

# A brief introduction to Statistics

*\*\*\*preliminary version\*\*\* use at your own risk*

P. Oliva

July 15, 2009

# Contents

<b>1</b>	<b>Fundamental definitions</b>	<b>3</b>
1.1	Conditioned probability . . . . .	6
1.2	Expected value and Cumulative Function . . . . .	7
1.3	Momentum of order $n$ th and Variance . . . . .	8
<b>2</b>	<b>Estimators and momenta</b>	<b>10</b>
2.1	Central estimators . . . . .	10
2.2	Moment-Generating Function . . . . .	12
<b>3</b>	<b>Probability distributions</b>	<b>13</b>
3.1	Binomial distribution . . . . .	13
3.2	Poisson distribution . . . . .	15
3.3	Gauss distribution . . . . .	16
3.3.1	More over likelihood . . . . .	17
3.4	Asymptotic behavior . . . . .	18
3.5	Further distributions . . . . .	19
<b>4</b>	<b><math>\chi^2</math> distribution</b>	<b>20</b>
4.1	Gamma function . . . . .	20
4.2	$\chi^2$ variable . . . . .	21
4.3	$\chi^2$ p.d.f. . . . .	21
4.4	$\chi^2$ test . . . . .	22
<b>5</b>	<b>Law of large numbers and the Central limit theorem</b>	<b>23</b>
5.1	The Čebyšev's inequality . . . . .	23
5.2	Law of large numbers . . . . .	24
5.3	Central limit theorem . . . . .	24
<b>6</b>	<b>Parameter Estimation and Maximum likelihood estimation (MLE)</b>	<b>25</b>
6.1	The maximum likelihood function . . . . .	26
6.2	Consistency, distortion, efficiency . . . . .	26
6.3	Estimators for the Gaussian . . . . .	28
6.4	Estimators for the Poisson . . . . .	28
6.5	Estimators for the Bernoulli . . . . .	29
6.6	Variance of the $s^2$ estimator . . . . .	29

# 1 Fundamental definitions

It is very difficult to give out a pure mathematical definition of probability which doesn't use itself the concept of probability. It is easy to see that the **combinatory definition** of probability (Laplace)

$$p = \frac{\# \text{ of favorable occurrences}}{\# \text{ of total possible outcomes in an equiprobable sample space}}$$

or the **frequentist definition**

$$p \approx \frac{\# \text{ of cases favorable}}{\# \text{ of total attempt}}$$

(when the number of runs is high and all the attempts are made in an equiprobable conditions), are both using themselves the concept of probability, asking for the equiprobability of a set of events because of the requested symmetry of the event-extraction mechanism<sup>1</sup>.

In order to avoid such a tautology we can surely consider in a very intuitive way the concept of *probability zero* that an event can happen. That is to say if an event never showed up in the whole history of the universe we say this event has zero probability to be seen<sup>2</sup>. This “frequentist view” is not accepted by all: there is another point of view called the “subjective” one, the so called Bayesian probability which interprets the concept of probability as “a measure of a state of knowledge”. We won't get through this distinctions but it's obvious that the second approach is *probably* the most used one in the real life when we decide what to do and what to bet.

Is not difficult to recognize that the followed method in every-day-life is the one so well described by philosopher Hume:

*“There is certainly a probability, which arises from a superiority of chances on any side; and according as this superiority increases, and surpasses the opposite chances, the probability receives a proportionable increase, and begets still a higher degree of belief or assent to that side, in which we discover the superiority. If a dye were marked with one figure or number of spots*

---

<sup>1</sup>This is called *principle of indifference* (also called principle of insufficient reason) and it's due to Laplace in contraposition with the principle of sufficient reason of Leibnitz.

<sup>2</sup>Also this concept is dangerous: while we would assign null probability to an event that we think is impossible (i.e. the sentence: *tomorrow the sun won't rise*) the opposite is not true! That is to say even if an event is practically “impossible” this doesn't mean *it has* zero probability. Think to all the things that normally happen everyday and try to figure out what was the probability that *yesterday* you would have given to the fact that *you are now* reading this line, dressed like you are now, at this precise moment, etc.. It is obvious that *something that has happened has probability one* and that's it. But if something with very very small probability should “never” happen then nothing would happen at all: the life is continuously made by an infinity of little things with very small probabilities that summed together comes to happen.

*on four sides, and with another figure or number of spots on the two remaining sides, it would be more probable, that the former would turn up than the latter; though, if it had a thousand sides marked in the same manner, and only one side different, the probability would be much higher, and our belief or expectation of the event more steady and secure. This process of the thought or reasoning may seem trivial and obvious; but to those who consider it more narrowly, it may, perhaps, afford matter for curious speculation.”*

In 1993 the International Organization for Standardization (ISO) published a “Guide to the expression of uncertainty in measurement” where we can read:

In contrast to this frequency-based point of view of probability, an equally valid viewpoint is that probability is a measure of the degree of belief that an event will occur. For example, suppose one has a chance of winning a small sum of money  $D$  and one is a rational bettor. One's degree of belief in event  $A$  occurring is  $p = 0.5$  if one is indifferent to this two betting choices: (1) receiving  $D$  if event  $A$  occurs but nothing if it does not occur; (2) receiving  $D$  if event  $A$  does not occur but nothing if it does occur. Recommendation INC-1 (1980) upon which this Guide rests implicitly adopts such a viewpoint of probability since it views expressions such as equation (E.6) as the appropriate way to calculate the combined standard uncertainty of a result of a measurement.

where the cited equation E.6 refers to systematic and statistical uncertainties propagation<sup>3</sup>.

The object of interest in the whole theory of probability is to study the casual events which are defined as phenomenons we can repeat as much times as we want (at least in principle), that are individually nonpredictable and mutually exclusive (they cannot occur at the same time, like having two values in one throw of a single dice or a coin and so on).

The ensemble of all the possible events is called the event space  $I$  and each subset  $A$ ,  $B$ , etc. of  $I$  is a certain group of possible results that we take as *the event* we look at. The empty set  $\emptyset$  and the set  $I$  itself are members of  $I$  being conversely the *impossible* event and the *sure* event. If is possible to fix a correspondence law that assign to each element of a subset  $A \in I$  one and only one real (or discrete) value  $x$ , then  $x$  is called a casual variable defined over  $A$ .

---

<sup>3</sup>the most recent update is ISO/IEC Guide 98-3:2008

In the most simple case we can label the possible events with a integer and positive number  $k = 1, 2, \dots, n$ . Let's consider the case those events are mutually exclusive, thus they cannot occur at the same time, and that for sure one event at least will occur. Following conditions then apply for the probability  $p_k$  of the event labelled with the integer  $k$ :

$$p_k \geq 0 \quad \forall k \quad ; \quad \sum_{k=1}^n p_k = 1$$

and for each subset  $A$  of positive integers we can associate the probability that one event of the subset will show up  $P(A) \equiv \sum_{k \in A} p_k$ . With this definition we can demonstrate the following fundamental properties being  $A$  and  $B$  subsets of  $I$ :

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(A) &\geq 0 \quad ; \quad P(\emptyset) = 0 \quad ; \quad P(I) = 1 \end{aligned} \quad (1)$$

where  $I$  is the set of all numbers  $n$ .

From this point is possible to build up a theory starting from the  $P(A)$  quantities; in general, in fact, the events can be labeled with a real number  $x$  so that the probability is

$$P(A) \equiv \int_{x \in A} dP(x) = \int_{x \in A} p(x) dx \quad (2)$$

where  $p(x)$  is a generalized function (i.e. could contain a Dirac delta function).

Let's now watch again the first equation of (1): is clear that we had to subtract the intersection between  $A$  and  $B$  in order to avoid a double count. If there are no ways  $A$  and  $B$  could ever be both true, is clear that the intersection  $A \cap B$  is zero. In this case we say that the two sets are *incompatible*:

$$P(A \cap B) = P(\emptyset) = 0 \quad (3)$$

This consideration (3) seems trivial but indeed is crucial in the whole theory which is based on the fact that the probability  $P$  of an event is between zero and one, and on the possibility to "cover" the entire event space with a series of non intersecting sets (that is to say it exists a complete selfexcluding, or incompatible covering sets)  $H_i$  so that  $\forall i \neq j$

$$H_i \cap H_j = \emptyset \quad ; \quad \bigcup_{i=1}^n H_i = I \quad (4)$$

It follows that  $P(A) = \sum_{i=1}^n P(A \cap H_i)$  and the intersection between the  $H_i$  and  $A$  is exactly the whole set  $A$ ,  $\forall A$  chosen.

## 1.1 Conditioned probability

Now we define<sup>4</sup>

$$P(A \cap H) = P(H)P(A|H) \Rightarrow P(A|H) \equiv \frac{P(A \cap H)}{P(H)} \quad (5)$$

which holds only if  $P(H) \neq 0$ . The symbol  $P(A|H)$  means the probability that  $A$  will show up once we are aware that  $H$  is true (which is different from saying that  $H$  causes  $A$  or  $H$  occurs before  $A$ ). Naturally we must have that  $P(A|A) = 1$  whatever event  $A$  we choose. Note also that the conditioning symbol  $|$  is in general not symmetric.

