. There are many possible definitions of correlation and each may be better adapted to some problem. We here discuss some of them.

Let us take two sets of data of the same length, $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$. The index i that labels the elements in each set is such that the determins some natural

[1] From the Greek, means 'almost measurement'.

order. For example, it may represent a discrete time and the elements x_i and y_i may be two different observables measured at these time steps.

The correlation between the two sets X and Y is then defined as

$$C_{xy} = \frac{1}{\sigma_x \sigma_y} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \tag{9}$$

with the averages μ_x and μ_y, and the standard deviations, σ_x and σ_y defined above. C_{xy} takes values between -1 (complete decorrelation) and 1 (complete correlation).

Another example can be drawn from physics: a magnetic system modelled by what is called the Ising model. Within this model the magnetic system is represented by spins (see the Quantum Mechanics course!) on a lattice. Each spin is a little vector that can only point up and down and is hence represented by a variable s that takes values ± 1 (for up and down, respectively). Each spin is labelled by an index i that represents the site it occupies on the lattice. If the lattice is cubic and d dimensional and one then has $n = L^d$ spins in the system, with L the linear length of the d dimensional cube. the set we want to study is then $X = \{s_1, \ldots, s_n\}$. The magnetized state is represented by a configuration such that the magnetization density, $m = n^{-1} \sum_{i=1}^{n} s_i$, takes a non-zero value while the paramagnetic state is such that $m = 0$. Note that the magnetization density m is just the average of the set X, $m = \mu_x$. Now, one can imagine that the configuration evolves in time, meaning that each spin changes its configuration as time evolves. And one may be interested in comparing the configurations of the total system at different times. This is achieved by computing the correlation (9) where the X is the set of n spin values at one time, say t_1, and the Y is the set of n spin values at another time, say t_2. In other words, $X = \{s_1(t_1), s_2(t_1), \ldots, s_n(t_1)\}$, $Y = \{s_1(t_2), s_2(t_2), \ldots, s_n(t_2)\}$ and

$$C_{xy} = \frac{1}{\sigma(t_1)\sigma(t_2)} \sum_{i=1}^{n} (s_i(t_1) - m(t_1))(s_i(t_2) - m(t_2)) \ . \tag{10}$$

<u>*Covariance.*</u> The numerator in (9) is called the covariance of x and y.

3.1 Properties

The average of a set of data is simplify 'translated' by a uniform translation of all the elements in the set. More precisely, if μ_x is the average of a set $\{x_1, \ldots, x_n\}$, then $\mu_x + c$ is the average of the set $\{x_1 + c, \ldots, x_n + c\}$ with c a constant.

The standard deviation of a set of data is not modified by a uniform translation of all the elements in the set. More precisely, if σ_x is the standard deviation of the set $\{x_1, \ldots, x_n\}$, then σ_x is also the standard deviation of the set $\{x_1 + c, \ldots, x_n + c\}$ with c a constant.

The correlation between two sets of data remains unchanged a uniform translation of the two sets by the same constant.

Another type of data transformation that we shall encounter is the multiplication by a constant. The main properties that we shall use are:

The average of a set of data that has been multiplied by a constant is simply the constant times the original average. The mean-square deviation of a set of data that has

been multiplied by a constant is equal to the constant times the orpiginal mean-square displacement. Finally, the correlation between two sets of data that have been multiplied by the same constant remains unchanged.

(See Appendix A for the proofs of these results.)

3.2 An example

A set of data X is represented by the bar diagram shown in Fig. 2.

The average is given by $\mu_x = \frac{1}{34}(1 \times 1 + 2 \times 3 + 3 \times 4 + 4 \times 6 + 5 \times 4 + 6 \times 3 + 7 \times 2 + 8 \times 1 + 9 \times 0 + 10 \times 0 + 11 \times 0 + 12 \times 0 + 13 \times 4 + 14 \times 0 + 15 \times 6)$ that is equal to $245/34 = 7.21$ and is shown with a green vertical line.

The median is given by $x_{median} \equiv \frac{1}{2}(y_{17} + y_{18})$ for the ordered ensemble Y. This yields $x_{median} = (4+5)/2 = 4.5$. Note that the peaks at 13 and 15 push the average to higher values than the median.

There are two modes, $x = 4$ and $x = 15$.

The mean-square-displacement or variance is $\sigma_x^2 \equiv \langle x^2 \rangle - \mu_x^2$. The second term is simply $7.5^2 = 56.25$. The first term reads: $\frac{1}{34}(1^2 \times 1 + 2^2 \times 3 + 3^2 \times 4 + 4^2 \times 6 + 5^2 \times 4 + 6^2 \times 3 + 7^2 \times 2 + 8^2 \times 1 + 9^2 \times 0 + 10^2 \times 0 + 11^2 \times 0 + 12^2 \times 0 + 13^2 \times 4 + 14^2 \times 0 + 15^2 \times 6)$ that yields $\frac{1}{34}(1 + 12 + 36 + 96 + 100 + 108 + 98 + 64 + 556 + 1350) = 2421/34 \approx 71.2$. Thus, $\sigma = \sqrt{71.2 - 51.98} = 4.38$. This value is shown in the figure with two horizontal blue arrows with this modulus traced at half the height of the plot and starting at the location of the mean.

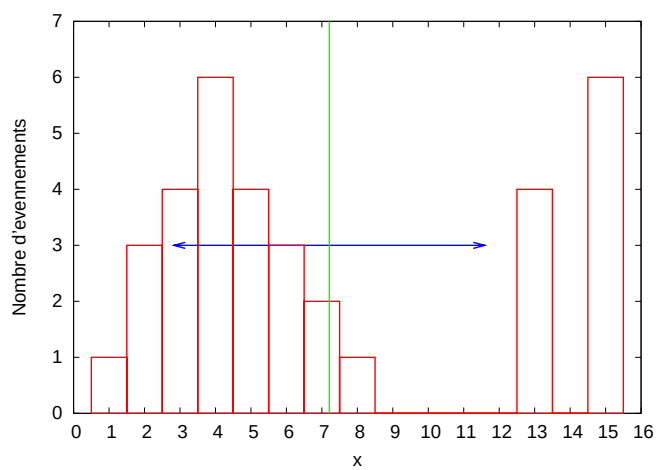

Figure 2: Bar diagram representing a set of data. The average is indicated with a vertical line. The median is at $x = 4.5$. The horizontal line shows 2σ at half length.

4 Probability theory

We shall develop the theory of probabilities using a relevant example: tossing a coin. Not surprisingly, this is a very well-chosen example since the theory of probabilities has largely developed to try to win in gambling.

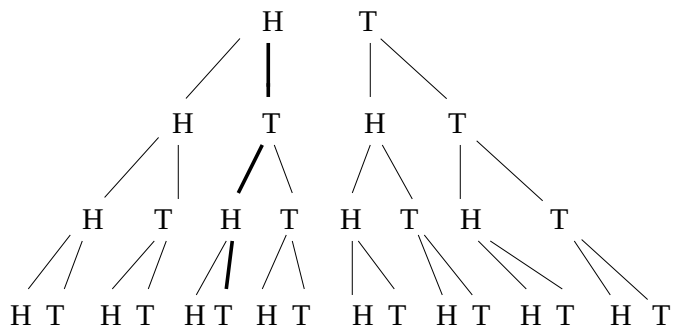

Figure 3: A tree diagram representing all possible results of tossing a coin n times. In bold, one particular outcome.

4.1 Tossing a coin

If one tosses a coin it is practically impossible to predict if it will fall heads or tails (try to solve Newton's equations for the coin and air...) However, repeating the experiment *many times* one can predict general patterns of the head-tail sequence. This is a general feature of *random events*: one is unable to predict the outcome of a single experiment but one can predict some general features of a sequence of them. Note that the list with the results of tossing the coin many times constitutes an example of raw data as the ones that we manipulated in the previous Section.

The sequence of tosses can be represented by a "two-leave" tree, as the one in Fig. 3. The first level in the tree represents the two possible outcomes of the first toss. Each node branches in the two possible outcomes of the second toss; thus, the second level of the tree has four elements. The constructions follows in this way until the n-th level that corresponds to the n-th toss and has 2^n elements. The outcome of a single experiment, say HTHT is a *walk* on the tree as the one shown in bold face on Fig. 3.

4.1.1 Frequency and probability

If we toss an unbiased coin we expect all paths to be equally likely. Why? This is a very important *assumption*! How can we justify it?

Imagine that we toss the same coin a very large number of times, $n \to \infty$, we count the number of H and Ts that we obtain, and we construct the *probability or large N limit frequency*:

$$P(H) \equiv \lim_{n\to\infty} f_n(\#\mathrm{H}) = \lim_{n\to\infty} \frac{\#\mathrm{H}s}{n} \,. \tag{11}$$

$$P(T) \equiv \lim_{n\to\infty} f_n(\#\mathrm{T}) = \lim_{n\to\infty} \left(\frac{n - \#\mathrm{H}s}{n}\right) = 1 - P(\mathrm{H}) \,. \tag{12}$$

This definition quantifies the subjective idea 'what do we expect to get after tossing a coin?' If the coin is not biased, if n is sufficiently large, we should get one half of Hs and one half of Ts. Thus, for $n \to \infty$:

$$p \equiv P(H) = \frac{1}{2} = 1 - P(T) \qquad q \equiv P(T) = \frac{1}{2} \,. \tag{13}$$

If, instead, the coin is biased, H (or T) will appear more frequently. Then $p > \frac{1}{2}$ and $q < \frac{1}{2}$ or *viceversa*.

4.2 "Ideal experimental' definition of probability

Based on the toin coss example it is natural to define in general the *probability of an event* as the value taken by the frequency in the limit $n \to \infty$:

$$P(\mathcal{E}) \equiv \lim_{n\to\infty} f_n(\mathcal{E}) \tag{14}$$

where $\mathcal{E}$ denotes the event we are interested in, e.g. getting 3Hs after tossing $n \to \infty$ coins.

Let us mention that one is used to listen to weather reports in which people talk about the "probability of having rain tomorrow". In this case, the definition of probability as a limiting procedure is much less clear and becomes much more subjective. We stick to objective cases here where the above definition can be safely applied.

But, it is actually not necessary to go to such extreme example to realize that sometimes the definition above may be ambiguous. This problem has been discussed in the past. Kolmogorov gave ra rigurous deifinition of probability that we shall not discuss here. We shall simply say that it is based on using the list of properties that we discuss below as a definition of probability.

4.3 Properties

We now list a series of properties of probabilities. Some of them are obvious and follow simply from the frequency-based definition given above. Some other are not and constitute new definitions.

Semi-positive definite. It is clear that the probability of an event is a quantity that can take only positive or zero values:

$$P(\mathcal{E}) \geq 0 \ . \tag{15}$$

Bound. It is also clear from its definition that the probability of an event is bounded by one:

$$P(\mathcal{E}) \leq 1 \ . \tag{16}$$

Normalization. Note that the sum of the number of occurrences of each event is equal to the total number of possible event. Thus, the sum of the frequencies over the events is normalised to one, and the sum of probabilities is also normalised to one:

$$\sum_{\mathcal{E}} P(\mathcal{E}) = \sum_{\mathcal{E}} \lim_{n\to\infty} f_n(\mathcal{E}) = 1 \tag{17}$$

(assuming that one can exchange sum and limit).

Probability of the complementary event. If the probability of occurrence of an event, $\mathcal{E}$, is $P(\mathcal{E})$, the probability of non-occurrence of the same event (its *complementary event*, $\overline{\mathcal{E}}$) is

$$P(\overline{\mathcal{E}}) = 1 - P(\mathcal{E}) \ . \tag{18}$$

This is just a consequence of the normalization of the probabilities.

For example, when we introduced the coin toss problem before we said that when tossing a normal coin one finds H one half of the times and T the other half of the times. In toher words, we *assumed* that the probability of getting H in a toss is 1/2 and the probability of finding the complemetary event T is $1 - 1/2 = 1/2$ too.

Addition principle. For any pair of events $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$P(\mathcal{E}_1 \bigvee \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \bigwedge \mathcal{E}_2) \ . \tag{19}$$

The symbol $\bigvee$ represents the logical (not exclusive) "or" while the symbol $\bigwedge$ represents the logical "and". The meaning of the logical or is that the event $\mathcal{E}_1 \bigvee \mathcal{E}_2$ is true whenever $\mathcal{E}_1$ is true, $\mathcal{E}_2$ is true or both $\mathcal{E}_1$ and $\mathcal{E}_2$ are true. The meaning of the logical and is that the event $\mathcal{E}_1 \bigwedge \mathcal{E}_2$ is true only if $\mathcal{E}_1$ and $\mathcal{E}_2$ are true.

Let us illustrate this property with an example. Imagine that one is playing with two dices, a red and a blue one (to make them distinguishable). What is the probability for getting 1 with the red one or 2 with blue one?

This problem is represented mathematically as follows. The event $\mathcal{E}_1$ is getting 1 with the red dice. The event $\mathcal{E}_2$ is getting 2 with the blue dice. There are $6\times6 = 36$ possible outcomes of throwing the two dices, i.e. getting the pairs $(1,1); (1,2); \ldots; (2,1); (2,2); \ldots; (6,6)$ where the first element in each pair is the result of the red dice and the second one is the result of the blue dice. The event $\mathcal{E}_1 \bigvee \mathcal{E}_2$ is true in the cases $(1,1)$;$(1,2)$;$(1,3)$;$(1,4)$;$(1,5)$;$(1,6)$; $(2,2)$;$(3,2)$;$(4,2)$;$(5,2)$;$(6,2)$, so there are 11 true realizations and, since the dices are not biased, after $n \to \infty$ repetitions of the experiment "throwing the two dices" we expact to find 11/36 successful tries. Thus, the probability of this event is $P(\mathcal{E}_1 \bigvee \mathcal{E}_2) = 11/36$.

Now, the number of positive outcomes for the event $\mathcal{E}_1$ is just 6. The same applies to the event $\mathcal{E}_2$. The sum of these two numbers is 12 and it does not coincide with result above. The reason why is that the event $(1,2)$, that solves the problem, is counted twice. Once in the number of positive outcomes for $\mathcal{E}_1$ and once in the number of positive outcomes for $\mathcal{E}_2$. One needs to correct this double counting and this is why there is the last term in (19). Taking this into account, one finds that the right-hand-side of this equation predicts $P(\mathcal{E}_1 \bigvee \mathcal{E}_2) = (6 + 6 - 1)/36 = 11/36$ which is the correct result.

Particular case: mutually exclusive events. When the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive, $\mathcal{E}_1 \bigwedge \mathcal{E}_2 = 0$ and the formula (19) simplifies to

$$P(\mathcal{E}_1 \bigvee \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) \ . \tag{20}$$

Working again with dices, mutually exclusive events are 'getting 1 and 2 with one dice after one throw. There are plenty of other examples of this sort.

Joint probability. One calls in this way the probability of simultaneous occurrence of a number of events,

$$P(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n) \ . \tag{21}$$

The comma means here the same as the symbol $\bigwedge$ in (19).

