INFN-PISA
April 6, 2005

Methods of Statistical Data Analysis in High Energy Physics.

T. Del Prete

Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Italy

Scuola di Perfezionamento in Fisica Nucleare
Universitá degli Studi di Pisa

DRAFT

Foreword

E' non mi e' incognito come molti hanno avuto e hanno
opinione che le cose del mondo sieno in modo governate dalla fortuna
e da Dio, che gli uomini, con la prudenza loro, non possino correggerle,
anzi non vi abbino rimedio alcuno; e per questo potrebbero judicare che
non fossi da insudare molto nelle cose, ma lasciarsi governare dalla sorte
N. Machiavelli, Il Principe, Cap XXV, Par. 1

These notes have been written for a course in Probability and Statistics of the "Scuola di Perfezionamento in Fisica Nucleare" of the University of Pisa in the Academic Year 1996-1997. The aim of the course was to introduce the students to the methods used in the analysis of High Energy Physics experiments.

To maintain a good readability most of the mathematical part has been sacrificed, leaving only what we think is essential in understanding the subject and to avoid the impression that statistics be just a set of rules, like those for cooking.

We have tried to include many examples and problems, mostly from High Energy Physics which the students have been urged to solve.

We have tried to cover both the frequency theory and the Bayesian approach, with the attempt to clarify the philosophical questions underneath the two schools.

Statistics methods are particularly useful in those experiments where we want to extract the maximum of information from few events. This happens often in frontier physics. Young physicists are sometimes unprepared to solve problems like setting upper limits on average rates in counting experiments. In the literature the frequency theory and Bayesian schools offer recipes which sometimes are contradictory. In these cases it is important to have a deeper understanding of the subject to make our own judgment.

The distinction between "statistical" and "systematic" effects is another intriguing subject to the young physicists approaching the data analysis. It is probably here where the Bayesian methods are more used. However the principle of coverage, often not fully appreciate by the Bayesian school, must be hold and care must be used in quoting the results. We have tried to discuss these methods with examples always from High Energy Physics.

Reference to standard textbooks is always needed whenever the subject becomes intriguing. At the end of these lectures we present a bibliography which is by far not complete, but we hope at least useful.

Pareva fosse dato
variare a piacimento
il testo; che mutevoli
fossero in quel libro
le pagine e le parti.
Cosi' malgrado il nero
lavoro delle sorti
erano nel perpetuo avvenimento
davvero quelle carte.
Invariabile era solo
l'opera dell'aria
che le sfoglia, le gira,
le consuma, in altro
insieme con se stessa le transmuta.
Cosi' pareva, cosi' era.

(M.Luzi, 1999)

Introduction

When we describe the results of an experiment we may be lead to a model in which the inputs determines exactly the experiment's outcome. There are easy examples of this description: The acceleration of a body under a constant force or the motion of planets in their orbits. In these cases the model which describes the experiment is called *deterministic*. The laws of Classical Physics are deterministic.

There are cases in which a deterministic description is not adequate. Consider the outcome of an experiment which measure the live time of a radioactive material. The lifetimes of different atoms are not the same, rather they are distributed according to an exponential law. There is an intrinsic impossibility to foresee which will be the time in which an atom will disintegrate.

Consider the outcome of an experiment consisting in tossing a coin. Though *in principle* if we would know exactly the position, spin, velocity etc. of the coin we could know if the coin will land head or tail, in practice, the available knowledge of the toss is insufficient and the coin will land head or tail almost at random.

The description of an experiment (*model*) in which the exact knowledge of inputs does not permit the calculation of the outcomes is called *non deterministic*. In this case, though the outcomes are unpredictable, we can observe statistical regularities. The relative frequencies of the outcomes (*probabilities*) can be computed by a mathematical description of the experiment (*model*).

The theory of Probability deals with the models while Statistics is concerned with the description of the experimental results.

Probability and Statistics are two different sciences:

- The theory of Probability is a branch of pure mathematics. It is based on axioms and definitions and then is deduced deductively. The neatest approach is based on set theory, measure theory and Lebesgue integration.
- The theory of Statistics is a branch of applied mathematics and is essentially inductive, since we argue from the observation of events back to a hypothesis.

The name "statistics" (from the Latin *status*) was coined in the XVIII century to indicate that part of political science that organizes of economical and demographical data. Methods were refined in the course of the time to cope with different and more complex problems.

- **EXAMPLE of a problem in Probability**
  Find the probability of observing n heads out of N throws when tossing a coin, once we know that the probability p of landing heads. The solution is given by the binomial distribution and is:
  $$P(n) = \binom{N}{n} \cdot p^n \cdot (1-p)^{N-n} \tag{1}$$
- **EXAMPLE of a problem in Statistics**
  A coin is tossed N times and it falls heads n times. What can we say on the unknown

probability p of landing heads? The answer cannot be as unambiguous as in the previous example. We know that the probability is a bounded number:

$$0 \leq p \leq 1$$

If $n > 0$ also the probability will be larger than zero and if $n < N$, the probability will be less than one. Intuition tells us that the most likely value of p will be $p = \frac{n}{N}$. (We will discuss later what "the most likely" means, for the moment we stick to our intuition). We will ask next which are the intervals of p that are likely to include the true value of p.

$$p_1 \leq p \leq p_2 \tag{2}$$

The larger we make the difference $p_2 - p_1$ the more certain we are to include p in the interval. On the other side the larger the interval the less information we have on p. There is always uncertainty at work in Statistics. Our aim is to measure this uncertainty. More precisely we will have to compute the probability that the true value (though unknown) of the probability is comprised in the interval $p_2 - p_1$.

It is also intuitive that the result of the previous problem of probability will help in finding the solution to the Statistics problem.

Chapter 1

Probability

...
5. The probability of any event is the ratio between the value at which an expectation depending on the happening of an event ought to be computed and the value of the thing expected upon its happening.
6. By chance I mean the same as probability.
...
T. Bayes (1763)

Probability theory is nothing but common sense reduced to calculation. (Laplace 1818)

1.1 Definitions

The probability is a number that quantifies the happening of a random event. The concept has been made more precise in the course of time, but *no definition is yet universally accepted.*

- **Combinatorial** (Laplace) An event A can occur in N distinct and equally likely ways and n of these are the *favorable* cases. Then the probability is defined as:

$$P(A) = \frac{n}{N} \tag{1.1}$$

- **Frequency theory** (R. Von Mieses) We observe an event A occurring n times in N trials, the probability is defined as the limit:

$$P(A) = \lim_{N\to\infty} \frac{n}{N} \tag{1.2}$$

- **Subjective** (F.P. Ramsey, B. De Finetti). The probability of an event A to occur (or that an event will occur) is the measure of one's belief in its occurring. In this definition the probability loses any objective content. This is the so called *Modern* definition. The power and limitations of this approach will be discusses at length later.

- **Axiomatic** (A.N. Kolmogorov) The previous definitions of probability are based on the experience and practical instances. There exists a mathematical theory of probability which is based on the concept of the measure on the set of elementary events in an abstract space. This theory is the basis for all the previous definitions.

1.1.1 Combinatorial

A random event can occur in N distinct and equally likely ways. n of these events have an attribute A. The probability that the outcome has the attribute A is $\frac{n}{N}$.

As an example: a dice is thrown. The probability of obtaining 2 is 1/6 for a perfect dice. In fact there are six possible, equally likely events and only one is the good one.

This definition is based on the principle of *insufficient reason* or *indifference principle*: if there are no reasons to assign different probabilities to different events these are to be considered equally likely (Laplace).

The definition holds the suspect to contain circular references and in addition is rather limited and some problems cannot be formulated in the combinatorial definition. For instance: What is the probability that a man of forty will die within the year? In fact we cannot define the possible and favorable cases.

It is also not always possible to define unambiguously the possible cases. In addition the definition of possible cases is somehow conventional. Consider for instance an experiment where we toss two dices.

Outcome						
2	1,1					
3	1,2	2,1				
4	1,3	2,2	3,1			
5	1,4	2,3	3,2	4,1		
6	1,5	2,4	3,3	4,2	5,1	
7	1,6	2,5	3,4	4,3	5,2	6,1
8	2,6	3,5	4,4	5,3	6,2	
9	3,6	4,5	5,4	6,3		
10	4,6	5,5	6,4			
11	5,6	6,5				
12	6,6					

Table 1.1: *Possible outcome when tossing two dices.*

In table 1.1 the first raw is the outcome of the throw while the second lists the number of possible dice configurations. Assume we want to compute the probability the outcome is 3. We can argue in various ways:

- The number of possible outcomes is 11, 3 is just one out of the 11 possibilities so the probability to get 3 is:
$$P(3) = \frac{1}{11} \qquad\qquad (WRONG)$$
.
- The number of possible ways in which the dices can land is $6 \times 6 = 36$. There are two distinct ways to get 3, the probability then is
$$P(3) = \frac{2}{36} = \frac{1}{18} \qquad\qquad (CORRECT)$$
.
- Are the two possibilities 1,2 and 2,1 really distinct? If they are **not** the probability is $P(3) = 1/21$. For macroscopic dices they are distinct: we can throw the first, read the

result and then toss the second. Or we can color dices with different colors. In quantum statistics they cannot be distinguished. (And in fact Boltzmann statistics is different from Bose statistics.)

Example 1.1 *A player of die asked Galileo Galilei why in the game with three dices the sum 10 was more likely than 9, while the number of rolls whose sum is 10 and 9, as Table 1.2 shows, is the same. Galilei noticed that the result* $1+3+6=10$ *could be obtained in six different*

Outcome	dye 1	dye 2	dye 3
9	1	2	6
	1	3	5
	1	4	4
	2	2	5
	2	3	4
	3	3	3
10	1	3	6
	1	4	5
	2	2	6
	2	3	5
	2	4	4
	3	3	4

Table 1.2: *Rolls that sum to 9 and 10.*

ways, the number of permutation of three dices, while $1+4+4=9$ *in three different ways and* $3+3+3=9$ *only in one way. Therefore the previous list was not exhaustive since the distinct permutation of dices were not counted. If we properly count all the independent permutations of the dices we obtain 27 combinations to sum to 10 and 25 to sum 9. Since the total number of possible cases is* $6\times6\times6=216$*, the probabilities are* $P_9=25/216$ *and* $P_{10}=27/216$*.*

Exercise 1.1 *What is the probability that tossing* n *times a dice we get at least once 6?*

Exercise 1.2 *An urn contains* m *balls.* a *balls are white. We take a block of* n *balls. What is the probability that* k_1 *are white and* $k_2=n-k_1$ *are black?*

Exercise 1.3 *On a plane we draw a set of parallel and equidistant lines. A needle is let fall onto the plane. If the length of the needle is* l *and the distance between lines is* $a>l$*, which is the probability that the needle will cross one line? (This problem is one of the first on geometrical probabilities and was proposed in 1733 by Buffon who solved it and published the solution in 1777.)*

Exercise 1.4 *What is the probability that the length of a chord drawn randomly on a circle of radius* R *is smaller than* R*? (There is NO unique solution to this problem since the problem lacks of some specification. What is it? Try different assumption on how you draw a random chord.)*

1.1.2 Frequency Theory

The frequency theory of probability is based on the empirical observation that the frequency of occurrence of a random event (E) shows a remarkable regularity. In spite of the irregular behavior of the individual events, their frequency, in a long sequence of random experiments, performed in uniform conditions, is rather stable and seems to converge to a constant value as the number of experiments tends to infinity. This limit is called *probability* of the event.

The mathematical abstraction of this empirical fact leads to the concept of probability as a non negative measure of an event (which is changed to an abstract member of a σ algebra) P_E.

The frequency theory of the probability offers an operative way of measuring P_E by computing the frequency of event $f = \nu/n$ in a sequence of n experiments. In this way we can verify any mathematical model of a random process.

The theory relies on two concepts: the *random event* and the possibility of performing *long run of experiments* in uniform conditions, if not in practice, at least in principle. These concepts, though looking innocent, are subtle and far reaching.

Random Processes

Though the definition seems innocent and quite reasonable, it is appropriate to examine in some details the assumptions that are implicit in this theory.

First of all we have to examine the meaning of “random” which somehow escapes any precise definition. Let us make some examples to clarify the subject.

When we flip a coin, a classical example of Bernoulli trials, if we know its starting position, velocity, angular momentum, we are able to compute exactly the coin trajectory. If we know also the elasticity of the table where the coin is going to land, we could compute for each flip which would be the result. Then we have a perfectly defined deterministic problem and we leave no space to the game of chance. But usually we do not control all the initial conditions and we know very little of the exact characteristics of the table where the coin is going to bounce. A very little change of the initial conditions has a dominating influence on the result. The macroscopic result is an unpredictable result from each coin flip. The same sort of considerations apply to the other games of chance.

The same situation may arise even in the case of deterministic laws, if the laws are very complicate and not well known. The number of moon eclipses during a year can be predicted with absolute certainty by an astronomer but to an uncultivated person the variations of the number of eclipses from year to year would appear fluctuating as the sequences in a game of chance.

In the field of biology, medicine etc. the arguments of the two previous examples suggest that it is impossible to make exact predictions. Quantities like the life span of an individual, the number of deaths for car accidents in the next weekend, the number of people attending a conference etc. will be subject to “random fluctuations” similar to those observed in the game of chances.

If the observation aims to the measurement of a physical constant, even if the method of measurement and the external conditions are kept constant, we know, by experience that the repetition of the observation will yield different results. This effect is ascribed to the action of very many small disturbing factors which affect the specific measurement in a way that the experimenter cannot control. These effects fluctuate from measurement to measurement in an unpredictable way, introducing an “error” in the measurement. In the same way, small, uncontrollable variation of the production process during the manufacture of an article, would result in a final product whose property are different from unit to unit in an unpredictable way.

In the type of random experiments described above it is not possible to predict the outcomes of individual experiments because of the irregular fluctuations which escape any exact calculation. However the situation changes if we study a sequence of experiments. In spite of the irregular results of each experiment, the average of a long run sequence of real experiments performed in uniform conditions, shows remarkable regularities.

When we throw a die, the outcome of each experiment is unpredictable, but the frequency with which a given outcome occurs becomes rather stable, when the number of experiments becomes very large. This is the statistical regularity which constitutes the empirical basis of the statistical theory.

Historically this remarkable behavior of the frequency was first observed in the field of the games of chance (coin, dices, cards, etc.). The frequency of a given result seemed to converge to the same precise value when the game was repeated very many times. This empirical fact, in the hands of Pascal, Fermat, Huygens, James Bernoulli, in the middle of the seventeenth century gave the origin to the first theory of Probability.

Soon after the same regularity was observed in the field of demographic data. The long run stability of the frequency ratio is a general characteristic of random experiments performed under uniform conditions.

The Population and the Sample

In some cases, mainly when we are concerned with observations on individuals from human population, this regularity is interpreted by considering the observed individuals as *samples* from a very large *parent population*. Consider a population of a large city made of N individuals. We are concerned with a characteristic of the individuals, as their height. We observe n individuals (a sample of individual from the population of the city) and we measure a number n_h of individuals with their height in a given range (h_1, h_2). The frequency $\nu = n_h/n$ will change if we repeat the observation on another sample of n persons. However if the sample size increases, the frequency will tend to the true value $\nu^* = n_H/N$ when we sample the whole population of the city.

We may consider a limiting case when the size of the population N tends to infinity. This is a mathematical abstraction, but helps in figuring out what happens of the relative frequency in a sample. In a finite sample the frequency will be always subject to fluctuations since no sample will saturate the whole population, but it is reasonable to assume that the relative fluctuations of the frequency will decrease when the sample size increases and eventually tend to the *true* value.

Mathematical Formalism

The theory of Probability originated from the problems connected with the game of chances. In these cases the results that, a priori, are possible are arranged in a number of cases (the six sides of a die, the 52 cards of a pack etc.). Laplace was the first to frame explicitly the fundamental principle of equally possible causes. In any kind of observation it is possible to divide the outcomes in equally possible cases and the probability is the ratio between the favorable cases and the total number of possible cases.

This combinatorial definition of the Probability has a serious flaw. In fact no rule is given to decide if two cases are equally possible or not. Moreover it is not clear how to extend the definition to those observations not belonging to the games of chance.

Many authors have attempted to replace the old definition with more flexible ones. The mathematical work in this field was largely influenced by the tendency to built a mathematical theory on axiomatic basis. Von Mieses (1931) defined the probability as the limit of the frequency ratio. The existence of such a limit is the first axiom of his theory.

A different school accept the observation of the constancy of the frequency but avoids the postulate of the existence of the limit and introduces the Probability as a number associated with an event. The mathematical theory proceeds with axioms which express the rules to operate with those numbers, idealizing the observed properties of the frequencies (Kolmogorov, 1933).

The two schools, though differing in the mathematical basis, agree on the assumption that Probability is a theory of phenomena which exhibit statistical regularities of the type discussed above. Both schools agree that the mathematical concept of probability is realized empirically in the frequency and hence they are open to empirical verification of the mathematical assumptions.

A completely different point of view is expressed by a more general definition of Probability as the *degree of belief* (Keynes (1921), Jeffreys (1939), De Finetti (1974)). According to this theory any proposition has a numerically measurable probability. This applies, not only to what we have called random events, but to any hypothesis, proposition etc, that does not have a certainty content. Therefore proposition like:

- The Ergodic hypothesis is true.
- The absolute value of the charge of electrons and protons are equal.
- I will not fail tomorrow's exams
- etc.

have a probability number associated with it. However, since no objective rule can be constructed to express one's opinion on such a type of statements, this theory of Probability is called *subjective*. Since the support of the frequency interpretation to probability is lost, it is not clear if such probabilities can be empirically verified. This is the price the has to be payed to extend the concept of Probability outside the domain of random events.

The repeated Samples

The previous discussion has clarified the role that the samples have in the frequency theory. A sample is what is needed to proceed from the observation to the population from which the sample was drawn. The properties of the population are the object of the scientific measurement. This is especially true in Statistics where we infer from the sample to the population. The population assumes a sort of reality: it is the abstract object from which the measured sample was drawn. The interpretation of probability as the limit of the frequency has given a degree of reality to those samples that have not been measured but are virtually drawn from the population by virtual experiments. This is the standard practice of MonteCarlo methods to compute probabilities. To understand better all the implications of these assumptions that we have made we will discuss in very general terms an example of *statistical inference*. Even if this anticipates a subject that will be discussed further on in these lectures, we believe that analysis of this example is needed now to discuss a typical aspect of the frequency theory.

We shall start by assuming that an experiment has been performed resulting in n measurements of the quantity X. We have a *sample of* n from a population which we will assume to be completely specified by the probability distribution $F(X)$ of the variable under observation x.

To be specific, let us assume that the experiment is the measurement of the decay rate of a radioactive sample; the specific observation has provided a rate R. The population is the ensemble of all possible measurements of the rate, Poisson distributed with mean μ, an infinite population.

From our measurement R we want to make an assessment of the proposition: *the mean decay rate is* μ. This is the hypothesis H_0 whose probability we have to estimate from the single measurement that we have done.

We have done a single measurement and it might be practically impossible to repeat many times the experiment to construct the empirical frequency. However, at least in principle, we can repeat the measurement, as many times as we like. This virtual sampling from the Poisson population of average μ, would make possible to construct a sampling frequency whose limit is the Poisson distribution. From the sampling frequency we can compute the fraction of experiments whose outcome is larger than R, our measured value. Actually is not necessary to perform these experiments, since the frequency can be computed from the Poisson probabilities:

$$f\big(r > R \big| \mu\big) = P\big(r > R \big| \mu\big) = \sum_{n>R} Poiss(n \mid \mu) = \epsilon$$

if ϵ is very small (say 10^{-3}) then we can conclude that we have observed a very rare event, which would occur only once in a thousand experiments, or the model is wrong. On this basis we can decide if the hypothesis H_0 is true.

All this is very plausible. However we are implicitly giving a degree of reality to virtual (and never made) experiments on which we have built our conclusions. This is coherent, and in fact a logical conclusion, from the definition of Probability as the frequency ratio. All our interest was focused on the population and the real measurement is a mere *accident*, only an instance of the possible samples that can be drawn from the population.

1.2 Axiomatic Theory of the Probability

As we have discussed above, it is due to A. Kolmogorov the construction of a mathematical model which describes phenomena showing statistical regularities. His theory is constructed on sets which gather the outcomes of the experiment. We will begin by refreshing the simplest rules of the set theory.

1.2.1 Sets

Assume that in a given experiment the possible outcomes are:

$$\{e_1 \ e_2 \ \cdots \ \}$$

For instance if we throw a dice the possible outcomes are:

$$\{1 \ 2 \ 3 \ 4 \ 5 \ 6\}$$

(Here we assume for simplicity that the possible outcomes are numerable.) Each e_i is called *an elementary event*. The collection of all e_i is called *sample space*.

An *event* is a subset of S, i.e. the ensemble of elementary events that share the same property. In the previous example an event is a roll of a die whose results is even. However the set of all possible subsets of S is mathematically not free of ambiguities (in the case of continuous spaces) and we need a more subtle concept, as we will see in the next section.

If we toss a coin the sample space is simply:

$$\{H \quad T\}$$

A subset A of the sample space S is indicated by:

$$A \subseteq S$$

is a collection of events comprised in S. (By definition S is a subset of itself.) The set of elements $x_i \in S$ which are NOT in A is a subset of S, it is called the complement of A and is indicated as $\overline{A}$.

Example 1.2 *In a throw of two dices the sample space of the outcomes is:*

$$S \equiv \{2 \ 3 \ \cdots \ 12\}$$

The set of even outcomes and the set of odd outcomes:

$$E \equiv \{2\ 4\ 6\ 8\ 10\ 12\}$$

$$O \equiv \{3\ 5\ 7\ 9\ 11\}$$

are subsets of S. In addition E is the complement of O.

Example 1.3 *Let us throw a coin 10 times. An elementary event is a sequence of ten heads and tails (an ntuples):*

$$e_i = \{T\ T\ H\ \cdots\}$$

$$S \equiv \{All\ possible\ distinct\ sequences\ of\ 10\ H\ or\ T\}$$

If we call:

$$A_i \equiv \{set\ of\ sequences\ with\ the\ i-th\ throw\ gave\ H\}$$

then the complement to A_i is:

$$\overline{A_i} \equiv \{set\ of\ sequences\ with\ the\ i-th\ throw\ gave\ T\}$$

An useful tool to visualize these concepts and illustrate the algebra of sets is provided by the Venn diagrams (Fig. 1.1).