Now it's said two subsets  $A$  and  $B$  to be independent if

$$P(A \cap B) = P(A) \cdot P(B) \quad (6)$$

which is pretty much different from the disjoint condition (4)! If now  $A$  and  $B$  are independent, using (6) and (5), we find that  $P(A|B) = P(A)$  and  $P(B|A) = P(B)$ . If instead we find that, in example,  $P(A|B) > P(A)$  or  $P(A|B) < P(A)$ , than we say that  $A$  and  $B$  are conversely positively and negatively correlated.

Let's now be  $H_k$  a complete selfexcluding ensemble in the sense of (4). We can thus write for the probability  $P(E)$  that

$$P(E) = P(1 \cap E) = P\left(\bigcup_{k=1}^n H_k\right) \cap E = \sum_{k=1}^n P(H_k \cap E) = \sum_{k=1}^n P(H_k) \cdot P(E|H_k)$$

where we used identity (5) in the last equivalence; the above equation holds also only for the arguments  $P(H_k \cap E) = P(H_k) \cdot P(E|H_k)$ , so that we can write, using (5) again,

$$P(H_k|E) = \frac{P(H_k)P(E|H_k)}{P(E)} \Rightarrow \frac{P(H_k|E)}{P(H_k)} = \frac{P(E|H_k)}{P(E)} \quad (7)$$

for the single  $k$ . This is called **Bayes' theorem**. Note that the value of denominator  $P(E)$  does not depend on  $k$  and can be easily found by normalizing  $\sum_k P(H_k|B) = 1$ . If we want to identify the  $H_k$  with the all the possible (selfexcluding) causes for  $E$  to happen, we can read the (7) as the probabilities of the causes of event  $E$ . But we'll see it in more detail later.

In the real life is useful to use this point of view:

$$P(E) = \frac{\sum_k P(H_k) \cdot P(E|H_k)}{\sum_k P(H_k)}$$

---

<sup>4</sup>To be honest this is not a definition but a consequence of a definition or even a theorem. Nobody really uses in practice the (5) as a definition because nobody want the situation in which both  $P(H)$  and  $P(A \cap H)$  are to be evaluated everytime. Here for sake of simplicity we don't want to go into rigorous definitions and theoretic details.

whit  $\sum_k P(H_k) = 1$  that allows us to see the probability of the event  $E$  as the average over its conditioned probabilities of the events belonging to a complete class (the class  $H_k$ ) weighted with the probabilities of the conditioning. Then the version that probably you will use of the Bayes' theorem is the one we can write when a complete class of hypothesis  $H_k$  (with the (4) valid) is considered, then

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{\sum_k P(E|H_k) \cdot P(H_k)} \quad (8)$$

Different forms of the Bayes' theorem give us different ways of interpreting it: equation (7) tells us that  $P(A_k)$  is modified by the  $B = true$  hypothesis of the same factor  $P(B)$  is modified by the  $A_k = true$  one. On the other side, equation (8) tells us that the *a posteriori* probability  $P(H_i|E)$  that  $H_i$  will happen under the hypothesis  $E = true$  is proportional to the *a priori* probability  $P(H_i)$  that is the initial probability conditioned by all the other preliminary hypothesis  $H_0$  excluding  $E$  event (we should write  $P(H_i|H_0)$  indeed); the proportion between the *a posteriori* and the *a priori* probabilities is the factor  $P(E|H_i)$  also called likelihood which gives a measure of how  $E$  is likely to happen because of  $H_i$ . This likelihood relies on the knowledge of all the other hypothesis  $H_0$  (other than  $E$ ).

So in a short way we can say that the Bayes' theorem gives us this proportion: (*a posteriori prob.*)  $\propto$  (likelihood)  $\cdot$  (*a priori prob.*). Thus, to have zero final probability is sufficient that the likelihood is zero while to have the final probability equal to 1 is necessary that the likelihood is different from zero *and that the a priori probability is 1* (so to consider definitive a theory is needed to consider it definitive since the very beginning which is of course not acceptable. It follows that any theory will be temporary until the end of humankind).

## 1.2 Expected value and Cumulative Function

Before to procede we must give out some indispensable definitions of several values we can obtain from the series or from the integrals (in case of a continue variable) of functions depending from our stochastic variable  $x$  which can assume the values  $x_k = (x_1, x_2, \dots, x_k)$  with probabilities  $P_k \equiv P(x_k)$  or  $x \in [a, b]$  and *probability density function*  $f(x)$  (from now on p.d.f.) for a continue variable. The p.d.f. function being defined in a complete class of events must satisfy the following conditions:

$$0 \leq f(x) \leq 1$$

$$P(x \in A \cup x \in B) = f(x \in A) + f(x \in B)$$

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

First of all we define the **expected value** of  $x$  (or simply the population mean) being

$$E[x] \equiv \sum_{k=1}^n x_k P_k$$

$$E[x] \equiv \int_a^b x f(x) dx \quad (9)$$

for the discrete and continuous case respectively (this value is by convention indicated with the letter  $\mu$ ). This definition holds if integral (or series) is absolutely convergent in (9). The quantity  $f(x)dx$  in the continue variable case is called *probability density* and gives the probability to find the variable  $x$  within the interval  $[x, x + dx]$ ; is clear that  $\int_A P(x)dx = 1$  if the region of integration  $A$  refers to the entire range of variable  $x$ . Is also trivial to extend the function  $f(x)$  outside the eventual range of the  $x$  variable defining  $f(x) = 0$  outside this range in a way that is always possible to write

$$E[x] \equiv \int_{-\infty}^{+\infty} x f(x) dx$$

It may be useful to define another function that gives you the probability to find your variable with a value less than or equal to  $x$ , the so-called **cumulative function**

$$F(x) \equiv \int_{-\infty}^x f(x') dx' \quad (10)$$

so that the probability density function is simply  $f(x) = \frac{\partial F(x)}{\partial x}$  (if  $F(x)$  is well-behaving).

The cumulative function has several useful properties, in example

$$0 \leq F(x) \leq 1$$

$$\lim_{x \rightarrow -\infty} F(x) = 0 \quad \lim_{x \rightarrow \infty} F(x) = 1$$

$$\lim_{\varepsilon \rightarrow 0} F(x + \varepsilon) = F(x)$$

### 1.3 Momentum of order $n$ th and Variance

Generalizing the definition (9) we can define the general  $n$ th momentum around the value  $x_0$  (not necessary 0 or  $\mu$ ) as

$$M_{x_0}^n[x] \equiv E[(x - x_0)^n] = \int_{-\infty}^{+\infty} (x - x_0)^n f(x) dx \quad (11)$$

then it's clear that the mean as defined in (9) is nothing but the 1st momentum around  $x_0 = 0$ ,  $M_0^1[x]$ . Another very important momentum is the

2nd momentum around the expected value,  $M_{E[x]}^2$ , (the momenta around the expected value are called *central* momenta, we will discuss more about momenta in the next section), the so called **variance of  $x$** :

$$\begin{aligned} \text{Var}[x] &= \sum_k (x_k - E[x])^2 P_k \\ \text{Var}[x] &= \int_{-\infty}^{+\infty} (x - E[x])^2 f(x) dx \end{aligned} \quad (12)$$

for both discrete and continuous case, moreover it's very useful to note that one can calculate the variance very easily, time to time, by only evaluating the expected value of the  $x^2$  ( $E[x^2]$ ), having already calculated the  $E[x]$  because

$$E[(x - E[x])^2] = \sum_k (x_k^2 - 2E[x]x_k + E[x]^2)P_k = E[x^2] - E[x]^2 \quad (13)$$

It's easy to prove the variance has nice properties such

$$\text{Var}[a] = 0 \quad \forall a = \text{const.}$$

$$\text{Var}[ax] = a^2 \text{Var}[x]$$

and so forth. The second one can be generalized to any linear combination with constant coefficients:  $\text{Var}[\sum_i a_i x_i] = \sum_i a_i^2 \text{Var}[x_i]$ . The demonstration is trivial but quite boring and we leave it to the reader.