Independence principle. If two events can occur independently of the realisation of the other, the probability of the simultaneous occurrence of them is simply the product of the individual probabilities:

$$P(\mathcal{E}_1, \mathcal{E}_2) = P(\mathcal{E}_1)P(\mathcal{E}_2) \ . \tag{22}$$

The generalisation to n independent events is straightforward:

$$P(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n) = P(\mathcal{E}_1)P(\mathcal{E}_2)\ldots P(\mathcal{E}_n) \ . \tag{23}$$

As an example one can imagine playing with dice and cards and asking about events that are associated to the dice and to the cards, independently.

Conditional probability. The probability of the occurrence of an event event $\mathcal{E}_1$ conditioned to the occurrence of another event $\mathcal{E}_2$ is

$$P(\mathcal{E}_1|\mathcal{E}_2) = \frac{P(\mathcal{E}_1, \mathcal{E}_2)}{P(\mathcal{E}_2)} \ . \tag{24}$$

Clearly, if $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent events, $P(\mathcal{E}_1|\mathcal{E}_2) = P(\mathcal{E}_1)$

In the TDs we shall see plenty of examples that illustrate these properties.

4.4 Discrete and continuous random variables

A careful experimentalist performs a measurement many times in identical conditions, then calculates the frequency of each result and from them, in the limit of a very large number of measurements (that in real life are never infinite!) *estimates* the probability of each result. We call *random variable* the result of the experiment.

For instance, in the coin toss problem the random variable, let us call it x, takes only two possible values, H and T. We can associate these two exclusive results with the numbers 0 and 1 and then call the random variable x *bimodal*. Bimodal random variables are *discrete* since they can only take values on a discrete set $(0, 1$ in this case).

Other random variables can take values on continuous sets, as the real numbers, and are hence called *continuous* random variables. Imagine that we look at the absolute value of the velocity v of a particle moving within a gas. In principle, this value can be any real number (most probably bounded...) and hence v is a continuous random variable.

When dealing with continuous random variables we have to be more precise about what we mean by the probability of an event. In this case, the quantity that is well defined is the probability for the event to take values within a given interval. In the case of the velocity of a particle in the gas, we ask what is the probability that the particle has an absolute velocity comprised between v_1 and $v_1 + \Delta v$. In the limit in which Δv is infinitesimally small ($\Delta v \to dv$) this allows us to define the *probability density* as the probability for the random variable to take values within the infinitesimal interval of length dv starting at v_1:

$$p(v)dv \tag{25}$$

We use lower cases to denote probability densities and upper cases to indicate probabilities.

4.5 Characteristics of random variables

In the same way as in Sect. 3 we characterised a list of numbers using a series of definitions (average, median, histograms, etc.), here we define similar quantities to characterise the behaviour of random variables.

Expected value or mean. The expected values of a discrete and a continuous random variable are

$$E(x) = \sum_x xP(x) \ , \qquad E(x) = \int dx \ p(x)x \ , \tag{26}$$

respectively.

Variance. The variance of a discrete and a continuous random variable are defined as

$$\sigma_x^2 \equiv \sum_x (x - E(x))^2 P(x) \qquad \sigma_x^2 \equiv \int dx \ p(x)(x - E(x))^2 \ , \tag{27}$$

respectively.

Momenta The k-th momentum of a random variable is defined as

$$\mu^{(k)} \equiv \sum_x x^k P(x) \qquad \mu^{(k)} \equiv \int dx \ x^k \, p(x) \ . \tag{28}$$

Indeed, knowing all momenta one can reconstruct the functional form of P or p.

Probability distribution. The set of probabilities associated to a random variable is called a *probability distribution.* The probability distributions may be represented by tables but it is far more convenient to represent them with formulæor diagrams that generalize the frequency plot in Fig. 1-right. These are just plots with $P(x)$ in the y-axis and x in the x-axis.

Cumulative probability. With the probability distribution of a discrete random variable one constructs a cumulative probability that is simply the probability for the random variable to take a value that is larger than some chosen one:

$$F(x_1) = \sum_{x \geq x_1} P(x) \ . \tag{29}$$

Similarly, for a continuous random variable,

$$F(x_1) = \int_{x_1}^{\infty} dx \ p(x) \ . \tag{30}$$

5 Probability distributions

In this Section we present some probability distributions that appear recurrently in the study of physical, biological, and sociological systems among others.

5.1 The binomial

A simple question one can pose now is: how many Hs do we expect after a finite number n of tosses?

5.1.1 By "definition"

Let us focus on the case $p = q = \frac{1}{2}$. The first way to compute the searched probability $P_n(h)$ is by using the frequency-based definition:

$$P_n(h) = \lim_{\overline{n}\to\infty} \frac{\#\text{ results with } h \text{ Hs in a the seq. of length } n}{\text{total } \#\text{ results}} \tag{31}$$

where $\overline{n}$ is the number of repetitions of the experiment 'constructing a sequence – made of Hs and Ts and of length n – with the result of throwing an unbiased coin n times.

For an unbiased coin each sequence should appear and equal number of times (taking care of the order of appearences of Hs and Ts).

The total number of sequences one can construct is 2^n since at each position in the sequence one has the choice between two results (H or T). The total number of sequences, among the 2^n possible ones, that have h Hs is a number that shall compute below.

5.1.2 Using the properties of the probability

We have already stressed that it is more precise to use the properties of the probabity (more than its frequency-based definition). Let us now follor this route to derive a formula for $P_n(h)$. We shall not assume $p = q = 1/2$ but derive the formula in the general case:

$$P(H) = p\ , \qquad P(T) = q = 1 - p\ . \tag{32}$$

The dientity $p + q = 1$ follows from the normalization of the probability.

First, let us use the property of independence between throws of the coin to write:

$$P(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n) = P(\mathcal{E}_1)P(\mathcal{E}_2)\ldots P(\mathcal{E}_n)\ , \tag{33}$$

where $\mathcal{E}_1$ is, for example, getting H in the first throw, $\mathcal{E}_2$ is, for example, getting T in the second throw, and so on. Thus, the probability of *any* sequence with h Hs and $n - h$ Ts is:

$$p^h q^{n-h}\ . \tag{34}$$

This means that any two sequences with the same number of Hs and Ts has the same probability, irrespectively of the order in which the Hs and the Ts appear.

What we want to compute is the probability to get h Hs and $n - h$ Ts. This means that *all* sequence with h Hs and $n - h$ Ts are successful results. In order to take care of this we shall use the property of additivity of the probabilities:

$$P(\mathcal{E}_1 \bigvee \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \bigwedge \mathcal{E}_2)\ . \tag{35}$$

Now, the event $\mathcal{E}_1$ and $\mathcal{E}_2$ represent getting sequence number one (that is a successful outome of the experiment) and sequence number two (that is another successful outome of the experiment). Thus is two say $\mathcal{E}_1$ and $\mathcal{E}_2$ are two different sequences with h Hs and $n - h$ Ts that differ only in the order in which these appear. These two events are mutually exclusive (meaning that I either get one or the other but I cannot get the two of them simultaneously). Thus, $P(\mathcal{E}_1 \bigwedge \mathcal{E}_2) = 0$.

Finally, since all sequences with h Hs and $n-h$ Ts are successful outcomes of the experiment, and each fo this sequences is equiprobable, the probability I am looking for is

$$P_n(h) = \sum p^h(1-p)^{n-h} = \text{number of seq. of length } n \text{ with } h \text{ Hs } p^h(1-p)^{n-h} \,. \quad (36)$$

And we are again confronted to computing the number of sequences of length n with h Hs. Note that when $p = q = 1/2$ we recover the result in (31).

5.1.3 The counting argument

In this section we shall give an argument to compute the missing number in (**??**).

If $n = 1$ (first level in the tree) we have two possible outcomes, H and T. The number of sequences of length $n = 1$ with 0 Hs is one and the number of sequences of length $n = 1$ with 1 H is also one. These are only two possible results for a sequence of length $n = 1$.

Let us now take $n = 2$. In this case we have four possible results. These are TT, TH, HT, HH and, since we are only interested in the number of Hs obtained and not in the order in which they appear the second and third cases coincide (look at the second level in the tree in Fig. 3). Thus, we have

event we want	outcome that leads to the event	#such outcomes
0 H	TT	1
1 H	HT & TH	2
2 H	HH	1

Let us now go on and check what happens if $n = 3$.

event we want	outcome that leads to the event	#such outcomes
0 H	TTT	1
1 H	HTT, THT & TTH	3
2 H	HHT, THH & HTH	3
3 H	HHH	1

So, what happens if n becomes really large? Can we generalize this counting argument? Yes, indeed.

To get 0Hs we need to get all Ts and there is only one path that leads to this result for all values of n.

Now, to get 1H and $(n-1)$Ts we have several possibilities. We can get one H in the first toss and then all Ts. We can get one T, then one H and subsequently $(n-2)$ Ts, etc. Generalizing, we see that we have n choices of getting one H and $(n-1)$ Ts, that correspond to the location on the sequence (level of the tree) where the H arrives.

We are now in a position to generalize what we have just done for one H to the appearance of two Hs. First, we have n choices to locate the first H. Next, we are left with $(n-1)$ choices to locate the second H in the sequence. Thus, the number of possible

ways to get two *distinguishable* Hs is $n(n-1)$. However, all the Hs are *identical* meaning that nothing distinguishes the first H from the second one. This means that we have overestimated the number of ways in which we can get two Hs by, indeed, their *permutations*. It is clear that the number of ways in which one can order two elements is two [say they were distinguishable, then we can create the two sequences (H_1,H_2) and (H_2,H_1)]. Thus, the actual number we are looking for is $n(n-1)/2$ where the 2 in the denominator comes from the fact that the two Hs are not distinguishable.

Let us try the case with 3Hs before proposing a general formula. In this case, it is clear that we have n choices for the first H, $n-1$ choices for the second one and $n-2$ choices for the third one. Now, the number of correct outcomes is not just simply $n(n-1)(n-2)$ since, as before, we cannot distinguish between which H comes first, second and third. This means that we need to divide by the number of possible permutations of three elements. And this is equal to 3×2 [the six cases being (H_1,H_2,H_3), (H_1,H_3,H_2), (H_2,H_1,H_3), (H_2,H_3,H_1), (H_3,H_1,H_2), (H_3,H_2,H_1)]. Finally, we have $n(n-1)(n-2)/(3\times 2)$.

We now note a number of features in the particular cases we discussed above that allow us to derive the general equation we are looking for. First, let us look at the numerators. They are a product of factors n, $n-1$, etc. and the product stops at $n-(h-1)$ where h is the number of Hs. As for the denominator, it is also a product that starts with h and ends with 1. If we accept this argument, we find that the number of events leading to h appearances of H is

$$\frac{n(n-1)(n-2)\ldots(n-h+1)}{h(h-1)(h-2)\ldots 1} = \frac{n!}{h!(n-h)!} \equiv \left(\begin{array}{c} n \\ h \end{array} \right)$$

where we have defined the *factorial*

$$n! \equiv n \times (n-1) \times (n-2) \ldots 2 \times 1 \ , \tag{37}$$

and the last symbol represents the *combinatorial number*.

We finally have

$$P_n(h) = \left(\begin{array}{c} n \\ h \end{array} \right) p^h (1-p)^{n-h} \ .$$

5.1.4 Discussion

The form of the bimodal distribution and its evolution with n is shown in Figs. 4 ($p = q = 1/2$) and 8($p \neq q$). Note that when n incrases the figures look more and more peaked about their maximum that occurs at $\approx n/2$. This statement can be proven rigorously. The proof is slightly difficult and uses some mathematical results that the students may not know yet. Still, for the sake of completeness, we discuss it in Appendix B.

With a simple calculation one proves

$$E(h) = np \ , \qquad \sigma_h^2 = np(1-p) \tag{38}$$

(see Appendix C). Note that $rv_h = \sigma_h/E(h) \propto 1/\sqrt{n}$ and this tends to 0 when $n \to \infty$. This means that the distributions look more and more smooth and peaked when n increases.

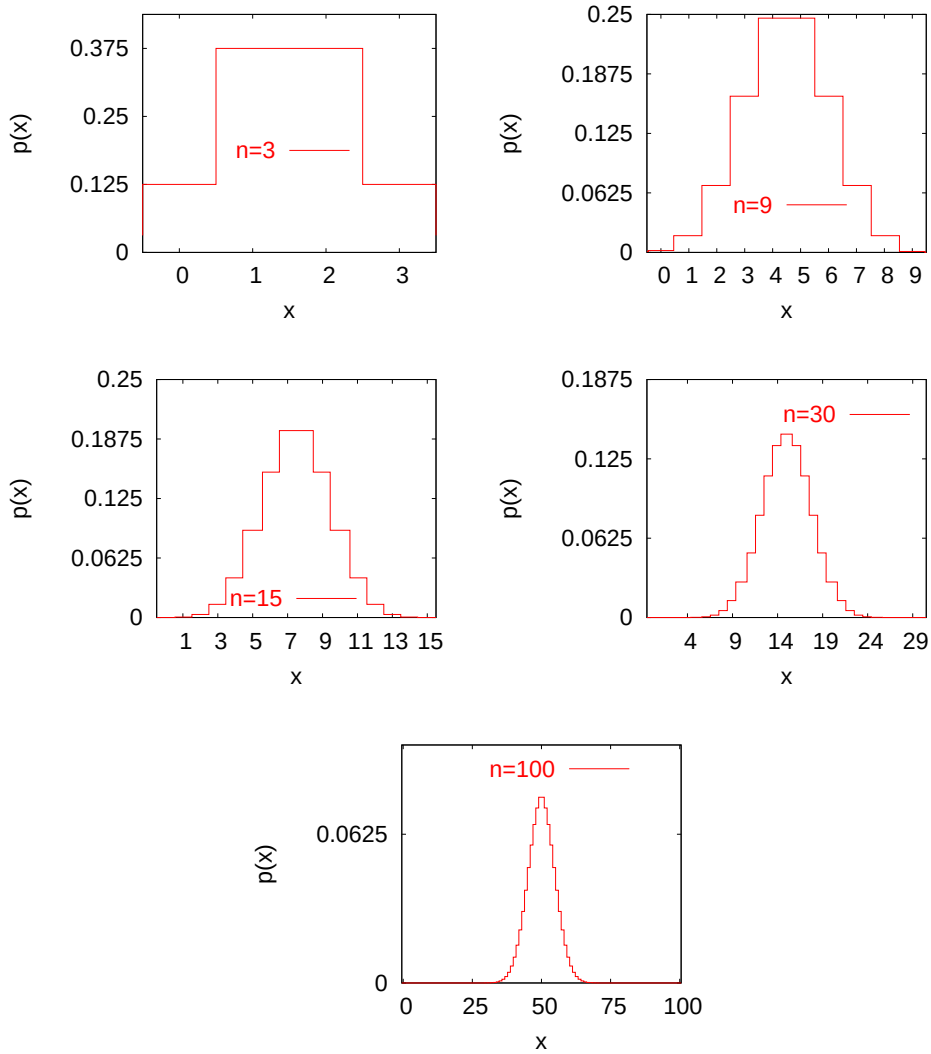

Figure 4: Evolution with n of the binomial distribution function. Note how the form is more and more continuous as n increases. One sees from these plots that the location of the maximum is at $\approx n/2$ while the width of the curve at half height is $\approx 2\sigma_h \approx \sqrt{n}$.