Let us consider two subsets of the sample space, A and B:

$$A \equiv \{set\ of\ outcomes\ with\ property\ A\} \subseteq S$$

$$B \equiv \{set\ of\ outcomes\ with\ property\ B\} \subseteq S$$

We define *union* of the two sets A and B:

$$A \cup B \equiv \{set\ of\ outcomes\ with\ either\ property\ A\ OR\ B\}$$

and the *intersection* of A and B:

$$A \cap B \equiv \{set\ of\ outcomes\ with\ property\ A\ AND\ B\}$$

The complement, union and intersection are shown in Fig.1.1.

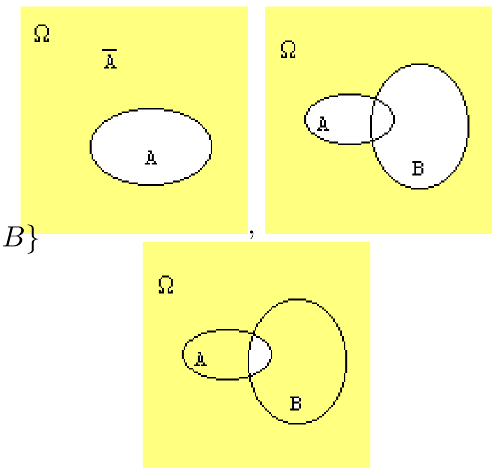


Figure 1.1: *Venn diagrams.*

Example 1.4 *If we toss a coin 100 times, an event is an n-tuple:*

$$x \equiv \{e_1 \cdots e_{100}\} \qquad e_i \equiv H\ or\ T$$

the set of all events such that the i-th toss has given head:

$$A_i \equiv \{x \subseteq S : x_i = H\}$$

$$\bigcup_n A_n \equiv \{events\ with\ at\ least\ one\ head\}$$

$$\bigcap_n A_n \equiv \{one\ event\ with\ all\ heads\}$$

Exercise 1.5 *In an experiment we toss a coin three times.*

- *Write down the sample space S.*
- *Write down the subset A consisting of events with H at the first toss.*
- *Write down the subset B consisting of events with T at the first toss.*
- *Write down the intersection $A \bigcap B$.*
- *Write down the subset A_2 consisting of events with H at the second toss.*
- *Write down the intersection $B \bigcap A_2$*

Example 1.5 *Let us verify the relation:*

$$A = (A \cap B) \cup (A \cap \overline{B})$$

where A and B are any subsets of the sample space. We can verify the relation with a truth table (Tab 1.3). In the table the fist two columns exhaust all the possible assignments of $x \in S$. The last column shows that only when $x \in A$ it belong also to $(A \cap B) \cup (A \cap \overline{B})$

$x \in A$	$x \in B$	$x \in A \cap B$	$x \in A \cap \overline{B}$	$x \in (A \cap B) \cup (A \cap \overline{B})$
T	T	T	F	T
T	F	F	T	T
F	T	F	F	F
F	F	F	F	F

Table 1.3: *Truth table for the relation $(A \cap B) \cup (A \cap \overline{B})$. The first two columns exhaust all possibilities for the values of x.*

Exercise 1.6 *Draw the Venn diagram to verify the previous relation.*

Exercise 1.7 *Verify with truth table and Venn diagrams the two following distributive laws:*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Exercise 1.8 *Verify with the truth table and Venn diagram the following relation:*

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

Its generalization:

$$\overline{A \cup B \cup C \cdots} = \overline{A} \cap \overline{B} \cap \overline{C} \cdots$$

is straightforward and is called DeMorgan's theorem.

1.2.2 Axiomatic Definition of Probability

Kolmogorov solved the problem of defining consistently the probability of occurrence of an event by using the set theory.

Let us start from the set of elementary outcomes: $S \equiv \{sample\ space\}$. We will call *event class* E the class of all subsets of S.

- S is in E, S is an event.

- If A is in E, also $\overline{A}$ is in E.
- If $A_1, A_2 \cdots$ are in E then also their union $(\bigcup_n A_n)$ is in E.
- If $A_1, A_2 \cdots$ are in E then also their intersection $(\bigcap_n A_n)$ is in E. (This is, in fact, a consequence of the two previous properties.)

An event class is also called a *sigma algebra* or *a sigma field* or a *Borel field*. The elements of E are called *events*.

On this structure (the sample space S and the event class E) we define, for each A in E a non negative measure $P(E)$ called *probability* with the following properties (axioms):

- for any A in E
$$0 \leq P(A) \leq 1$$
- The probability of occurrence of any event is one:
$$P(S) = 1.$$
- If sets A_i of elementary events have **NO** element in common, then the probability of events belonging to any A_i
$$P(\bigcup_i A_i) = \sum_i P(A)$$
(countable additivity).

1.2.3 The Laws of Probability

With the use of the axioms and the properties of set theory it is easy to prove the following theorems:

Theorem 1.1 *For any subset A of S*

$$P(\overline{A}) = 1 - P(A)$$

Proof 1.1 *In fact $1 = P(S) = P(A) + P(\overline{A})$, since $S = A \cap \overline{A}$ and $A \cap \overline{A} = 0$.* ∎

Theorem 1.2 *If $B \subseteq A$ then $P(B) \leq P(A)$*

Proof 1.2 *We can write $A = (A \cap B) \cup (A \cap \overline{B})$ and since $B \subseteq A$, $A = B \cup (A \cap \overline{B})$, the union of two disjoint sets, from which: $P(A) = P(B) + P(A \cap \overline{B}) \geq P(B)$.* ∎

Theorem 1.3 *For any two subsets of S (A and B):*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof 1.3 *We can write $B = (A \cap B) \cup (\overline{A} \cap B)$, the union of two disjoint sets, therefore $P(B) = P(A \cap B) + P(\overline{A} \cap B)$. We can also write: $A \cup B = (A \cup \overline{A}) \cap (A \cup B) = A \cup (\overline{A} \cap B)$ (distributive law), the sum of two disjoint sets. Then we have: $P(A \cup B) = P(A) + P(\overline{A} \cap B)$, from which the theorem follows.* ∎

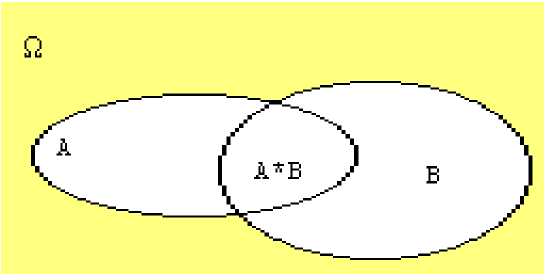


Figure 1.2: *Adding probabilities.*

With a Venn diagram this theorem is intuitive: Fig. 1.2 shows the two subsets A and B. If we add: $P(A) + P(B)$ the probability of the intersection $A \cap B$ is counted twice.

Example 1.6 *In the roll of two dices success is defined if at least one lands even. What is the probability of success? We have to compute $P(E_1 \cup E_2)$. If the dices are fair we have $P(E_{1,2}) = 1/2$ and, since the events E_1 and E_2 are independent:*

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= P(E_1) + P(E_2) - P(E_1) \times P(E_2) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4} \end{aligned}$$

The rule holds even if one die is loaded to give $P(E_1) = 1$, as can be easily verified.

Exercise 1.9 *Prove the theorems.*

Theorem 1.4

$$P(\bigcup_i A_i) \leq \sum_i P(A_i)$$

Theorem 1.5

$$P(A \bigcap B) \geq 1 - P(\overline{A}) - P(\overline{B})$$

Theorem 1.6

$$P(\bigcap_i A_i) \geq 1 - \sum_i P(\overline{A_i})$$

Example 1.7 *To explain theorem 3 let us take a card from a pack of 52.*

$$\begin{aligned} P(King \cup Club) &= P(King) + P(Club) - P(King\ of\ Club) \\ &= 4/52 + 13/52 - 1/52 = 16/52 \end{aligned}$$

The law can be generalized:

$$\begin{aligned} P(e_1 \cup e_2 + \cdots) &= \sum_i P(e_i) - \sum_{i<j} P(e_i \cap e_j) + \\ &+ \sum_{j<i<k} P(e_i \cap e_j \cap e_k) - \cdots + (-1)^{m+1} P(e_1 \cap \cdots e_m) \end{aligned}$$

In particular if $e_1 \cap e_2 = 0$, (the two sets have no common members) then:

$$P(e_1 \cup e_2) = P(e_1) + P(e_2)$$

Example 1.8 *We roll three dices. As in the previous example we define success if at least one die lands even. Which is the probability of success?*

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P(E_1 \cup (E_2 \cup E_3)) \\ &= P(E_1) + P(E_2 \cup E_3) - P(E_1 \cap (E_2 \cup E_3)) \\ &= P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_2) - P(E_1 \cap (E_2 \cup E_3))) \end{aligned}$$

But, since:

$$E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3)$$

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_2) - \\ &- P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3) \end{aligned}$$

The expression can be further simplified by the independence of the three events. If all the dices are fair we get $P(E_1 \cup E_2 \cup E_3) = 17/27$. The rule is true also if one die is loaded to give always even. In this case we obtain $P(E_1 \cup E_2 \cup E_3) = 1$. Draw also a Venn diagram to convince yourself.

Example 1.9 *Assume that we want to detect a π^0. π^0 are unstable particles that are produced in the collision of energetic hadron and decay, with a mean life $\tau \approx 8 \cdot 10^{-17}$sec. The most probable decay mode is:*

$$\pi^0 \to \gamma + \gamma$$

Let us assume that we can produce such π^0 and an apparatus exists which is able to detect γ. If the efficiency to detect a single γ is ϵ, what is the efficiency to detect a π^0? Let us consider a few cases.

- *We want to detect both photons. The detection of γs are two independent measurements*[1] *therefore:*

$$P(2\gamma) = P(\gamma \cap \gamma) = P(\gamma) \cdot P(\gamma) = \epsilon^2$$

- *The event of detecting only one photon has, instead the probability:*

$$\begin{aligned} P(1\gamma) &= P(\gamma_1) \times P(\overline{\gamma_2}) + P(\gamma_2) \times P(\overline{\gamma_1}) \\ &= P(\gamma_1) \times (1 - P(\gamma_2)) + P(\gamma_2) \times (1 - P(\gamma_1)) \\ &= 2\epsilon(1 - \epsilon) \end{aligned}$$

- *The event of detecting no photons has a probability:*

$$P(0\gamma) = P(\overline{\gamma_1}) \times P(\overline{\gamma_2}) = (1 - \epsilon)^2$$

These cases have exhausted all the possible situations. The sum of the probabilities is one:

$$\begin{aligned} P(2\gamma) + P(1\gamma) + P(0\gamma) &= \epsilon^2 + 2\epsilon(1 - \epsilon) + (1 - \epsilon)^2 \\ &= \epsilon^2 + 2\epsilon - 2\epsilon^2 + 1 - 2\epsilon + \epsilon^2 = 1 \end{aligned}$$

The probability of detecting a π^0 might also be defined as the probability of detecting at least one of the γs:

$$\begin{aligned} P(\geq 1\gamma) &= P(\gamma_1 \cup \gamma_2) = \\ &= P(\gamma_1) + P(\gamma_2) - P(\gamma_1 \cap \gamma_2) = \\ &= \epsilon(2 - \epsilon) \end{aligned}$$

So we can choice what we define "detect a π^0" and use the corresponding efficiency.

[1] This might be not so clear. Consider a π^0 produced with momentum $\vec{p}$. The energy of each photon is not fixed but is uniformly distributed from a minimum to a maximum value. The kinematics fix a constrain so that when one photon has the maximum energy, the other has the minimum energy. If the probability to detect a photon depends of the photon energy (as it is usually the case) then in a very asymmetric decay we detect more likely the energetic photon and less likely the other photon. This might induce a dependence in the detection probability. However the event of detecting or not a γ is independent of the detection of the other. A γ which has a low energy can be paired to a high energy one, but could be the decay of a low energy π^0. The independence, as you see, is a rather subtle concept.

1.2.4 Conditional Probability

Suppose A and B are subsets of the sample space S, having probabilities $P(A)$ and $P(B)$. Suppose further that we are interested only in the elements of A. We want to express the probabilities *relative* to the sample space A. This new probability will be called *conditional probability*. The Venn diagram (Fig.1.3) illustrates the meaning of conditional probability. By inspecting the different areas in the diagram we write:

$$
\begin{aligned}
P(A) &= \frac{N_A}{N} \\
P(B) &= \frac{N_B}{N} \\
P(A \cap B) &= \frac{N_C}{N} \\
P(B|A) &= \frac{N_C}{N_A} \\
P(A|B) &= \frac{N_C}{N_B}
\end{aligned}
$$

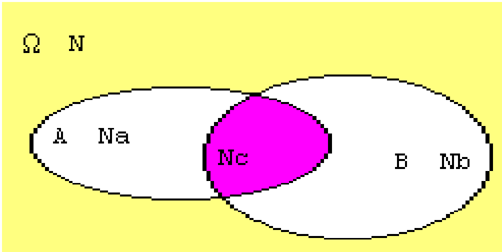


Figure 1.3: *Conditional probability, definition of the elements.*

$P(A \cap B)$ is the measure of the overlap region of A and B relative to Ω. $P(B|A)$ is the measure of the overlap relative to A, while $P(B|A)$ is the overlap relative to B. $P(B|A)$ is the probability of B given A (*conditional probability*).

These probabilities satisfy:

$$
\begin{aligned}
P(A \cap B) &= P(B|A) \cdot P(A) \\
P(A \cap B) &= P(A|B) \cdot P(B) \\
P(A|B) &= \frac{P(A \cap B)}{P(B)}
\end{aligned}
$$

All probabilities are in fact always conditional probabilities, being subject to some condition, often not explicitly stated.

If the knowledge that the event E_1 has occurred does not affect the probability of event E_2, E_2 is said to be *independent* of E_1. Then:

$$
\begin{aligned}
P(E_2|E_1) &= P(E_2) \qquad \textit{and also} \\
P(E_1 \cap E_2) &= P(E_1) \cdot P(E_2)
\end{aligned}
$$

Example 1.10 *A card is drawn from a pack of 52.*

$$
\begin{aligned}
P(A) &= P(\textit{an ace}) = \frac{4}{52} \qquad \textit{and} \\
P(B) &= P(\textit{a spade}) = \frac{13}{52}
\end{aligned}
$$

Since $P(ace\ of\ spade) = 1/52 = P(A) \cdot P(B)$, the events are independent. However, if we add a joker to the pack:

$$
\begin{aligned}
P(A) &= \frac{4}{53} \qquad \textit{and} \\
P(B) &= \frac{13}{53} \qquad \textit{but} \\
P(A \cap B) &= \frac{1}{53} \qquad \textit{while} \\
P(A) \cdot P(B) &= 1/53 - 1/53^2
\end{aligned}
$$

The two events (A = ace, B = spade) are no longer independent. The argument is that if we know that a card is an ace, then it cannot be a joker and we have added information to the problem if it is a spade or not.

Example 1.11 *We roll two dices. What is the probability that the sum is 9, if the first die gave 5? Let us call F_5 the event in which the first die landed 5 and S_9 the event in which the sum gave 9. Let us write down the sample spaces.*

$$\begin{aligned} S &= \{(1,1),(1,2),\cdots(6,6)\} \\ F_5 &= \{(5,1),(5,2),(5,3),5,4),(5,5),(5,6)\} \\ S_9 &= \{(3,6),(4,5),(5,4),6,3)\} \\ S_9 \cap F_5 &= \{(5,4)\} \end{aligned}$$

hence:

$$P(F_5) = \frac{6}{36} \qquad\qquad P(S_9 \cap F_5) = \frac{1}{36}$$

$$P(S_9 \mid F_5) = \frac{P(S_9 \cap F_5)}{P(F_5)} = \frac{1}{6}$$

Example 1.12 *The lifetime of a radioactive atom is described by the law:*

$$P(t) \propto e^{-t/\tau}$$

Assume we have prepared the system at time $t = 0$. The probability that the atom is not yet decayed at time x is: $P(t > x) \propto exp(-x/\tau)$. What is the probability that the atom will not decay in the following y seconds?

- *Event A: the atom still exists at time x.*
- *Event B: the atom still exists at time $x + y$.*

Note that $B = A \cap B$. We have to compute the conditional probability:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)} \propto exp(-y/\tau)$$

The probability of surviving the next y seconds is the same as the probability of surviving the first y seconds.

Exercise 1.10 *Consider two events A and B. We know that:*

$$P(A) = 0.1$$

$$P(B) = 1.0$$

are they independent?

Exercise 1.11 *Consider two events A and B which are mutually exclusive. Are they independent?*
(Hint: what is $P(A \cap B)$?)

Exercise 1.12 *Verify that if A and B are independent events, then:*

$$P(A \cup B) = 1 - P(\overline{A} \cap \overline{B}) = 1 - P(\overline{A}) \cdot P(\overline{B})$$

Exercise 1.13 *Which relation exists, if any, between $P(A \mid B)$ and $P(A \mid \overline{B})$?*

Exercise 1.14 *Prove the following relation:*

$$P(A \cap B \cap C) = P(A \mid B \cap C) \cdot P(B \mid C) \cdot P(C)$$

and generalize to:

$$P(\cap_i A_i)$$

Exercise 1.15 *In a high energy experiment two independent programs are used to identify the class of events A. The first program has found N_1 events and the second program N_2 events. N_{12} events were found by both programs. Which is the detection efficiency of the two programs?*

Exercise 1.16 *Capture efficiency of light in a optical fiber. The readout of a plastic scintillator is sometimes realized with optical fibers which capture a small fraction of light, shift its frequency and emits it uniformly so that a small fraction is channeled by the fiber and taken to a light detector. The readout can be made of one, two or three fibers. Assume that N photons are emitted by the scintillator and that a single fiber can shift only m photons. What would be the light shifted in the three configurations?*

1.2.5 Marginal Probability

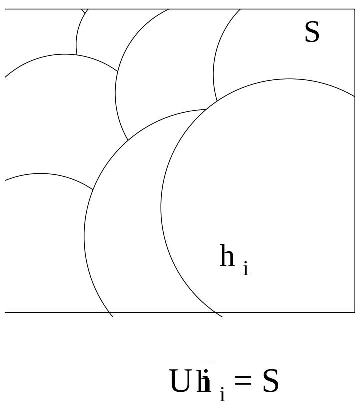


Figure 1.4: *Marginal probability.*

Let us consider an ensemble of subsets of S $\{H_i, i = 1 \cdots m\}$ such that they cover the whole sample space S and are mutually disjoint (see figure 1.4):

$$\begin{aligned} S &= \cup_i H_i \\ H_i \cap H_j &= 0 \qquad if \quad i \neq j \end{aligned}$$

We can then write:

$$A = A \cap \cup_i H_i = \cup_i A \cap H_i$$

Since the H_i are mutually disjoint sets we have:

$$P(A) = \sum_i P(A \cap H_i) = \sum_i P(A \mid H_i) P(H_i)$$

1.2.6 Bayes' Theorem

We have seen how to compute the conditional probability $P(A \mid B)$, i.e. the probability of A given B, or the probability of A knowing that B has occurred. We now face the problem of computing the *inverse* probability: given A what is the probability of B: $P(B \mid A)$.

The solution of this problem is due to Rev. T. Bayes at the beginning of 17th century and mathematically it is a simple manipulation of conditional and marginal probability. However the interpretation of the theorem is subtle and points to fundamental concepts at the basis of the probability theory.

Assume that the sample space Ω is divided among n mutually exclusive subsets B_i.

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

If A is also a set belonging to Ω then:

$$P(A \cap B_i) = P(B_i \mid A) \cdot P(A) = P(A \mid B_i) \cdot P(B_i)$$

therefore:

$$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{P(A)}$$

But:

$$P(A) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(A \mid B_j) \cdot P(B_j)$$

Finally:

$$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_{j=1,n} P(A \mid B_j) \cdot P(B_j)}$$

This is Bayes theorem.

Example 1.13 *Each of three urns U_1, U_2 and U_3 contain two coins. The first urn contains two gold coins, the second urn one gold and one silver coin and the third two silver coins. We choose an urn at random (we do not know which is urn 1).*

A indicates the event of picking up a gold coin. The the conditional probabilities: $P(A|U_i)$ are the probabilities that given the urn U_i we pick up the gold coin. $P(U_i)$ is the probability of selecting the urn i. $P(U_i|A)$ is the a posteriori probability that given the observation of a golden coin we have chosen the urn i.

- *$P(U_i)$ are the prior probabilities,*
- *$P(U_i|A)$ are the posterior probabilities (improved information on which is urn 1) and*
- *$P(A|U_i)$ is called Likelihood.*

We have to know the prior probabilities before we can apply Bayes theorem. Since the urns are selected at random, it seems reasonable to put: $P(U_i) = 1/3$ (Bayes postulate, see later).

In addition $P(A|U_1) = 1$, $P(A|U_2) = 1/2$ and $P(A|U_3) = 0$ by construction. Now we can compute the posterior probabilities:

$$\begin{aligned}
P(U_1|A) &= \frac{1 \cdot 1/3}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3} \\
P(U_2|A) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{3} \\
P(U_3|A) &= 0
\end{aligned}$$

The urn that we have randomly selected and in which we have found a gold coin has the larger probability of being U_1. Our chances of getting the second gold coin would be double if we draw the second coin out of the same urn. The observation has changed the prior probabilities.

To apply Bayes theorem we need to know the prior probabilities. Suppose we are not told the method to select the urns in the previous problem, then we would not know $P(U_i)$. How do we compute the prior probabilities that correspond to our ignorance? It seems that there is no satisfactory answer to this question.

Example 1.14 *A box contains 3 balls either black or white. We take a ball, it is white. We replace the ball and extract a second ball; again white. We replace the ball and make a third extraction; again white. What is the probability that all the 3 balls in the box are white? Let us call:*

- *H: the 3 balls in the box are white.*
- *E: the 3 independent extractions give 3 white balls.*

In this example H is not a random variable, rather it is a fact: the composition of the box and $P(H)$ is either 1 or 0 depending on which balls were put in the box.

We can recover somehow the probability concept by imagining that we have chosen the actual box among a collection of boxes with different mixtures of white and black balls. $P(H)$ is then the probability of having chosen the box with 3 white balls. From this discussion it is clear that the problem is lacking some data: how was made the collection of boxes and how do we chose among them?