Finally we can extend to the case of a variable  $\vec{x} = (x_1, x_2, \dots, x_n)$  with dimension  $n$ . The expected value is still

$$E[\vec{x}] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \vec{x} f(\vec{x}) dx_1 dx_2 \dots dx_n$$

and the variance

$$\text{Var}[\vec{x}] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} (\vec{x} - E[\vec{x}])^2 f(\vec{x}) dx_1 dx_2 \dots dx_n$$

In the two dimensions case is often needed to calculate the **covariance**

$$\begin{aligned} V_{xy} &= E[(x - E[x])(y - E[y])] = E[xy] - E[x]E[y] = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f(x, y) dx dy - E[x]E[y] \end{aligned} \quad (14)$$

the matrix  $\text{cov}[a, b]$  is by construction symmetric in  $a$  and  $b$  and the diagonal elements are the variances  $\text{Var}[a]$ ,  $\text{Var}[b]$ , etc. To give an estimation of how much two random variables are correlated is used the **correlation coefficient**

$$\rho = \frac{V_{xy}}{\sqrt{\text{Var}[x]\text{Var}[y]}} \quad (15)$$

## 2 Estimators and momenta

We already saw that given a set of values from any stochastic variable  $x$  we can define the expected value and the variance only if we know (or we hypothesizes) the p.d.f.; in the real life instead, there are cases in which the p.d.f. is not known. In those cases we can always calculate some estimators to have some information about the behavior of our unknown p.d.f.. Thus we define, given a set of numerical values  $x_k$  from our variable, the following **central estimators**:

### 2.1 Central estimators

*The mode* is the value  $\tilde{x}$  that is repeated more often than any other in a distribution. Is not always possible to define it (think to a uniform distribution like the 6-faces dice which have probability 1/6 each one to show up in a throw).

*The arithmetic mean* is the most important central tendency estimator and also the most famous. Typically it is indicated with  $\bar{x}$  and it is defined as

$$\bar{x} \equiv \frac{1}{N} \sum_{k=1}^N x_k \quad (16)$$

Between all the other properties we underline two essential qualities of the arithmetic mean: the sum of the deviations, of all the values of  $x$ , from their arithmetic mean, is zero

$$\sum_{j=1}^N (x_j - \bar{x}) = 0$$

and it's very easy<sup>5</sup> to prove that the arithmetic mean is the value that minimize the sum of the squared deviations between the values and any other value  $\langle x \rangle$  whatever defined

$$\text{minimum} \left\{ \sum_{j=1}^N (x_j - \langle x \rangle)^2 \right\} = \sum_{j=1}^N (x_j - \bar{x})^2$$

*The median* the median  $\hat{x}$  is that value splitting the higher half of a sample, a population, or a probability distribution, from the lower half. It's possible to prove (but we don't do it here) that the median minimize the sum of the absolute deviations between the values  $x_i$  and any

---

<sup>5</sup>  $\sum_{j=1}^N (x_j - \langle x \rangle)^2 = \sum [(x_j - \bar{x}) + (\bar{x} - \langle x \rangle)]^2 = \sum (x_j - \bar{x})^2 + \sum (\bar{x} - \langle x \rangle)^2 \geq \sum (x_j - \bar{x})^2$

other value  $\langle x \rangle$

$$\text{minimum} \left\{ \sum_{j=1}^N |x_j - \langle x \rangle| \right\} = \sum_{j=1}^N |x_j - \hat{x}|$$

**The geometric mean** indicates the central tendency of a set of numbers.

It is similar to the arithmetic mean except that instead of adding the set of numbers the numbers, those are multiplied and then the  $n$ th root of the resulting product is taken.

$$\bar{x}_G \equiv \left( \prod_{k=1}^N x_k \right)^{\frac{1}{N}}$$

This mean is use in example to test a laboratory balance with two harms and a weighing pan suspended from each arm: if you measure your unknown mass by placing it in one pan, and standard masses are added to the other pan until the beam is as close to equilibrium as possible, you will have a weight  $M_1$  which includes the balance's systematic error. To take care of the systematic error you have to weight again your object using the other pan switching the standards and the unknown mass each other. Then you will get a second value  $M_2$  for your unknown mass  $M_x$ . The right value is then the geometric mean  $M_x = \sqrt{M_1 M_2}$ .

**The harmonic mean** The harmonic mean  $\bar{x}_H$  is defined as

$$\bar{x}_H \equiv \left( \frac{1}{N} \sum_{k=1}^N \frac{1}{x_k} \right)^{-1}$$

and typically is appropriate for situations when the average of rates is desired.

There are other kinds of mean but we won't go into details for the previous defined ones are the most important and the most used. In the symmetric and monomodal distributions the mode, the median and the mean are coincident while for asymmetric distributions they are different. An empirical relation holds for distributions which are not too much irregular:

$$\bar{x} - \tilde{x} \simeq 3(\bar{x} - \hat{x})$$

which is useful for a quick check. Moreover is always true that  $\bar{x}_H \leq \bar{x}_G \leq \bar{x}$ .

As already said, the arithmetic mean is the most important because for a large enough entries holds the following:

$$\lim_{N \rightarrow \infty} \bar{x} = E[x]$$

and this happen when the probability is viewed as the limit of the frequencies for high numbers. Of course in general  $\bar{x}$  and  $E[x]$  are different being  $E[x]$  the theoretical mean value *over the entire population* while  $\bar{x}$  is the mean over a particular subset of the population, namely *the sample*. Naturally the expected value of the mean is equal to to the expected value ( $E[\bar{x}] = E[x]$ , it directly follows by the definition).

## 2.2 Moment-Generating Function

We want to recall now the definition (11) for the generic momentum: it is nice to know that there is a function which generates all the momenta, and is called the Moment-Generating function  $G(t)$ . If we define

$$G(t) = E[\exp(xt)] \quad (17)$$

as the generating function then we can easily find out that if it exists on an open interval around  $t = 0$ , then the coefficients of the MacLaurin series are all the momenta around zero of the p.d.f.:

$$\begin{aligned} G(t) &= E[e^{xt}] = 1 + E[xt] + \frac{1}{2}E[x^2t^2] + \dots + \frac{1}{k!}E[x^kt^k] + \dots = \\ &= 1 + t \int_{-\infty}^{+\infty} xf(x)dx + \dots + \frac{t^k}{k!} \int_{-\infty}^{+\infty} x^k f(x)dx + \dots = \\ &= M_0^0[x] + tM_0^1[x] + \frac{t^2}{2}M_0^2[x] + \dots \end{aligned}$$

so that in general we can calculate

$$\left. \frac{d^k}{dt^k} G(t) \right|_{t=0} = E[x^k] \equiv M_0^k[x]$$

and more generally

$$\left. \frac{d^k}{dt^k} E[\exp((x - E[x])t)] \right|_{t=0} \equiv M_{E[x]}^k[x] \quad (18)$$

It's useful to note that a p.d.f that is symmetric respect the expected value has all its central momenta (around the expected value) of odd order (1, 3, 5, ...) equal to zero, for this reason sometimes is helpfull to check wheather the 3rd central momentum (called the *skewness*) vanishes or not, to estimate the asymmetry of a given p.d.f. while the next momentum, the 4th central momentum (called kurtosis -  $\kappa\nu\rho\tau\acute{o}\varsigma$  in greek means curved, bent, round -) is useful to estimate how much the p.d.f. is flat or not, that is to say is a measure of the peakedness of the p.d.f.. The normal distribution that we will meet in the next section has by convention skewness and kurtosis coefficients equal to zero, so that the normal distribution is taken as standard for all the others.

### 3 Probability distributions

In this section we will see the most important p.d.f.s conversely the Binomial (Bernoulli), the Poisson and the Gauss distributions which are the most used ones. We will also see minor distributions that are somehow important for the applications.

#### 3.1 Binomial distribution

Also known as Bernoulli distribution is used when we deal with a dicotomic variable (which can only take two values, true and false, 0 and 1, etc.) with probabilities  $p$  and  $(1 - p)$  constant during the experiment (what is called a Bernoulli variable). The results of each run of the experiment are independent from the previously obtained results (they are not correlated).

The probability that in  $N$  repeated runs of the same experiment the wanted event  $A$  with probability  $p$  will show up  $n$  times against the probability  $(1 - p)$  to have the other event  $\neg A$  (all we choose to be our non-event) is

$$P_{N,p}(n) = \binom{N}{n} p^n (1 - p)^{N-n} \quad (19)$$

where

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

The normalization is simply verified:

$$\sum_{n=0}^N \frac{N!}{(N - n)!n!} p^n (1 - p)^{N-n} = 1 \quad (20)$$

because the terms are nothing but those from the develops of the binomy  $[p + (1 - p)]^N = 1$ .

The expected value of the (19) is

$$\begin{aligned} E[n] &= \sum_{n=0}^N n P_{N,p}(n) = \sum_{n=0}^N \frac{N!n}{(N - n)!n!} p^n (1 - p)^{N-n} = \\ &= Np \sum_{n=0}^{N-1} \frac{(N - 1)!}{(N - n)!(n - 1)!} p^{n-1} (1 - p)^{N-n} = \\ &= Np \sum_{k=0}^{N-1} \frac{(N - 1)!}{(N - k - 1)!k!} p^k (1 - p)^{N-k-1} = \\ &= Np \sum_{k=0}^{N-1} \binom{N - 1}{k} p^k (1 - p)^{(N-1)-k} = Np \end{aligned}$$

where  $k \equiv n - 1$  and the last equivalence holds because of the normalization (20). The variance is easily calculated using (13):

$$Var[n] = E[n^2] - E[n]^2 = E[n^2] - N^2p^2$$

so that we have to calculate  $E[n^2]$ , which is

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

where again we called  $k = n - 1$ .