The bimodal distribution is a 'two-parameter' one. n and p are the parameters controlling its form.

To summarize, we have just studied a problem, the tossing of a coin and we have derived an equation that gives us the probability of occurrence of an event (h Hs) in a series of n experiments. This formula is valid, obviously, only for this problem (and others that can be mapped onto it). Even if it is just an example, it is illuminating and it allows us to define the concept of probability.

We can now construct a figure with $P_n(h)$ similarly to the ones we constructed for experimental data in the beginning of this course. In the present case we work with a theoretical curve, determined by Eq. (38), that we shall use to *model* experimental data. In a few words, the modeling procedure goes as follows: given a set of data one wants to characterise, one presents them in a frequency plot and one draws, on top of the experimental frequency curve, the theoretical one. Due to the experimental errors

that we discuss in Sect. 8.1, and that are impossible to avoid, the curves cannot match perfectly well. One then has to determine whether the deviations are significant or not. If they are, the theoretical prediction has to be discarded. If they are not one can keep it. We shall come back to this problem later.

5.1.5 Some examples: the random walk and others

A typical physical realization of the toss coin problem is the random walk problem that we shall study in great detail in Transport Phenomena suring the second semester. Take a one dimensional lattice with spacing a and a drunken walker that can occupy the sites on this lattice. On knows that in each time step the walker moves right one half of the times and it moves left the other half. After n time steps, the walker might have taken h steps to the right and $n-h$ steps to the left. $P_n(h)$ represents the probability of this composed event. This rather simple problem has many application in physics and biology and it is at the basis of the theory of diffusion.

The bimodal disctribution applies to problems that can be cast in the form of a yes or no answer. For example, given a population of n patients one can wonder what is the probability that h among them have asthma if one knows that the probability of each patient having asthma is p. This is given by $P_n(h)$ with parameters n and p.

5.2 The Poisson distribution

The distribution

$$P(x) = \frac{\mu^x}{x!} e^{-\mu} \,, \tag{39}$$

with μ a parameter and x a discrete random variable, taking values $x = 0, 1, 2, \ldots$, is called Poisson probability distribution function (PDF). See Fig. 5.

With a simple calculation one proves that the parameter μ is the expected value of the random variable, $E(x) = \mu$ and $\sigma_x^2 = \mu$.

Poisson PDF can be obtained as the limit of the binomial distribution when n is very large, p is small and np is kept fixed.

Two examples of random variables described by the Poisson distribution are the following.

Take a gas confined to a volume V with average density ρ. If one divides the total volume into small boxes with equal volume v, the local density fluctuations from box to box. The number fo particles in each little volume v is described by a Poisson distribution.

Another example is the one of the local connectivity in the so-called Erdos-Renyi random graph with average connectivity c. The number of links reaching a vertex fluctuate according to a Poisson distribution with parameter $\mu = c$.

5.3 The normal or Gaussian distribution

The formula

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,, \tag{40}$$

where x takes all values on the real axis and μ and σ^2 are two parameters, represents the Gaussian probability density. One easily checks, by direct integration, that the expected value of x, $E(x)$, equals μ and its variance, σ_x^2, equals σ^2. See Fig. 6.

The Gaussian probability density can also be viewed as a limit of the binomial distribution. In this case one keeps p finite and takes n to infinity.

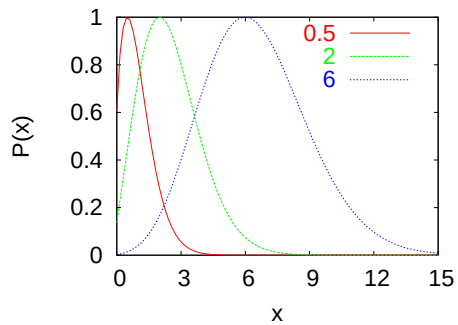

Figure 5: Poisson distribution for three values of μ given in the key.

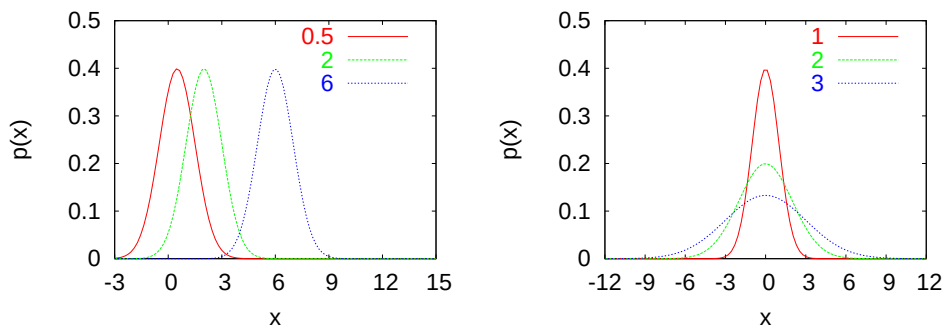

Figure 6: Gaussian probability density. Left: fixed $\sigma^2 = 1$ and three values of μ given in the key. Right: fixed $\mu = 0$ and three values of σ given in the key.

The Gaussian distribution in its *normal form* has zero mean and unit variance. Given a generic Gaussian distribution, the normal form is achieved by defining $y = (x - \mu)/\sigma$ and transforming $p(x)$ into $p(y)$:

$$p(y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}} \,. \tag{41}$$

The cumulative probability of a random variable with a Gaussian distribution are tabulated or can be computed numerically. They are of great use in Inferential Statistics.

5.4 Approximating the binomial

The calculation of the Poisson and Gaussian distributions is much simpler than the calculation of the binomial. Of course, these distributions are not identical.

A rule of thumb tells that the Poissonian approximation to the binomial is rather accurate when

$$n \geq 20 \qquad p \leq 0.05 \,. \tag{42}$$

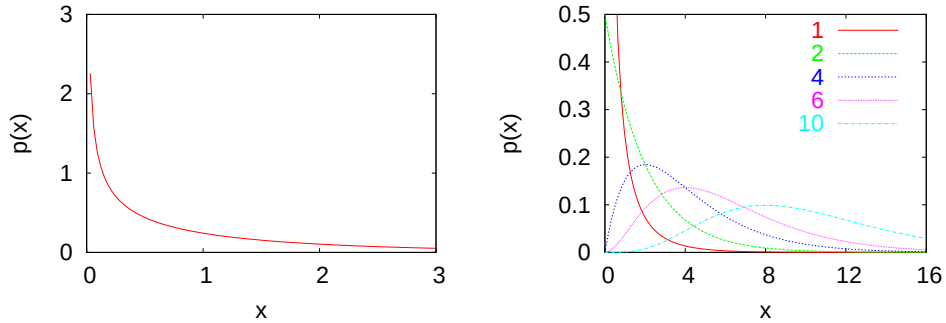

Figure 7: Left: the χ^2 probability density with one degree of freedom. Right: comparison between the χ^2 probability density for n degrees of freedom, using n values given in the key.

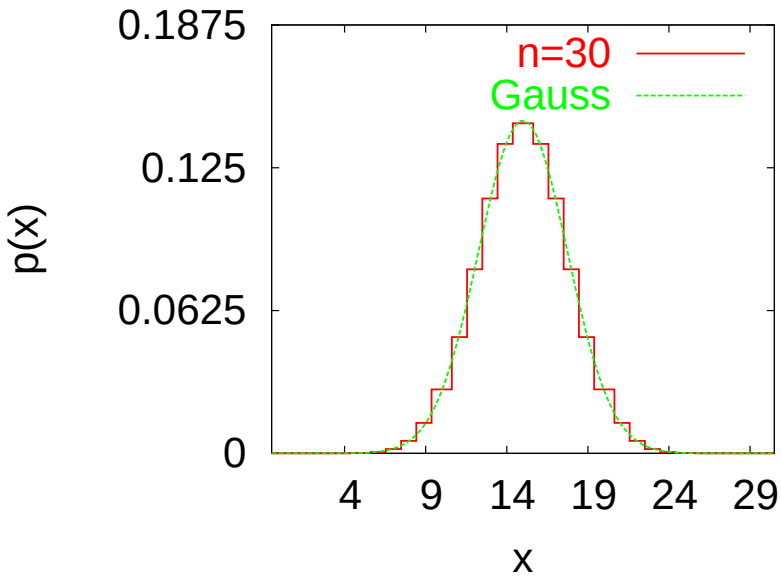

Figure 8: The binomial compared to the Gaussian.

Let us check this statement with an example. Take a binomial distribution with $n = 70$ and $p = 0.02$. The probability of three successes is

$$P(h = 3, n = 70) = \frac{70!}{67!\,3!}\,0.02^3(0.98)^{67} = 0.1151\ . \tag{43}$$

while the Poissonian approximation is

$$P(\mu = np = 1.4, x = 3) = \frac{1.4^3}{3!}\,e^{-1.4} = 0.1128\ , \tag{44}$$

a very good result.

The binomial distribution is discrete while the Gaussian is continuous. To approximate the former by the latter one has to be precise what one really means. Thus, the binomial probability of having h will be approaximated by the Gaussian cumulative probability of

finding the continuous variable x within the interval $[h-0.5, h+0.5]$:

$$P(h,n) \approx \int_{h-0.6}^{h+0.5} dx\, p(x) \ . \tag{45}$$

Let us test this hypothesis with an example. Suppose that 15% of the cars coming out of an assembly plant have some defect. In a delivery of 40 cars what is the probability that exactly five cars hace defects? The actual answer is given by the binomial formula

$$\frac{40!}{35!5!}\, 0.15^5 0.85^{35} = 0.1692 \ . \tag{46}$$

But the calculation of the factorials is quite complicated. What about the Gaussian approximation to this result? The mean and average are of the binomial $\mu = np = 40 \times 0.15 = 6$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.15 \times 0.85} = 2258$. Thus, the Gaussian approximation is

$$\int_{4.5}^{5.5} dx \frac{1}{\sqrt{2\pi}2258}\, e^{-\frac{-(x-6)^2}{2\times 2258^2}} = 0.1583 \tag{47}$$

and, again, this result is quite close to the exact one. Note that in this case, $p = 0.15$ is not small (it is not smaller than 0.05 as in the above example where we used the Poissonian approximation).

5.5 The χ^2 distributions

Suppose that y is a continuous random variable with a Gaussian normal distribution and define $x = y^2$. x is also a continuous random variable, semi-positive definite, and its distribution is called χ^2. Changing variables explicitly (see Appendix E) one finds

$$p(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}}\, e^{-\frac{x}{2}} \ , \tag{48}$$

see Fig. 7. One finds that the mean and variance of the χ^2 distribution are 1 and 2, respectively.

The χ^2 random variable defined above is actually called a chi-square variable with one degree of freedom. One can construct χ^2 variables with n degrees of freedom using

$$x_n = y_1^2 + y_2^2 + \ldots + y_n^2 \tag{49}$$

where y_i, $i = 1, \ldots, n$ are independent random variables each one distributed according to a Gaussian normal form. We shall call χ_n^2 one such random variable. The general chi-square density function is

$$p(x) = \frac{1}{\mathcal{N}}\, x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}} \ , \qquad \mathcal{N} = \int_0^\infty dx\, x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}} \tag{50}$$

The expectation and variance of a χ_n^2 variable are n and $2n$, respectively. In Fig. 7-right we compare the χ^2 probability density for n degrees of freedom using several increasing values of n. We note from the figure that the curve for $n = 10$ resembles a Gaussian distribution. This is a consequence of the *central limit theorem* that we shall study later.

The χ^2 distribution plays a very important role in statistical inference.

5.6 The t distribution

The Student t-distribution is also very important in statistical inference. It was introduced by "Student" a pseudonym that a Guinness employee used to publish papers on his personal research. Given y a random variable with a normal Gaussian distribution, and z a chi-square variable with m degrees of freedom, then

$$x \equiv \frac{y}{\sqrt{z/m}} \tag{51}$$

has the t-probability density. The derivation of this distribution is very hard (we shall not do it here) and the final formula for it is very complicated. The form it takes looks like

$$g(x) = c\left(1 + \frac{x^2}{m}\right)^{-(m+1)/2} \tag{52}$$

The t-distribution is symmetric about zero where it takes its maximum, and it resembles the Gaussian form apart from the fact that its tails are higher.

6 Two central results in probability theory

6.1 Law of large numbers

We have already mentioned that if we toss a coin a large number of times, the number of Hs found will be close to $n/2$. This result is a particular form of the *law of large numbers.*

Take a random variable x and measure it n times, i.e. draw the numbers $x_1, x_2, \ldots, x_n$. Construct the average of these numbers, as done in Sect. 3:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{53}$$

The law of large numbers tell us that the probability for this value to be different from the expected value of x, $E(x)$, tends to zero when $n \to \infty$:

$$P(|x - \mu_x|) > \epsilon) \to_{n\to\infty} 0 \ . \tag{54}$$

This statement can be rigorously proven. We shall not discuss the proof here. It can be found in the literature.

6.2 Central limit theorem

The *central limit theorem* states that in the large n limit the random variable constructed with the sum of independent, identically distributed random variables,

$$y \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \ , \tag{55}$$

has a Gaussian distribution with expectation value equal to the expectation value of the original random variable, $E(x)$, and variance given by σ_x^2/n. In other words, y is

a continuous random variable (even if x might have been a discrete one) distributed according to

$$p(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\, e^{-\frac{(y-E(y))^2}{2\sigma_y^2}} = \sqrt{\frac{n}{2\pi\sigma_x^2}}\, e^{-\frac{n(y-E(x))^2}{2\sigma_x^2}} \ . \tag{56}$$

Note the importance of this theorem for developing experiments. An experiment consists of the measurement of an observable. A single measurement does not make sense since the result found will be subject to many sources of noise. The result of an experiment only has a statistical sense. If one repeats the measurement n times and constructs the average of the results, $x_1, \ldots, x_n$, the central limit theorem ensures that the y will be a normal distributed random variable with expected value $E(x)$ and variance σ_x^2/n. Incresing the number n one then reduces the width of the Gaussian and for sufficiently large n one is sure to approach the actual expected value of the observable x with the average of the data, y.

7 Statistical estimation

In the last part of the course we have discussed problems in which we knew the probabilities of certain random variables and draw conclusions from them. In many cases of practical interest one has a population to characterise with a measurement that is quantified by a random variable. This random variable (discrete or continuous) is distributed according to some probability distribution. Unfortunately, one does not know this probability distribution but needs to *infer* it from partial measurements performed on samples. Statistics tells us with which *confidence* we can estimate the behaviour of the population from that of the sample.

But, many questions arise about the composition of the samples. What size should one choose? In which manner should one choose the members of the sample? What conclusions about the population can be drawn from the ones obtained from the sample? With which degree of confidence can these conclusions be taken?