To be more specific, let us imagine two reasonable possibilities:

A) We make a collection of boxes each with a different composition:

$$\begin{aligned} B_1 &= \{W\ W\ W\} \\ B_2 &= \{W\ W\ B\} \\ B_3 &= \{W\ B\ B\} \\ B_4 &= \{B\ B\ B\} \end{aligned}$$

B) We consider that, since we have two possibilities for each ball (B or W), then we have 8 (2^3) possible combination among which to choose:

$$\begin{aligned} B_1 &= \{W\ W\ W\} \\ B_2,\ B_3,\ B_4 &= \{W\ W\ B\}\ \{W\ B\ W\}\ \{B\ W\ W\} \\ B_5,\ B_6,\ B_7 &= \{W\ B\ B\}\ \{B\ B\ W\}\ \{B\ W\ B\} \\ B_8 &= \{B\ B\ B\} \end{aligned}$$

You can find some analogy with the problem of the two dices discussed above. Here, however, the problem is substantial: you have no way to decide which of the two collections is the correct one. Both are reasonable and it is to the problem to specify the collection.

Assume that the problem specifies the second collection, then we can work out the solution of the problem. Let us call B_i the event of having selected the box i. Bayes theorem reads ($H = B_1$ and E = 3 white balls):

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{\sum P(E \mid B_i) \cdot P(B_i)}$$

Now: $P(E \mid H) = 1$ and $P(H) = \frac{1}{8}$.

$$\sum P(E \mid B_i) \cdot P(B_i) = \frac{1}{8} + (\frac{2}{3})^3\frac{3}{8} + (\frac{1}{3})^3\frac{3}{8} + 0$$

and the net result is:

$$P(H \mid E) = \frac{1}{2}$$

Exercise 1.17 *what is the probability if collection A was specified?*

Example 1.15 *A beam of high energy particles is composed of pions and electrons in the known ratio: $\frac{N_e}{N_\pi} = 10^{-3}$ Our experiment is interested only in electron induced reactions. To this purpose we have made a Trigger, i.e. an arrangement of particle detectors which has a larger efficiency to detect electrons than pions (Fig. 1.5).*

Let us assume that our Trigger is such that the efficiencies are:

$$\epsilon_e = 0.98 \qquad \epsilon_\pi = 0.03$$

What is the probability that the incoming particle is an electron if the trigger has fired?

$$
\begin{aligned}
A &= event: \ the\ beam\ particle\ is\ an\ electron \\
B &= event: \ the\ beam\ particle\ is\ a\ pion \\
P(A) &= 10^{-3} \\
P(B) &= 1 - 10^{-3} \\
P(T \mid A) &= \epsilon_e = 0.98 \\
P(T \mid B) &= \epsilon_\pi = 0.03 \\
P(A \mid T) &= \frac{P(T \mid A) \cdot P(A)}{P(T \mid A) \cdot P(A) + P(T \mid \pi) \cdot P(B)} = 0.032
\end{aligned}
$$

But A and B are really random events? A particle is either a pion or an electron. How can we talk of probabilities of real events? Note also that, in order to apply Bayes' theorem, we have taken $P(e)$ from the relative abundance of electrons in the beam. When the Trigger fires no other information was available.

Exercise 1.18 *Assume that we have doubled the trigger system. How would you use the information of the two triggers? What is $P(A \mid T)$ in this case? (Assume that the two systems are identical.)*

Example 1.16 *Epidemiological studies have established that 0.1% of the population of a state is affected by AIDS. A test exists which yields positive results in sick persons with a probability of 0.98, fails in only 2% of cases. The test, however, gives a positive result in a sane person in 3% of cases. You pass the test with positive result. What is the probability that you are really affected by AIDS?*

This example is, in fact identical to the previous one:

$$
\begin{aligned}
P(AIDS) &= 10^{-3} \\
P(\overline{AIDS}) &= 1 - 10^{-3} \\
P(T \mid AIDS) &= \epsilon_e = 0.98 \\
P(T \mid \overline{AIDS}) &= \epsilon_\pi = 0.03 \\
P(AIDS \mid T) &= \frac{P(T \mid AIDS) \cdot P(AIDS)}{P(T \mid AIDS) \cdot P(AIDS) + P(T \mid \overline{AIDS}) \cdot P(\overline{AIDS})} = 0.032
\end{aligned}
$$

Even if the test is positive you don't have to worry too much, it is very likely that you are sane.

$P(AIDS)$ is the key to understand this apparent paradox. You have used the value averaged on all the population and, since most of the persons are sane, we have found a probability very near to the probability that the test fails ($P(T \mid \overline{AIDS})$). If the test is positive you would have to repeat the test and use the first one as a new prior in Bayes formula.

However if a person undertakes an AIDS test, there are probably other symptoms or suspects, such that the prior to be used will be much larger that the average on all the population and the outcome of the test will be much more meaningful.

Also in this case we have seen that Bayes theorem depends critically on the prior. Often the prior depends not on objective facts but on the taste of the experimenter. The resulting probabilities are related not only on the measurements but also on the set of assumptions that have determined the prior.

1.3 Bayes Theorem and the Subjective Probability

Io iudico bene questo: che sia meglio essere impetuoso
che respettivo; perche' la fortuna e' donna, ed e' necessario, volendola
tener sotto, batterla e urtarla. E si vede che la si lascia piu' vincere
da questo che da quelli che freddamente procedono; e pero' sempre,
come donna, e' amica de' giovani, perche' sono meno rispettivi, piu' feroci,
e con piu' audacia la comandano.
N. Machiavelli, Il Principe, Cap XXV, Par. 8

The Bayesian interpretation of *probability* is not based on the empirical evidence of random events and the stability of their frequency in a long sequence of experiments; rather it is epistemological and considers propositions on which we do not have certainty. Though there is not certainty, still a proposition can appear more or less plausible. The probability is a measure of this plausibility and it is expressed as a real number that we can take in the interval $(0, 1)$.

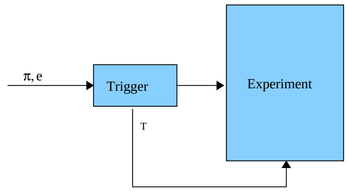


Figure 1.5: *A trigger system.*

Both interpretations of probability where present at the dawn of this science, in the middle of seventeenth century. The very name of probability has its etymology in the Latin word *probus*: praised by a wise person and seems to indicate the cognitive meaning of the word.

The best way to understand this concept of probability, as opposed to the statistical one, is to follow the same line of reasoning of Bayes. Rev. Bayes knew the answer to the problem: In a Bernoulli trial of N with probability p what is the probability of n successes? The calculation can be done if p is known and leads to the Bernoulli formula $P(n \mid p, N)$. But, if p is unknown and, in the experiment we have measured n, what can be said of p? The symmetry of Bayes theorem suggests the answer:

$$P(p \mid n, N) \propto P(n \mid p, N) \cdot P(p)$$

However there are two conceptual problems:

- How can we talk of $P(p \mid n, N)$, since p is a constant?
- What is $P(p)$ (the prior)?

Bayes could give a meaning to $P(p \mid n, N)$: since p is unknown, it is subject to probabilistic analysis. This idea was already elaborated by Leibniz, among the others. $P(p \mid n)$ is a way of quantifying our personal believe on the value of p, before and after the measurement. Different persons can have different opinions and hence different $P(p \mid n)$. p is not a random event, it is an unknown parameter whose value we try to evaluate based on our opinion and on the result of the experiment. In this scheme no frequency interpretation of $P(p)$ is possible, and no verification by the experience.

The problem that stopped Bayes from publishing his theorem was how to convey our knowledge of p in the mathematical expression of $P(p)$. The greatest difficulty concerns the case where we do not have any information on p. The naive solution to consider all the possibilities as equally probable (an uniform distribution) leads often to inconsistencies, as he soon realized. A simple way to show the key of the inconsistency is the following. Assume that a parameter θ can take any of the values: $\theta_1, \theta_2, \cdots, \theta_k$ with $k \geq 3$. If nothing is new on θ then it appears natural to assume $P(\theta) = 1/k$. Let us now define a new variable ϕ such that $\phi = 1$ if $\theta = \theta_1$ and 0 in any other cases. Being ignorant on θ we are also ignorant on ϕ. However the two priors:

$$P(\theta) = \frac{1}{k} \qquad \text{and} \qquad P(\phi) = \frac{1}{2}$$

are clearly inconsistent.

In fact this problem is one of the most difficult of the Bayesian statistics and after more than two centuries still lacks of a clean solution.

The use of Bayes theorem is the basis of Bayesian statistics. We have proved the theorem in the framework of axiomatic theory of probability, when the terms in the equation are random events. Let us consider an example where Bayes theorem is used in a more general sense.

Assume that we have made an experiment to measure the lifetime of an unstable particle. This experiment has yielded the value t_1. We know that the probability density is:

$$P(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

Here τ is the particle's lifetime. We use Bayes theorem to invert this probability:

$$P(\tau \mid t_1) = \frac{P(t_1 \mid \tau) \cdot P(\tau)}{\int P(t_1 \mid \tau) P(\tau) d\tau}$$

which is now interpreted as the probability density of τ, given the measurement t_1 and the prior knowledge $P(\tau)$. The quantity $P(t_1 \mid \tau)$ is called *likelihood* of the measurement.

- τ is a parameter and has a precise, though unknown value. We cannot speak, logically, of probability distribution of τ. There are cases where we can safely speak of frequency distribution of a parameter. In Physics this is not usually the case, parameters as the mass of the neutrino, coupling constants etc, are considered constants of the Nature, fixed even if unknown. As such they cannot be subject to probability statements. However $P(\tau \mid t_1)$ can be considered a probability function of our rational belief on the value of τ. In this way we can give sense to the machinery of Bayes theorem which takes the meaning of a way of improving of out knowledge on the parameter. Once the measurement is made our knowledge is increased and the *posterior* probability $P(\tau \mid t_1)$ will become the prior of the next measurement and so on.

- Before we can use Bayes theorem, we have to know the prior $P(\tau)$. This means that we need a prescription of how to translate our prior knowledge (prior to the actual measurement) into a precise mathematical form: $P(\tau)$. There are many cases where $P(\tau)$ is known, from previous experiments, but what to do if we are the first to measure the lifetime of this particle? We cannot use a flat prior in the interval $(0, \infty)$ since neither the prior or the posterior would be normalizable. We could use, as a guide, the common sense that tells us that lifetime τ is comprised in a more restricted interval, for instance $\tau \in (10^{-19}, 10^{19})$ *seconds*. However another physicist could have a different opinion and choose a different interval and obtain a different result.

It is precisely this use of Bayes theorem that is at the basis of the subjective definition of probability and of the Bayesian statistics. It is different from the frequency theory of statistics, but completely sound, once the two above facts are accepted. This point of view was used by Laplace, Poisson and Poincare', among the others, till the beginning of the twentieth century. It was the intrinsic ambiguity of the form of the prior probability and the consequent subjectivity of the results that lead Neyman, the Pearsons, Fisher and others to introduce the concepts of frequency in the definition of probability and to start the classical or frequency theory of statistics.

1.3.1 Bayes Postulate

If we have no knowledge with which to estimate the prior probability density, we can measure our complete ignorance by assuming an equally likely prior:

$$P(U_i) = Constant$$

This is Bayes postulate in its simplest form. Bayes himself was so doubtful of the validity of his postulate that he did not publish his work during his life.

The problem of the prior is even more complicate if the possible U_i are not events but parameters, or hypothesis, that we have to estimate from data. The parameters, as the composition of each urn, are not random variables but fixed parameters in the classical approach to statistics.

Bayes theorem is the basis of the *Subjective* definition of probability: the posterior probability measures one's belief in different possible values of the parameter, after the knowledge gained in the measurement. The prior probability measures one's belief on the parameters before the measurement is made.

There is a school of statistics called *Bayesian* or *modern* which is based on these assumptions. In Bayesian statistics one assumes that the parameters have not fixed values but are described by a probability distribution, as any random variable. (Or, in softer terms, one's rational belief on the values of the parameters can be described by a probability distribution). This distribution is changed (updated) after each measurement.

The criticism to Bayesian interpretation is that any physicist will have different degrees of belief and the conclusions drawn from a measurement will be subjective. The defense is that in planning and analyzing any experiment there is a large number of assumptions which are shared among scientists. In designing, for instance, an experiment to measure the mass of neutrinos, we consider only the eV range of masses. This, and many other assumptions, are the *priors* shared by the community of scientists which work in the field. After all information is made explicit, all physicists should agree on prior probabilities.

As an example of the kind of difficulty in the use of Bayes postulate, let us consider the decay of a unstable particle. The data can be described in terms of the lifetime τ or the decay constant λ (=$1/\tau$). We could assume for prior probability $P(\tau) = cost$. Or $P(\lambda) = cost$ if we are more interested on λ. But (as we will see better later on):

$$P(\tau) = P(\lambda) \mid \frac{d\lambda}{d\tau} \mid = \lambda^2 P(\lambda)$$

The prior probability functions are different for the two variables, therefore point and interval estimation will be different. We have a very unscientific situation of two physicists who reach different conclusions using the same data.

This is how Kendall, a modern frequency theory expert in statistics, describes the status of the art:

The principal difficulties arising out of Bayes' postulate appear from the standpoint of the frequency theory of probability. If we adopt the approach in which probability is a measure of attitude of mind, it is reasonable to take prior probabilities to be equal when nothing is known to the contrary, for the mind holds them in equal doubt. The frequency theory, however, would require the states of events corresponding to various θ_i (the unknown parameters) to be distributed with equal frequency in some population from which the actual θ_i has emanated, if Bayes' postulate is to be applied. This appears to some statisticians, though not to all, to be asking too much of the universe. The postulate is one of the crucial points in the theory of probability. Many adherents of the subjective school accept it. Many of those of the frequency school reject it. (.....) There is still so much disagreement on this subject that one cannot put forward any set of view points as orthodox. One thing, however, is clear: anyone who rejects Bayes' postulate must put something *on its place. The problem which Bayes attempted to solve is supremely important in scientific inference and it scarcely seems possible to have any scientific thought at all without some solution, however intuitive and however empirical to the problem. We are constantly compelled to assess the degree of credence to be accorded to hypotheses on given data; the struggle for existence, in Thiele's phrase, compels us to consult the oracles. Various substitutes for Bayes' postulate have been proposed. Some of these are put forward as solution to the problem in particular classes of cases; such are the principles of least squares and the minimum chi-square. There is one principle, however, of general application, that of maximum likelihood. [13]*

1.3.2 Bayes and Frequency Theory

Let us consider the example discussed above where we measure the lifetime of unstable atoms. The experiment yielded a sequence of results: $\{t_1, t_2 \cdots t_n\}$. The lifetime of the substance is τ whose value is unknown. Here the data are repeatable, since I can continue the measurements as long as I want, and the frequency has a well defined meaning. It seems that in this example both schools of probability would agree at least on the conceptual scheme. Unfortunately this is not the case.

In Frequency theory τ is a fixed, though unknown, parameter. Since it is a parameter we cannot make any probability statements on it, as it would be a random variable. On the contrary $t_1 \cdots$ are random variables, distributed with a density:

$$t \sim f(t \mid \tau)$$

All the probability statements are derived from the distribution of data for a given parameter τ. Those values of τ that would produce an unlikely value of the data have to be rejected. The probability is an intrinsic property of the system. In an ontological way we could say that the system exists, independently of the measurements that have been done; experiments can be repeated at will and their outcomes represents the real system. In this conceptual scheme the samples represents the real system and we can speak of *population* and *samples*.

This point of view is considered very distorted and even confusing in Bayesian statistics. In Bayesian statistics τ is unknown, hence subject to probabilistic statements. On the contrary data, once observed, are not any longer random variables, but fixed data. The prior $P(\tau)$ and the likelihood $f(x \mid \tau)$ are both subjective distributions that express the experimenter's beliefs about τ, prior to observing the data. There is no such a thing as *absolute* probability: *probability does not exist (de Finetti).* In this conceptual scheme there is no such a thing a *population* or *samples* since there is no connection between frequency and probability.

1.3.3 The Bayesian Mathematics

Bayes theorem has been deduced using Kolmogorov axioms. However Kolmogorov mathematical model was constructed on the frequency theory of probability and it is unclear if the subjective definition fits in this frame. The problem was solved by Cox [58] in 1946. He was able to put in an axiomatic form the Bayesian probability and to show that from these basis he could deduce the same rules as deduced from Kologorov axioms.

The Bayesian probability is a measure of plausibility and is expressed by a real number (there is no need of complex numbers). The probability is a real number comprised between zero and one, associated to a proposition, and expresses how much we believe in it. Cox looked for the rules that these numbers should obey to keep the logical consistency and coherence with plausible reasoning. He started from two axioms:

- If we specify something on a proposition A to be true, then we have also specified something on how much A is false.

- If we have specified how much A is true and also how much B is true, given that A is true, the we must have also specified how much we believe that A and B are true.

Starting from these two axioms Cox was able to derive the two relations:

$$P(A \mid I) + P(\overline{A} \mid I) = 1 \qquad \textit{sum rule}$$

$$P(A \cap B \mid I) = P(A \mid B \cap I) \cdot P(B \mid I) \qquad \textit{product rule}$$

I is the background information, relevant in assessing the probability of A or B. For instance the probability that it will rain in the next hour depends on the temperature, on the humidity, on the latitude etc. Absolute probabilities do not exist.

Starting from these relation it is possible to derive the axioms of Kolmogorov theory, hence all the theorems that we have discussed are also valid in the Bayesian theory, although with a different meaning.

Least Informative Probabilities

While in the frequency theory probabilities reflect a status of nature (there exist absolute probabilities), in the Bayes theory probabilities are assigned, not measured, since they reflect our belief on the propositions. The assignment of probabilities that make use of the minimum amount of information is what is called *Least Informative Probabilities (LIP)*. This is clearly the case if nothing is known on the proposition.

The use of uniform probabilities (following Bayes postulate) though often used leads to inconsistencies, as we have discussed above. Hence a variety of methods are used that make use of reasonable arguments.

- If the probability concerns a bounded parameter, whose scale is not known (i.e. the length of a fish that could be a whale or a sardine), the LIP should be invariant in a change of scale like $\theta \to k \times \theta$,

  $$P(\theta \mid I)d\theta = P(k \cdot \theta \mid I)d(k \cdot \theta) = k \cdot P(k \cdot \theta \mid I)d\theta$$

  hence:

  $$P(\theta \mid I) \propto \frac{1}{\theta}$$

  This is the so called Jeffreys' prior.

- If, on the contrary, a parameter is unbounded and we are interested in the location of the parameter, then the LIP should be invariant under the transformation $\theta \to \theta + a$. In words

  $$P(\theta \mid I)d\theta = P(\theta + a \mid I)d(\theta + a) = P(\theta + a \mid I)d\theta$$

  hence:

  $$P(\theta \mid I) = \textit{constant}$$

  and we recover Bayes prescription.

Testable Information

It can happen that, though we do not know the distribution of a parameter X, still some information exists. Typical problem is that the only the average ($E(X)$) or the variance ($E(X - \mu)^2)$) is known. The problem is to find the probability density of X ($f(x)$) which is *maximally noncommittal* with respect to the missing information.

Jaines (1957) suggested that we should maximize the *Shannon Information Entropy*:

$$S = -\int_{\Omega} f(x) \cdot \log f(x)\ dx$$

under the constraints posed by the problem.

Example 1.17 *Let us assume that we know only the expectation value (μ) of a variable x, bounded from below, what can we say on its density? Let us call the density $f(x)$. According to Jaines we should maximize the quantity:*

$$S = -\int_{0}^{\infty} f(x) \log f(x) dx$$

under the constraints:

$$\begin{aligned} 1 &= \int_0^\infty f(x)dx \\ \mu &= \int_0^\infty x f(x)dx \end{aligned}$$

Or find the function $f(x)$ which maximizes the form:

$$T = \lambda_0 \big(1 - \int_0^\infty f(x)dx\big) + \lambda_1 \big(\mu - \int_0^\infty x f(x)dx\big) - \int_0^\infty f(x) \log f(x)dx$$

This is a problem of Variational Calculus which is solved by Euler equation:

$$\frac{\partial T}{\partial f} - \frac{d}{dx}\frac{\partial T}{\partial f'} = 0$$

where f' is the derivative of f with respect to x. In our case this leads to the equation:

$$\int_0^\infty \big(\lambda_0 x + \lambda_1 + \log f + 1\big) dx = 0$$

which has the solution, once the normalization and the mean value are kept into account:

$$f(x) = \frac{e^{-x/\mu}}{\mu}$$

i.e. the exponential distribution.

Exercise 1.19 *A die is known to give an average result of 4.5.*

$$\sum_i i \cdot P_i = 4.5$$

where p_i is the probability of each result. Clearly the assumption of uniform probability is not compatible with this information. The die is not fair. Find the least informative values of p_i.

1.3.4 Another Example of Bayesian Reasoning

We will come to the consequences of the theorem and to Bayes Postulate later on in these notes. For the time being let make an example. Assume that:

- A is "the pulse height in a shower counter" and
- B is "the particle is an electron"

We all know that an electron leaves large pulses in a shower counter and we can compute (by EGS MonteCarlo) the probability that an electron will leave a pulse larger than a given threshold. This is the direct probability $P(A \mid B)$. Now, by the inverse probability $P(B \mid A)$ one would quantify the probability that a particle is an electron if it leaves large ionization in the shower counter.

The last sentence is meaningless in classical statistics: a particle is an electron or not, there is no probability in it.

However, given a pulse height in the shower counter we could bet on the fact that the particle that has reached the counter is an electron. Further measurements on the particle could improve our confidence on identification of the particle as an electron. From this point of view, this improper probability also changes along with the measurement. Probability loses any absolute meaning and becomes a measure of our knowledge of the facts. These are the basic postulates of the *subjective* definition of probability, or Bayesian or modern school of statistics.

1.3.5 One more Remark on Bayes Theorem

It is important to stress the different interpretations that are given by the two schools of statistics to the same statement, which sometimes is the origin of some confusion. Bayes theorem has different meaning in Bayesian and Frequency theory of statistics. The theorem reads (the same notations as in the previous sections):

$$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_i P(A \mid B_i) \cdot P(B_i)}$$

In Bayesian statistics $P(B_i)$ is the prior density of our degree of belief on the possible occurrences of B_i (prior of B_i). $P(A \mid B_i)$ is the likelihood of A given B_i and $P(B_i \mid A)$ is the posterior probability of B_i once our knowledge is increased by A's occurrence.

In classical statistics if B_i is considered what causes A, then $P(B_i)$ is the relative frequency of occurrence of B_i. $P(A \mid B_i)$ is the relative frequency of A given B_i, while $P(A \mid B_i)$ is the relative frequency with which B_i is the cause of A.