If now we note that  $\sum_{n=0}^{N-1} \frac{(N-1)!(k+1)}{(N-1-k)!k!} p^n (1-p)^{N-1-k} = E[k+1]$  we can rename again  $k \equiv n$  without any problems just to write  $E[n^2] = NpE_{(N-1)}[n+1] = Np \cdot [E_{(N-1)}[n]+1] = Np[(N-1)p+1] = N^2p^2 + Np(1-p)$  (where  $E_{(N-1)}$  means that the summatory is from 1 to  $N-1$ ) so that

$$Var[n] = Np(1-p)$$

If we would have used the Moment-generating function we'd have wrote  $G(t) = \sum_{n=0}^N P_{N,p}(n)e^{tn}$  and, by factorizing  $pe^t$ , it would have been easy to see that  $G_n(t) = (pe^t + q)^N$  where  $q \equiv (1-p)$ . Then we would had only to calculate

$$E[n] \equiv M_0^1 = \left. \frac{dG(t)}{dt} \right|_{t=0} = Np$$

$$M_0^2 = \left. \frac{d^2G(t)}{dt^2} \right|_{t=0} = Np + N(N-1)p^2 \Rightarrow Var[n] = M_0^2 - (M_0^1)^2 = Npq$$

The Bernoulli distribution (19) can be generalized to a multinomial distribution where the possible results are more than two. Let's say we have  $k$  events with  $p_1, p_2, \dots, p_k$  probabilities. Then (if we take care that  $\sum_{i=0}^k p_i = 1$ ) we can write

$$P_{N,p_j}(n_j) = \frac{N!}{\prod_{j=1}^k n_j!} \prod_{j=1}^k p_j^{n_j}$$

with the only condition  $\sum n_j = N$ . We do not prove the expected value and the variance, we only give them:

$$E[n_j] = Np_j \quad ; \quad Var[n_j] = Np_j(1-p_j)$$

In the special case of three different results ( $i, j$  and everything else) we find that the covariance  $V_{ij} = E[(n_i - E[n_i])(n_j - E[n_j])] = -Np_i p_j$ .

### 3.2 Poisson distribution

If we take a binomial distribution and we push the limit for  $N \rightarrow \infty$  and  $p \rightarrow 0$  with the condition  $Np = \text{const.} \equiv a$  (that's very important!), then the binomial distribution tends to

$$P_a(m) = \frac{a^m}{m!} e^{-a} \quad (21)$$

which is called the Poisson distribution with the constant  $a > 0$  called the parameter of Poisson law. The variable in the Poisson distribution is a casual discrete one, expressing the probability to see a certain number of events in a determined period of time if those events are belonging to a common identical distribution and are uncorrelated each other.

The (21) describes a casual discrete variable that assumes only non negative integer values. We must now prove what we said about the limit: remembering the condition  $Np = \text{const.}$  in the limit we write

$$P_a(m) = \frac{a^m}{m!} e^{-a} = \lim_{N \rightarrow \infty} \binom{N}{m} p^m (1-p)^{N-m}$$

let's massage the Bernoulli one: if  $Np = a$  then  $p = a/N$

$$\begin{aligned} & \frac{N!}{(N-m)!m!} \left(\frac{a}{N}\right)^m \left(1 - \frac{a}{N}\right)^{N-m} = \\ &= \frac{a^m}{m!} \frac{N(N-1)(N-2)\cdots(N-m+1)}{N^m} \frac{\left(1 - \frac{a}{N}\right)^N}{\left(1 - \frac{a}{N}\right)^m} = \\ &= \frac{a^m}{m!} \left(1 - \frac{a}{N}\right)^N \frac{\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right)}{\left(1 - \frac{a}{N}\right)^m} \end{aligned}$$

now let's remember that  $\lim_{N \rightarrow \infty} \left(1 - \frac{a}{N}\right)^N = e^{-a}$  and pushing the limit we see that all the members in the productory in the second part of the last result will tend to one, so that

$$\lim_{N \rightarrow \infty} \frac{a^m}{m!} \left(1 - \frac{a}{N}\right)^N \frac{\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right)}{\left(1 - \frac{a}{N}\right)^m} = \frac{a^m}{m!} e^{-a} \equiv P_a(m)$$

and our prove is made.

Now let's check out that the (21), as all the others probability functions, is normalized: if we sum over we have

$$\sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a} = e^{-a} \left(1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \cdots\right) = e^{-a} e^a = 1$$

as we expected.

Now we calculate the expected value:

$$E[k] = \sum_{k=0}^{\infty} k \frac{a^k}{k!} e^{-a} = e^{-a} \sum_{k=1}^{\infty} a \frac{(a)^{k-1}}{(k-1)!} \equiv ae^{-a} \sum_{j=0}^{\infty} \frac{a^j}{j!} = ae^{-a} e^a = a$$

and for the variance we first calculate  $E[n^2]$  with the same procedure and we subtract  $E[n^2] - E[n]^2$ ; we write then

$$\begin{aligned} E[k^2] &= \sum_{k=0}^{\infty} k^2 \frac{a^k}{k!} e^{-a} = ae^{-a} \left\{ \sum_{j=0}^{\infty} j \frac{a^j}{j!} + \sum_{j=0}^{\infty} \frac{a^j}{j!} \right\} = \\ &= ae^{-a} [ae^a + e^a] = a^2 + a \Rightarrow \text{Var}[n] = a^2 + a - a^2 = a \end{aligned}$$

Then both the expected value and the variance are equal to the parameter  $a = Np$ . We won't do the calculations but it's easy to use for the same result the moment generating function  $G(t) = E[e^{mt}] = e^{-a} e^{ae^t}$  and the derivatives respect  $t$  calculated in  $t = 0$  shall give you the moments around 0.

### 3.3 Gauss distribution

The Gaussian distribution is the most important ever. It is the limit law which all the other distribution we've seen tends to for high numbers. We will encounter soon a law that ensure you to find a gaussian distribution under certain general conditions as soon as the entries become very large.

The Gaussian, or *normal*, distribution rules out a continue casual variable and has the following form:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (22)$$

where the coefficient  $A \equiv \frac{1}{\sqrt{2\pi\sigma^2}}$  is the normalizing factor so that  $\int_{-\infty}^{\infty} f(x)dx = 1$ . We will now discuss the meaning of the two parameter  $\mu$  and  $\sigma$ , but before we want to notice the following facts: the (22) is symmetric respect to  $\mu$  ( $f(x-\mu) = f(\mu-x)$ ),  $f(x)$  has a maximum for  $x = \mu$  (in fact the first derivative is proportional to  $(x-\mu)$  which vanishes for  $x = \mu$ ),  $f(x = \mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$ , the flex points are for  $x = \mu \pm \sigma$ , etc.

Now we want to see what meaning  $\mu$  and  $\sigma$  do have: the expected value is

$$E[x] = \int_{-\infty}^{+\infty} x f(x) dx = A \left\{ \sqrt{2}\sigma\mu \int_{-\infty}^{+\infty} e^{-t^2} dt + (\sqrt{2}\sigma)^2 \int_{-\infty}^{+\infty} t e^{-t^2} dt \right\}$$

having used the substitution  $t = (x - \mu)/\sqrt{2}\sigma$ . The second integral is over a symmetric interval with an odd integrand function so it vanishes while the first integral is the Euler one and it's equal to  $\sqrt{\pi}$ . Thus we find that

$$E[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{2\pi}\sigma\mu = \mu$$

To calculate the variance as always we must pass through the calculation of  $E[x^2]$ . Using the same substitution  $t = (x - \mu)/\sqrt{2}\sigma$  we find that

$$E[x^2] = \frac{1}{\sqrt{\pi}} \left\{ 2\sigma^2 \int_{-\infty}^{+\infty} t^2 e^{-t^2} dt + \mu^2 \int_{-\infty}^{+\infty} e^{-t^2} dt + 2\sqrt{2}\sigma\mu \int_{-\infty}^{+\infty} t e^{-t^2} dt \right\}$$

as before, the last integral vanishes because of its symmetry, while the second one is equal again to  $\sqrt{\pi}$  and the first to  $\sqrt{\pi}/2$ . Therefore we find that

$$Var[x] = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Of course we could have obtained the same results by using the moment generating function  $G(t) = E[e^{xt}] = e^{t\left(\mu + \frac{\sigma^2 t}{2}\right)}$  (obtained after having put  $\xi = \frac{x-\mu}{\sqrt{2}\sigma}$  and having solved the integral), but we won't give here the calculation left for the willing student.