A number of receipts tell us how to estimate several properties of the unknown probability distribution (e.g. its mean μ and its variance σ^2) using the results of measurements on a sample. Using these receipts one calculates *statistics* which are just functions of the items in the sample that give us clues as to how is the PDF we are looking for. When the statistics are used to estimate the value of an unknown quantity they are call *estimators*. We review them here. In addition, one wants to quantify

$$P(|\text{estimator} - \text{property of population}| < \epsilon) \ . \tag{57}$$

7.1 Properties of estimators

<u>Random character.</u> In general, any estimator is a random variable itself since it depends on the elements taken to form the sample.

One can define many estimators of a given quantity. Among them one is the best, or more adequate for the problem we have at hand. In the following we list a number of properties that a good estimator should have.

Maximum likelihood. The maximum likelihood estimator is the one that has the maximum probability of getting close to the associated property of the PDF of the population.

Consistency. Imagine that one uses a sample of size 10 and gets a certain result for the estimator. Let us now take a sample of size 100, for which we get another result for the estimator. The consistency property means that, as intuitively expected, the result from the larger sample will be closer (in probability) to the actual result of the property of the population. In mathematical terms:

$$P(|\text{estimator}(n=10) - \text{property}| < \epsilon) > P(|\text{estimator}(n=100) - \text{property}| < \epsilon) \ . \quad (58)$$

(Un)bias. Remember that an estimator is a random variable and hence characterised by a PDF that we call $q(\text{estimator})$. An unbiased estimator is such that its expectation value equals the parameter we are trying to estimate:

$$E(\text{estimator}) = \int d\text{estimator}\ q(\text{estimator})\,\text{estimator} = \int dx\ p(x)\,x \ . \quad (59)$$

Variance. Being a random variable, the estimator has an associated variance. One prefers estimators with a small variance to estimators with a large variance.

$$\sigma^2_{\text{estimator}} = \int d\text{estimator}\ q(\text{estimator})\,\text{estimator}^2 - (\int d\text{estimator}\ q(\text{estimator})\,\text{estimator})^2 \ . \quad (60)$$

Note that sometimes one uses estimators that do not satisfy all these properties simultaneously. In practical applications one needs to choose a sample and investigate it (i.e. ask every member of a sample who is he or she going to vote for). Some definitions may lead to very precise results but be impractical (i.e. choose a sample with half the size of the population in an election poll). Thus, one needs to make some compromises and choose the best estimator for the concrete problem at hand. Sometimes it is better to choose a biased estimator with a smaller variance than an unbiased estimator with a large variance (just because by choosing the sample we maybe unlucky and get a very bad result).

In the equations above we assumed that the estimator is a continuous random variable. Similar forms can be written for discrete ones.

7.2 Some useful estimators.

Estimator of the mean It is clear that the best estimator of the mean is

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \ . \quad (61)$$

This is just the average of the results obtained examining the sample.

Estimator of the variance The best estimator of the variance is

$$\overline{\sigma}^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ . \quad (62)$$

The denominator $n-1$ is not so obvious! but it turns out that one needs to use $n-1$ instead of the more natural choice n to get an unbiased estimator of the variance.

7.3 Confidence interval of the mean

Once we have chosen an estimator of the mean of a population, we have to realise that this itself is a random variable with a certain PDF. How is this PDF characterised, i.e. which are its mean and variance, and which form does this PDF take?

Let us calculate the estimator of the mean using samples of size n. The estimator of the mean is a sum over n random variables, divided by n. Thus, the expectation value of the random variable "estimator of mean" is

$$E(\overline{x}) = \frac{1}{n}\sum_{i=1} E(x) = E(x) \ , \tag{63}$$

i.e. it is equal to the expectation value of the population. NB $E(\overline{x}) \equiv \int d\overline{x} q(\overline{x})\overline{x}$ while $E(x) = \int dx p(x) x$.

What about the variance of the random variable $\overline{x}$? We have shown that the variance of a random variable z built with the sum of two random variables, x and y, is equal to the sum of the variances, $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$, $z = x + y$. Now, in our case, each element in the sum is equal to x/n. And we also showed that the variance of x/n, $\sigma_{x/n}^2$, equals the variance of x divided by n^2: $\sigma_{x/n}^2 = \sigma_x^2/n^2$. Hence,

$$\sigma_{\overline{x}}^2 = (\sigma_{x/n}^2 + \ldots + \sigma_{x/n}^2) = n\sigma_{x/n}^2 = n\sigma_x^2/n^2 = \sigma_x^2/n \tag{64}$$

We are now going to exploit these results, together with the central limit theorem, to predict the confidence interval of the estimator (61). The central limit theorem tells us that, when n is sufficiently large, $\overline{x}$ is Gaussian distributed with mean equal to the mean of the population, and variance equal to the variance of the population divided by n:

$$\mu_{\overline{x}} = \mu_x \ , \qquad \sigma_{\overline{x}}^2 = \frac{\sigma_x^2}{n^2} \ , \qquad p(\overline{x}) = \frac{1}{\sqrt{2\pi\sigma_{\overline{x}}^2}} \exp\left(-\frac{(\overline{x}-\mu_{\overline{x}})^2}{2\sigma_{\overline{x}}^2}\right) \ . \tag{65}$$

This result is central to the inference of the population mean from the sample mean. For any pair of values x_1 and x_2 we know which is the probability that a sample mean lies within the interval $[x_1, x_2]$, $P(x_1 \leq \overline{x} \leq x_2)$. We also know which is the probability for the distance between the sample mean and the average of the sample mean, normalised by the standard deviation, to lie within an interval, i.e. we know $P(\frac{|\overline{x}-\mu_{\overline{x}}|}{\sigma_{\overline{x}}} \leq x_1)$. Now, since $\mu_{\overline{x}} = \mu_x$, we know

$$P(\frac{|\overline{x}-\mu_x|}{\sigma_{\overline{x}}} \leq x_1) \ . \tag{66}$$

These values are given in Tables that one can find in any textbook on Statistics or they are very easy to calculate with, for example, SciLab. Inverting this reasoning, we can now read this expression as yielding the probability that the population mean is at most at a distance x_1 from the sample mean:

$$P(|\mu_x - \overline{x}| \leq x_1\sigma_{\overline{x}}) \ . \tag{67}$$

Let us explain what we mean with an example: A new drug lowers the heart rates by varying amounts with standard deviation of 2.49 beats/minute ($\sigma_{\overline{x}}$). Imagine that a sample of 50 persons averages a drop of 5.32 beats/min (the sample mean $\overline{x}$). What is

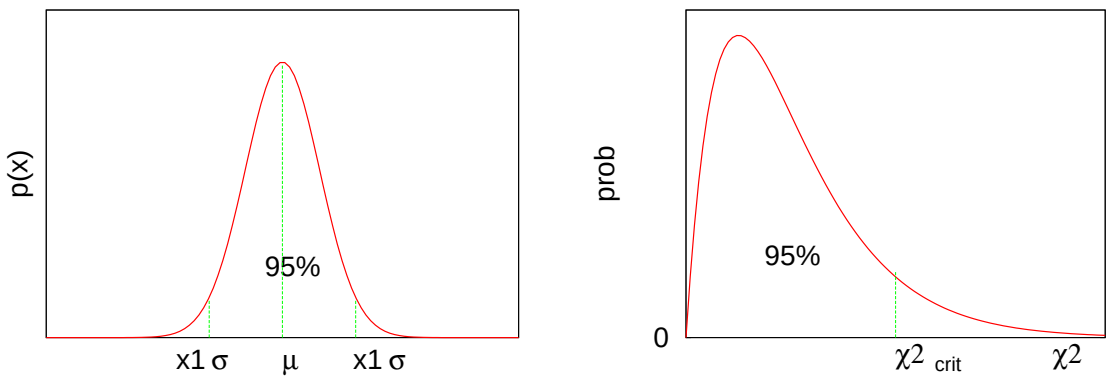

Figure 9: Left: A Gaussian curve. The area under the curve between the two values $x = \mu_x \pm x_1\sigma_x$ represents the 95% of the total probability when $x_1 = 1.96$. See the example in the text. Right: The χ^2 distribution and the definition of χ^2_{crit}.

the mean lowering at a 95% confidence level? The standard deviation of the sample mean is $\sigma_{\overline{x}} = \sigma_x/\sqrt{n} = (2.49 \text{ beats/min})/\sqrt{50} = 0.352$ beats/min. The mean of the sample is $\overline{x} = 5.32$ beats/min.

Since the PDF of sample mean is a Gaussian, we know that $P(|\overline{x}-\mu_x|/\sigma_{\overline{x}} \leq x_1)$ contains 95% of the area of the PDF if $x_1 = 1.96$ (tabulated data). Thus, the extreme values of μ_x that fall within the selected interval are $\overline{x} \pm x_1 \times \sigma_{\overline{x}} = (5.32 \pm 1.96 \times 0.352)$ beats/min, i.e. we can be 95% sure that μ_x is between 4.63 and 6.01 beats/min. See Fig. 9-left.

7.4 Back to variance

Note that, in general, we do not know σ_x^2, the variance of the population and this is why we need to estimate it. We then use Eq. (62) to do it.

8 Analysis of experimental data

In this Section we explore several questions: why is there noise in an experiment, how to describe an experimental distribution, how to determine if two measurements are correlated (one is a consequence of the other) or not, and how can one describe a set of correlated experimental data.

8.1 Sources of noise

Why do the results of an experiment change from one measurement to another? One can identify three rather obvious sources of 'noise' that imply this difference.

The first one is that the experimental conditions change from one measurement to the other. For example, if the experimental set-up is in a room at temperature T from one measurement to the next small variations of the temperature are possible and these might affect the experimental result. A careful experimentalist must try to diminish the effect of the external parameters as much as possible. But to eliminate them completely is impossible.

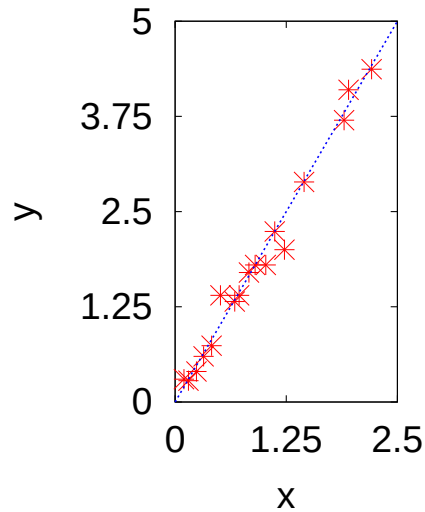

Figure 10: A set of data. The line represents a linear fit.

One is then forced to measure quantities that are not really constant in time and hence fluctuate from measurement to measument. The best description of the data is then given by a probability distribution, the probability distribution to find such or such result.

The exact from of the searched probability distribution can only be obtained performing an infinite number of experiments. This, of course, is not feasible. Thus, one is forced to estimate the form of the prbability distribution from a finite number of measurements. This impossibility gives rise to the *statistical error*. The way to reduce this error is simple: just do more measurements. But after a certain point this is impossible (too long and expensive) and not really necessary.

Finally, there are the *systematic errors* induced by the fact that the measurement is done with the help of some apparatus that itself might introduce errors and deviate the intrinsic distribution function of the observable. One can correct this induced 'mistake' by knowing in as much detail as possible the way in which the apparatus works. But again, this is very hard.

8.2 Linear regression

Usually in sciences we want to determine a functional dependence out of a set of data presented as a "two-column" table of length n with, in each row $i = 1, \ldots, n$, the result of the measurements for the pairs (independent variable x_i, dependent variable y_i). How can one derive the functional form

$$y = f(x) \tag{68}$$

from the set of pairs (x_i, y_i) that will certainly deviate (due to noise in the experiment, etc.) from the ideal case $y_i = f(x_i)$?

First, one traces a scatter plot with a point for each pair in the data set. After some training, just by eye inspection one can guess which is the functional form that yields the best function describing the data.

Take for instance the data represented in Fig. 10. It is clear that the relation between

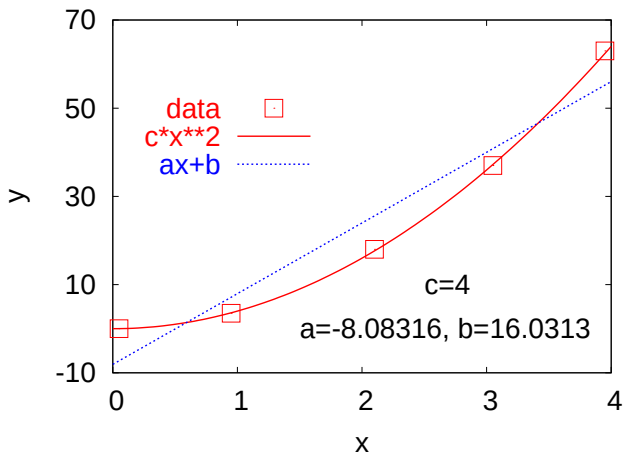

Figure 11: A set of data. One line represents a linear fit and the otherone a quadratic fit (given in the key) that clearly represents the data more accurately.

these points is linear:

$$y = f(x) = ax + b \tag{69}$$

But, what are the values of the parameters a (slope) and b (vertical intercept)? This is not obvious to guess. In order to find these parameters we wish to minimise the distance between the points and the line. The question is, how do we define the distance? The most natural definition is

$$\text{error}(a,b) \equiv \sum_{i=1}^{n} d_i^2 \equiv \sum_{i=1}^{n} \left[y_i - (ax_i + b)\right]^2 \ , \tag{70}$$

and this is the quantity we want to minimize with respect to a and b:

$$\left.\frac{\partial \text{error}(a,b)}{\partial a}\right|_{a^*,b^*} = 0 \qquad \left.\frac{\partial \text{error}(a,b)}{\partial b}\right|_{a^*,b^*} = 0 \tag{71}$$

These coupled equations (together with the check of minimization on the second derivatives) yield the searched values a^*, b^*.

Knowing the best line that fits the data is not enough to know that a line is indeed the best fit to this data. Even if not expected from eye inspection, a quadratic curve could yield, in principle a better fit to the data in Fig. 10. A polynome with order equal to the number of points will obviously yield a perfect fit. However, fits with a very large number of parameters are meaningless. The quality of th fit can be quantified with the 'goodness-of-fit' method that we discuss below.

8.3 More general fitting

In general, the data will not be described with a *linear function* but a more general form will be needed. One then proposes a functional form that – from eye inspection – may describe the data. This functional form will, in general, depend on a few parameters. The *best* description of the data occurs when the distance between the theoretically curve

and the experimental data is minimized. In more technical terms, if $f(x; a_1, \ldots, a_k)$ is the proposed function with $a_1, \ldots, a_k$ the unknown parameters, the optimal description of the data is obtained for values of the parameters such that the distance

$$d(a_1, \ldots, a_k) \equiv \sum_{i=1}^{n} [f(x_i; a_1, \ldots, a_k) - y_i^{exp}]^2 \tag{72}$$

takes its minimum value. These values are fixed by the conditions

$$\left. \frac{\partial d(a_1, \ldots, a_k)}{\partial a_j} \right|_{a_k^*} = 0 \,, \quad \forall j = 1, \ldots, n \,, \tag{73}$$

$$\left. \frac{\partial^2 d(a_1, \ldots, a_k)}{\partial a_j^2} \right|_{a_k^*} > 0 \,, \quad \forall j = 1, \ldots, n \,. \tag{74}$$

For example, the data shown in Fig. 11 are better described by a quadratic law than by a linear regression.