Summary of this Chapter

	Rules of Set Theory
Distributive Laws	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
De Morgan	$\overline{A \cup B} = \overline{A} \cap \overline{B}$ $\overline{A \cap B} = \overline{A} \cup \overline{B}$
	Rules of Probability Theory
$\forall\, A \subseteq S$	$P(A) = 1 - P(\overline{A})$
$\forall\, B \subseteq A$	$P(B) \leq P(A)$ $P(\cup_i A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots$ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(\cup_i A_i) \leq \sum_i P(A_i)$ $P(\cap_i A_i) \geq 1 - \sum_i P(\overline{A_i})$ $P(A \cap B) \geq 1 - P(\overline{A}) - P(\overline{B})$ $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
Independence	$P(A \mid B) = P(A)$ $P(A \cap B) = P(A) \cdot P(B)$ $P(A \cup B) = 1 - P(\overline{A}) \cdot P(\overline{B})$
Bayes Theorem	$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum P(A \mid B_i) \cdot P(B_i)}$

Chapter 2

Probability Density and its Moments

A random variable [1] can be discrete or continuous. If the outcome of a measurement can have a finite or at least a numerable number of values we have a discrete variate. Once we give the probability of occurrence of each event $P(e_i)$, we have totally specified the problem. We have seen examples of discrete variates in the previous chapter: the outcome of a toss of dices or the number of atoms, in a radioactive sample, which decays in a fixed interval of time. In the first case the possible outcomes are finite and can be enumerated directly. In the second case the number of possible outcomes is very large, ideally infinite,

A word on terminology used in the mathematical literature; the term *distribution function* is used for a monotonic increasing function from 0 to 1, as the argument runs from the lower limit to the higher limit. In statistic literature this type of function is called *cumulative distribution function*. The name *density function* defines a function whose integral is one. We will made a distinction among: *Normal distribution*, meaning the type of probability model, *Normal density function* which is the probability density and *Normal distribution function* which is the cumulative function of the Normal density.

If the outcome of a measurement is a continuous number, we define a probability density, i.e. the probability dP that the outcome of the measurement of the variate X is found to be between x and $x + dx$.

$$dP = f(x)\ dx = dF$$

(In this case $f(x)$ is the density and $F(x)$ the distribution.) An example of continuous probability density is given by the time at which a atom, from a radioactive substance, decays. Once the origin of times is defined the probability of decay in an interval dt is:

$$dP = \frac{1}{\tau} e^{-t/\tau}\ dt \to \frac{dP}{dt} = \frac{1}{\tau} e^{-t/\tau}$$

In both cases the problem is completely specified by the function P: the probability (or the probability density) of each event.

2.1 Probability moments

Many of the important properties of probability density functions can be summarized by a few key quantities:

[1]The vocabulary which is used in statistics, is sometimes confusing. Here a few definitions.

The outcome of an experiment will be an *event* and a collection of events constitutes the *sample*. The totality of outcomes from an experiment is called *population*. (We draw a sample from a population.) We will consider *faithful* samples, such that from the sample we can determine the properties of the population. In *random sample* every possible sample has the same probability of being selected. In the same sense we think of a measurement as being *unbiased*. A *random event* is a member of the population, its numerical value is a *random variable* or *variate*.

- expected value (or *mean*): $E(x)$ which is defined as:

$$\mu = E(x) = \sum_i x_i \cdot P(x_i) \qquad\qquad E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

for discrete and continuous random variables respectively.
In the case of a function of a random variable, the expected value is defined as:

$$E\big(h(x)\big) = \sum_i h(x_i) \cdot P(x_i) \qquad\qquad E\big(h(x)\big) = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

for discrete and continuous random variables.
E is a linear operator, therefore if

$$G(x) = f(x) + h(x) \qquad \rightarrow \qquad E(G) = E(f) + E(h)$$

Similarly if $G(x) = k \cdot h(x)$, with k constant, then

$$E(G) = k \cdot E(h)$$

- *Variance*

$$\sigma^2 = Var(x) = E((x-\mu)^2) = E(x^2) - \mu^2$$

This variable (or its square root (σ), called standard deviation) measures the dispersion of the outcomes.

Exercise 2.1 *Prove that* $Var(k \cdot x) = k^2 \cdot Var(x)$

- Skewness

$$\gamma_1 = \frac{E(x-\mu)^3}{\sigma^3}$$

This variable measures the asymmetry of the *p.d.f.*, it is negative if $f(x)$ has longer tails to the left of the mean.

2.2 The Moment Generating Function

More generally the *moments* of a random variable are defined as follows:
The n^{th} moment about the origin is:

$$\mu'_n = E(x^n)$$

The n^{th} moment about the mean is:

$$\mu_n = E\big((x-\mu)^n\big)$$

The complete set of moments, under very general conditions, determines the probability density.

Theorem 2.1 *If two probability densities f_1 and f_2 have the same moments and if the difference $f_1 - f_2$ can be expanded in power of x, then $f_1 = f_2$.*

Proof 2.1 *Assume that we can write:*

$$f_1(x) - f_2(x) = \sum_i c_i x^i$$

Then:

$$\begin{aligned}
0 \leq \int_{-\infty}^{+\infty} (f_1(x) - f_2(x))^2 \, dx = \int_{-\infty}^{+\infty} (f_1(x) - f_2(x)) \cdot (\sum_{i=0,\infty} c_i x^i) \, dx \\
= \sum_{i=0,\infty} c_i \int_{-\infty}^{+\infty} (f_1(x) - f_2(x)) x^i \, dx \\
= \sum_{i=1,\infty} (c_i(\mu'_{i,1} - \mu'_{i,2})) = 0
\end{aligned}$$

hence $f_1(x) = f_2(x)$ ∎

The moment generating function of the *p.d.f.* $f(x)$ is:

$$\Phi_x(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} \cdot f(x) dx$$

(In the case of discrete *p.d.f.* the integral is replaced by a sum.) Note that: $\Phi_x(0) = 1$.

The importance of *m.g.f.* stems from the following consideration: if the function $e^{x\,t}$ can be expanded in series, the expected value of $e^{x\,t}$ is:

$$\begin{aligned}
\Phi_x(t) = E(e^{xt}) &= E(1 + xt + \frac{(xt)^2}{2!} + \frac{(xt)^3}{3!} + ...) = \\
&= 1 + \mu'_1 \cdot t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \cdots \\
&= \sum_{n=0}^{\infty} \frac{E(x^n)}{n!} t^n
\end{aligned}$$

The moments are the coefficients of the terms $\frac{t^n}{n!}$. An alternative way is:

$$\mu'_n = \frac{\partial^n}{\partial t^n} E(e^{xt}) \bigg|_{t=0}$$

Exercise 2.2 *Prove the previous relation.*

Similar relations can be defined also for the moment generating function about the mean.

2.3 An important Theorem

Theorem 2.2 *Suppose that X and Y are independent random variables whose m.g.f. are $\phi_X(t)$ and $\phi_Y(t)$ respectively. The random variable:*

$$Z = a \cdot X + b \cdot Y + c$$

(with a, b and c real numbers) has a m.g.f.:

$$\phi_Z = e^{ct} \cdot \phi_X(at) \cdot \phi_Y(bt)$$

Proof 2.2 *In fact:*

$$\begin{aligned}
\phi_Z(t) &= E(e^{(aX+bY+c)t}) = \\
&= \int dx \int dy\ e^{(ax+by+c)t} \cdot f_X(x) \cdot f_Y(y) \qquad \textit{(Independence of X and Y)} \\
&= e^{ct} \cdot \left(\int f_X(x) e^{axt}\ dx \right) \cdot \left(\int f_Y(y) e^{byt}\ dy \right) = \\
&= e^{ct}\ \phi_X(at)\ \phi_Y(bt)
\end{aligned}$$

■

This property can easily be generalized to the case of the sum of any number of independent random variables. Assume X_i are independent random variables with probability density function $f_i(x_i)$, then the variable $Z = \sum_{i=1,n} X_i$ has a *m.g.f.* which is the product of the *m.g.f.* of each variable:

$$\begin{aligned}
\phi_{\sum X_i}(t) &= E\left(e^{t \sum X_i}\right) = \\
&= \int \int \cdots \int e^{t \sum x_i} \cdot \prod f_i(x_i) dx_i = \\
&= \prod \phi_{X_i}(t)
\end{aligned}$$

If all the variables have the same probability density function:

$$\phi_{\sum X_i}(t) = (\phi_X(t))^n$$

Summary of this Chapter

Assume: $x \sim f(x)$	
Define Moments:	$\begin{cases} \mu_n' &= E\big(x^n\big) = \int_{-\infty}^{+\infty} x^n f(x)dx \\ \mu_n &= E\big((x-\mu)^n\big) = \int_{-\infty}^{+\infty} (x-\mu)^n f(x)dx \end{cases}$
The Moment Generating Function	$\Phi_x(t) = E\big(e^{tx}\big) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$
	$\begin{cases} \Phi(t)_x &= \sum_{n=0,\infty} \frac{E\left(x^n\right)}{n!} t^n \\ \mu_n' &= \frac{\partial^n \Phi_x(t)}{\partial t^n}\Big\|_{t=0} \end{cases}$
If x, y are independent variables then $z = a\ x + b\ y + c$ has a M.G.F.:	$\Phi_z(t) = e^{ct} \cdot \Phi_x(at) \cdot \Phi_y(bt)$
In particular if	$z = \sum_{i=1,n} x_i$ (x_i all independent)
then:	$\Phi_z(t) = \big(\Phi_x(t)\big)^n$

Chapter 3

Discrete Distributions

An event E which can have different outcomes is called random event. When E is a numeric quantity it is called random variable or *variate*. If the number of outcomes is finite, we associate to each e_i a probability P_i:

$$Pr(E = e_i) = P_i$$

(The number of outcomes can also be a countable infinite.) Examples of discrete distributions are:

- Binomial.
- Poisson.
- Compound Poisson.
- Bose Einstein etc.

We will only go briefly through those which are more common in the practice of High Energy Physics data analysis.

3.1 Bernoulli Trials

The simplest and certainly most important elementary experiment in probability is the process of randomly toss a coin. The outcome can be *Head* or *Tail* with a given probability p and $1-p$ respectively.

3.1.1 Definition

The abstraction of this basic process is the Bernoulli Trials: a sequence of random trials that satisfy the following assumptions:

- Each trial has two possible outcomes (Success or Failure);
- Trials are independent, the outcome of a trial has no influence on the outcome of another trial;
- Each trial has a probability of success (p) constant through the experiment.

The sample space of a single trial is formed by the two states $S = \{S, F\}$. The sample space of n Bernoulli trial has 2^n states, a succession of n symbols, S and F; each state represents one possible outcome of the compound experiment. Since the trials are independent, probabilities multiply:

$$P(\{SFFSSSS....F\}) = p(1-p)(1-p)pppp...(1-p)$$

Familiar examples are the toss of a fair coin (Success is Head) with probability $p = 1/2$ or the roll of a die (Success is Ace) with probability $p = 1/6$.

3.1.2 Stochastic assumption

The Bernoulli scheme is a theoretical model and experience only will prove if it is suitable to describe specific observations. Though it looks evident, there are cases where intuition is against it. The man of the street who plays late Lotto numbers with the argument that late numbers are more likely to occur, in fact denies the independence of successive trials. After a sequence of 15 Heads in the tossing of a coin, Tail is not more likely to occur on the 16th trial. If such was the case we have to abandon the stochastic independence of successive trials and accept that Nature has some sort of memory. The philosopher K. Marbe[1] was a supporter of this line of thought and found many supporters. Bernoulli or Marbe schemes cannot be proved or disproved on the basis of logic. Marbe scheme has to be abandoned because it is not supported by empirical evidence.

A Bernoulli trial process can be described by a sequence of Indicators: a random variable which takes the values 1 and 0 to denote Success and failure.

$$I_1, I_2 \cdots$$
$$P(I_1 = 1) = p \quad , \quad P(I_1 = 0) = 1 - p$$

Exercise 3.1 *Prove that*

$$P(\{I_l = i_l \;, l = 1, n\}) = p^k (1-p)^{n-k} \qquad k = \sum_{l=1}^{n} i_l$$

Exercise 3.2 *Show that, if $U_1, U_2 \cdots$ are independent random variables, each with uniform distribution in $[0,1]$ and $p \in [0,1)$, the event $U_i = u_i < p$ is a Bernoulli trial process.*

3.2 Binomial Distribution

Binomial, or Bernoulli distribution, applies when the random variable has only two outcomes, as when we toss a coin. Let E be the result of a toss. There are only two possible outcomes: head and tail.

Assume p is the probability of a success (say head). For N trials, the probability that the first r tosses are heads and the following $N - r$ tosses are tails is:

$$p^r \cdot (1-p)^{N-r}$$

This is only one out of the possible ways to get r successes out of N attempts. We have to multiply the previous expression by the number of times we can get the same result in any order. The total probability is:

$$P(r, N) = \binom{N}{r} \cdot p^r \cdot (1-p)^{N-r} \qquad (r = 0, 1, \cdots, N)$$

This is the binomial or Bernoulli distribution. See Fig. 3.1.

---

[1]K. Marbe, Die Gleichfomigkeit in der Welt, Munich, 1916

Exercise 3.3 *prove that:*

$$\sum_{r=0}^{N} P(r) = 1$$

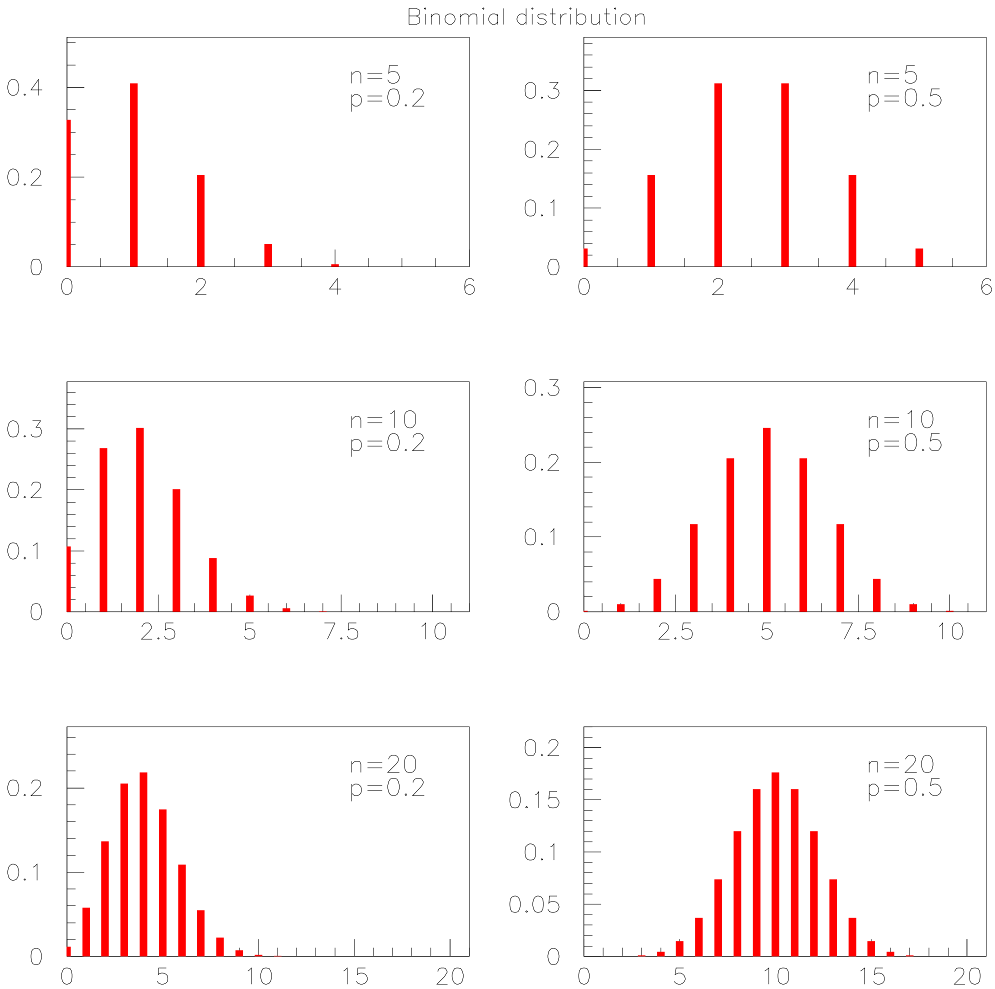


Figure 3.1: *Binomial distribution as a function of r for various p and N.*

The average of the Binomial distribution is:

$$\begin{aligned}
\mu &= E(r) = \sum_{r=1}^{N} r \binom{N}{r} p^r (1-p)^{N-r} = \\
&= \sum_{r=1}^{N} r \frac{N!}{r! \cdot (N-r)!} p^r (1-p)^{N-r} = \\
&= Np \sum_{r=1}^{N} \frac{(N-1)!}{(r-1)! \cdot (N-r)!} p^{r-1} (1-p)^{N-r} = \\
&= Np \sum_{s=0}^{K} \binom{K}{s} p^s (1-p)^{K-s} = \\
&= Np
\end{aligned}$$

Note that the first sum was started from 1, since the term with $r = 0$ would not contribute. In the last passage we have changed the variables ($s = r - 1$ and $K = N - 1$).

Exercise 3.4 *Prove that the variance is* $Var(r) = Np(1-p)$.

Exercise 3.5 *Prove that the skewness is* $\gamma_1 = Np(1-p)(1-2p)$.

Exercise 3.6 *Consider the function of* r*:* $h(r) = \frac{r}{N}$. *Prove that*

$$E\left(h(r)\right) = p \qquad\qquad Var\left(h(x)\right) = \frac{p(1-p)}{N}$$

In statistics we have to estimate p *from the measurements. We have obtained* n *successes in* N *trials. To infer the value of* p *we use the function* $h = n/N$*, a statistics, i.e. a function of the random variable* n *which we use to infer* p*. The variable* h *is also a random variable, since it is function of a random variable. Important properties that a statistics should have are:*

- *it is an unbiased estimate of* p *since its expected value is* p*.*
- *It is consistent since its variance tends to zero when* $N \to \infty$*.*
- *The accuracy increases as the standard deviation decreases i.e. with square root of* N*.*

The Moment generating function can also be easily computed:

$$\begin{aligned}
\Phi_r(t) &= \sum_{r=0}^{N} e^{rt} B(r, N, p) = \\
&= \sum_{r=0}^{N} \binom{N}{r} \cdot (p \cdot e^t)^r \cdot (1-p)^{N-r} = \\
&= (p \cdot e^t + 1 - p)^N
\end{aligned}$$

Example 3.1 *Let us perform a random experiment making* n *Bernoulli trials* $\{I_1, I_2 \cdots I_n\}$ *and consider the variable:*

$$X_n = \sum_{j=1}^{n} I_j$$

that is the number of Successes in n trials. If K is the subset of $S = \{1, 2 \cdots n\}$ made of only k elements, then:

$$P\left((I_j = 1\ , j \in K) \cap (I_l = 0\ l \in S - K)\right) = p^k (1-p)^{n-k}$$

as we have proved previously. The number of subsets made of k elements from a set of n elements is

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

and we obtain again the Binomial distribution in a more formal way.

3.3 Geometrical Distribution

Assume that we are doing a Bernoulli experiment (a Bernoulli trial), measuring a binomial variate, but we are not interested in the number of successes in N trials, but in the *number of trials before the first success.*

Example 3.2 *How many times should we try an exam, before passing it? (if the outcome of an exam is a random, Bernoulli variable! However, if this is the case, you should change the teacher!)*

In this case the random variable is the number of trial n. If the probability of success in the Bernoulli trial is p then the probability to have success at the n^{th} trial is given by the joint probability of having $n-1$ failures followed by one success. Since the individual tests are independent, the joint probability is the product:

$$P(n \mid p) = (1-p)^{n-1} p$$

The random variable n can vary from 1 to ∞. It is easy to show that the distribution is normalized to 1, since:

$$\sum_{n=1,\infty} p(1-p)^{n-1} = p \sum_{s=0,\infty} (1-p)^s = \frac{p}{1-(1-p)} = 1$$

The moment generating function is:

$$\begin{aligned}
\Phi_n(t) =& E(e^{tn}) = \sum_{n=1,\infty} e^{tn} (1-p)^{n-1} p \\
=& p \sum_{s=0,\infty} e^{t(s+1)} (1-p)^s = pe^t \sum_{s=0,\infty} \left(e^t(1-p)\right)^s = \\
=& \frac{p}{e^{-t} - (1-p)}
\end{aligned}$$

From the moment generating function it is possible to compute:

$$E(n) = \frac{1}{p} \qquad Var(n) = \frac{1-p}{p^2}$$

Exercise 3.7 *Prove the previous relations.*

Example 3.3 *St. Petersburg paradox. In a game of cards the probability of success is p. The rules of the game are simple: we can bet any amount of money at even stake: if we win we get what we have bet, if we loose we pay that amount. We decide to use the following strategy, known as* the martingale strategy*:*

- *we start betting c;*
- *if we loose, we double the bet;*
- *we stop at the first success.*

Let us call W the winning and L the losses. Since each game is independent of any other game, the possible outcomes are only two and the probability of success is constant, we are in the Bernoulli trials scheme.

The amount of losses in $n-1$ games is:

$$\begin{aligned} L &= c + 2c + \cdots 2^{n-1}c \\ &= c\frac{2^n - 1}{2-1} = c(2^n - 1) \end{aligned}$$

The losses increase as a power of the number of games played, but at the game n we win and stop. The win at the n game is $W = c \cdot 2^n$ therefore the net gain, at the end is $W - L = c$. It seems that we have found a perfect strategy: the net gain is not a random variable, we always gain the money with which we started and the gain is even independent of p. However let us compute the average amount of money that we have to bet to be sure to win the last game. The random variable n follows the geometrical distribution:

$$\begin{aligned} E(L) &= E(c(2^n - 1)) \\ &= c\left[p\sum_{n=1}^{\infty} 2^n (1-p)^{n-1} - 1\right] \\ &= c\left[2p\sum_{s=0}^{\infty} \big(2(1-p)\big)^s - 1\right] \end{aligned}$$

The series is convergent only if $2(1-p) < 1$, i.e. if $p > 1/2$;

$$\begin{aligned} E(L) &= \infty \qquad p \leq \tfrac{1}{2} \\ &= \tfrac{c}{2p-1} \qquad p > \tfrac{1}{2} \end{aligned}$$

This strategy is flawed even if the game is fair.