We finish here by underline an important variable we can always assign: by choosing  $z = \left(\frac{x-\mu}{\sigma}\right)$  we will have a normalized gaussian with expected value zero and variance one. This is the standard variable and is used mainly because the integral from the cumulative function is possible only via numerical calculation and is tabulated. So to compare one's special case to the tables is needed to transform it with this standard variable.

### 3.3.1 More over likelihood

Let's now say that we want to measure the expected value of a random variable. As we saw the likelihood function for a measurement result  $x$  is  $f(x|\mu)$  and very often the model we would take for our measurement process is the gaussian distribution so that

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

so now given a measure  $x$  and the hypothesis over  $\mu$  we can write

$$f(\mu|x) \propto f(x|\mu) \cdot f_0(\mu)$$

where the function  $f_0(\mu)$  takes into account our state of uncertainty regarding the value  $\mu$  having all the informations available at that moment except for the knowledge over the occurrence of experimental data  $x$ . Thus we can decide wheather  $\mu$  is the right value for the expected of the gaussian model we've hypothesized or not, having different  $\mu_1, \mu_2, \dots$  under test. We will get more details on this procedure later in the next sections.

### 3.4 Asymptotic behavior

Now it's time to see how the Bernoulli and the Poisson functions tends to a Gaussian asymptotically. We start with the Bernoulli (19) one: let's remind the De Moivre-Stirling formula

$$x! \simeq \sqrt{2\pi} \cdot x^x \sqrt{x} e^{-x} \quad (23)$$

and let's put  $\xi = n - Np$ , we have then

$$\begin{aligned} P_{N,p}(n) &\equiv P_{N,p}(\xi + Np) = \binom{N}{\xi + Np} p^{\xi + Np} (1-p)^{N - \xi - Np} = \\ &= \frac{N!}{(Nq - \xi)!(\xi + Np)!} p^{\xi + Np} q^{Nq - \xi} \end{aligned}$$

where as always  $q = (1 - p)$ . using now the Stirling approximation (23) we obtain (the reader won't be disappointed if we jump some easy arithmetic passages where we used that  $\exp[-N + N(p + q)] = 1$ )

$$P_{N,p}(\xi) \simeq \frac{1}{\sqrt{2\pi Npq}} \left(1 + \frac{\xi}{Np}\right)^{-Np - \xi - 1/2} \left(1 + \frac{\xi}{Nq}\right)^{-Nq + \xi - 1/2} \equiv \frac{1}{\sqrt{2\pi Npq}} K_p K_q$$

Now we can consider the  $\ln(K_p)$  and  $\ln(K_q)$  developing in MacLaurin series ( $\ln(1 \pm x) = \pm x \mp \frac{x^2}{2} \pm \frac{x^3}{3} + \dots$  valid only if  $x \ll 1$ ) obtaining (again some passage is skipped)

$$\ln(K_p) \simeq -\xi - \xi^2 \left(\frac{1}{2Np}\right)$$

$$\ln(K_q) \simeq \xi - \xi^2 \left(\frac{1}{2Nq}\right)$$

so the sum of the two is  $\ln(K_p) + \ln(K_q) = -\frac{\xi^2}{2N} \left(\frac{p+q}{pq}\right) = -\frac{\xi^2}{2Npq}$  and then

$$P_{N,p}(\xi) \approx \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{\xi^2}{2Npq}} \quad \text{for } N \gg 1$$

which is the Gauss distribution if we identify  $Np = \mu$  and  $Npq = \sigma^2$ .

Let's roll now for the Poisson distribution. Again we need to set  $\xi = n - a$  so the (21) become

$$P_a(\xi + a) = \frac{a^{\xi + a}}{(\xi + a)!} e^{-a}$$

and when the parameter  $a$  grows large we can again use the Stirling approximation

$$P_a(\xi) \simeq \frac{1}{\sqrt{2\pi}} \frac{a^{\xi + a}}{(\xi + a)^{\xi + a + 1/2}} \simeq \frac{1}{\sqrt{2\pi a}} \left(1 + \frac{\xi}{a}\right)^{-\xi - a - 1/2} e^{\xi} \equiv \frac{1}{\sqrt{2\pi a}} K_a K_e$$

Now we can do the same as before developing with MacLaurin the  $\ln(K_a) \simeq -\xi - \frac{\xi^2}{2a}$  while the  $\ln(K_e) = \ln(e^\xi)$  is simply  $\xi$ . Summing the two will give us  $-\frac{\xi^2}{2a}$  so that we can write

$$P_a(\xi) \approx \frac{1}{\sqrt{2\pi a}} e^{-\frac{\xi^2}{2a}} \quad \text{for } a \gg 1$$

which is again the Gauss distribution if we identify  $\mu = \sigma^2 = a$ .

### 3.5 Further distributions

The following distributions are somehow less important but still necessary to know. First we can define the **uniform distribution** that is the simplest kind of continuous variable distribution, assigning the same probability density to all the possible values of the casual variable within a certain range.

The uniform distribution can be written as

$$f_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (24)$$

The expected value and the variance are trivial to calculate

$$E[x] = \int_a^b \frac{x}{b-a} dx = \frac{1}{2}(a+b)$$

$$Var[x] = \int_a^b \frac{[x - \frac{1}{2}(a+b)]^2}{b-a} dx = \frac{1}{12}(b-a)^2$$

Similar to the uniform distribution is the **triangular** one: this distribution also called the *Simpson distribution* is useful when the degrees of belief linearly decreases from the central value of the distribution  $x_0$ . If we have that the certain event is to find the variable inside the range  $x_0 \pm \Delta$ , then the triangular distribution standard deviation is  $\sigma = \frac{\Delta}{\sqrt{6}}$  (compare it with the uniform's one  $\sigma = \frac{2\Delta}{\sqrt{12}} = \frac{\Delta}{\sqrt{3}}$ ).

The next distribution is the **exponential distribution**:

$$f_\xi(x) = \frac{1}{\xi} e^{-x/\xi} \quad (25)$$

where again the calculations for  $E[x]$  and  $Var[x]$  are trivial

$$E[x] = \frac{1}{\xi} \int_0^{+\infty} x e^{-x/\xi} dx = \xi$$

$$Var[x] = \frac{1}{\xi} \int_0^{+\infty} (x - \xi)^2 e^{-x/\xi} dx = \xi^2$$

A distribution most often used to model resonances in high energy physics is the **Breit-Wigner** (or **Cauchy**) distribution

$$f_{\Gamma, x_0}(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2} \quad (26)$$

where  $\Gamma$  is the width of a resonance for the particle with mass  $x_0$ . This peculiar distribution is also of pedagogic interest because its expectation value is not defined, moreover this distribution has no variance or higher moments defined. On the contrary its median and mode are well defined and are both equal to  $x_0$ . Thus, the hypothesis of finite variance in the central limit theorem cannot be dropped. The general form of the (26) is the **Cauchy** distribution, defined as

$$f(x) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2} \quad (27)$$

where  $x_0$  is the *location parameter* (where the peak of the distribution is) and  $\gamma$  is the *scale parameter* which specifies the half-width at half-maximum (HWHM).

## 4 $\chi^2$ distribution

A special kind of p.d.f is the  $\chi^2$  distribution. This distribution is so important that deserves a separate section. Before to define it we need to introduce briefly another function, the **Euler Gamma function**, because we will use it as soon as we start to define the  $\chi^2$  variable.

### 4.1 Gamma function

We define then the Gamma function as

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt \quad x > 0 \quad (28)$$

where the integral converges.

By integrating by parts we can see that

$$\Gamma(x) = \left[ \frac{e^{-t} t^x}{x} \right]_0^{\infty} + \int_0^{+\infty} \frac{t^x}{x} e^{-t} dt = \frac{1}{x} \Gamma(x+1)$$

so in general we can write recursively that

$$\begin{cases} \Gamma(1) = 1 \\ \Gamma(x+1) = x\Gamma(x) \end{cases} \quad (29)$$

furthermore we can note that, given an integer  $n$

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-1) = n(n-1)\cdots\Gamma(1) = n!$$

We won't demonstrate it but it also results that  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$  and in general for the semi-integers

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{\sqrt{\pi}}{2^n} [(2n-1)(2n-3)(2n-5)\cdots 7 \cdot 5 \cdot 3 \cdot 1] \quad (30)$$

## 4.2 $\chi^2$ variable

Now we are ready for introducing the  $\chi^2$  variable: given  $\nu$  independent variables  $x_\nu$  coming from  $\nu$  gaussian distributions (with  $\mu_1, \dots, \mu_\nu$  and  $\sigma_1, \dots, \sigma_\nu$  each one) we define

$$\chi^2 \equiv \sum_{k=1}^{\nu} \left( \frac{x_k - \mu_k}{\sigma_k} \right)^2 \equiv \sum_{k=1}^{\nu} \chi_k^2 \quad (31)$$

where  $\nu$  is also called the *degrees of freedom* and they represent the number of values that at the end are free to vary.