In general, one uses rather simple functions to describe experimental data since one expects them to arise from hopefully simple theoretical laws. We stress once again that a fit with a very large number of adjustable parameters is meaningless.

8.4 χ^2 analysis for goodness-of-fit

One interesting problem is to try to describe some observation with a theoretical form. But, of course, the matching between the experimental data and the theoretical curve will never be exact. The question then arises: how do we know if the observed differences are relevant and tell us that the hypothesis is wrong or they are just normal "noise" associated to the measurement? There are ways to quantify the deviations between experimental observation and the theoretical prediction that allow us to decide whether we accept the "fit" or not. Let us discuss one such method called the χ^2 analysis.

The strategy goes as follows. The measurement is done in one sample out of many possible ones. The question we want to ask is: what is the chance that the results of the measurement differ from the theoretical ones as observed? If this probability is small, then we have to discard the hypothesis. If it is large, then we can accept the hypothesis.

How do we define the discrepancy between the experimental observation and theoretical prediction? By constructing the χ^2 variable

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i^{theor} - y_i^{exp})^2}{y_i^{theor}} \,. \tag{75}$$

If χ^2 takes a small value we may suspect that the hypothesis is correct. Instead, if it is large, we may suspect it is wrong. But how large is large and how small is small? We shall choose a critical value χ^2_{crit} as a criterium to discern between acceptable and non acceptable hypotheses.

As already found with the estimator of the mean and the estimator of the variance, the χ^2 variable is itself a random variable since it is computed from a given sample. Extending the central limit theorem, as it name suggests, one shows that χ^2 is distributed according to a χ^2-distribution, see Sect. 5.5, with *df* number of degrees of freedom (that depends

on n; naively $df = n$ but in practice $df = n - 1$ with n the number of data points). It is important to remember that the form of the χ^2 distribution depends on df. The critical value χ^2_{crit} is determined by the degree of confidence one wants, i.e. for a degree of confidence expressed as a percentage that corresponds to the area under the χ^2-PDF, we find the χ^2_{crit} that limits this area. An example is given in Fig. 9.

In practice then what we do is the following. We accept thr hypothesis – the data is described by the theoretical predicition – with a 95% confidence level if the calculated value χ^2 falls below the χ^2_{crit} associated to having 95% of the weight of the total probability to the left of this value. Otherwise, we reject the hypothesis and we look for an alternative theory.

8.5 χ^2-analysis for independence

Imagine we have two observations, say x and y, and we want to test if one variable is affected by the other one (a question related to finding a functional relation between one set of data and the other).

The idea is to fill in a table x_i, y_i with the data assuming they are independent, then fill in the same table with the actual data, construct the χ^2 function (taking the table constructed under the assumption of independence as the theoretical one, and the other as the observed one) and check if, for the corresponding value of df the χ^2 obtained is acceptable or not. In the first case, the data are independent, in the second they are not.

1 Basic notions: details

1.1 The variance

The variance can be written in two equivalent ways:

$$\sigma_x^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2 = \langle\, x^2 \,\rangle - \mu_x^2 \tag{76}$$

The second equality is very simple to prove by explanding the square:

$$\begin{aligned}
\sigma_x^2 &\equiv \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2+\mu_x^2-2x_i\mu_x\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2+\mu_x^2\frac{1}{n}\sum_{i=1}^{n}1-2\mu_x\frac{1}{n}\sum_{i=1}^{n}x_i \\
&= \langle\, x^2 \,\rangle+\mu_x^2-2\langle\, x \,\rangle^2 \\
&= \langle\, x^2 \,\rangle-\mu_x^2
\end{aligned}$$

It is important to know how to handle this type of sums.

1.2 Effect of the addition of a constant

If μ_x is the average of the set $\{x_1,\ldots,x_n\}$, then μ_x+c is the average of the set $\{\tilde{x}_1,\ldots,\tilde{x}_n\} = \{x_1+c,\ldots,x_n+c\}$ with c a constant. To prove this statement, let us call $\mu_{\tilde{x}}\,\rangle$ the average of the set $\{\tilde{x}_1,\ldots,\tilde{x}_n\}$. Then

$$\mu_{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_i+c) = \frac{1}{n}\sum_{i=1}^{n}x_i + c\,\frac{1}{n}\sum_{i=1}^{n}1 = \mu_x + c\,.$$

The standard deviation is unchanged by a uniform translation of the data. Let us call $\sigma_{\tilde{x}}^2$ the variance of the set $\{\tilde{x}_1,\ldots,\tilde{x}_n\}$. Then

$$\begin{aligned}
\sigma_{\tilde{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i-\mu_{\tilde{x}})^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i+c-(\mu_x+c)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2 \\
&= \sigma_x^2\,.
\end{aligned}$$

The correlation between two sets of data remains unchanged by a uniform translation of the two sets of data by two (possibly different) constants ($x_i \to x_i+c_1$ and $y_i \to y_i+c_2$ with $c_1 \neq c_2$). Using the same notation as above, the proof goes as follows:

$$C_{\tilde{x}\tilde{y}} = \frac{1}{\sigma_{\tilde{x}}\sigma_{\tilde{y}}}\,\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i-\mu_{\tilde{x}})(\tilde{y}_i-\mu_{\tilde{y}}) \tag{77}$$

$$= \frac{1}{\sigma_x\sigma_y}\,\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) \tag{78}$$

$$= C_{xy} \tag{79}$$

1.3 Effect of the multiplication by a constant

The average of a set of data that has been scaled by the same constant c is just $\mu_{\tilde{x}} = c\mu_x$. Proof:

$$\mu_{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n}(cx_i) = c\,\frac{1}{n}\sum_{i=1}^{n}x_i = c\mu_x\ .$$

The mean-square deviation of a set of data that has been scaled by the constant c is $\sigma_{\tilde{x}} = c\sigma_x$. Proof:

$$\begin{aligned}\sigma_{\tilde{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mu_{\tilde{x}})^2 = c^2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \\ &= c^2\sigma_x^2 \quad \Rightarrow \quad \sigma_{\tilde{x}} = c\sigma_x\ .\end{aligned}$$

The correlation of two sets of data that have been multiplied by two (possibly different constants c_1 and c_2 is not modified by this operation. Proof:

$$C_{\tilde{x}\tilde{y}} = \frac{1}{\sigma_{\tilde{x}}\sigma_{\tilde{y}}}\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mu_{\tilde{x}})(\tilde{y}_i - \mu_{\tilde{y}}) \tag{80}$$

$$= \frac{1}{c_1c_2\sigma_x\sigma_y}\,c_1c_2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \tag{81}$$

$$= C_{xy}\ . \tag{82}$$

2 The binomial

2.1 The maximum of $f_n(f)$

The way to calculate the maximum of $f_n(h)$ on the natural numbers is the following. Let us first take the ln of $f_n(h)$. Since the ln is a monotonic function, the maximum of $\ln f_n(h)$ is located at the maximum of $f_n(h)$. Working with the $\ln f_n(h)$ is convenient since it is a smoother function than $f_n(h)$. Note that if n is finite, the set of values that this function can take is discrete. However, since we are interested in studying the large n limit we can assume that we can take a continuum limit and then differentiate the function $\ln f_n(h)$ with respect to h. We then have

$$\frac{\partial}{\partial h}\ln f_n(h) = -\frac{\partial}{\partial h}\ln\left((n-h)!h!\right) \tag{83}$$

If we assume that $h \gg 1$ and $n - h \gg 1$ we can approximate the factorial using Stirling's formula

$$\ln h! \sim h\ln h - h \quad \text{if } \ h \gg 1\ . \tag{84}$$

Now, after some straightforward manipulations of the derivatives we find that the extremum condition

$$\left.\frac{\partial}{\partial h}\ln f_n(h)\right|_{h_{max}} = 0 \tag{85}$$

leads to

$$h_{max} = \frac{n}{2} \tag{86}$$

as expected (one can test the stability of this extremum and check that it is indeed a maximum).

2.2 Derivation of mean and mean-square displacement

There are two ways of deriving these results and both are instructive.

The first set of proofs rely on manipulating the sums. The mean of a binomial is

$$\begin{aligned}
E(h) &= \sum_{h=0}^{n} h \, \frac{n!}{h!(n-h)!} \, p^h \, (1-p)^{n-h} \\
&= \sum_{h=1}^{n} \frac{n!}{(h-1)!(n-h)!} \, p \, p^{h-1} \, (1-p)^{(n-1)-(h-1)} \\
&= p \sum_{h=1}^{n} \frac{n(n-1)!}{(h-1)![(n-1)-(h-1)]!} \, p^{h-1} \, (1-p)^{(n-1)-(h-1)} \\
&= pn \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} \, p^k \, (1-p)^{(n-1)-k} \\
&= pn \, .
\end{aligned}$$

The average of h^2 is

$$\begin{aligned}
E(h^2) &= \sum_{h=0}^{n} h^2 \, \frac{n!}{h!(n-h)!} \, p^h \, (1-p)^{n-h} \\
&= p \sum_{h=1}^{n} h \, \frac{n(n-1)!}{(h-1)![(n-1)-(h-1)]!} \, p^{h-1} \, (1-p)^{(n-1)-(h-1)} \\
&= pn \sum_{k=0}^{n-1} (k+1) \, \frac{(n-1)!}{k!(n-1-k)!} \, p^k \, (1-p)^{(n-1)-k} \\
&= pn(E(k)+1) \\
&= pn[p(n-1)+1] \, .
\end{aligned}$$

Then, the mean-square displacement is

$$\begin{aligned}
\sigma_h &\equiv \sqrt{E(h^2) - (E(h))^2} = \sqrt{pn[p(n-1)+1) - (pn)^2} \\
&= \sqrt{(pn)^2 - p^2 n + pn - (pn)^2} = \sqrt{pn(1-p)} \, .
\end{aligned}$$

The second set of proofs is more general and is based on the use of generating functional. The mean of a binomial is

$$E(h) = \sum_{h=0}^{n} h \, \frac{n!}{h!(n-h)!} \, p^h \, q^{n-h}$$

where, for convenience, we called q the factor $(1-p)$ This expression can also be written as

$$E(h) = p \frac{\partial}{\partial p} \sum_{h=0}^{n} \frac{n!}{h!(n-h)!} \, p^h \, q^{n-h} \Big|_{q=1-p}$$

that, using Newton's binomial formula, is equal to

$$E(h) = p\frac{\partial}{\partial p}(p+q)^n\Big|_{q=1-p} = pn(p+q)^{n-1}\Big|_{q=1-p} = pn\ .$$

Similarly,

$$\begin{aligned}
E(h^2) &= \sum_{h=0}^{n} h^2\,\frac{n!}{h!(n-h)!}\,p^h\,(1-p)^{n-h} \\
&= p^2\frac{\partial^2}{\partial p^2}\sum_{h=0}^{n}\frac{n!}{h!(n-h)!}\,p^h\,q^{n-h}\Big|_{q=1-p} + E(h) \\
&= p^2\frac{\partial^2}{\partial p^2}(p+q)^n\Big|_{q=1-p} + np \\
&= p^2 n(n-1)(p+q)^{n-2}\Big|_{q=1-p} + np \\
&= n(n-1)p^2 + np\ .
\end{aligned}$$

The mean-square displacement is then

$$\begin{aligned}
\sigma_h &\equiv \sqrt{E(h^2)-(E(h))^2} \\
&= \sqrt{n(n-1)p^2+np-(np)^2} = \sqrt{-np^2+np} = \sqrt{np(1-p)}\ .
\end{aligned}$$

3 The Poisson distribution

3.1 The average and mean-square displacement

The average of the Poisson distribution is

$$E(x) = \sum_{x=0}^{\infty} x\,\frac{\mu^x}{x!}\,e^{-\mu} = \sum_{x=1}^{\infty} x\,\frac{\mu\mu^{x-1}}{x!}\,e^{-\mu} = \mu\sum_{x'=0}^{\infty}\frac{\mu^{x'}}{x'!}\,e^{-\mu} = \mu$$

The average of x^2 is

$$\begin{aligned}
E(x) &= \sum_{x=0}^{\infty} x^2\,\frac{\mu^x}{x!}\,e^{-\mu} = \sum_{x=1}^{\infty} x^2\,\frac{\mu\mu^{x-1}}{x!}\,e^{-\mu} = \mu\sum_{x'=0}^{\infty}(x'+1)\,\frac{\mu^{x'}}{x'!}\,e^{-\mu} \\
&= \mu(\mu+1)
\end{aligned}$$

and the mean-square displacement is

$$\sigma_x = \sqrt{E(x^2)-(E(x))^2} = \sqrt{\mu(\mu+1)-\mu^2} = \sqrt{\mu}\ .$$

3.2 The mode

When μ is large, the mode of the Poisson distribution coincides with its average. Indeed, the mode is the value of x where the maximum is attained:

$$\frac{\partial}{\partial x}P(x)\Big|_{x_{mode}} = 0\ .$$

This derivative can be easily calculated using Stirling's expression for the factorial. This approximation is valid whenever the mode appears at large values of x. We shall use it as a working hypothesis and then check from the result when it holds. Thus,

$$\frac{\partial}{\partial x}P(x) = e^{-\mu}\frac{\partial}{\partial x}\left[\frac{\mu^x}{x!}\right]$$

The best way to compute this derivative is to note that

$$g(x) = \ln f(x) \quad \Rightarrow \quad \frac{dg(x)}{dx} = \frac{1}{f(x)}\frac{df(x)}{dx} \quad \Rightarrow \quad \frac{df(x)}{dx} = f(x)\frac{dg(x)}{dx} \tag{87}$$

Taking $f(x) = \mu^x/x!$ and $g(x) = \ln f(x) = \ln \mu^x - \ln x! = x\ln\mu - \ln x! \approx x\ln\mu - x\ln x + x$ Then

$$\begin{aligned}\frac{df(x)}{dx} &= f(x)\frac{dg(x)}{dx} = \frac{\mu^x}{x!}\,\frac{d[x\ln\mu - x\ln x + x]}{dx} \\ &= \frac{\mu^x}{x!}\,[\ln\mu - \ln x - 1 + 1] = \frac{\mu^x}{x!}\,[\ln\mu - \ln x]\end{aligned}$$

The derivative of the pdf reads

$$\frac{\partial}{\partial x}P(x) = e^{-\mu}\frac{\partial}{\partial x}\left[\frac{\mu^x}{x!}\right] = e^{-\mu}\frac{\mu^x}{x!}\,[\ln\mu - \ln x]$$

that vanishes at

$$x_{mode} = \mu\ . \tag{88}$$

4 The Gaussian distribution

The mode of the Gaussian is equal to its mean. The mean and variance of this pdf are μ and σ^2, respectively. 2σ characterizes the width of the pdf. Indeed, it is equal to the distance between the two values of x that correspond to inflection points x_{if} and it is approximately equal to the width at half height. The inflection points are defined from

$$\left.\frac{d^2p(x)}{dx^2}\right|_{x_{ip}} = 0 \tag{89}$$

and after a simple calculation one finds $x_{ip} = \mu \pm \sigma$. The maximum height is $p(x=\mu) = (2\pi\sigma^2)^{-1/2}$. The values of x where the half height is realized are given by

$$p(x_{hh}) = \frac{1}{2\sqrt{2\pi\sigma^2}} \quad \Rightarrow \quad x_{hh} = \mu \pm \sigma\sqrt{2\ln 2}\ . \tag{90}$$

The width is then $\Delta x \equiv x_{hh}^+ - x_{hh}^- = \sqrt{2\ln 2}\sigma$. Note that the factor $\sqrt{2\ln 2} \approx 1.17$ is also quite close to 2.