3.4 Pascal Distribution or Negative Binomial

This problem is a generalization of the previous one. We are doing a Bernoulli trial experiment and we want to know how many times we have to repeat the experiment before counting r successes. If p is the probability of the Bernoulli trial, then the probability we are looking for is the joint probability of having $r-1$ successes in $n-1$ trials and the probability of having a success at the n^{th} trial. Again each trial is independent and:

$$P(n \mid r, p) = \left(\begin{array}{c} n-1 \\ r-1 \end{array}\right) p^{r-1}(1-p)^{n-1-(r-1)}p = \left(\begin{array}{c} n-1 \\ r-1 \end{array}\right) p^r (1-p)^{n-r}$$

n is the random variables while r and p are parameters of the distribution. The possible values of n are integers from r to ∞. The moment generating function can be computed as in the case

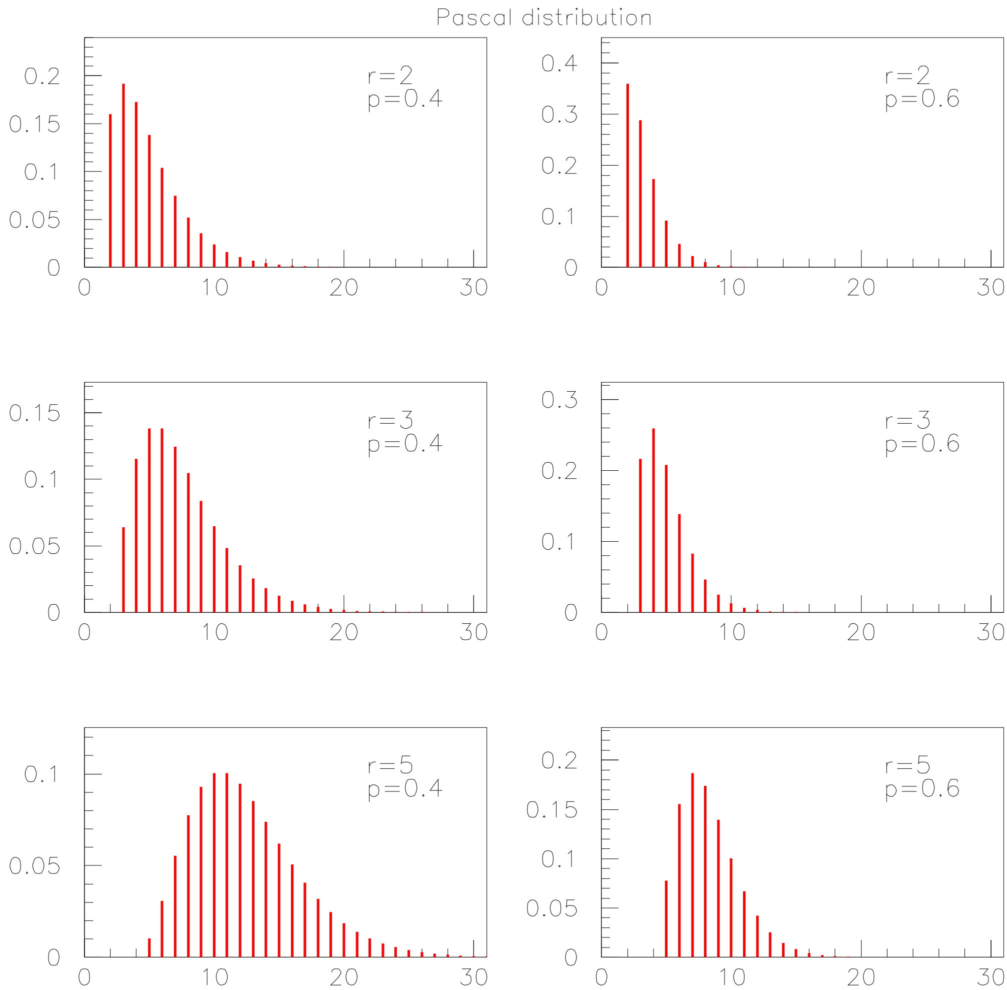


Figure 3.2: *The Pascal distribution.*

of the geometric distribution:

$$\Phi_n(t) = E(e^{tn}) = \sum_{m=k,\infty} e^{mt} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} p^k (1-p)^{m-k}$$

$$= e^{tk} p^k \sum_{s=0,\infty} \begin{pmatrix} s+k-1 \\ k-1 \end{pmatrix} \left((e^t(1-p) \right)^s$$

and, using the identity:

$$\left(\frac{1}{1-x} \right)^k = \sum_{s=0,\infty} \begin{pmatrix} s+k-1 \\ k-1 \end{pmatrix} x^s$$

we obtain

$$\begin{aligned}\Phi_n(t) =& e^{tk}p^k\left(\frac{1}{1-e^t(1-p)}\right)^k\\ =& \left(\frac{p}{e^{-t}-(1-p)}\right)^k\end{aligned}$$

From the moment generating function it is possible to compute:

$$E(n) = \frac{k}{p} \qquad Var(n) = \frac{k(1-p)}{p^2}$$

Exercise 3.8 *Prove the previous relations.*

Example 3.4 *The Banach matches boxes. (This example is inspired by the reference to Banach smoking habits, made by Steinhaus in an address honoring Banach.) A certain mathematician always carries one match box in his left pocket and one in his right pocket. When he wants a match he randomly selects a pocket. Suppose that initially each box contains N matches. At the moment he discovers that the selected box is empty, the other box may contain $r = 0, 1, 2, \cdots$. What is the probability u_r that the other box contains r matches? This problem can be solved with Pascal distribution, even if it is more difficult to recognize that this distribution applies. Let's call "success" the event: left pocket is chosen. The left pocket will be found empty when the right pocket has r matches only when $N-r$ "failures" precede the $(N+1)^{st}$ "success". The total number of trial is $2N-r+1$ and the number of successes is $N+1$. The Pascal probability gives us the probability to find in the right pocket r matches:*

$$P(2N-r+1 \mid N+1, p) = \binom{2N-r}{N} p^{N+1}(1-p)^{N-r}$$

We have to multiply by two the previous expression since the same argument holds for the other pocket as well. Since $p = \frac{1}{2}$, we have:

$$u_r = \binom{2N-r}{N} 2^{r-2N}$$

3.5 Poisson Distribution

Assume:

- The occurrence of an event at a time t is independent of the history before t. Its probability is constant, independent of t.
- The probability of one event in dt is

  $$P(one\ event\ in\ (t, t+dt)) = \mu \cdot\ dt$$

  constant and independent of t.
- The probability of two or more events in dt is zero.

The probability of zero events in $t + dt$ is:

$$P_0(t+dt) = P_0(t)\cdot(1-\mu\cdot dt)$$

from which:

$$\frac{P_0(t+dt) - P_0(t)}{dt} = -\mu \cdot P_0(t)$$

Letting dt go to zero and integrating:

$$P_0(t) = \mu e^{-\mu t}$$

This is the *exponential* distribution for t, the waiting time between events. Or the probability that in a Poisson process, in the interval t no event has occurred.

The probability of one event only in the time $t+dt$ is given by the sum of the two possibilities: one event in $(0,t)$ and none in $(t, t+dt)$ or no events in $(0,t)$ and one event in $(t, t+dt)$:

$$P_1(t+dt) = P_1(t) \cdot (1 - \mu \cdot dt) + P_0(t) \cdot \mu dt$$

We get the equation:

$$\frac{dP_1}{dt} = -\mu P_1(t) + \mu \cdot e^{-\mu t}$$

whose solution is:

$$P_1(t) = \mu t \cdot e^{-\mu t}$$

The probability of r events in the time $t + dt$ is obtained in a similar way:

$$P_r(t+dt) = P_r(t) \cdot P_0(dt) + P_{r-1}(t) \cdot P_1(dt)$$

yielding:

$$\frac{P_r(t+dt) - P_r(t)}{dt} = -\mu P_r(t) + \mu P_{r-1}(t)$$

or:

$$\frac{dP_r}{dt} = -\mu \cdot P_r(t) + \mu \cdot P_{r-1}(t)$$

whose solution is:

$$P_r(t) = \frac{(\mu t)^r}{r!} \cdot e^{-\mu \cdot t}$$

Exercise 3.9 *verify that the above function satisfy the differential equation.*

Another derivation will be given, since it provides more insight to this important distribution. A process (for instance the decay of a radioactive source) in the time interval t has an average rate μt. Let us divide the interval t in a large number (N) of time intervals $\Delta t = t/N$. In each of these intervals the probability of observing one event is p.

The probability of r successes in N trials is (Binomial distribution):

$$P(r, N) = \left(\begin{array}{c} N \\ r \end{array} \right) \cdot p^r \cdot (1-p)^{N-r}$$

The probability of observing one event, in each interval Δt, is $p = \mu t/N$, hence:

$$\begin{aligned}
P(r,N) &= \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} = \\
&= \frac{N!}{r!(N-r)!} \cdot (\frac{\mu t}{N})^r \cdot (1 - \frac{\mu t}{N})^{N-r} = \\
&= \frac{N(N-1)\cdots(N-r+1)}{N^r} \frac{(\mu t)^r}{r!} \frac{(1-\frac{\mu t}{N})^N}{(1-\frac{\mu t}{N})^r} = \\
&= 1(1-\frac{1}{N})\cdots(1-\frac{r-1}{N}) \frac{(\mu t)^r}{r!} \frac{(1-\frac{\mu t}{N})^N}{(1-\frac{\mu t}{N})^r} \\
&\rightarrow \frac{(\mu t)^r}{r!} e^{-\mu t} \qquad When\ N \rightarrow \infty
\end{aligned}$$

Fig. 3.3 shows the Poisson distribution as a function of r for some values of μ

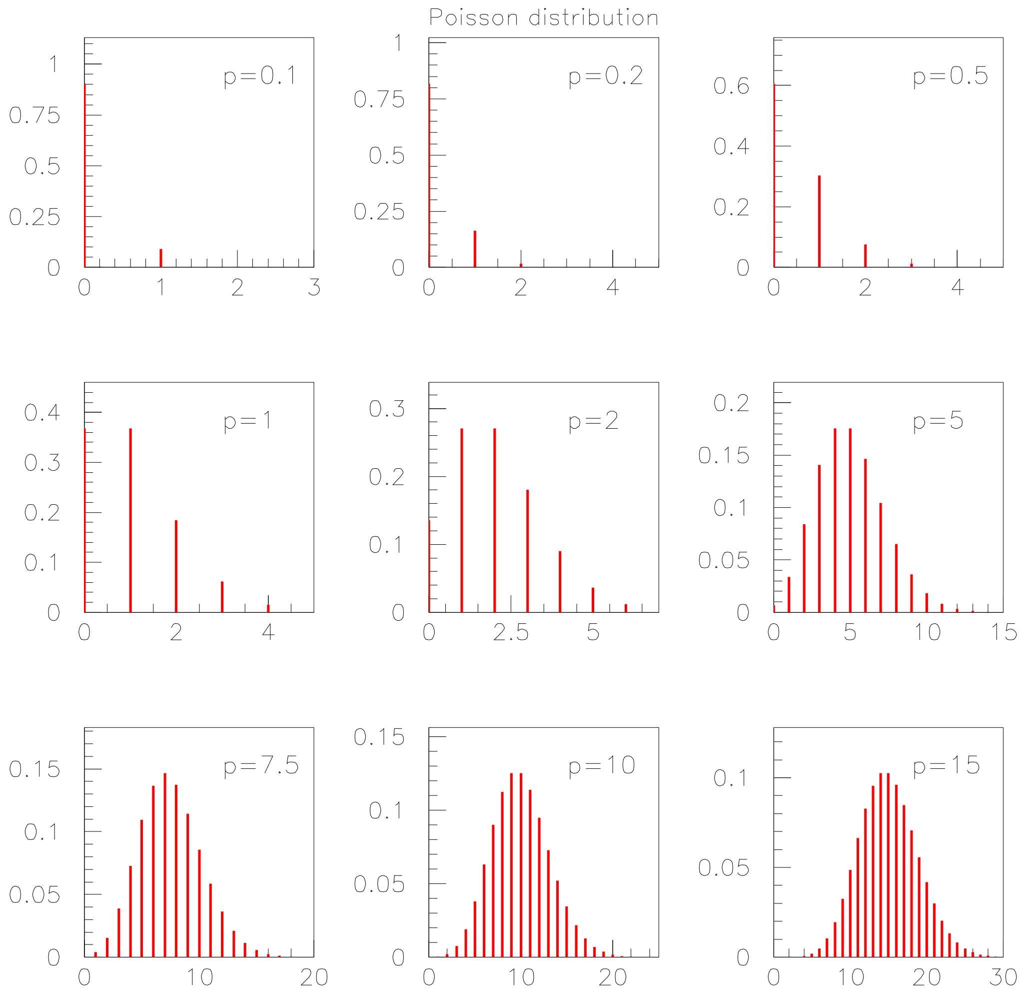


Figure 3.3: *Poisson distribution as a function of r for various* $p = \mu$.

Exercise 3.10 *prove that in the Poisson distribution:*

$$E(x) = \mu \qquad\qquad Var(x) = \mu$$

The Moment Generating Function is, for the Poisson distribution:

$$\begin{aligned} \Phi_{Poisson}(t) &= \sum_{r=0}^{\infty} e^{rt} \frac{\mu^r}{r!} \cdot e^{-\mu} = \\ &= e^{\mu \cdot (e^t - 1)} \end{aligned}$$

Exercise 3.11 *find the m.g.f. about the mean and derive the mean and variance.*

Example 3.5 *The quark search was based on the measurement of specific ionization* (μ) *of a charged track in a sensitive medium. Since the specific ionization is proportional to the square of* *the charge of the particle, a quark with charge $q = 1/3$ would ionize $1/9$ of a pion or an electron with the same kinetic energy. Assume that the experiment is performed with a photographic emulsion in which a pion would have a ionization density $\mu = 9$ grains/mm. A quark would ionize one grain per mm. Since the process of ionization in emulsion can be approximated by a Poisson process, the probability that a pion would ionize 1 grain per mm is:*

$$P(1 \mid \mu = 9) = \frac{1}{1!} 9^1 e^{-9} = 10^{-3}$$

and we would observe one (fake!) candidate quark in thousand pion tracks. This probability would considerably be reduced if only the emulsion would be twice more sensitive with a specific ionization of 18 grains per mm for pions. In fact in this case we would have:

$$P(2 = \frac{18}{q^2} \mid \mu = 18) = 2.5\ 10^{-6}$$

Example 3.6 *The probability to puncture a tire is one per year. If the event can be modeled by a Poisson process then the probability to puncture precisely once is $P(1 \mid \mu = 1) = 37\%$, while the probability to puncture twice is $P(2 \mid \mu = 1) = 18\%$ etc.*

Exercise 3.12 *Under which assumption can the Poisson distribution be used in the previous example?*

Exercise 3.13 *A cosmic ray experiment consists in recording the number of muons crossing the apparatus in an interval of time $\Delta t = 1$ minute. The experiment lasted $T = 51$ hours and recorded in total $N = 19728$ events.*

- *Which is the average number of cosmic ray events in one minute?*
- *In how many intervals of one minute do we expect to measure zero events?*
- *In how many intervals do we expect to measure 1 event and 6 events?*

Exercise 3.14 *We have set up an experiment at an electron accelerator to detects photons. The counter which detect photons is a Geiger counter with the following characteristics:*

- *Dead time $T_D = 300\mu sec$. (Dead time is the time, after having recorded a count, in which the counter remains insensitive.)*
- *Efficiency $\epsilon = 2\%$.*
- *Transverse dimensions $\Delta x\ \Delta\ y = 2 \times 10\ cm^2$.*

We know that the photon beam consists of bursts of γ which last $50\ \mu$ sec and a repetition rate of 180 bursts per second. The Geiger counter has measured 3600 counts/minute. What is the intensity of the photon beam?

Exercise 3.15 *In an experiment at a $e^+\ e^-$ collider, the frequency of signal events is A, while the background of cosmic rays has a rate B.*

- *Verify that the probability $P(n)$ that between two successive signal events we have n background events is:*
$$P(n) = \frac{R^n}{(1+R)^{n+1}} \qquad R = \frac{B}{A}$$
- *Plot $P(n)$ for $R = 0.1,\ 0.5,\ 1.,\ 5$*

- *Verify that* $\sum_{n=0,\infty} P(n) = 1$

Exercise 3.16 *It is common practice in high rate experiments to "prescale" the events. To prescale by a factor n means that, after the current event we will skip $n-1$ events and accept only the nth one. Assume for simplicity that we prescale of a factor 2, i.e. we take only the second event. If the event rate is constant and events are random, compute the probability $P(t_1 < t < t_2)$ that the time between two accepted events is comprised between t_1 and t_2. Compute the probability that the time interval between two accepted events is less than t.*

3.5.1 The sum of Poisson distributions

Assume a counter records the radioactive decays of a source with an average rate R_s per unit time in surroundings where the background rate is R_b. The counter measures *the sum* of source and background rates, each being Poisson distributed.

The probability that the counter records r counts is made by the sum of probabilities that the counter records 0 signal events and r backgrounds, plus the probability that one signal and $r-1$ backgrounds are recorded etc. :

$$\begin{aligned}
P(r \mid R_s \cdot t, R_b \cdot t) &= \sum_{b=0}^{r} P(r-b \mid R_s t) \cdot P(b \mid R_b t) \\
&= \sum_{b=0}^{r} \frac{(R_s t)^{(r-b)}}{(r-b)!} \cdot e^{-(R_s t)} \frac{(R_b t)^b}{b!} \cdot e^{-(R_b t)} \\
&= \frac{e^{-(R_s+R_b)t}}{r!} \sum_{b=0}^{r} \frac{r!}{(r-b)!b!} (R_s t)^{r-b} (R_b t)^b \\
&= \frac{e^{-(R_s+R_b)t}}{r!} ((R_s+R_b)t)^r = P(r \mid (R_s+R_b)t)
\end{aligned}$$

Poisson distributions add.

Example 3.7 *The same result can be found with the technique of the moment generating function. In fact if*

$$\begin{aligned}
r_1 &\sim Poisson(r \mid \mu_1) \\
r_2 &\sim Poisson(r \mid \mu_2)
\end{aligned}$$

are two independent Poisson variables and we are interested in the distribution of $s = r_1 + r_2$. We have:

$$\begin{aligned}
\Phi_s(t) =& \Phi_{r_1}\Phi_{r_2} = e^{\mu_1(e^t-1)} \cdot e^{\mu_2(e^t-1)} \\
=& e^{(\mu_1+\mu_2)(e^t-1)}
\end{aligned}$$

Hence:

$$s \sim Poisson(s \mid \mu_1 + \mu_2)$$

With this method it is easy to generalize the result. If:

$$s = \sum_i r_1 \qquad with \qquad r_i \sim Poisson(r \mid \mu_i)$$

then

$$s \sim Poisson(s \mid \sum \mu_i)$$

3.5.2 Poisson Distribution Measured by Inefficient Counters

A random rate is measured by a counter whose efficiency is p. Assume a random process is Poisson distributed with mean R. To register r counts in the time t at least r particles must have crossed the counter. The probability is obtained by adding all the probabilities to have r counts.

If n particles cross the counter the probability to register r counts has a Binomial distribution, therefore:

$$\begin{aligned}
P(r) &= \sum_{n=r}^{\infty} P(n \mid R) \cdot B(r,n,p) = \\
&= \sum_{n=r}^{\infty} (\frac{n!}{r!(n-r)!} \cdot p^r \cdot (1-p)^{n-r}) \cdot (\frac{1}{n!} \cdot R^n \cdot e^{-R}) = \\
&= \frac{1}{r!}(pR)^r e^{-R} \sum_{n=r}^{\infty} \frac{1}{(n-r)!}((1-p)R)^{n-r} = \\
&= \frac{1}{r!}(pR)^r e^{-R} \sum_{n=0}^{\infty} \frac{1}{n!}((1-p)R)^{n} = \\
&= \frac{1}{r!}(pR)^r e^{-R} e^{(1-p)R} = \frac{1}{r!}(pR)^r e^{-pR}
\end{aligned}$$

which is nothing but a Poisson distribution with average pR.

3.5.3 Compound Poisson

This is an example of a distribution which is built with elementary distributions and is used to describe processes like the ionization of heavy particles or the queuing in a pipeline where various process take place.

Let r_1, r_2,...r_n be a set of n independent Poisson variables with common mean value μ and let n, the number of variables, be also a Poisson variable with mean ν. We want to find the distribution of the sum of the n variables:

$$P(r) = \sum_{n=0}^{\infty} P(r = \sum_{i=1}^{n} r_i) \cdot P(n, \nu)$$

I.e. the Poisson probability to have n variables times the joint probability for the n variables to sum up to r. From the theorem of addition of Poisson variables the sum is Poisson distributed with mean $\nu \cdot \mu$. Hence:

$$P(r) = \sum_{n=0}^{\infty} (\frac{1}{r!}(n\mu)^r e^{-n\mu}) \cdot (\frac{1}{n!}\nu^n e^{-\nu})$$

This is the compound Poisson distribution.