Now we can calculate the expected value of the  $\chi^2$  p.d.f. without knowing the p.d.f. yet only by the definition (31): in fact for a discrete variable the general  $k$ th momentum  $M_0^k$  is equal (take definition (11) in the discrete case) to  $M_0^k[x] = \sum_{j=1}^{\nu} x_j^k p(x_j)$  so in our case we simply calculate

$$E[\chi^2] = \sum_{j=1}^{\nu} E\left[\frac{(x_j - \mu_j)^2}{\sigma_j^2}\right] = \sum_{j=1}^{\nu} \frac{1}{\sigma_j^2} E[(x_j - \mu_j)^2] = \sum_{j=1}^{\nu} \frac{1}{\sigma_j^2} \cdot \sigma_j^2 = \nu$$

therefore, the expected value of the  $\chi^2$  is equal to the number of its degrees of freedom.

## 4.3 $\chi^2$ p.d.f.

Finally we show the  $\chi^2$  p.d.f. which is very hard to remember

$$f(\chi^2) = \frac{(\chi^2)^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}} \Gamma(\nu/2)} e^{-\frac{\chi^2}{2}} \quad 0 < \chi^2 < \infty \quad (32)$$

The important thing to note out from the (32) is that for  $\nu \leq 2$  the function decrease monotonically when the  $\chi^2$  increase (for  $\nu = 2$  is a negative exponential) while for  $\nu > 2$  the (32) shows a maximum in  $\chi^2 = \nu - 2$

The variance of the  $\chi^2$  can be calculated remembering that  $Var[\sum_{k=1}^{\nu} \chi_k^2] = \sum_k Var[\chi_k^2] = \nu Var[\chi_k^2]$  so, if we remember that the variable itself is defined such to have the expected value equal to one (is the standard gaussian variable to the power of two) we can write for the single addendum

$$Var[\chi_k^2] = E[(\chi_k^2 - E[\chi_k^2])^2] = E[(\chi_k^2)^2] - (E[\chi_k^2])^2 = E[(\chi_k^2)^2] - 1$$

so we have to solve the integral

$$E[\eta^2] = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} \eta^{3/2} e^{-\frac{\eta}{2}} d\eta$$

where of course  $\eta = \chi_k^2$ .

Using the equivalence (30) in the integral we can write

$$\frac{1}{\sqrt{2\pi}} \int_0^{+\infty} \eta^{3/2} e^{-\frac{\eta}{2}} d\eta = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} \eta^{\frac{5}{2}-1} e^{-\frac{\eta}{2}} = \Gamma\left(\frac{5}{2}\right) \frac{2^{5/2}}{\sqrt{2\pi}} = 3$$

and finally

$$Var[\chi^2] = \nu(3 - 1) = 2\nu$$

It is possible to demonstrate that the  $\chi^2$  distribution tends to a normal one in the  $\nu \rightarrow \infty$  limit (demonstration left to the student<sup>6</sup>).

#### 4.4 $\chi^2$ test

The probability that the  $\chi^2$  variable with  $\nu$  degrees of freedom will assume a numerical value greater than  $\chi_{\nu,\alpha}^2$ , where  $\alpha$  is the *significance level* (usually the quantity  $1 - \alpha$  is called confidence level, so if we choose  $\alpha = 0.05$  the confidence level will be the 95%), is given by

$$P(\chi^2 > \chi_{\nu,\alpha}^2) = \int_{\chi_{\nu,\alpha}^2}^{+\infty} f(\chi^2) d\chi^2 = \alpha$$

so that the cumulative function gives the confidence level

$$F(\chi_{\nu,\alpha}^2) = \int_0^{\chi_{\nu,\alpha}^2} f(\chi^2) d\chi^2 = 1 - \alpha$$

If we have a value to be measured (let's say  $\mu$ , i.e. the height of your desk) and we take several measurements (described by a continuous variable  $x$  taking the  $n$  values  $\{x_i \pm \sigma_i\}$ ,  $i = 1, \dots, n$ ), we can build the  $\chi^2 = \sum_{k=1}^{\nu} \left(\frac{x_k - \mu}{\sigma_k}\right)^2$  as in (31). We can now ask what is the probability to obtain worse values than our set from the measurement process. If the data are independents than this probability is the productory of the singular probabilities to have a worse measure. If we assume that each measure process is dominated by a gaussian distribution (which is almost always a very good approximation) then this probability is

$$P = \left(\frac{1}{2\pi}\right)^{n/2} \int \dots \int e^{-(y_1^2 + y_2^2 + \dots + y_n^2)} dy_1 \dots dy_n$$

---

<sup>6</sup>the momentum-generating function for the  $\chi^2$  is  $G(t) = (1 - 2t)^{-\frac{\nu}{2}}$  so if one uses the reduced variable  $Z = \frac{\chi^2 - \nu}{\sqrt{2\nu}}$  is easy to see that the MacLaurin series for  $\ln(G(Z))$  describes an exponential when the limit  $\nu \rightarrow \infty$  is performed. This exponential is equal to the momentum-generating function of the standard gaussian.

with the condition  $y_1^2 + y_2^2 + \dots + y_n^2 > \chi^2$ , where the  $y_k$  variables are the standard gaussian variables. Note that the  $\chi^2$  we built is linked to the distance of our measurements from the expected value (theoretical) which comes from an hypothesis we have made (i.e. our first bet on the desk height). In this case the degrees of freedom are the same as the number of data we took  $\nu = n$ . Let's now take the case in which the same data  $\{x_i \pm \sigma_i\}$  are coming from a theoretical law (not a constant anymore like the desk height). If this law has several parameters let's say  $k$  parameters  $\{\lambda_1, \dots, \lambda_k\}$ , then is clear that we have  $k$  relations between the data and the unknown parameters so the number of degrees of freedom is now  $\nu = n - k$  because this is the number of really independent measurements we have.

It is clear then, that if our data are good with our hypothesis the experimental  $\chi^2$  value will not be so much different from the number  $\nu$ , the degrees of freedom. On the other way if the  $\chi^2$  we obtain is too much large compared to  $\nu$ , than the data are not compatible with our theoretical hypothesis. Be aware that if the  $\chi^2$  is too much small compared to  $\nu$  this means that our data are too much close to the theoretical values and this is also not good: it could mean that we probably overestimated our standard deviations  $\sigma_i$ .

## 5 Law of large numbers and the Central limit theorem

The fundamental concept about the Law of large numbers (LLN) is that in the limit  $N \rightarrow \infty$  the (16) calculated from the set  $x_i$  of independent casual variables ( $P(x_1, x_2, \dots, x_N) = P(x_1)P(x_2) \dots P(x_N)$ ) tends to the expected value (9). Note that this is valid only if  $\int dx f(x)|x| < \infty$ .

In its weak form, the LLN states that  $\forall \varepsilon > 0$  the probability  $P(|\bar{x} - E[x]| > \varepsilon)$  tends to zero when  $N \rightarrow \infty$ . There is a strong version of this law but we won't go into details for sake of simplicity.

### 5.1 The Čebyšev's inequality

To prove the LLN we must pass through the Čebyšev's inequality:

$$P(|X - \mu| \geq K\sigma) \leq \frac{1}{K^2} \quad (33)$$

where  $X$  is a random variable with any p.d.f. which has the expected value  $E[X] = \mu \equiv \int xf(x)dx$  and the variance  $Var[X] = \sigma^2 \equiv \int (x - \mu)^2 f(x)dx < \infty$ .

To prove (33) we only need to write

$$\sigma^2 > \int_{|x-\mu|>K\sigma} (x - \mu)^2 f(x)dx > K^2 \sigma^2 \int_{|x-\mu|>K\sigma} f(x)dx$$

thus it must be true that

$$\int_{|x-\mu|>K\sigma} f(x)dx < \frac{1}{K^2}$$

quod erat demonstrandum.

Of course the power of the (33) is that is valid for any p.d.f. though the estimation itself is really pessimistic (in fact most of the time we have that  $P(|X - \mu| \geq K\sigma) \lll \frac{1}{K^2}$ ).

## 5.2 Law of large numbers

A direct consequence of (33) is (remember the long-term stability of the mean of a random variable) that for large numbers the *sample mean*  $\bar{x}$  will tend to approach the expected value  $E[x]$ , the average over the population. In fact if we apply the (33) to the sample mean  $\bar{x}$  (let's say over a sample of  $n$  values) we find that

$$P(|\bar{x} - E[\bar{x}]| \geq K\sigma_{\bar{x}}) = P\left(|\bar{x} - E[x]| \geq K\frac{\sigma_x}{\sqrt{n}}\right) \leq \frac{1}{K^2}$$

having used  $E[\bar{x}] = E[x]$  and  $Var[\bar{x}] = \frac{Var[x]}{\sqrt{n}}$ . Let's now take  $\epsilon = K\sigma_x/\sqrt{n}$ , then

$$P(|\bar{x} - E[x]| \geq \epsilon) \leq \frac{\sigma_x^2}{n\epsilon^2} \Rightarrow P(|\bar{x} - E[x]| \leq \epsilon) \geq (1 - \frac{\sigma_x^2}{n\epsilon^2})$$

that is to say that  $\forall \epsilon > 0 \exists n$  so that  $\lim_{n \rightarrow \infty} P(-\epsilon \leq |\bar{x} - E[x]| \leq +\epsilon) = 1$  which is the definition of  $E[x]$  as limit of  $\bar{x}$  for large numbers Q.E.D.