5 Change of variables

Let $p(x)$ be the probability density of a continuous random variable x. Consider now a new random variable y that is defined as $y = f(x)$. The probabiliy distribution of y, that we call $\pi(y)$ should verify:

$$\pi(y)dy = p(x)dx \ . \tag{91}$$

From this requirement one derives the functional form of $\pi(y)$:

$$\pi(y) = p(x)\ \mathcal{J} = p(x)\ \left|\frac{dx}{dy}\right| = p(x)\left|\frac{1}{df(x)/dx}\right| \ , \tag{92}$$

where $\mathcal{J}$ is the Jacobian of the change of variables and $dy/dx = df(x)/dx$ and $dy/dy = 1/(df(x)/dx))$.

Statistiques TD 1 :

Statistique descriptive

Le but de ce TD est d'utiliser les outils présentés dans le cours (histogrammes, moyennes, variances,...) pour analyser un ensemble de données.

1. Les notes obtenues par un groupe d'étudiants à l'examen final de Statistiques sont : 7, 8, 8, 8, 9, 9, 9 ,9 , 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 12, 12, 12, 13, 13, 13, 14, 20.

   (a) Déterminer les valeurs que la variable "notes" peut prendre.
   Est-ce un ensemble borné ?

   (b) Tracer l'histogramme associé à cette série de données.
   On va maintenant déterminer ses propriétés de façon qualitative.
   Que peut-on remarquer, à première vue, sur le rapport entre la moyenne et la médiane ?
   Déterminez le mode et comparez-le à la moyenne.
   Quelle sera la valeur de l'écart-type ? Grande ou petite par rapport à la moyenne ?

   (c) Calculer le mode, la moyenne, la médiane et l'écart-type et vérifier si les prédictions qualitatives précédentes sont valables.

2. Après effectuer une mesure on trouve les résultats répresentés dans la Fig. 12

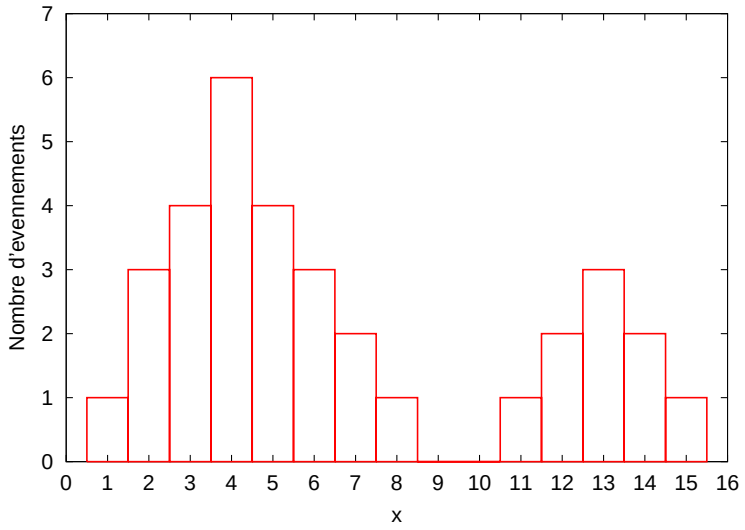

Figure 12: Répresentation graphique des résultats d'une mesure.

   (a) Calculer la moyenne et la variance.

(b) Déterminer la médiane et le mode.

(c) Donner un exemple de mesure donnant ce type de résultat.

(d) Montrer la moyenne, le mode et la médiane sur l'histogramme.

(e) Tracer une ligne horizontale répresentant l'écart type. Comment se compare-t-il à la largeur du diagramme à mis-hauteur ?

3. Les salaires de cinq employés sont 1000, 1500, 3000, 3200 et 4000 Euros par mois.

(a) En général, que peut-on dire sur l'intervalle de variation des salaires ?

(b) Quel est le salaire moyen ?

(c) Quel est le salaire médian ?

(d) Sans faire de calculs explicites, croyez-vous que les deux résultats précédents seront modifiés si tous les salariés reçoivent une augmentation de 200 Euros ?

(e) Calculer les nouvelles moyenne et médiane et vérifier si vos prédictions sont exactes.

(f) Répéter les deux questions précédentes si l'augmentation est de 10% pour chaque salarié.

(g) Calculer la variance des salaires avant et après les augmentations. Comment est modifiée la variance ?

(h) Calculer la moyenne, médiane et variance si on ajoute un sixième employé avec un salaire de 1000 Euros par mois au premier ensemble (avant augmentation). Comment sont modifiés les résultats par rapport aux valeurs originelles ?

(i) Quel est l'effet de l'augmentation sur la forme de l'histogramme des salaires ? Utiliser plusieurs tailles pour les bins et discuter le résultat.

4. On veut étudier les propriétés statistiques de deux ensembles $X = \{1, 3, 6, 10\}$ et $Y = \{1, 2, 3, 5\}$.

(a) Calculer la moyenne (qu'on appelera μ_X) et l'écart type moyen (qu'on appelera σ_X) de l'ensemble X.

(b) Calculer la moyenne (qu'on appelera μ_Y) et l'écart type moyen (qu'on appelera σ_Y) de l'ensemble Y.

(c) Calculer la correlation entre les ensembles X et Y.

(d) Construire l'ensemble Z des différences entre les éléments de X et Y.

(e) Calculer la moyenne (μ_Z) et l'écart type moyen (σ_Z) de Z.

(f) Exprimer μ_Z en fonction de μ_X et μ_Y.

(g) Exprimer σ_Z^2 en fonction de σ_X^2 et σ_Y^2.

(h) Comment sont modifiées la moyenne μ_X et la variance σ_X^2 si on transforme l'ensemble X en $X' = \{x_i + c\}$ avec c une constante ? (voir notes du cours.)

(i) Comment sont modifiées la moyenne μ_X et la variance σ_X^2 si on transforme l'ensemble X en $X' = \{cx_i\}$ avec c une constante ? (voir notes du cours.)

Statistiques TD 2 :

Notions de probabilité

Le but de ce TD est de vous familiariser avec les probabilités. Nous commençons par revoir quelques propriétés générales.

1. Propriétés générales.

   (a) *Événements complémentaires.*
   Si la probabilité d'un étudiant de réussir un concours est 0.3, quelle est la probabilité de faillite du même étudiant dans le même concours ?

   (b) *Principe d'addition, événements mutuellement exclusifs.*
   Les probabilités de Jean, Marie et Pierre d'obtenir un poste de professeur à l'université sont 0.5, 0.3 et 0.2, respectivement. Quelle est la probabilité que Jean ou Marie obtiennent ce poste ?

   (c) *Principe d'addition général.*
   La probabilité d'une compagnie d'obtenir un contrat pour construire une maison est 0.5. La probabilité que la même compagnie obtienne un contrat pour paver une rue est 0.3. La probabilité pour qu'elle obtienne les deux contrats simultanément est 0.13
   Quelle est la probabilité pour que la compagnie obtienne au moins un contrat ?

   (d) *Indépendance des événements.*

      i. La probabilité de tirer un 3 avec un dé est 1/6, la probabilité de tiré un 2 avec un autre dé est aussi 1/6. Quelle est la probabilité de tirer un 3 et un 2 avec les deux dés ?

      ii. Les deux événements de la question (a) sont-ils indépendants ?

      iii. Si les probabilités des événements indépendants $E_1, \ldots, E_n$ sont $p_1, \ldots, p_n$, quelle est la probabilité simultanée de ces n événements ?

2. Dans la suite nous jouerons aux dés pour continuer à vous familiariser avec les probabilités.

   (a) Quelle est la probabilité d'obtenir un nombre pair après le jet d'un dé ?
   Et celle d'obtenir un nombre impair ?

   (b) Quelle est la probabilité de *ne pas* obtenir un 6 ?

   (c) Quelle est la probabilité d'obtenir le même résultat si on jette deux dés ? (Plus précisément, on veut obtenir les couples $(1,1)$, $(2,2)$, $(3,3)$, $(4,4)$, $(5,5)$ ou $(6,6)$.)

   (d) Quelle est la probabilité d'obtenir au moins un 6 après jeter un dé deux fois ?

(e) Le chevalier de Méré, un grand joueur, a trouvé un jeu de dés gagnant. Son pari était qu'il obtiendrait au moins un 6 en quatre jets d'un dé. Expliquer pourquoi il s'agit bien d'un pari gagnant.

3. Et maintenant on joue aux cartes. Prenons un jeu de 52 cartes sans joker. Leurs couleurs sont pique, coeur, carreau et trèfle.

   (a) Quelle est la probabilité de tirer un coeur ?

   (b) Quelle est la probabilité de tirer un roi ?

   (c) Quelle est la probabilité de tirer le roi de coeur ?

   (d) Quelle est la probabilité de tirer un roi ou un coeur ? Trouver le résultat en utilisant deux arguments différents.

   (e) Les évennements 'tirer un roi' et 'tirer un coeur' sont-ils indépendants ?

   (f) Répondre la dernière question si l'on rajoute un joker au jeu de cartes. Expliquer le résultat trouvé.

4. Prenons un ensemble de variables aléatoires $x_1, \ldots, x_n$ distribuées selon $P_1(x_1), \ldots,$ $p(x_k)$,...,$P_n(x_n)$ (quelques unes de ces variables peuvent être continues comme par exemple, x_k, ave $p(x_k)$ sa densité de probabilité).

   Exprimer la valeur d'expectation, $E(y)$, et la variance, σ_Y, de la variable aléatoire construite avec la somme des précendentes, $y = x_1 + x_2 + \ldots + x_n$, en fonction de $E(x_1), \ldots, E(x_n)$ et $\sigma_{x_1}, \ldots, \sigma_{x_n}$.

   Imaginex maintenant que les x_i répresentent les valeurs de la même variables aléatoire x (avec valeur de expectation $E(x)$ et variance σ_X) obtenus après n expériences indépendantes. Quelle est la moyenne de y ? Quelle est sa variance ?

   Ce résulatat fait partie du théorème de la limite centrale qu'on étudiera plus tard.

Statistiques TD 3 : Lois de probabilité

Le but de ce TD est d'étudier plusieurs distributions de probabilité communes en physique, biologie, etc. Les questions en italiques devront être répondus à la maison. Quelques-unes d'entre elles nécessitent de SciLab.

1 La distribution uniforme

La densité de probabilité d'une variable aléatoire X distribuée uniformément sur l'intervalle $[-a/2; a/2]$ est donnée par :

$$f(X) = Cte \tag{93}$$

1. Calculer la valeur de la constante de cette densité de probabilité.
2. Calculer la valeur d'expectation de X ($E(x)$).
3. Calculer la variance de X (σ_X^2).
4. Calculer le moment d'ordre 3 de X.
5. Calculer enfin μ_4/σ^4, où μ_4 est le moment d'ordre 4 de X.
6. Que remarque-t-on par rapport aux moments pairs et impairs ? Trouvez un argument général pour montrer ce résultat.

Supposons que des particules peuvent tomber aléatoirement sur une surface carrée d'aire d^2. Trouver la valeur d'expectation et la variance de la densité de particules.

2 La distribution binomiale

1. Une variable aléatoire X suit une loi de probabilité binomiale si :

$$P(X = k) = \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k} \tag{94}$$

   où $0 \leq p \leq 1$ et $0 \leq k \leq n$. C'est la probabilité qu'un événement de probabilité p apparaissent exactement k fois lorsque l'on réalise n expériences indépendantes.

   (a) Est-ce une distribution discrète ou continue ?

   (b) *Prenez $p = 0.4$. A l'aide d'un logiciel graphique, tracez cette distribution pour $n = 1, 10, 20, 30, 40$. Discutez la forme des courbes obtenues.*

   (c) *Vérifier sur chaque courbe si la moyenne est bien donnée par $\mu = np$.*

   (d) *Vérifier sur chaque courbe si la variance est bien donnée $\sigma^2 = np(1-p)$.*

2. La probabilité qu'un client dépense plus de 50 Euros en faisant ses courses au supermarché est 1/3. On supposera qu'il y a 4 clients à la caisse.
   (a) Quelle est la distribution de probabilité pour le nombre de ces clients dépensant plus de 50 Euros ?
   (b) Calculer la moyenne de cette distribution.
   (c) Calculer la variance de cette distribution.
3. Dans une plaine, la probabilité de trouver de l'eau à moins de 100 mètres est 0.6. Six propriétaires vont percer des puits.
   (a) Quelle est la distribution de probabilité pour le nombre de puits au-delà de 100 mètres ?
   (b) Tracez cette distribution.
   (c) Calculer sa valeur d'expectation.
   (d) Calculer sa variance.
4. On suppose que la probabilité d'une ampoule d'être défectuese est 0.1. Quelle est la probabilité qu'exactement deux ampoules parmi quatre soient défectueuses ?
5. Un chef de service a remarqué que la probabilité que chaque employé arrive en retard est 0.125. Quelle est la probabilité qu'exactement un employé arrive en retard s'il y a six employés dans le service ? Quelle est la probabilité qu'exactement 2 employés du même service n'arrivent pas en retard ?

3 La distribution Poissonnienne

La distribution Poissonnienne est donnée par la formule :

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu} \tag{95}$$

1. Est-ce une distribution discrète ou continue ?
2. *A l'aide d'un logiciel graphique tracez cette distribution pour plusieurs valeurs de* μ, $\mu = 0.4, 4, 8, 12, 16$.
3. *D'après les graphes obtenus, quelle est la moyenne de cette distribution ?*
4. *Comparer les résultats des graphes précédents à une distribution binomiale de moyenne* $\mu (= np)$.
5. *Que peut-on conclure pour le rapport entre ces deux distributions ?*
6. Un des premiers exemples de cette distribution a été donné par les morts dûs aux coups de cheval dans l'armée allemande à la fin du XIXe siècle. Plus précisément, pendant la période de 20 ans allant de 1875 à 1894, il y a eu une moyenne de 0.7 mort par an dû à ce type d'accident dans les 14 bataillons de cavalerie.

(a) Calculer la distribution des morts en supposant qu'elle est donnée par une Poissonnienne.

(b) Quelle est la distribution numérique attendue pour le nombre de morts en fonction du nombre de groupes ? (Celle-ci correspond à multiplier les probabilités du point précédent par 20×14, c'est-à-dire, le nombre d'ans fois le nombre de bataillons.)