The moment generating function can be put in a closed form. In fact

$$\begin{aligned}
\Phi_r(t) = E(e^{rt}) &= \sum_{r=0,\infty}\sum_{n=0,\infty}\frac{1}{r!}(n\mu)^r e^{-n\mu}\frac{1}{n!}\nu^n e^{-\nu} e^{rt} \\
&= \sum_{n=0,\infty}\frac{1}{n!}\nu^n e^{-\nu}\sum_{r=0,\infty}\frac{1}{r!}\left(n\mu e^t\right)^r \\
&= \sum_{n=0,\infty}\frac{1}{n!}\nu^n e^{-\nu} e^{-n\mu} e^{n\mu e^t} = \sum_{n=0,\infty}\frac{1}{n!}\nu^n e^{-\nu} e^{n\mu(e^t-1)} \\
&= e^{-\nu}\sum_{n=0,\infty}\frac{1}{n!}\left(\nu e^{\mu(e^t-1)}\right)^n = e^{-\nu} e^{\nu e^{\mu(e^t-1)}} \\
&= e^{\nu(e^{\mu(e^t-1)}-1)}
\end{aligned}$$

With the moment generating function it is possible to compute the first central moments. In fact:

$$\begin{aligned}
\left.\frac{\partial\Phi}{\partial t}\right|_{t=0} &= \left. e^{\nu e^{\mu(e^t-1)}-1}\nu e^{\mu(e^t-1)}\mu e^t\right|_{t=0} = \nu\mu \\
\left.\frac{\partial^2\Phi}{\partial t^2}\right|_{t=0} &= \nu\mu\left.\left(e^{\nu e^{\mu(e^t-1)}-1}\nu e^{\mu(e^t-1)}\mu e^t + e^{\nu e^{\mu(e^t-1)}-1}e^{\mu(e^t-1)}\mu e^t + e^{\nu e^{\mu(e^t-1)}-1}e^{\mu(e^t-1)}e^t\right)\right|_{t=0} \\
&= \nu\mu(\nu\mu + \mu + 1)
\end{aligned}$$

hence

$$\begin{aligned}
E(r) &= \mu\nu \\
Var(r) &= \nu\mu(\mu+1)
\end{aligned}$$

Example 3.8 *In a quark search experiment [27] a positive result was claimed. The experiment consisted in exposing to cosmic ray a delayed expansion Wilson cloud chamber. The experiment recorded about 55000 exposures and in one exposure an abnormally low ionization particle was detected. The picture is shown in Fig 3.4. We have marked the low ionizing track (R) with an*

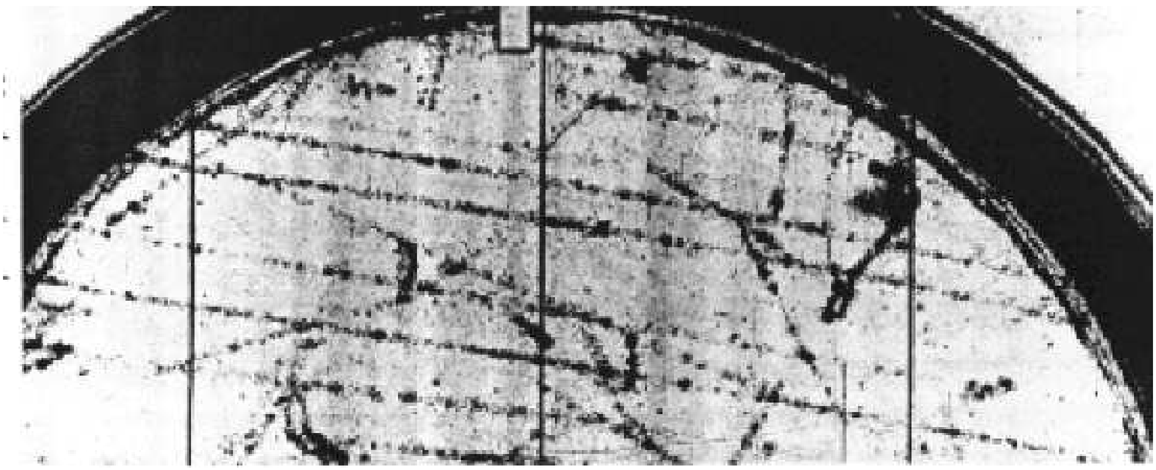


Figure 3.4: *Picture of a cloud chamber exposure where a claim for a successful search for quark was made.*

arrow. The event contains 5 tracks in addition to the "quark". The tracks are parallel and have *about the same transverse dimensions, proving that they are not late or early tracks. Counting droplets, in sections not containing δ rays, it was found that a normal track has 229 droplets per unit length, while track R only 110.*

McCusker argued as follows: the probability that a normal track has produced so low ionization (or less) is:

$$P(r \leq 110) = \sum_{i=0}^{110} \frac{229^i}{i!} \cdot e^{-229} \approx 10^{-18}$$

Therefore the low ionization cannot be due to extreme tails of standard processes. It must be produced by a different mechanism, like the passage of a low ionizing particle.

The argument is weak since many experimental effects are neglected but, for what concerns us, the statistics is wrong. As pointed out by Adair [28] the formation of droplets in a cloud chamber proceeds in two steps:

- *Primary knock on electrons are produced. Their number follows the Poisson statistics with average μ.*
- *The primary electron ionizes producing an average of $\nu = 4$ secondary electrons.*

The average number of secondary is about 4, if we require that the primary electrons are not visible in the cloud chamber, as in the case of the published analysis. Therefore the statistics of droplet formation is NOT a simple Poisson, but a compound Poisson with $\mu = 229$ and $\nu = 57$. If we consider only the fluctuation of the primary electrons we get a probability:

$$P(r \leq \frac{110}{4}) = \sum_{i=0} 28 \frac{57^i}{i!} \cdot e^{-57} \approx 5\ 10^{-5}$$

which is considerably larger than in McCusker analysis.

Exercise 3.17 *Compute the correct probability, by using the Compound Poisson statistics.*

3.6 The Bose-Einstein Distribution

This is the only quantum statistics that we will consider. In deriving the Binomial distribution we have multiplied the probability of r successes and $N - r$ failures by the number of ways in which this outcome could be obtained.

In a quantum system the $\begin{pmatrix} N \\ r \end{pmatrix}$ ways are indistinguishable and is not allowed to consider them individually, they are all the same. We obtain the Bose Einstein distribution:

$$P_{BE}(r, N) = K_N p^r (1 - p)^{N-r}$$

where K_N is a normalization constant. Calling:

$$\eta = \frac{p}{1 - p}$$

and letting $N \to \infty$ ($\eta < 1$) we have:

$$P(r)_{BE} = K\eta^r \qquad (r \geq 0)$$

The normalization is easily found:

$$1 = K \sum_{r=0,\infty} \eta^r = K \frac{1}{1 - \eta}$$

so that:

$$P(r)_{BE} = (1-\eta)\eta^r$$

The Moment generating function is:

$$M(t) = \sum_{r=0,\infty} (1-\eta)\eta^r e^{tr} = \frac{1-\eta}{1-\eta e^t}$$

A power series expansion leads to:

$$\begin{aligned} \mu &= \frac{\eta}{1-\eta} \\ Var(r) &= \mu^2 + \mu \end{aligned}$$

Exercise 3.18 *Prove the previous formulas.*

This shows that the fluctuations in the BE statistics are larger that in the Poisson case with the same mean. And this is a consequence of the quantum entanglement of the particles.

By using the expression for μ we can rewrite the BE distribution as:

$$P_{BE}(r) = \frac{1}{(1+\mu)(1+\frac{1}{\mu})^r}$$

Distribution	Form	E(r)	Var(r)	Momentum Generating Function
Binomial	$\begin{pmatrix} N \\ r \end{pmatrix} p^r(1-p)^{N-r}$	Np	$Np(1-p)$	$(pe^t+1-p)^N$
Geometric	$p^r(1-p)^{r-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{e^{-t}-1+p}$
Neg. Binomial	$\begin{pmatrix} r-1 \\ n-1 \end{pmatrix} p^N(1-p)^{r-N}$	$\frac{N}{p}$	$\frac{(1-p)m}{p^2}$	$(\frac{p}{e^{-t}-1+p})^k$
Poisson	$\frac{\mu^r e^{-\mu}}{r!}$	μ	μ	$e^{\mu(e^t-1)}$
Compound Poisson	$\sum_n(\frac{(n\mu)^r e^{-n\mu}}{r!})(\frac{\nu^r e^{-\nu}}{n!})$	$\mu\nu$	$\mu\nu(1+\mu)$	$e^{\nu(e^{\mu(e^t-1)}-1)}$
Bose-Einstein	$(1-\eta)\eta^r$	$\frac{\eta}{1-\eta}=\mu$	$\frac{\eta}{(1-\eta)^2}$	$\frac{1-\eta}{1-\eta e^t}$

Table 3.1: *Summary of most important discrete distributions.*

Chapter 4

Continuous Distributions

A random variable x can also assume continuous values. In this case we define the probability that x is comprised in a small interval dx:

$$f(x) \cdot dx = dF$$

$f(x)$ is called the probability density function (*p.d.f.*) and $F(x)$ its cumulative distribution or simply distribution.

4.1 Uniform Distribution

In this simple case the probability density is:

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

The distribution is very simple but also very important.

- A rod is measured to the nearest mm. The residual error x has an uniform distribution:

$$f(x) = 1. \qquad\qquad -0.5 < x < 0.5.$$

- The energy distribution of photons from the decay of mono-energetic π^0 has uniform distribution.
- Pseudo random number generators in (0,1) exist and are very often used in MC methods.

The moment generating function is:

$$\Phi_x(t) = E(e^{xt}) = \frac{1}{b-a}\int_a^b e^{tx} = \frac{1}{b-a}\frac{1}{t}\left(e^{bt} - e^{at}\right)$$

The average and variance are readily computed:

$$E(x) = \int_a^b \frac{x}{b-a}dx = \frac{b-a}{2}$$
$$E(x^2) = \int_a^b \frac{x^2}{b-a}dx = \frac{b^2+ab+a^2}{3}$$

hence

$$Var(x) = E(x^2) - (E(x))^2 = \frac{(b-a)^2}{12}$$

Exercise 4.1 *Compute, the average and variance of the uniform distribution using the m.g.f..*

4.2 The exponential distribution

Let us consider an ensemble of unstable atoms. At the time $t = 0$ we have n_0 atoms; as the time passes some of the atoms decay and the number of atoms still existing at the time t will change: $n(t)$. In fact the number of atoms that decay in the interval dt at the time t is equal to $n(t) \cdot dt/\tau$. $1/\tau$ is the probability that an atom decays in the unit time. We are assuming that:

- The decay probability (τ) is independent of the time.
- Each atom decays independently from the others.

With these assumptions, the number of atoms still alive at the time $t + dt$ is:

$$n(t + dt) = n(t) - n(t)\frac{dt}{\tau}$$

The equation can be easily solved yielding:

$$n(t) = n_0 \; e^{-t/\tau}$$

Hence the probability that an atom is still alive at the time t is:

$$P(T > t) = \frac{n(t)}{n_0} = e^{-t/\tau}$$

and the *p.d.f.*, i.e. the probability that one atom decays in dt at t is:

$$f(t) = \frac{d}{dt}P(T > t) = \frac{1}{\tau}e^{-t/\tau}$$

This example has shown a process determined by an exponential density. The basic assumption have been listed above. Not all decays are described by an exponential function. If a decay can cause the decay of other atoms, as it happens in chain processes (a fission reactor), then the instantaneous rate is not described by an exponential decay.

The Moment Generating Function is:

$$\Phi_t(s) = E\left(e^{ts}\right) = \int_0^\infty \frac{1}{\tau}e^{st-t/\tau}dt = \frac{1}{\tau}\int_0^\infty e^{(s\tau-1)/\tau} = \frac{1}{1-s\tau}$$

We can now compute the mean and the variance of the exponential density:

$$\begin{aligned}
\bar{t} &= \left.\frac{\partial \Phi_t(s)}{\partial s}\right|_{s=0} = \left.\frac{\tau}{(1-s\tau)^2}\right|_{s=0} = \tau \\
\overline{t^2} &= \left.\frac{\partial^2 \Phi_t(s)}{\partial s^2}\right|_{s=0} = \left.\frac{2\tau^2}{(1-s\tau)^4}\right|_{s=0} = 2\tau^2 \\
Var(t) &= \overline{t^2} - \bar{t}^2 = \tau^2
\end{aligned}$$

4.3 Normal Distribution

It is probably the most frequently used distribution. It is bell shaped and extends from minus to plus infinity. The probability density is:

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

The distribution is completely specified by μ and σ (Fig 4.1). The Standard Normal distribution

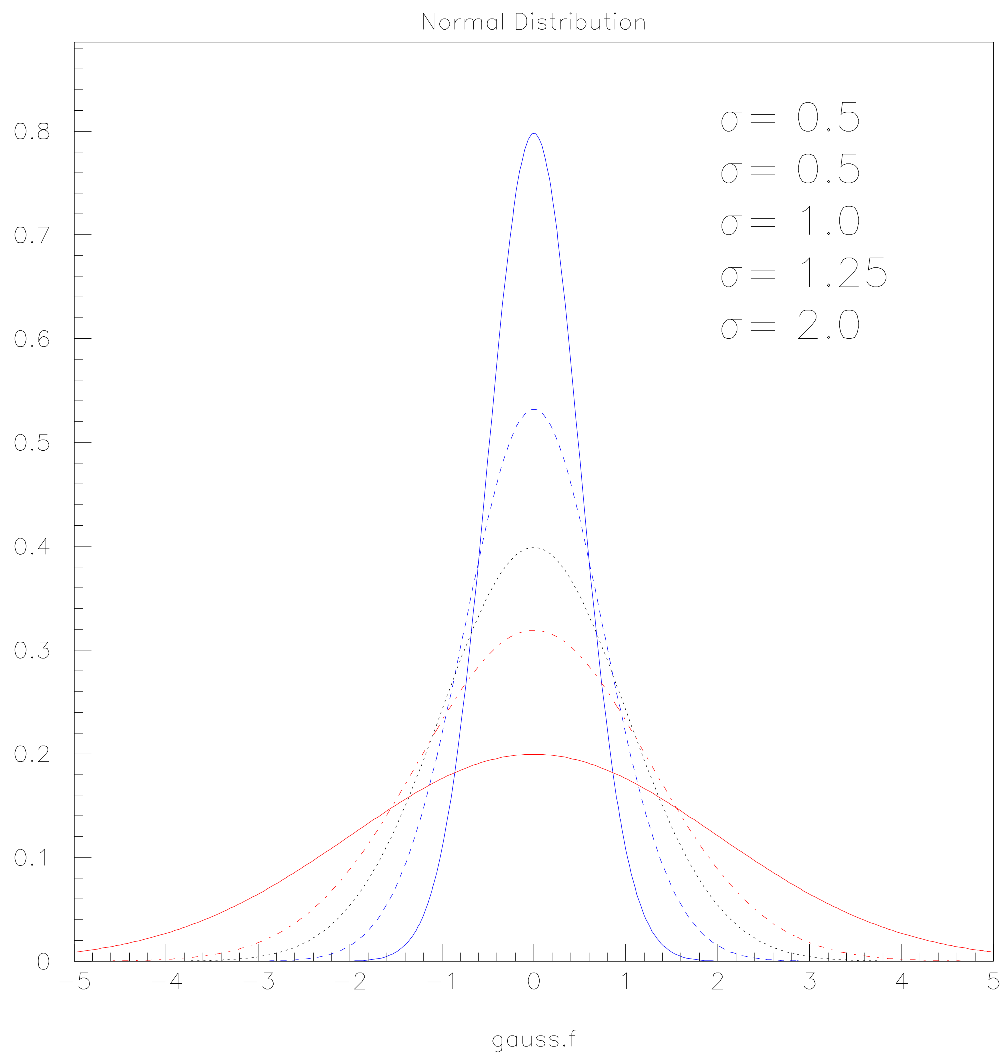


Figure 4.1: *The Normal p.d.f. with $\mu = 0$ as a function of x for some values of σ.*

is a special case where $\mu = 0$ and $\sigma = 1$. The probability density is bell shaped and Figure 4.2 shows some scaling parameters.

For a Standard Normal *p.d.f.*, the cumulative distribution is (See also Fig 4.3 and Tab 4.1 and 4.2) is:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

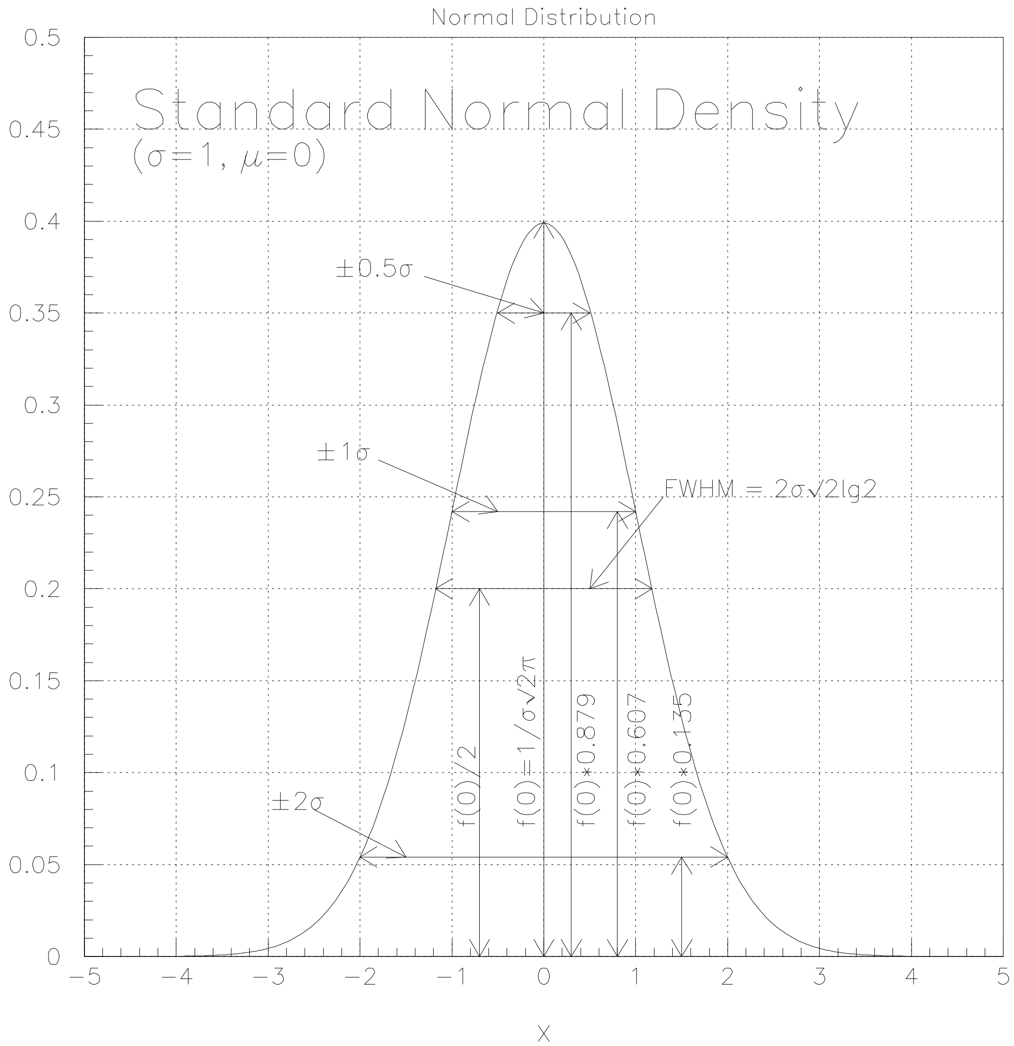


Figure 4.2: *The shape of the Normal p.d.f..*

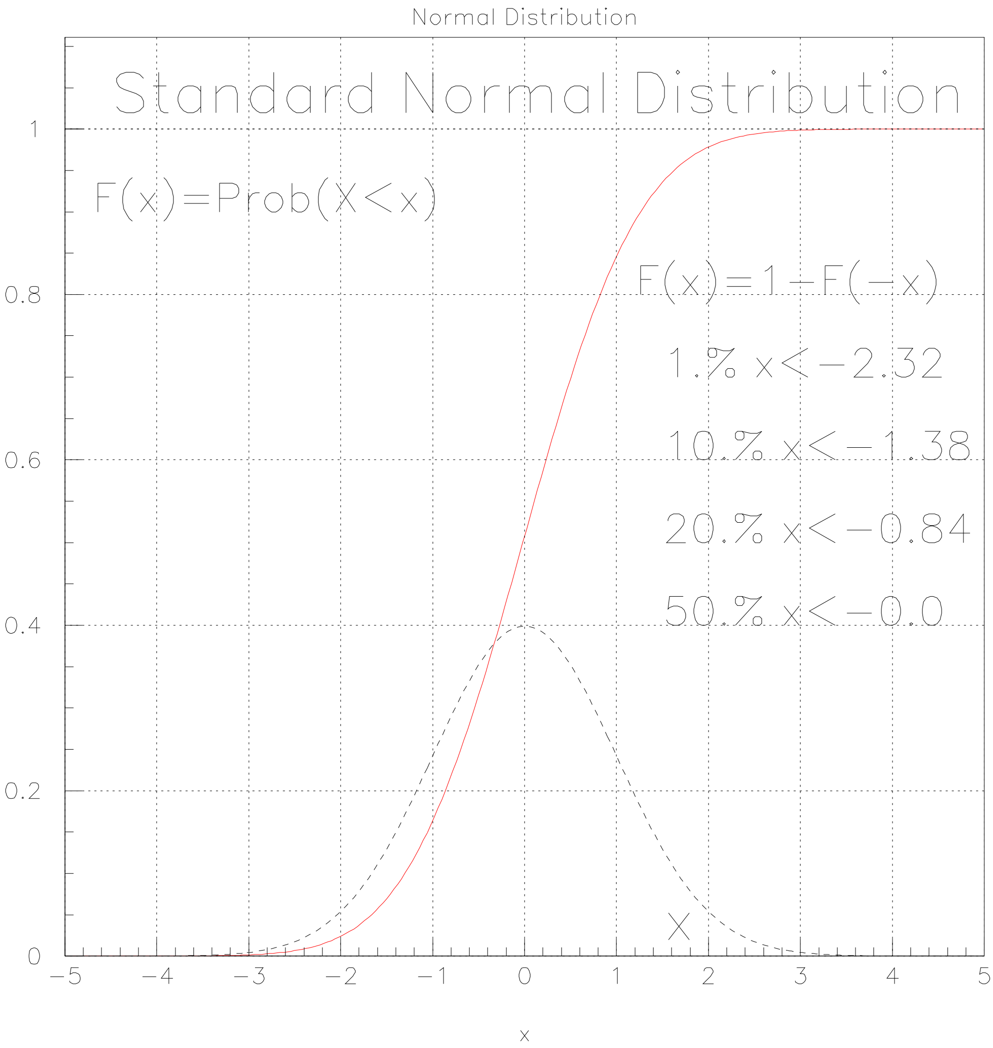


Figure 4.3: *The cumulative function for a standard normal p.d.f..*

x	$\Delta x = 0$	0.020	0.040	0.060	0.080	0.100	0.120	0.140	0.160	0.180
0.000	0.5000	0.5080	0.5160	0.5239	0.5319	0.5398	0.5478	0.5557	0.5636	0.5714
0.200	0.5793	0.5871	0.5948	0.6026	0.6103	0.6179	0.6255	0.6331	0.6406	0.6480
0.400	0.6554	0.6628	0.6700	0.6772	0.6844	0.6915	0.6985	0.7054	0.7123	0.7190
0.600	0.7257	0.7324	0.7389	0.7454	0.7517	0.7580	0.7642	0.7703	0.7764	0.7823
0.800	0.7881	0.7939	0.7995	0.8051	0.8106	0.8159	0.8212	0.8264	0.8315	0.8365
1.000	0.8413	0.8461	0.8508	0.8554	0.8599	0.8643	0.8686	0.8729	0.8770	0.8810
1.200	0.8849	0.8888	0.8925	0.8962	0.8997	0.9032	0.9066	0.9099	0.9131	0.9162
1.400	0.9192	0.9222	0.9251	0.9279	0.9306	0.9332	0.9357	0.9382	0.9406	0.9429
1.600	0.9452	0.9474	0.9495	0.9515	0.9535	0.9554	0.9573	0.9591	0.9608	0.9625
1.800	0.9641	0.9656	0.9671	0.9686	0.9699	0.9713	0.9726	0.9738	0.9750	0.9761
2.000	0.9772	0.9783	0.9793	0.9803	0.9812	0.9821	0.9830	0.9838	0.9846	0.9854