## 5.3 Central limit theorem

So far, nothing was said on the p.d.f. of the variable  $x$  when we have large numbers. It is possible to see that under certain conditions not only the *sample mean* tends to the expected value but also the p.d.f tends to a normal (gaussian) distribution, irrespective of the shape of the original distribution function. This is known as the Central limit theorem (CLT). We want now see what conditions are needed to prove this sentence.

Let's take  $n$  independent variables  $x_1, \dots, x_n$  coming from  $f_1, \dots, f_n$  p.d.f. each one with a given expected value and variance, namely  $E[x_i]$  and  $Var[x_i]$ ,  $i = 1, \dots, n$ . Suppose  $E[x_i] < \infty$  and  $Var[x_i] < \infty \forall i$  and suppose that they are also about the same order of magnitude  $\forall i$ . Than is possible to prove that the variable  $X = \sum_i a_i x_i$  built as a linear combination with whatever constant coefficients  $a_i$ , has a distribution which tends to a normal one for large numbers (CLT); that is to say if we put

$$\mu = E[X] = \sum_{i=1}^n a_i E[x_i] \quad \sigma^2 = Var[X] = \sum_{i=1}^n a_i^2 Var[x_i]$$

then for large numbers the p.d.f of variable  $X$  is

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

We don't give a prove for this general case for sake of simplicity. We will only demonstrate a special case, that is when the  $n$  independent variables  $x_1, \dots, x_n$  comes *all from the same p.d.f.*  $f(x)$  (with no restrictions on the form of  $f(x)$ ) with expected value  $\mu$  and variance  $\sigma^2$ . First we build the variable  $X = \sum_i x_i$  (so  $a_i = 1 \forall i = 1, \dots, n$ ), second we notice that the moment generating function must be  $G_X(t) = g_{x_1}(t) \cdots g_{x_n}(t) = (g_x(t))^n$  where, developing till the second order we have

$$g_x(t) = \int_{-\infty}^{+\infty} e^{xt} f(x) dx \Rightarrow g_x(t) \simeq g_x(0) + \frac{\partial g_x(0)}{\partial t} t + \frac{\partial^2 g_x(0)}{\partial t^2} \frac{t^2}{2}$$

with (remember the (18))

$$g_x(0) = 1 \quad \frac{\partial g_x(0)}{\partial t} = \mu \quad \frac{\partial^2 g_x(0)}{\partial t^2} = \sigma^2 + \mu^2$$

finally, to the second order we write

$$\begin{aligned} G_X(t) &\simeq \left\{ 1 + \mu t + (\sigma^2 + \mu^2) \frac{t^2}{2} \right\}^n = \left\{ 1 + \frac{E[X]}{n} t + \left( \frac{Var[X]}{n} + \frac{E^2[X]}{n^2} \right) \frac{t^2}{2} \right\}^n \approx \\ &\approx \left\{ 1 + \frac{E[X]}{n} t + \frac{Var[X]}{2n} t^2 \right\}^n \end{aligned}$$

where the last approximation is valid for we are evaluating the behavior for large  $n$ . Now we need to take the following position:

$$k = \frac{E[X]}{n} t + \frac{Var[X]}{2n} t^2$$

so the limit  $n \rightarrow \infty$  turns into the limit for  $k \rightarrow 0$

$$\lim_{k \rightarrow 0} (1 + k)^{\left\{ \frac{1}{k} \left[ E[X] t + \frac{Var[X]}{2} t^2 \right] \right\}} = e^{\left( E[X] t + \frac{Var[X]}{2} t^2 \right)}$$

and this is exactly the moment generating function of a gaussian distribution Q.E.D. (note that the Fourier transformation of a gaussian is still a gaussian).

## 6 Parameter Estimation and Maximum likelihood estimation (MLE)

The general procedure we want now to illustrate helps to find, from a finite number of observational data, informations about the whole event space

of a certain random variable  $x$ . Let's assume we have a set of independent values  $(x_1, \dots, x_n)$  from a known p.d.f  $f(x; \lambda_k)$  where the parameters  $\lambda_k$  are not known. Is there a way to give an estimation of these parameters? The answer is affirmative; moreover, there is more than one method to estimate the  $\lambda$  parameters. We will see here only the most famous method called the Maximum Likelihood (ML).

## 6.1 The maximum likelihood function

Let's consider the simplest case of only one parameter ( $f(x; \lambda)$ ) and let's define the following estimator:

$$L(\lambda) = \prod_{k=1}^n f(x_k; \lambda) \quad (34)$$

This is called the maximum likelihood function. It results that the best estimation for  $\lambda$  is the one that maximize the maximum likelihood function (34):

$$\frac{\partial L(\lambda)}{\partial \lambda} = 0 \quad ; \quad \left. \frac{\partial^2 L(\lambda)}{\partial \lambda^2} \right|_{\lambda=\lambda^*} < 0$$

where  $\lambda^*$  is the value that makes the first derivative of  $L(\lambda)$  vanish. Of course it could be annoying to take the derivative of  $L(\lambda)$  function defined in (34) because of the productory, so it's a better good idea to consider its logarithm  $\ln L(\lambda) = \sum_{k=1}^n \ln f(x_k; \lambda)$  for the logarithm is a crescent monotonic function and doesn't affect the maximum search. Naturally, nothing changes if  $\lambda$  is a vector  $\vec{\lambda} = (\lambda_1, \lambda_2, \dots)$  and we only need to consider the  $n$  equations

$$\frac{\partial}{\partial \lambda_k} \ln L(\vec{\lambda}) = 0 \quad (k = 1, 2, \dots, n)$$

but the solution is easy only if the  $\lambda_k$  appears linearly, otherwise numerical solution can be found anyways.

The maximum likelihood function being a function of casual variables is a casual variable itself having its own p.d.f. and all the attributes. Now a nice property of an estimator which is obtained with the Maximum likelihood method is that the p.d.f of the maximum likelihood function tends, for large numbers, to a normal distribution with expected value and variance

$$E[\lambda] = \lambda^* \quad Var[\lambda] = \left[ -\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2} \right]^{-1} \Bigg|_{\lambda=\lambda^*} \quad (35)$$

## 6.2 Consistency, distortion, efficiency

An estimator such the (34) should have some properties to be a good one. The most important properties for a good estimator are

- *consistency* the estimator which gives an estimation  $\lambda^*$  that converges to the expected value of the population parameter when the sample grows up ( $\lim_{n \rightarrow \infty} \lambda^* = E[\lambda]$ ), is said to be a consistent estimator. We've seen that the arithmetic mean and the standard deviation are consistent estimators of respectively the expected value and the variance of the population. In general if we use as estimator the maximum likelihood function (34) we will have only consistent estimations.
- *distortion* if  $\lambda^*$  is a consistent estimation of the true parameter  $\lambda$  every other estimation  $f(n)\lambda^*$  with  $\lim_{n \rightarrow \infty} f(n) = 1$  is a consistent estimation. To choose the best one we must ask that not only for large numbers the expected value of the estimation gives the parameter, but also for all values of  $n$ . Note that the estimator for the variance

$$S^2 = \frac{1}{n} \sum_k (x_k - \bar{x})^2$$

is consistent but is distorted:

$$\begin{aligned} E[S^2] &= E \left[ \frac{1}{n} \sum_{k=1}^n [(x_k - \mu) - (\bar{x} - \mu)]^2 \right] = \\ &= E \left[ \frac{1}{n} \sum_{k=1}^n \left[ (x_k - \mu) - \left( \frac{1}{n} \sum_{k=1}^n x_k - \mu \right) \right]^2 \right] = \\ &= \frac{1}{n} E \left[ \sum_{k=1}^n (x_k - \mu)^2 \right] - \frac{1}{n^2} E \left[ \sum_{k=1}^n (x_k - \mu)^2 \right] = \\ &= \frac{1}{n} n \sigma^2 - \frac{1}{n^2} n \sigma^2 = \left( 1 - \frac{1}{n} \right) \sigma^2 \end{aligned}$$

and this distortion (we see that for  $n = 1$  the variance even vanishes) is removed when we use the correct estimator

$$s^2 = \frac{1}{n-1} \sum_k (x_k - \bar{x})^2 \quad \Rightarrow \quad E[s^2] = \frac{n}{n-1} \left( 1 - \frac{1}{n} \right) \sigma^2 = \sigma^2$$

- *efficiency* When we have an estimator which is consistent and not distorted we still have to check if the estimation variance is the smallest or not because we surely prefer between several estimation which are all consistent and not distorted, the one that has the smallest variance. This is said to be the efficient estimation. In example the best estimator for the expected value is the sample mean, in fact its variance is  $Var[\bar{x}] = \sigma/n$  while the variance for the median, let's say, is  $Var[\hat{x}] = \frac{\pi\sigma^2}{2n} > \frac{\sigma^2}{n}$ .