(c) De façon qualitative, peut-on dire que ces prédictions sont précises, si les vrais nombres des morts ont été les suivants :

0 morts	144 groupes	1 mort	91 groupes
2 morts	32 groupes	3 morts	11 groupes
4 morts	2 groupes	5 morts	0 groupes

(96)

On reverra cette question de façon quantitative dans le TD **4 ou 5**.

4 La distribution exponentielle

On considere une population d'insectes soumise à l'observation permanente. La durée de vie est une variable aléatoire que l'on suppose dontinue et positive et dont la densité est donnée par :

$$f(x) = \alpha e^{-\alpha x} \tag{97}$$

avec $\alpha > 0$.

1. Quels valeurs peut prendre la variable x ?

2. Déterminer si cette fonction peut-répresenter une densité de probabilité.

3. Calculer la valeur d'expectation de x, $E(x)$, pour $p(x) = f(x)$.

4. On constate que, sur 1000 insectes, 273 vivent au delà de 100 unités de temps. En déduire une valeur approchée de $E(x)$.

5. On étudie dans cette question la fecondité d'un insecte vivant pendant un intervalle de temps fini $[t, t + \Delta t]$. On découpe cette intervalle en un nombre entier n d'intervalles de longueur δt. On fait les hypothèse suivantes :

   La probabilité que l'insecte considéré ponde un œuf au cours de l'un quelconque des intervalles δt est $p\delta t$ (avec $0 < p < 1$).

   La probabilité qu'il y ait plus d'un œuf pondu pendant l'un quelconque des intervalles δt est négligéable si δt est choisi suffisament petit.

   Les événements survenants sur deux intervalles δt distincts sont indépendants.

   (a) Trouver la lois de distribution de N, le nombre totale d'œufs pondus dans la période $[t, t + \Delta t]$. Exprimer le résultat avec une formule mathématique.

   (b) Que dévient cette lois quand $n \to \infty$, $\delta t \to 0$ et p reste fini ? Donner l'expression mathématique du résultat. Discuter le résultat.

5 Une variation de la distribution exponentielle

Soit la fonction $f(x) = axe^{-\alpha x}$, définie pour des valeurs de x réels et positifs.

1. *Que sont les conditions que les paramètres a et α doivent satisfaire pour que cette fonction soit une densité de probabilité ?*
2. *Calculer sa moyenne et l'emplacement de son maximum. Sont ces deux valeurs les mêmes ?*
3. *Que peut-on dire sur les propriètés de symétrie de cette densité de probabilité ?*
4. *Relier la valeur de expectation de x et de $x + \lambda$ avec λ une constante.*
5. *Relier la valeur de expectation de x et de λx avec λ une constante.*

6 La distribution Gaussienne ou normale

La densité de probabilité d'une variable Gaussienne est donnée par la formule suivante :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{98}$$

1. Est-ce une distribution discrète ou continue ?
2. *A l'aide d'un logiciel graphique tracez cette distribution pour plusieurs valeurs de μ et σ^2 : $\mu = 0.4, 4, 8, 12, 16$ et $\sigma^2 = 0.24, 2.4, 4.8, 7.2, 9.6$.*
3. *Comment peut-on relier la distribution Gaussienne à la distribution binomiale ? Donnez la réponse graphiquement.*
4. Quelle est la moyenne de la distribution normale ? *Vérifiez ce résultat graphiquement.*
5. Si $\mu = 0$, montrer que la moyenne est zéro.
6. Calculer sa variance et *vérifiez le résultat graphiquement.*
7. Montrer que les points d'inflexion de $p(x)$ sont situés à $x - \mu = \pm\sigma$.

7 La distribution χ^2

1. Trouver la loi de distribution de variable $Y = X^2$ si X est une variable distribuée de façon normale, avec $\mu = 0$ et $\sigma^2 = 1$.

There are many excellent textbooks that describe the Theory of Probability and statistics. Just a very partial list of texts, most of them in French, is the following:

Statistiques TD 4 : Estimation statistique

Le but de ce TD est de commencer à vous familiariser avec les techniques statistiques proprement dites. Encore une fois, les questions en italiques sont pour la maison.

1. Le théorème de la limite centrale.

   (a) *Prenez 200 numéros téléphoniques de l'annuaire et rajoutez les 4 dernières chiffres de chaque. Tracez un histogramme avec les 200 résultats de cette opération.*

   Quelle distribution doit cet histogramme approcher pour grand nombre de numéros téléphoniques ?

   Quelle est la moyenne et la variance de cette distribution ?

2. On considère un échantillon aléatoire de taille n d'une population donnée. La mesure de la variable X sur cet ensemble donne : x_1, x_2, x_3, x_4, ..., x_n.

   Estimer la moyenne et l'écart-type de X.

   Peut-on prévoir exactement le résultat de la même mesure sur un autre éxcantillon de taille n ? Et ses moyenne et écart-type, sont-ils de variables aléatoires ?

3. Une usine produit des piles avec une durée de vie qui a une variance égale à 5.76 mois carrés. En prenant un échantillon de 64 piles on trouve une durée de vie moyenne égale à 12.35 mois.

   (a) Calculer l'écart-type de la durée de vie des piles.

   (b) Calculer la l'écart-type de la durée de vie moyenne estimée à partir de l'échantillon.

   (c) On veut estimer l'intervalle de confiance à 90% de la durée de vie moyenne pour toutes les piles produites dans cette usine.

   Trouver la distance à la moyenne (mesurée en unité de l'écart-type) qui contient 90% du poids d'une distribution normale. Utiliser ce résultat pour calculer l'intervalle de confiance. Donner les limites de cette intervalle en nombre de mois.

   (d) Quel aurait été l'intervalle de confiance à 90% si l'échantillon contenait 100 piles ?

4. On veut déterminer combien d'heures par semaine un adolescent passe devant la télévision. Aprés avoir interviewé 500 personnes on trouve $\sum_{i=1}^{N=500} x = 16475$ heures et $\sum_{i=1}^{N}(x_i - \overline{x})^2 = 48907$ heures^2.

   (a) Calculer la moyenne et l'écart-type de cet échantillon.

   (b) Quel est l'estimateur pour l'écart-type de la population ?

   (c) Calculer l'écart-type de la moyenne des echantillons de cette taille.

(d) Calculer, avec un intervalle de confiance de 98%, le nombre d'heures que les adolecents passent devant la télévision.

5. On essaie de déterminer la quantité moyenne de polluant qu'une usine déverse dans une rivière par jour. On a besoin d'un estimateur precis à 50 grammes près et on souhaite un niveau de confiance de 95%.

   (a) Si quelques mesures indiquent que la variance est $\sigma^2 = 21800$ grammes^2, exprimer l'estimateur de la moyenne en fonction du nombre de jours où l'on effectuera des mesures. On appellera ce nombre de jours n.

   (b) Trouver une borne inférieure pour n.

6. Un nouveau processus de génération de pierres precieuses a donné après son premier test six pierres de poids : 0.43, 0.52, 0.46, 0.49, 0.56 carats.

   (a) Quelle est la taille de l'échantillon ? Est-elle petite ou grande ? Quelle distribution doit-on utiliser pour estimer les propriétés de la population ?

   (b) Estimer la moyenne et la déviation standard de la population.

   (c) Trouver l'intervale de confiance à 90% pour le poids moyen des pierres générées avec cette procedure.

Statistiques TD 5 : χ^2 et Régressions linéaires

Le but de ce TD est de vous montrer comment utiliser la distribution du χ^2 pour quantifier la représentation de données par une distribution et les régressions linéaires pour relier une quantité (la variable indépendante) à une autre (la variable dépendante).

1. Un biologiste prétend que quatre espèces de mouches doivent apparaître dans les proportions 1 : 3 : 3 : 9. La mesure d'un échantillon donne : 226, 746, 733, 2277 mouches de chaque type.

   (a) Quel est le nombre de mouches attendu pour chaque espèce ?

   (b) Calculer $\chi^2 \equiv \sum_i (\text{nombreobservé}_i - \text{nombreattendu}_i)^2/\text{nombreattendu}$.

   (c) Combien de dégrés de liberté a-t-on ?

   (d) Quel est la valeur critique de la distribution χ^2 pour accepter l'hypothèse du biologiste avec un risque de 10% ?

   (e) Doit-on rejeter son hypothèse ?

2. La table ci-dessous montre le résultat d'une étude hospitalière du nombre de patients qui ont survecu un nombre d'années donné à un type de cancer.

Nombre d'années	0	1	2	3	Plus
Nombre de patients	60	110	125	88	67

   On veut tester si, avec un risque de 2.5%, ces données peuvent être décrites par une distribution Poissonienne de paramètre $\mu = 2.1$.

   (a) Calculer les probabilités Poissoniennes pour $k = 0, 1, 2, 3, \text{Plus}$.

   (b) À l'aide de ces prédictions Poissoniennes, calculer le nombre de patients parmi 450 supposés survivre $0, 1, 2, 3$ ans et plus.

   (c) Calculer le χ^2 pour ces données.

   (d) Quel est le nombre de dégrés de liberté pour ce problème ?

   (e) Quel est la valeur critique de χ^2 pour le niveau de risque ?

   (f) Doit-on rejeter l'hypothèse ?

3. La table suivante décrit le revenu et la consommation observés aux États Unis entre 1981 et 1995

An	Revenu	Consommation	*An*	Revenu	Consommation
1981	2200.2	1941.3	1989	3894.5	3594.8
1982	2347.3	2076.8	1990	4166.8	3839.3
1983	2522.4	2283.4	1991	4343.7	3975.1
1984	2810.0	2492.3	1992	4613.7	4219.8
1985	3002.0	2704.8	1993	4789.3	4454.1
1986	3187.6	2892.7	1994	5018.8	4698.7
1987	3363.1	3094.5	1995	5307.4	4924.3
1988	3640.8	3349.7			

(a) Tracer ces données sur un graphe avec Revenu en abscisse et Consommation en ordonnée (*nuage de points ou scatterplot*).

(b) Y a-t-il une corrélation entre ces données ?

(c) On essaiera de représenter ces données par une régression linéaire, en demandant que la somme des carrés des erreurs soit minimale. Trouver les équations qui déterminent la pente et l'ordonnée à l'origine ?

(d) Calculer la régression linéaire qui représente le mieux les données (au sens des moindres carrés).

(e) Tracer la droite trouvée sur le nuage de points et vérifier qu'elle est une bonne approximation des données.

(f) Calculer la précision de la régression linéaire,

$$r^2 \equiv 1 - \frac{\text{Somme erreurs carrés}}{\text{Somme (Consommation-Consommation moyenne)}^2}$$

(g) Tracer un graphe avec les residus. A quoi resemble-t-il ?

(h) Admettons qu'on peut utiliser le résultat de la régression linéaire pour predire le futur. Estimer la consommation si le revenu est égal à 6000.

4. Repetez l'analyse des points (d)-(g) pour l'ensemble de données suivant :

1	10.000	4	12.597	7	15.869	10	19.990
2	10.800	5	13.605	8	17.138	11	21.589
3	11.664	6	14.693	9	18.509	12	23.316

Conclure.

Sujets de suivis de Statistiques

Statistiques

1. Étude de séries temporelles.
2. Sondages d'opinion. À quelques jours du 2e tour d'une élection présidentielle, une agence de sonda ges publie les résultats suivants :
   - Candidat A : 49%
   - Candidat B : 51%

   Enquête réalisée auprès de 1000 personnes.

   Que vous inspire ces résultats ? En particulier, êtes-vous capables de détermine r l'incertitude liée à ces estimations ?
3. Tests médicaux Une grande entreprise pharmaceutique vient de produire un nouveau test capable d e diagnostiquer une certaine maladie. La prévalence de cette maladie sur l'ensemble de la population (c'est-à-dire la probabilité d'occurence de la maladie) est de 1%. Lors des essais, le test a fo urni les résultats suivants:
   - Probabilité que le test soit positif si le patient est malade : P(P|M) = 0,99 ;
   - Probabilité que le test soit positif si le patient est sain : P(P|S) = 0 ,01 ;

   Pensez-vous que ce test fournisse un bon diagnostic de la maladie ?
4. Démographie (numérique) L'institut national d'études démographique (INED) est chargé d'étudier la démogr aphie française. Le but de ce sujet est d'étudier un des jeux de données fournis par cet institut : il s'agit du nombre d'hommes et de femmes survivants à l'âge x pour 100000 français(es). Vous chercherez no tamment à déterminer la propabilité de décès à l'age x. Vous essaierez aussi d'obtenir la loi de probabi lité de la durée de vie des français(es).
5. Indice boursier (numérique) On étudiera à l'aide des outils statistiques présentés pendant le cours la série temporelle de l'indice CAC40 (indice de la bourse de Paris). On cherche ra notamment à caractériser la distribution de probabilité de gain en fonction de la durée du placement.

G. Saporta, *Probabilités, analyse des données et statistique*, Édition s Technip.

Histoire de sciences

1. Statistiques. Travail de K. Pearson et Biometrika.
2. Statistiques. Travail de R. Fisher.
3. Statistiques. Travail de Student.

4. Probabilité. Texte de Galileo.
5. Probabilité. Travail de Laplace.
6. Probabilité. Travail de Bernoulli.

Mathématiques

1. Prouvez le théorème de la limite centrale.
2. Discutez la limite Gaussienne de la loi binomiale.
3. Étude des lois des extrêmes.
4. Étude des lois de probabilité plus connues.
5. Marches au hasard, propriétés et applications.

Biologie

1. La modélisation la plus simple d'un polymère est donnée par une marche aléatoire sur un réseau. On étudiera cette modélisation et on donnera les expressions pour le rayon de giration et la distance entre les deux bouts de la chaîne. Touver une amelioration importante de ce modèle.

   P-G de Gennes, *Scaling concepts in polymer physics*, Cornell Univ. Press (1979).

2. On étudiera l'expérience de Luria-Delbrück montrant que les bactéries resistantes aux virus ou/et antibiotiques sont le résultat de mutations. (**Sujet difficile**.)

   G. B. Benedek and F. Villars, *Physics with illustrative examples from medicine and biology*, AIP Press & Springer, 2000.