Table 4.1: *Values of $F(x)$ the cumulative function of a standard normal p.d.f.*

p	Δp= .00000	.01000	.02000	.03000	.04000	.05000	.06000	.07000	.08000	.09000
.00000	$-\infty$	-2.32635	-2.05375	-1.88079	-1.75069	-1.64485	-1.55477	-1.47579	-1.40507	-1.34075
.10000	-1.28155	-1.22653	-1.17499	-1.12639	-1.08032	-1.03643	-.99446	-.95417	-.91537	-.87790
.20000	-.84162	-.80642	-.77219	-.73885	-.70630	-.67449	-.64334	-.61281	-.58284	-.55338
.30000	-.52440	-.49585	-.46770	-.43991	-.41246	-.38532	-.35846	-.33185	-.30548	-.27932
.40000	-.25335	-.22755	-.20189	-.17637	-.15097	-.12566	-.10043	-.07527	-.05015	-.02507
.50000	.00000	.02507	.05015	.07527	.10043	.12566	.15097	.17637	.20189	.22755
.60000	.25335	.27932	.30548	.33185	.35846	.38532	.41246	.43991	.46770	.49585
.70000	.52440	.55338	.58284	.61281	.64334	.67449	.70630	.73885	.77219	.80642
.80000	.84162	.87790	.91536	.95417	.99446	1.03643	1.08032	1.12639	1.17499	1.22653
.90000	1.28155	1.34075	1.40507	1.47579	1.55477	1.64485	1.75069	1.88079	2.05375	2.32635

Table 4.2: *Values of $x(F)$ the inverse cumulative function of a standard normal p.d.f.*

The average of a Normal distribution is:

$$E(x) = \int_{-\infty}^{\infty} y \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \mu$$

as can be easily proved by changing the variable $w = (y - \mu)/\sigma$. In the same way we can prove:

$$Var(x) = \int_{-\infty}^{\infty} (y-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \sigma^2$$

Exercise 4.2 *Show that in a Normal distribution the* $HWHM = \sigma \cdot \sqrt{2\log 2} \approx 2.36 \cdot \sigma$. *(HWHM means Half Width at Half Maximum and is an useful statistics, easier to read out of a plot than the standard deviation.)*

The Moment Generating Function is:

$$\begin{aligned} \Phi_x(t) &= E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= e^{\mu \cdot t + \frac{(\sigma \cdot t)^2}{2}} \end{aligned}$$

Exercise 4.3 *Derive the mean and variance from the m.g.f..*

Exercise 4.4 *Prove that* $\mu_{2n+1} = 0$ *and* $m_{2n} = \frac{(2n)! \cdot \sigma^{2n}}{2^n \cdot n!}$

Exercise 4.5 *If x is a normally distributed variable with mean μ and variance σ^2, show that the m.g.f. of the function $y = ax + b$ is:*

$$\Phi_{ax+b}(t) = e^{(a\cdot\mu+b)\cdot t + \frac{(a\cdot t\cdot\sigma)^2}{2}}$$

4.4 The Gamma Distribution

The Gamma distribution depends on two parameters: α and β. Its density is:

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha) \cdot \beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-\frac{x}{\beta}} \qquad (x > 0)$$

where:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} \cdot e^{-y} dy$$

with the property: $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$ and $\Gamma(n) = (n-1)!$ if n is integer.

The gamma function generates a variety of shapes, depending on the value of α ($\alpha > 0$). Figure 4.4 shows a plot of this function for some values of α. The parameter β is only a scaling factor. The χ^2 function with ν degrees of freedom is a particular case of the gamma function with $\beta = 2$ and $\alpha = \frac{\nu}{2}$. When α is integer, the Gamma function is also called Erlangen distribution.

The *m.g.f.* is:

$$\begin{aligned} \Phi_x(t) &= E(e^{xt}) = \frac{1}{\Gamma(\alpha) \cdot \beta^\alpha} \cdot \int_0^{\infty} y^{\alpha-1} \cdot e^{-y/\beta} \cdot e^{yt} dy \\ &= \frac{1}{\Gamma(\alpha) \cdot \beta^\alpha} \cdot \int_0^{\infty} y^{\alpha-1} \cdot e^{-y(\frac{1-\beta t}{\beta})} dy \\ &= \frac{1}{(1-\beta t)^\alpha} \cdot \underbrace{\frac{1}{\Gamma(\alpha) \cdot (\frac{\beta}{1-\beta t})^\alpha} \cdot \int_0^{\infty} y^{\alpha-1} \cdot e^{-y(\frac{1-\beta t}{\beta})} dy}_{=1} \\ &= \frac{1}{(1-\beta t)^\alpha} \end{aligned}$$

Since the integral is the normalization of the $\Gamma(x, \alpha, \frac{\beta}{1-\beta t})$ distribution, which is 1.

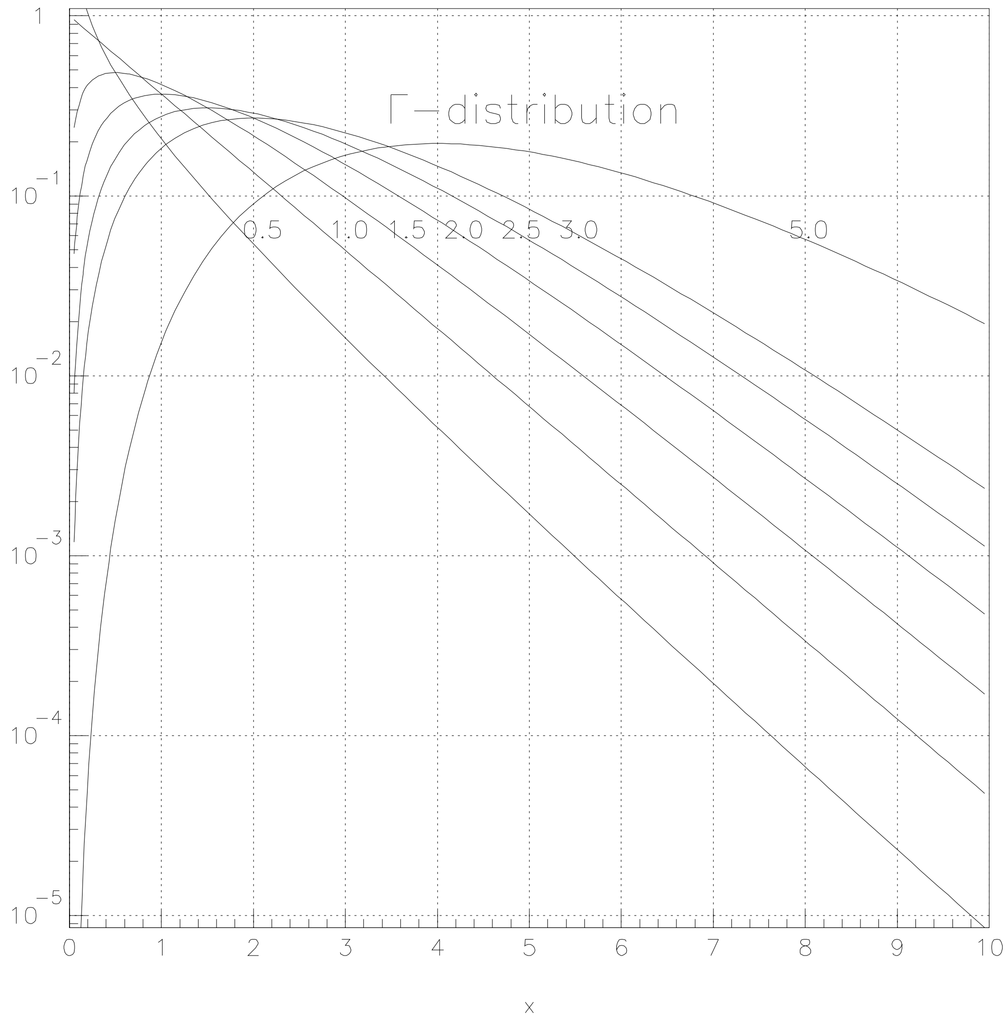


Figure 4.4: *Shapes of the Gamma function for α= 0.5, 1., 1.5, 2, 2.5, 3 and 5.*

Exercise 4.6 *Verify, using the moment generating function, that the mean and variance of the gamma distribution are:*

$$E(x) = \alpha \cdot \beta \qquad Var(x) = \alpha \cdot \beta^2$$

4.5 Cauchy Distribution

This is a rather unusual distribution whose most remarkable property is that it has infinite variance.

Suppose a gun, placed at 1 m from a wall, rotates at constant speed and once in a revolution at a random time a shot is fired. The distribution of shots on the walls is:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$$

In fact:

$$f(\theta) = d\theta/\pi \qquad (-\pi/2 < \theta < \pi/2)$$

and since

$$\theta = tan^{-1}(x) \qquad d\theta = dx/(1+x^2)$$

Therefore

$$f(x)dx = d\theta/\pi = \frac{1}{\pi}\frac{1}{1+x^2}$$

Exercise 4.7 *Prove that this distribution is properly normalized, and that its average and the variance are infinite.*

4.6 The χ^2 Distribution

Mathematically is only a special case of the Gamma distribution, but, because of its practical uses, it deserves a special section of its own. In fact later in these notes we will prove that the sum of the squares of k variables distributed normally, with zero mean and unit variance: $u = \sum_{i=1,k} x_i^2$ has a *p.d.f.*:

$$\chi^2_k(u) = \frac{1}{2^{k/2}\Gamma(k/2)} u^{k/2-1} e^{-u/2}$$

This is just a Gamma distribution with $\alpha = k/2$ and $\beta = 2$.

Fig. 4.5 shows the shape of this distribution for some values of k and as a function of u. The MGF is:

$$\Phi_u(t) = (1-2\,t)^{-k/2}$$

(see the MGF of the Gamma distribution). The MGF provides easily the mean and Variance:

$$E(u) = k \qquad Var(u) = 2k$$

It can also be shown that the distribution tends to a Normal distribution as $k \to \infty$. More precisely the variables:

$$z_1 = \frac{u-k}{\sqrt{2k}} \qquad z_2 = \sqrt{2u} - \sqrt{2k-1}$$

tend to be distributed as $N(0,1)$ when $k \to \infty$.

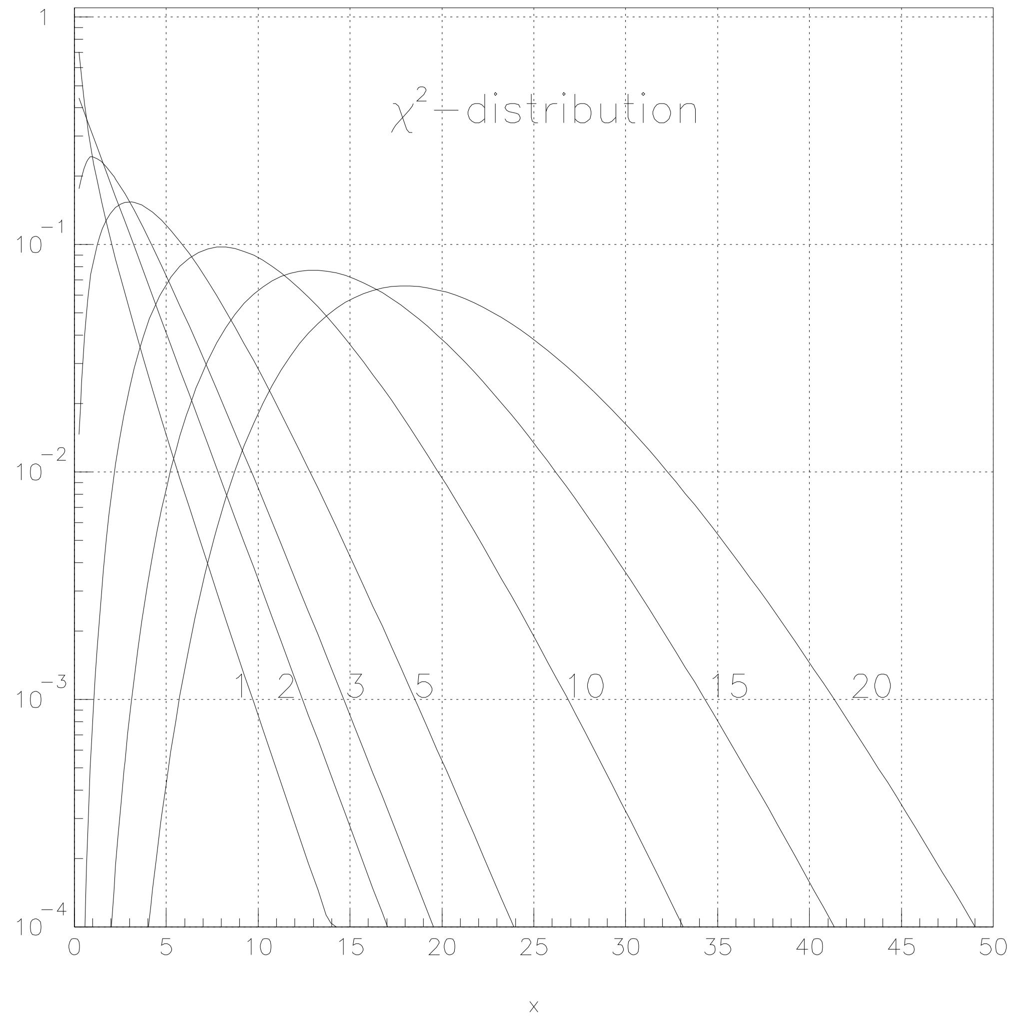


Figure 4.5: *Shapes of the Chisquare function for k= 1, 2, 3, 5, 10, 15 and 20.*

Distribution	*Form*	$E(r)$	$Var(r)$	*Momentum Generating Function*
Uniform	$f(x \mid a,b) = \frac{1}{b-a} \quad a \leq x \leq b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{1}{b-a}\frac{e^{bt}-e^{at}}{t}$
Exponential	$f(t \mid \tau) = \frac{e^{-t/\tau}}{\tau} \quad t \geq 0$	τ	τ^2	$\frac{1}{1-t\tau}$
Normal	$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ	σ^2	$exp(\mu t + \frac{(\sigma t)^2}{2})$
Gamma	$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$	$\frac{1}{(1-\beta t)^\alpha}$
Cauchy	$f(x \mid \mu) = \frac{1}{\pi(1+(x-\mu)^2)}$			

Table 4.3: *Summary of most important discrete distributions.*

Chapter 5

Distribution of a Function of a Random Variable

We want now to discuss how to compute the *p.d.f.* of a function of a random variable. We will start with the simple case of a discrete variable and then we will discuss the case of continuous *p.d.f.*.

5.1 Discrete Variable

We will assume first that there is a one to one correspondence between the values of a variable x and the function $h(x)$. In other words the function $h(x)$ is one valued. From the definition of probability it follows that the probability for x to have a particular value x_0 is the same as the probability for $h(x)$ to have the value $h(x_0)$, i.e.:

$$P\left(h(x) = h_0\right) = P\left(h(x) = h(x_0)\right) = P(x = x_0)$$

Example 5.1 $h(x) = cos(x)$ *(see Tab.5.1)*

x	0	$\pi/4$	$\pi/2$
$P(x)$	1/6	2/6	3/6
$h(x)$	1	0.71	0
$P(h)$	1/6	2/6	3/6

Table 5.1: *How to compute the probability of a function of a discrete random variable.*

In the case that there are several x_i which we have the same value of h, we have to sum the probabilities:

$$P(h(x) = h_0) = \sum P(x = x_i)$$

Example 5.2 $h(x) = cos(x)$, *see Tab. 5.2.*

5.2 The Case of one continuous Variable

By definition $f(x)dx = P(x \leq X \leq x + dx)$, i.e. $f(x)dx$ is the probability to find the value of X in the interval $(x, x + dx)$. In the case of the variable $h = h(x)$ the probability density is defined as:

$$g(h)dh = P(h \leq H \leq h + dh)$$

x	$-\pi/2$	$-\pi/4$	0	$\pi/4$	$\pi/2$
$P(x)$	1/35	3/35	6/35	10/35	15/35
$h(x)$	0	0.71	1.	0.71	0
$P(h)$	16/35	13/35	6/35		

Table 5.2: *The case of discrete random variable for a multi valued function*

To compute $g(h)$, let us assume first that $h(x)$ is a one-to-one function of x. We will find a small range of values of h which are produced by a small range of values of x with probability $f(x)dx$.

Let

$$x = x(h)$$

be the inverse of the function h. By substitution we find

$$f(x)\ dx = f\left(x(h)\right)\ dh$$

We can now compute the differential dx as a function of h. We have to take care of its sign, since the *p.d.f.* are NON negative by definition.

$$dx = \left|\frac{\partial x(h)}{\partial h}\right|\ dh$$

Combining the two equations we get:

$$f(x)\ dx = f(x(h)) \cdot \left|\frac{\partial x(h)}{\partial h}\right|\ dh$$

so that:

$$g(h) = f(x(h)) \cdot \left|\frac{\partial x(h)}{\partial h}\right|$$

Example 5.3 $h(x) = cos(x)$. *Assume that the p.d.f. of x is*

$$f(x) = a + bx \qquad (0 \le x \le \pi/2)$$

with $a = 1/\pi$ and $b = 4/\pi^2$.

$$x = ArcCos(h) \qquad\qquad \frac{dx(h)}{dh} = -\frac{1}{\sqrt{1-h^2}}$$

We then get ($0 \le h \le 1$):

$$g(h)\ dh = f(x(h)) \cdot \frac{\partial x}{\partial h} = \frac{a + b \cdot ArcCos(h)}{\sqrt{1-h^2}}\ dh$$

In the case there are several values of x which give the same value $h(x)$, we have to sum the probabilities, as we did in the discrete case (see Fig.5.1):

$$g(h)\ dh = \sum_{h(x_i)=h} f(x_i(h)) \cdot \left|\frac{\partial x_i}{\partial h}\right|$$

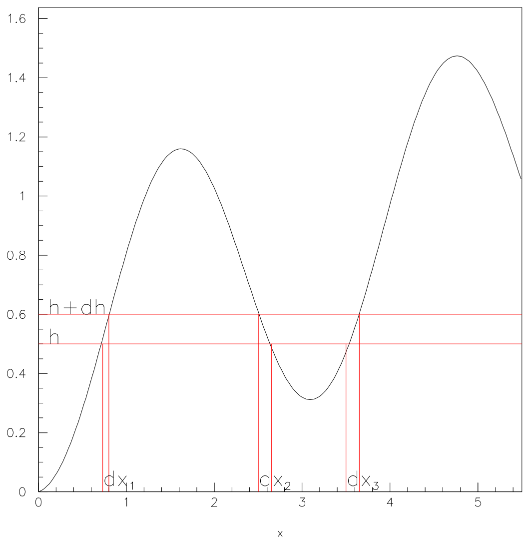


Figure 5.1: *Multi-valued function. To the same interval (h,h+dh) corresponds the intervals (x_1,x_1+dx_1), (x_2,$x_2 + dx_2$)....*

Example 5.4 *Assume that the variable X has a Standard Normal p.d.f:*

$$X \sim f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

let us compute the p.d.f. of the function $h = x^2$. The function has two inverses: $x_1 = \sqrt{h}$ and $x_2 = -\sqrt{h}$. The absolute values of the derivatives are:

$$\left| \frac{\partial x_{1,2}}{\partial h} \right| = \frac{1}{2\sqrt{h}}$$

hence, using the previous rules, we get:

$$\begin{aligned} g(h)dh &= \left(f\big(x_1(h)\big) \left| \frac{\partial x_1}{\partial h} \right| + f\big(x_2(h)\big) \left| \frac{\partial x_2}{\partial h} \right| \right) dh \\ &= \frac{1}{\sqrt{2\pi}} \left(e^{-h/2} \frac{1}{2\sqrt{h}} + e^{-h/2} \frac{1}{2\sqrt{h}} \right) dh \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{h}} e^{-\frac{h}{2}} dh \end{aligned}$$

The function can be cast in the form:

$$g(h) = \frac{h^{\alpha-1} e^{-h/\beta}}{\Gamma(\alpha)\beta^\alpha} \qquad (\alpha = 1/2, \quad \beta = 2)$$

The density is of the family of the Γ.