### 6.3 Estimators for the Gaussian

Let's now calculate the best estimators of the expected value and variance for the gaussian, the poissonian and the binomial distributions, to fix in mind what we just learned. Assuming that we have a set of results  $x_1, x_2, \dots, x_n$  from a gaussian with  $E[x] = \mu$  and  $Var[x] = \sigma^2$ . We want to see what are the best estimators using our sample. Writing the maximum likelihood function is easy:

$$L(\mu, \sigma^2) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k - \mu)^2}{2\sigma^2}}$$

and now let's take the logarithm of it

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} (\ln(2\pi) + \ln \sigma^2) - \sum_{k=1}^n \frac{(x_k - \mu)^2}{2\sigma^2}$$

to find

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = 0 \Rightarrow \mu = \frac{1}{n} \sum_{k=1}^n x_k$$

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2) = 0 \Rightarrow \sigma^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2$$

Please notice that if the variances were different the estimation for  $\mu$  would have been

$$\mu = \frac{\sum_{k=1}^n \frac{x_k}{\sigma_k^2}}{\sum_{k=1}^n \frac{1}{\sigma_k^2}}$$

and its variance  $Var[\mu] = (\sum_{k=1}^n \frac{1}{\sigma_k^2})^{-1}$  obtained from the (35).

### 6.4 Estimators for the Poisson

Let's do the same for a sample  $n_1, n_2, \dots, n_N$  Poisson distributed:

$$L(a) = \prod_{k=1}^N \frac{a^{n_k}}{n_k!} e^{-a} \Rightarrow \ln L(a) = -aN + \sum_{k=1}^N \frac{a^{n_k}}{n_k!}$$

so that

$$\frac{d}{da} \ln L(a) = -N + \frac{1}{a} \sum_{k=1}^N n_k = 0 \Rightarrow a = \frac{1}{N} \sum_{k=1}^N n_k$$

which is again the average as we already know. The variance can be found via (35)

$$Var[a] = \left[ -\frac{\partial^2 \ln L(a)}{\partial a^2} \right]^{-1} \Big|_a = \left( \frac{a^2}{\sum_{k=1}^N n_k} \right) = a/N$$

## 6.5 Estimators for the Bernoulli

Same thing as before, but this time the productory is not needed because we have only one term having found in the experiment  $n$  positive events over  $N$  attempts so  $L(p) \equiv P_{N,p}(n)$  and we can write the logarithm of it directly

$$\ln L(p) = \ln \frac{N!}{(N-n)!n!} + n \ln p + (N-n) \ln(1-p)$$

so that

$$\frac{d}{dp} \ln L(p) = \frac{n}{p} - \frac{N-n}{1-p} = 0 \Rightarrow p = n/N$$

and the variance

$$\text{Var}[p] = \left[ -\frac{\partial^2 \ln L(p)}{\partial p^2} \right]^{-1} \Big|_p = \frac{p(1-p)}{N}$$

if we repeat the experiment  $K$  times obtaining  $n_1, n_2, \dots, n_M$  favorable results each time over  $N$  attempts, then we need the productory and it's easy to find that  $p = \sum_{k=1}^M n_k / KN$ .

## 6.6 Variance of the $s^2$ estimator

We want to calculate the  $\text{Var}[s^2]$  where

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ &= \frac{1}{n-1} \left( n \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - 2n\bar{x} \left( \frac{1}{n} \sum_{i=1}^n x_i \right) + n\bar{x}^2 \right) = \\ &= \frac{n}{n-1} (\bar{x}^2 - 2\bar{x}^2 + \bar{x}^2) = \frac{n}{n-1} (\bar{x}^2 - \bar{x}^2) \end{aligned}$$

now we notice that

$$\begin{aligned} s^2 &= \frac{n}{n-1} (\bar{x}^2 - \bar{x}^2) = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \frac{1}{n^2} \sum_{i=1}^n x_i \sum_{j=1}^n x_j = \\ &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{1}{n(n-1)} \sum_{i=1}^n x_i \sum_{j=1}^n x_j \end{aligned}$$

we can split the second term of the above result in two (from now on if not differently specified the summatories are intended from 1 to  $n$ ):

$$s^2 = \frac{1}{n-1} \sum_i x_i^2 - \frac{1}{n(n-1)} \left( \sum_{i=j} x_i x_j + \sum_{i \neq j} x_i x_j \right)$$

where the last term will vanish when the expectation is taken.

The expected value is then

$$E[s^2] = \frac{1}{n-1}nE[x^2] - \frac{1}{n(n-1)}nE[x^2] = \left(\frac{n}{n-1} - \frac{1}{n-1}\right)E[x^2] \quad (36)$$

Let's now remember that  $Var[x] = E[x^2] - E[x]^2$  therefore  $Var[s^2] = E[s^4] - E[s^2]^2$ , so we only need to find the  $E[s^4]$ :

$$s^4 = \left[ \frac{1}{(n-1)} \sum_i x_i^2 - \frac{1}{n(n-1)} \sum_{i,k} x_i x_k \right]^2 =$$

$$\frac{1}{(n-1)^2} \left( \sum_i x_i^2 \right)^2 - \frac{2}{n(n-1)^2} \left( \sum_i x_i^2 \right) \left( \sum_i x_i \right)^2 + \frac{1}{n^2(n-1)^2} \left( \sum_i x_i \right)^4$$

Let's analyze the single members of the sum:

$$A \equiv \left( \sum_i x_i^2 \right)^2 = \left( \sum_i x_i^2 \right) \left( \sum_j x_j^2 \right) = \sum_i \left( x_i^2 \left[ \sum_{j=i} x_j^2 + \sum_{i \neq j} x_j^2 \right] \right) =$$

$$= \sum_i x_i^4 + \sum_{i \neq j} x_i^2 x_j^2 = \sum_{i=1}^n x_i^4 + 2 \sum_{i < j, 1}^{\frac{n(n-1)}{2}} x_i^2 x_j^2$$

Then, at the end we have

$$E[A] = nE[x^4] + n(n-1)E[x^2]^2 \quad (37)$$

We have to do the same with the second member

$$B \equiv \left( \sum_i x_i^2 \right) \left( \sum_i x_i \right)^2 = \left( \sum_i x_i^2 \right) \left[ \sum_k x_k^2 + \sum_{k \neq j} x_k x_j \right] =$$

$$= \sum_{i=1}^n x_i^4 + \sum_{i \neq j, 1}^{\frac{2 \cdot n(n-1)}{2}} (x_i^2 x_j^2 + x_i^3 x_j) + \sum_{i \neq j \neq k} x_i^2 x_j x_k$$

where the terms with odd power members vanishes when the expectation is made so that we have again

$$E[B] = nE[x^4] + n(n-1)E[x^2]^2 \quad (38)$$

Finally the last one:

$$C \equiv \left( \sum_i x_i \right)^4 = \left( \sum_a x_a \right) \left( \sum_b x_b \right) \left( \sum_c x_c \right) \left( \sum_d x_d \right) =$$

$$= \sum_{i=1}^n x_i^4 + \sum_{i \neq j} x_i^3 x_j + \sum_{i < j, 1}^{6 \frac{n(n-1)}{2}} x_i^2 x_j^2 + \sum_{i \neq j \neq k} x_i^2 x_j x_k + \sum_{a \neq b \neq c \neq d} x_a x_b x_c x_d$$

now again the second, fourth and last term are zero in the expectation so that we write

$$E[C] = nE[x^4] + 3n(n-1)E[x^2]^2 \quad (39)$$

Finally we have to sum

$$\begin{aligned} E[s^4] &= \frac{1}{(n-1)^2} E[A] - \frac{2}{n(n-1)^2} E[B] + \frac{1}{n^2(n-1)^2} E[C] = \\ &= \frac{1}{n} E[x^4] + \frac{(n^2 - 2n + 3)}{n(n-1)} E[x^2]^2 \end{aligned}$$

now we only need to subtract the square of (36) to find

$$\begin{aligned} Var[s^2] &= E[s^4] - E[x^2]^2 = \frac{1}{n} E[x^4] + \frac{(n^2 - 2n + 3)}{n(n-1)} E[x^2]^2 - E[x^2]^2 = \\ &= \frac{1}{n} \left[ E[x^4] - \frac{n-3}{n-1} E[x^2]^2 \right] \end{aligned}$$

Q.E.D.