Finances

1. Modèle de Black-Scholes.
2. Étude de la bourse.

Environnement

1. Feu de forêts.

Document SciLab préparé par Philippe Andrey

1 Calculer une moyenne

La fonction `mean` permet de calculer des moyennes de valeurs stockées dans des vecteurs et des matrices. Pour les matrices, il est possible de procéder en ligne ou en colonne:

```
--> moy = mean( x ); // moyenne des coefficients du vecteur x
--> moy = mean( m ); // moyenne des coefficients de la matrice m
--> moy = mean( m, 'c' ); // moy vecteur des moyennes des lignes
--> moy = mean( m, 'r' ); // moy vecteur des moyennes des colonnes
--> moy = mean( m(1,:) ); // moyenne de la première ligne
--> moy = mean( m(:,2) ); // moyenne de la deuxième colonne
```

2 Déterminer l'histogramme d'un ensemble de valeurs

La fonction `histplot` affiche un histogramme des valeurs stockées dans un vecteur. Le premier argument à passer à la fonction est soit le nombre de classes à utiliser soit un vecteur donnant les bornes de ces classes:

```
--> x = rand( 1, 100 ); // x vecteur de 100 valeurs comprises entre 0 et 1
--> histplot( 4, x ); // histogramme réalisé avec 4 classes
--> histplot( [0:0.2:1], x ); // histogramme des classes 0-0.2, 0.2-0.4, etc.
```

Si l'on souhaite faire plus qu'un simple affichage graphique de l'histogramme, il faut utiliser la fonction `dsearch` pour récupérer dans un tableau les effectifs des classes. Cette fonction prend deux arguments: le tableau des valeurs et le tableau des bornes des classes. On ne récupère en général que les deux premiers paramètres sur les trois qu'elle retourne, le deuxième étant précisément le tableau des effectifs:

```
--> x = rand( 1, 100 ); // x vecteur de 100 valeurs comprises entre 0 et 1
--> nclasses = 3;
--> bornes = linspace( 0.0, 1.0, nclasses+1 );
--> [idx,histo] = dsearch( x, bornes ); // histo tableau des effectifs des clas
ses
```

3 Déterminer les paramètres d'une droite de régression

On dispose de n couples de points dont les abscisses et les ordonnées sont stockées dans deux tableaux `x` et `y`, de taille une colonne et n lignes chacun. Les commandes ci-dessous permettent de déterminer et d'afficher la droite de régression pour ces données:

```
--> [a,b] = reglin( x, y );  // calcul des paramètres de la droite
--> plot2d( x, y, -1 );      // affichage des points sous la forme de croix
--> plot2d( x, a*x+b, 2 );   // tracé de la droite en bleu
```

Statistiques TP

License SVP 2004

Les générateurs de nombres aléatoires

Le but de ce TP est d'étudier les propriétés d'une variété de générateurs de nombres aléatoires. On verra que "les nombres aléatoires ne peuvent pas être générés avec des générateurs aléatoires" [Knuth, 81].

Conseils généraux

Le TP doit être préparé, c'est-à-dire que le polycopié doit avoir été lu et compris dans ces grandes lignes en arrivant à la séance. Des questions de préparation sont d'ailleurs à remettre à votre enseignant au début de la séance (elles seront notées 5/20).

La bibliographie à la fin du polycopié pourra aider à la préparation. Les étudiants pourront également, utiliser d'autres sources bibliographiques (trouvées, par exemple, sur internet) s'ils le desirent. Les étudiants sont vivement encouragés à préparer les codes numériques (en Scilab) avant le début de la séance.

Compte rendu

Un compte-rendu rédigé est à remettre à l'enseignant à la fin de la séance. Le compte-rendu doit présenter :

- la réponse aux questions de préparation (préparées à la maison).
- de manière concise, scientifique et claire ce que l'étudiant a fait pendant la séance.
- une discusion des problèmes rencontrés.
- Si vous avez le temps, l'étude associé à la section "travail complémentaire".

Notation

Questions de préparation : 5/20.
Travail pendant la séance et le reste du compte-rendu : 15/20.

1 Questions de préparation

1.1 Un générateur de nombres aléatoires purs

Donnez une méthode pour générer de nombres aléatoires purs binaires, c'est-à-dire une séquence $\underline{x} = (x_1, x_2, \ldots, x_n)$ avec $x_i = 0, 1$, pour $i = 1, \ldots, n$, en utilisant les résultats d'une expérience d'emission de particules par une source radioactive. Aide : utilisez les résultats de l'expérience étudiée en TP de Mesures Physiques pour répondre à cette question.

1.2 Les nombres pseudo aléatoires

Normalement on utilise les ordinateurs pour générer une séquence de nombres aléatoires. Les générateurs de nombres pseudo aléatoires sont des algorithmes qui calculent le i-éme nombre x_i en fonction des $i-1$-èmes nombres sortis précédemment. Clairement, ces séquences ne sont pas vraiment aléatoires.

1.2.1 Récurrences entières

On s'intéresse à générer une séquence de nombres aléatoires réels distribués de façon uniforme en [0,1].

La méthode la plus simple pour générer une séquence de nombres aléatoires entiers $\underline{x} \equiv (x_1, x_2, \ldots, x_n)$ est d'agir avec une fonction f sur le nombre précedent dans la séquence :

$$x_i = f(x_{i-1}) \ , \tag{99}$$

avec la condition initiale, ou *semence*, x_0. La fonction f doit incorporer l'arithmetique entière seulement. La séquence $\underline{x}$ est transformée en une séquence de nombres réels $\underline{r} \equiv (r_1, \ldots, r_n)$ avec $r_i \in [0,1]$ en divisant chaque entier par m, $r_i = x_i/m$, avec m la longueur du cycle (voir la suite).

1. Identifier les valeurs que chaque x_i peut prendre.

2. Si après k pas de la récurrence on trouve un x_k qui est déjà apparu ($x_k = x_i$ avec $i < k$) que se passe-t-il avec la suite de la séquence ?

Un des problèmes principales dans le développement d'un générateur de nombres aléatoires performant est de trouver une f avec un très long cycle. Une fois trouvé un tel générateur on doit vérifier que les nombres sont pseudo aléatoires.

3. Donner un exemple de séquence entière avec un cycle égal à 100. Aide : penser à utiliser la fonction module.

1.3 Le générateur module

Cet algorithme génère une séquence de nombres entiers $\underline{x}$ en utilisant la récurrence

$$x_i = (ax_{i-1} + c) \mod m \tag{100}$$

avec la *semence* x_0. Les trois *entiers* a, c et m sont des paramètres qui définissent la récurrence.

1.3.1 La longueur du cycle

La séquence générée par cet algorithme est périodique. La longueur du cycle dépend des valeurs des paramètres a, c et m ainsi que de la semence.

1. Donner une borne supérieure à la longueur du cycle.

2. Quelle valeurs va-t-on obtenir pour x_i si a, c et m sont pairs ? Dans ce cas, quelle est la borne supérieure à la longueur du cycle ?

3. Quelle est la séquence $\underline{x}$ si $x_0 = 0$ et $c = 0$? Quelle est la longueur du cycle ?

4. On essaie de générer une séquence aléatoire binaire (les éléments sont des 0 et 1 seulement) avec cet algorithme. Quel choix doit-on faire pour m ? Quelle séquence trouve-on ? Est elle aléatoire ?

   Il existent des générateurs de séquences binaires aléatoires bien plus performants mais on ne les discutera ici (voir [Knuth,81]).

A votre avis, les choix de la semence et, surtout, le choix des paramètres a, c et m sont-ils importants pour avoir un long cycle ?

2 Travail en séance

2.1 Le génerateur module : suite

1. Étudiez x_{i+1} en fonction de x_i pour le choix $a = 12$, $c = 0$ et $m = 143$. Quelle est la longueur du cycle ?

2. On peut prouver analytiquement que la séquence générée par cet algorithme a un cycle de longueur m si

   — c est relativement prime à m (cela veut dire que c et m n'ont pas des diviseurs communs appart le nombre 1).

   — $a - 1$ est un multiple de p pour chaque prime p qui est un diviseur de m.

   — $a - 1$ est un multiple de 4 si m est un multiple de 4.

   Donner un choix des paramètres qui vérifie ces contraintes pour les trois cas $m = 10, 16, 50$. Montrer les séquences générées par l'algorithme et vérifier qu'elles ont la longueur attendue.

2.2 Tests statistiques

Une fois choisis les paramètres pour avoir un cycle long on doit vérifier si les nombres trouvés sont *pseudo* aléatoires ou, en défaut, si la séquence a des corrélations importantes. Dans son ouvrage classique Knuth donne une série de tests à effectuer sur une séquence de nombres pour vérifier son caractère *pseudo* aléatoire et nous allons étudier quelquns d'entre eux.

1. Le *test de Kolmogorov-Smirnov* compare de façon qualitative la distribution obtenue numériquement à la distribution (uniforme dans le cas de l'algorithme module) attendue théoriquement.

   Soit ℓ la longueur du cycle et n un autre nombre entier plus petit ou égale à ℓ, $n \leq \ell$. On divise l'intervale [0,1] en n sousintervales de longueur $1/n$.

Comparer graphiquement le diagramme en fréquence trouvé numériquement, $F_n(x)$, à la distribution théorique, $F(x)$,

$$\begin{aligned} F_n(x) &= \frac{\text{nombre de } \; y_1, y_2, \ldots, y_n}{n} \\ F(x) &= \int_0^x dx P(x) = x \quad \text{for} \;\; x \in [0,1] \,, \end{aligned} \tag{101}$$

où $y_1, \ldots, y_n$ sont les valeurs calculés pour le choix de paramètres: $m = 10000, a = 11, c = 31$, et pour plusieurs valeurs de n, par exemple, $n = 10, 100, 1000$. Discuter le résultat de façon qualitative.

2. Le *test spectral* sert à étudier les propriétés de la distribution de probabilité jointe de t élḿents de la séquence. On a vérifié que tous les bons générateurs connus satisfaient ce test et tous les mauvais générateurs ne le passent pas.

   (a) Générez une séquence $\underline{x}$ de longueur $n = 1000$ en utilisant la récurrence $x_i = (137x_{i-1} + 187) \bmod m$. Diviser les nombres en pairs (x_{k-1}, x_k) et associer chaque pair à un vecteur en deux dimensions. Tracez un *scatter plot* sur le plan, avec chaque point correspondant à un couple. Qu'observe t-on ? Peut-on couvrir tous les points avec des lignes droites parallèles ?

   Repetez l'étude de la façon suivante : regroupez les nombres en triplets (x_{k-2}, x_{k-1}, x_k), associez chaque triplet à un vecteur en trois dimensions et tracez le *scatter plot* en trois dimensions.

   (b) Générez une séquence $\underline{x}$ de longueur $n = 10000$ en utilisant la récurrence (100). Diviser les 10000 nombres en 5000 groupes de 2 nombres chacun. Associez chaque couple de nombres

   On étudiera les choix suivants :

      i. $a = 77$, $c = 31$, $m = 10000$, $x_0 = 2$. Repetez pour d'autres valeurs de la semence, par exemple, $x_0 = 105, 21, 5778$.

      ii. $a = 173$, $c = 35$, $m = 10000$, $x_0 = 2$. Repetez pour d'autres valeurs de la semence.

      iii. $a = 101$, $c = 33$, $m = 10000$, $x_0 = 2$. Repetez pour d'autres valeurs de la semence.

      iv. Que remarque-t-on à simple vue ?

On serait facilement emmené à conclure que ce générateur est très mauvais. Pourtant, il faut remarquer que les nombres réels doivent être nécessairement tronqués par un ordinateur. Donc, un plot d'une distribution uniforme de nombres réels aura, elle-aussi, un "grain" périodique.

Une méthode pour vérifier si le grain est acceptable où non est de calculer la distance maximale entre toutes les familles de droites (plans, hyperplans) qui couvrent les point en deux (trois, plus hautes) dimensions. Si cette distance ne dépend pas de la dimension la séquence est *pseudo* aléatoire. Au contraire, si elle diminue avec la dimension de l'espace, la séquence a une périodicité.

2.3 Nombres aléatoires Gaussiens

Un algorithme pour générer des nombres *pseudo* aléatoires distribués de façon Gaussienne est le suivant :

i. Générer deux nombres *pseudo* aléatoires réels U_1 et U_2 distribués uniformément sur l'intervale [0,1].

ii. Transformer ces nombres en $V_1 = 2U_1 - 1$ et $V_2 = 2U_2 - 1$. V_1 et V_2 sont distribués uniforme ment en $[-1, 1]$.

iii. Calculer $S \equiv V_1^2 + V_2^2$. (S est une distance carrée.)

iv. Si $S \geq 1$ aller au point i. Autrement continuer.

v. Si $S = 0$ on prend $X_1 = X_2 = 0$. Autrement, on prend

$$X_1 = V_1\sqrt{\frac{-2\ln S}{S}}\,, \qquad X_2 = V_2\sqrt{\frac{-2\ln S}{S}}\,. \tag{102}$$

X_1 et X_2 ont une distribution Gaussienne avec moyenne zéro et déviation standard égale à un.

Utiliser ce générateur, tracer la distribution et comparer graphiquement à la lois théorique.

3 Travail supplémentaire

3.1 La séquence de Fibonacci

Évidement on peut envisager d'améliorer le générateur module de plusieurs façon. Une d'entre elles est d'allonger la récurrence Deux exemples sont

$$x_i = (x_{i-1} + x_{i-2}) \bmod m \qquad \text{récurrence de Fibonacci, et}$$
$$x_i = (x_{i-24} + x_{i-55}) \bmod m\ , i \geq 55$$

avec m pair et les premiers nombres $x_0, \ldots x_{54}$ entiers pas tous pairs.

Étudier ces deux séquences avec les tests KS et spectral. Montrer numériquement que la séquence de Fibonacci n'est pas du tout aléatoire tandis que la deuxième est bien plus *pseudo* aléatoire et a un cycle de longueur supérieure à m.

<u>Références :</u>

D. Knuth, *The art of computer programming* Vol. 2, 1981.
J. M. Thijssen, *Computational physics*, Cambridge Univ. Press.
M. P. Allen et D. J. Tildesley, *Computer simulations of liquids*, Oxford Univ. Science Publications.

D. P. Landau et K. Binder, *A Guide to Monte Carlo simulations in statitisal physics*, Cambridge Univ. Press.
M. E. J. Newman et G. T. Barkema, *Montecarlo methods in statistical physics*, Clarendon Press, Oxford Univ. Science Publications.

References

[1] M. Sternstein, *Statistics*, Barron's college review series, 1996.

[2] G. Benedek et F. M. H. Villars, *Physics with illustrative examples from medicine and biology, Statistical physics*, Biological physical series, AIP Press,

[3] D. Downing et J. Clark, *Statistics, the easy way*

[4] Couty, F., Debord, J., Fredon, D. Probabilités et statistiques : DEUG SV et ST, BTS biologiques et agricoles : résumés de cours, 157 exercices et problèmes corrigés (Masson , 1996).

[5] Combrouze, A. Dédé, A. Probabilités et statistiques. 1. cours, exercices et corrigés (Presses Universitaires de France , 1996).

[6] Roque, Jean-Louis Progresser et réussir en Mathématiques . [1] [Probabilités et statistiques] Probabilités et statistiques (EditeurParis : Espace Etudes Editions, 2000).

[7] Bouyer, J. Méthodes statistiques : médecine - biologie. Exercices corrigés (Editions ESTEM , 1999).

[8] Istas, Jacques Probabilités et statistiques : cours, exercices et problèmes résolus (Ellipses , 1999).

[9] Falissard, Bruno Comprendre et utiliser les statistiques dans les sciences de la vie (Masson, 1996).

[10] Harthong, Jacques. Probabilités & statistiques : de l'intuition aux applications (Diderot , 1996).

[11] Vigneron, C. Logak, E. Probabilités & statistiques. 1 (Diderot Editeur , 1995).

[12] Degrave, D. Degrave, C. Probabilités, statistiques (Bréal, 1995)

[13] A very complete website is http://www.statsoftinc.com/textbook/

Those of you interested in the histoty of science can have a look at
The lady tasting tea, D. Salsburg, (Owl books, 2001).