5.3 The Case of several Continuous Variables

The previous relations can be extended to the case of several variables. Suppose we have:

$$\{x_1, x_2, \cdots, x_n\} \sim f(x_1, x_2, \cdots, x_n)$$

that we want to transform to:

$$\{h_1, h_2, \cdots, h_n\} \qquad\qquad h_i = h_i(x_1, x_2, \cdots, x_n)$$

We have to find the inverse transformations:

$$x_i = x_i(h_1, h_2, \cdots, h_n)$$

The standard analysis shows that, in case the transformation is one-to-one:

$$g(h_1 \cdots h_n) = \sum_{\vec{h}(\vec{x_i})=h} \left(f(x_1(\vec{h}) \cdots x_n(\vec{h})) \cdot \left| \frac{\partial(x_1 \cdots x_n)}{\partial(h_1 \cdots h_n)} \right| \right)$$

where the last term is the absolute value of the Jacobian:

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial h_1} & \frac{\partial x_1}{\partial h_2} & \cdots & \frac{\partial x_1}{\partial h_n} \\ \frac{\partial x_2}{\partial h_1} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_n}{\partial h_1} & \cdots & \cdots & \frac{\partial x_n}{\partial h_n} \end{bmatrix}$$

Example 5.5 *Let us consider the simple case of only two variables, A and B with p.d.f. $P_A(a)$ and $P_B(b)$ and let us assume that they are not correlated. The joint p.d.f. is:*

$$P_{AB}(a, b) = P_A(a) \cdot P_B(b)$$

We have to compute the p.d.f. of a function of A and B: $C = f(A; B)$. We have to consider two new variables:

$$\begin{cases} c &= f(a, b) \\ g &= b \end{cases}$$

We have now to invert the previous relation in order to write:

$$\begin{cases} a &= A(c, g) \\ b &= g \end{cases}$$

The Jacobian is:

$$J = \frac{\partial(A, B)}{\partial(f, g)} = \begin{vmatrix} \frac{\partial A}{\partial c} & \frac{\partial A}{\partial g} \\ \frac{\partial B}{\partial c} & \frac{\partial B}{\partial g} \end{vmatrix} = \frac{\partial A}{\partial c}$$

The joint density of c and g is:

$$H(c, g) = P_A(a(c, g)) \cdot P_B(g)\,|J|$$

The density of c is obtained integrating away the variable g:

$$C(c) = \int_{-\infty}^{\infty} H(c, g)\; dg$$

Example 5.6 *Let us make an example of the procedure outlined in the previous example. Assume that*

$$\begin{cases} a &\sim N(0, 1) \\ b &\sim N(\mu, 1) \end{cases}$$

And we are interested in the function $c = a + b$.

$$\begin{cases} c &= a + b \\ g &= b \end{cases} \qquad \begin{cases} a &= c - g \\ b &= g \end{cases}$$

The Jacobian is one and the joint density is:

$$H(c,g) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}e^{-(c-g)^2/2}e^{-(\mu-g)^2/2}$$

The integral is performed squaring the exponent:

$$H(c,g) = \frac{1}{2\pi}exp\left[-\frac{1}{2}\left(\frac{(c-\mu)^2}{2} + \frac{(2g-(c+\mu))^2}{2}\right)\right]$$

and, integrating away g we obtain:

$$\begin{aligned} C(c) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} exp\left[-(g-\frac{c+\mu}{2})^2\right]\,dg\; exp\left[-\frac{(c-\mu)^2}{4}\right] \\ &= \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{(c-\mu)^2}{4}} \end{aligned}$$

which is, as we could have anticipated, a Normal distribution with mean μ and variance $1+1 = 2$. The procedure could become rather complicate. In many cases the complete calculation of the probability density is not needed; often only the mean and the variance are important and linearization or numerical calculation can solve the problem.

5.4 Linearized Approximation and the "Propagation of errors"

This section describes a method very often used in the practice when we want to compute the standard deviation of a function of a random variable.

In principle we should know or have derived the *p.d.f.* of the new variable (z) with the methods of the previous section and then compute:

$$\mu_z = E(z(x)) \qquad\qquad \sigma_z^2 = E((z(x)-\mu_z)^2)$$

However the procedure is often long and it is usually adequate to linearize the function and use the previous theorems.

Assume we have measured random variables

$$x_1, x_2, \cdots, x_n = \vec{x}$$

and we want to compute the variance of a function of these variables $y(\vec{x})$. We assume that we can perform a Taylor expansion around μ and truncate to the first term:

$$y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1,n}(x_i-\mu_i)\left.\frac{\partial y}{\partial x_i}\right|_{\vec{x}=\vec{\mu}} + \cdots$$

If we take the expectation value, the first order term vanishes:

$$\begin{aligned} E(y(\vec{x})) &= y(\vec{\mu}) + \cdots \\ Var(y(\vec{x})) &= E((y-E(y))^2) = \\ &= E(y(\vec{x})-y(\vec{\mu}))^2 = \\ &= \sum_{i=1,n}\sum_{j=1,n}\left.\left(\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right)\right|_{\vec{x}=\vec{\mu}} E((x_j-\mu_j)(x_i-\mu_i)) \end{aligned}$$

where, by definition:

$$E((x_j-\mu_j)(x_i-\mu_i)) = Var(\vec{x}) = \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) & \cdots \\ Cov(x_2,x_1) & Var(x_2) & \cdots \\ \cdots & \cdots & Var(x_n) \end{bmatrix}$$

is the covariance matrix of $\vec{x}$. The diagonal elements are the variance of the variables.

$$Var(y) = \sum_{i=1,n}\sum_{j=1,n} \frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\bigg|_{\vec{x}=\vec{\mu}} \cdot (Var(\vec{x}))_{i,j}$$

Usually the measured variable are uncorrelated and the variance matrix is diagonal. In the case of mutually independent variables the covariance matrix is diagonal and the previous formula becomes:

$$Var(y) = \sum_{i=1,n} (\frac{\partial y}{\partial x_i})^2\bigg|_{\vec{x}=\vec{\mu}} \cdot \sigma_i^2(\vec{x})$$

This is the well known formula for propagating the errors. In fact in most of the cases the x_i are uncorrelated (or their correlation is neglected). The general formula has to be used in the case the variables are correlated.

Now one should ask: when the linearization is allowed? There is no general answer. Figure 5.2 shows an example of what happens. The variable x is distributed as a Normal $(N(\mu, \sigma^2))$ and the variable $y = x^2$. A measure of the variability of x is the standard deviation σ. The function y can be linearized locally around μ (the line marked as *Lin. Appr.*). If, in the region of variability of x, the function y is well represented by the straight line, the linearization is good. This depend upon the value of σ as well as on the value of local curvature of the function.

Exercise 5.1 *Assume that x_1 and x_2 are uncorrelated random variables having variances $Var(x_1)$ and $Var(x_2)$ respectively. Compute the variance of the functions: $y = x_1 \cdot x_2$ and $y = \frac{x_1}{x_2}$*

The inverse of the covariance matrix is called the *weight matrix*.

Exercise 5.2 *Assume $x_1, x_2, \cdots, x_n$ are n measurements of the same quantity. It is convenient to consider the x_i as independent variables. Assume that all the measurements have the same variance $Var(x) = \sigma^2$. Compute the variance of the function $y = \frac{\sum x_i}{n}$.*

The previous formalism can be generalized to the case where a set of functions $\vec{y}$ is function of $\vec{x}$.

$$y_j(x_i), \qquad i = 1...n, \quad j = 1...m$$

Again we shall follow the approximate procedure by linearizing the functions by Taylor expansion about the mean.

$$y_j(\vec{x}) = y_j(\vec{\mu}) + \sum_i (x_i - \mu_i)\frac{\partial y_j}{\partial x_i}\bigg|_{\vec{x}=\vec{\mu}} \cdots$$

The variance of $\vec{y}$ is:

$$\begin{aligned} Var(y)_{k,l} &= E\left((y_k - E(y_k))(y_l - E(y_l))\right) = \\ &= \sum_{i=1,n}\sum_{j=1,n} (\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j})\bigg|_{\vec{x}=\vec{\mu}} E((x_i-\mu_i)(x_j-\mu_j)) = \\ &= \sum_{i=1,n}\sum_{j=1,n} (\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j})\bigg|_{\vec{x}=\vec{\mu}} Var_{i,j}(\vec{x}) \end{aligned}$$

It convenient to express these results in matrix notation.

$$\vec{y} = \vec{\mu_y} + S \cdot (\vec{x} - \vec{\mu}) + \cdots$$

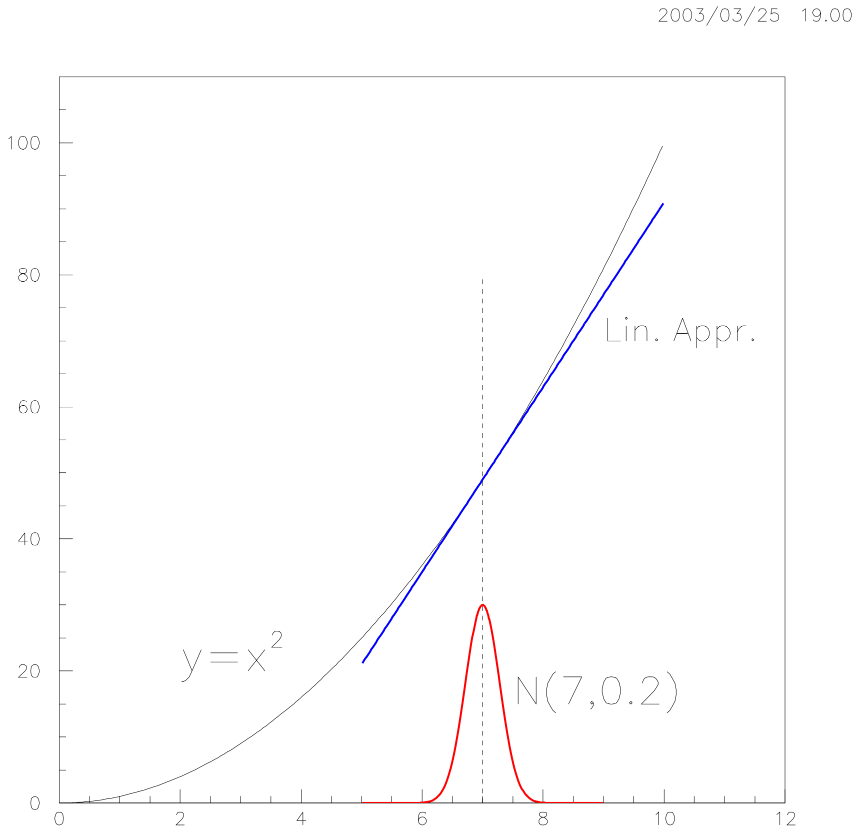


Figure 5.2: *An example of how the linearization works. The random variable x $N(\mu, \sigma^2)$ and the function of x is $y = x^2$.*

S is a matrix of the derivatives:

$$S_{k,l} = \frac{\partial y_k}{\partial x_l} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots \\ \cdots & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}_{m \times n}$$

and the law of propagation of errors becomes:

$$V(y) = SV(x)S^T$$

We have to stress that the formulas are approximate and their validity holds till we can neglect non linear terms of the functions. In the general case we have to use the full machinery by computing the *p.d.f.* of the new variable or use MC methods.

Summary of this Chapter

If: $x \sim f(x)$ and $h = h(x)$ then:	$h \sim g(h) = \sum_{h(x_i)=h} f\big(x_i(h)\big)\left\|\frac{\partial x(h)}{\partial h}\right\|_{x_i}$
This can be generalized: if $\{x_1, \cdots, x_n\} \sim f(x_1, \cdots, x_n)$ ($\vec{x} \sim f(\vec{x})$ in short) and $\begin{cases} h_1 & = f_1(\vec{x}) \\ \cdots & \\ h_n & = f_n(\vec{x}) \end{cases}$	$\vec{h} \sim \sum_{h(\vec{x_i})=\vec{h}} f\big(f(\vec{x_i}(\vec{h})\big)\left\|\frac{\partial\{x_1 \cdots x_n\}}{\partial\{h_1 \cdots h_n\}}\right\|_{x_i}$
Sometimes a linear approximation can be used: $h(\vec{x}) = h(\vec{\mu}) + \sum (x_i - \mu_i)\frac{\partial h}{\partial x_i}\big\|_{\vec{x}=\vec{\mu}} + \cdots$	$\begin{cases} E(h) & = h(E(\vec{x})) = h(\vec{\mu}) \\ Var(h) & = G_y \cdot Cov(x) G_y^T \end{cases}$
where	$\begin{cases} G_y\big\|_{i,j} & = \partial_{x_i} y_j \\ Cov(x)\big\|_{i,j} & = E\big((x_i - \mu_i)(x_j - \mu_j)\big) \end{cases}$
If the measurements are uncorrelated:	$\begin{cases} Cov(x)\big\|_{i,j} & = \sigma_i^2 \delta_{i,j} \\ Var(h) & = \sum \big(\partial_{x_i} h\big)^2 Var(x_i) \end{cases}$

Chapter 6

Joint Probability Distributions

6.1 Definitions

If X_1 and X_2 are discrete variables:

$$Pr((X_1 = r_1) \cap (X_2 = r_2)) = P_{r_1|r_2}$$

$P_{r_1|r_2}$ is the joint probability of the two variables r_1 and r_2 and must fulfill the unitarity:

$$\sum_{r_1} \sum_{r_2} P_{r_1|r_2} = 1$$

Example 6.1 *let X_1 be the column number and X_2 the raw number on a chess board. The probability that $X_1 = r_1$ and $X_2 = r_2$ if a square is chosen at random is:*

$$Pr((X_1 = r_1) \cap (X_2 = r_2)) = 1/64$$

A similar definition applies to continuous variables:

$$f(x_1, x_2)\, dx_1\, dx_2 = P((x_1 \le X_1 \le x_1 + dx_1) \cap (x_2 \le X_2 \le x_2 + dx_2)$$

The normalization is:

$$\int \int f(x_1, x_2)\, dx_1 dx_2 = 1$$

Example 6.2 *When we repeat an experiment n times, it is convenient to think of the results as n random variables. If the results are independent, and $g(x)$ is the p.d.f. of one measurement then the joint p.d.f. of the n measurements is:*

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} g(x_i)$$

6.2 The Correlation Coefficient

Consider now the case of continuous variables (X and Y). A correlation coefficient is defined as:

$$\rho = \frac{E((X - E(X)) \cdot (Y - E(Y)))}{\sigma_x \sigma_y}$$

The correlation coefficient is lies between -1 and +1.

This can be proved by considering the variance of a linear combination of the variables (for simplicity we consider central variables):

$$Var(aX+Y) = a^2 \cdot Var(X) + Var(Y) + 2 \cdot a \cdot E(XY) > 0$$

This inequality must be true for any a, therefore the discriminant must be negative:

$$E(XY)^2 - Var(X) \cdot Var(Y) < 0$$

from which the theorem follows.

If the correlation is 0 it means that the two variables are *uncorrelated*, this however does not mean that the two variables are *independent*.

As an example consider the two variables X and $Y = X^2$. Clearly X and Y are functionally related and dependent. However, if the *p.d.f.* of X is symmetric about 0 they are uncorrelated.

Exercise 6.1 *prove the previous statement. (Hint: what is E(XY)?)*

From two correlated variables X and Y we can construct uncorrelated variables. Let us consider the two variables: x and $z = y + a \cdot x$. The correlation coefficient is:

$$E\big((x-E(x))(z-E(z)\big) = E\big((x-E(x))(y+a\,x-E(y-a\,x)\big) = E\big((x-E(x))(y-E(y)\big)+a\,E\big((x-E(x))^2\big)$$

and the correlation between x and z vanishes if:

$$a = -\rho\frac{\sigma_y}{\sigma_x}$$

(ρ is the correlation coefficient of X and Y).

6.3 Correlated Variables

The random variables which are the outcome of a measurement are usually uncorrelated. However, sometimes, a correlation is introduced if a common process affects in the same way the measurements. Here we will discuss a simple example from a common laboratory practice.

Assume we analyze a charge spectrum made of two signals closely spaced in time, as shown in Fig. 6.1. The measurements are biased by a common noise which manifests as a low frequency shift of the pedestal, uncorrelated with the signals. (The pedestal is the average charge measured by our instrument when no input is present, i.e. the zero of our instrument.)

The measurement consists of measuring the individual signals and the pedestal. The two signals (Q_1 and Q_2) fluctuate independently because of the stochastic mechanisms at their origine (ionization losses etc.) but the two measured charges (q_1 and q_2) are correlated because of the baseline (b) fluctuation which is common to both.

If we would plot one against the other Q_1 and Q_2, we would observe something similar to figure 6.2 left, while the plot of q_1 against q_2 would look like figure 6.2 right. Notice that the dispersion of the measured charges is larger than the dispersion of the signal, due to the contribution of the pedestal noise.

The measurement can be improved if we measure also the value of the pedestal at the time of the signals and subtract, event by event, the pedestal from the measured charge. Suppose that the analysis is made by subtracting, event by event, the pedestal from the measured charges, and assume also that the *p.d.f.* of p, Q_1 and Q_2 is normal. The signals are obtained as:

$$Q_1 = q_1 - p \qquad\qquad Q_2 = q_2 - p$$

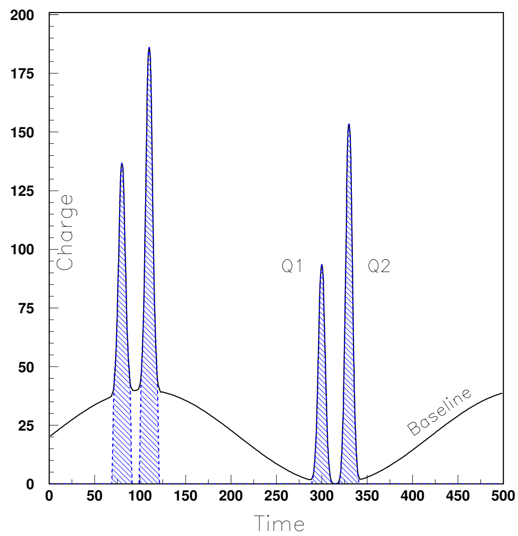
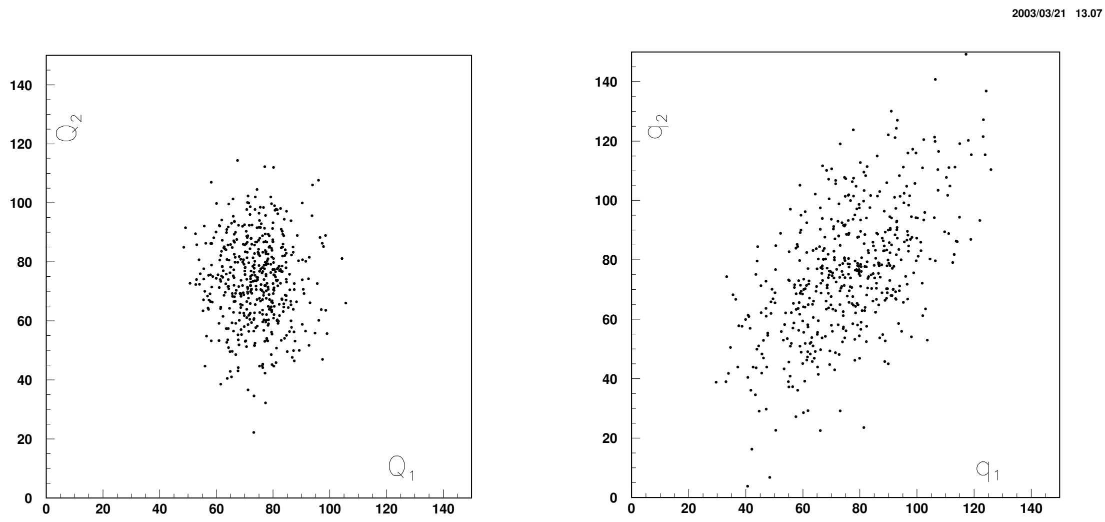


Figure 6.1: *Measurement of a charge with and ADC.*



,

Figure 6.2: *Correlation plot of the two uncorrelated signals (left) and the measured charges.*

It is reasonable to assume that p, Q_1 and Q_2 are uncorrelated, since they are produced by different mechanisms.

It is convenient to organize the measurements in matrix form:

$$\vec{X} = \begin{pmatrix} Q_1 \\ Q_2 \\ p \end{pmatrix} \qquad \vec{Y} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \qquad Var(\vec{X}) = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_p^2 \end{pmatrix}$$

To compute the covariance matrix of $\vec{Y}$ we will proceed as described in the previous chapter.

$$Var(\vec{Y}) = M \cdot Var(\vec{x}) \cdot M^T$$

The matrix of the derivatives is:

$$M = \frac{\partial \vec{Y}}{\partial \vec{X}} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

and a simple matrix algebra yields:

$$Var(\vec{Y}) = \begin{pmatrix} \sigma_1^2 + \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_2^2 + \sigma_p^2 \end{pmatrix}$$

Therefore the correlation coefficient is:

$$\rho(q_1, q_2) = \frac{\sigma_p^2}{\sqrt{\sigma_1^2 + \sigma_p^2}\sqrt{\sigma_2^2 + \sigma_p^2}}$$

As intuitive, the correlation vanishes if σ_p goes to zero.

It will be useful later to compute the weight matrix, i.e. the inverse of the variance.

$$W(\vec{Y}) = Var(\vec{Y})^{-1} = \begin{pmatrix} \frac{1}{(\sigma_1^2+\sigma_p^2)(1-\rho^2)} & \frac{-\sigma_p^2}{(\sigma_1^2+\sigma_p^2)(\sigma_2^2+\sigma_p^2)(1-\rho^2)} \\ \frac{-\sigma_p^2}{(\sigma_1^2+\sigma_p^2)(\sigma_2^2+\sigma_p^2)(1-\rho^2)} & \frac{1}{(\sigma_2^2+\sigma_p^2)(1-\rho^2)} \end{pmatrix}$$

The *p.d.f.* of q_1 and q_2 can be computed as described above. Since Q_1, Q_2 and p are uncorrelated and independent, their joint *p.d.f.* is the product of the three *p.d.f.* which we have assumed to be normal.

$$\begin{aligned} f(Q_1, Q_2, p) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(Q_1-\mu_1)^2}{2{\sigma_1}^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{(p-\mu_p)^2}{2{\sigma_p}^2}} \cdot \\ &\cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(Q_2-\mu_2)^2}{2{\sigma_2}^2}} \end{aligned}$$

Let first assume that the mean value of the pedestal is zero. We then change the variables:

$$\begin{aligned} q_1 &= Q_1 - \mu_1 + p \\ q_2 &= Q_2 - \mu_2 + p \\ P &= p \end{aligned}$$

(q_1, q_2 and p have zero average). We express q_1, q_2 and p as a function of Q_1, Q_2 and p and substitute in the expression of $f(Q_1, Q_2, p)$. The Jacobian is one. We then integrate away p to obtain the joint distribution of Q_1 and Q_2.

$$f(q_1, q_2, P) = \frac{1}{(\sqrt{2\pi})^3 \sigma_1 \sigma_2 \sigma_p} e^{-\frac{1}{2}(\frac{q_1^2}{\sigma_1^2} + \frac{q_2^2}{\sigma_2^2} + 2p(\frac{q_1}{\sigma_1^2} + \frac{q_2}{\sigma_2^2}) + p^2(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \frac{1}{\sigma_p^2}))}$$

The integral can be performed, remembering that:

$$\int_{-\infty}^{\infty} e^{-(ax^2+bx)}\, dx = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}}$$

And we get:

$$\int_{-\infty}^{\infty} f(q_1, q_2, p)\, dp = \frac{1}{(\sqrt{2\pi})^3} \frac{\sigma_m \sqrt{2\pi}}{\sigma_1 \sigma_2 \sigma_p} e^{-\frac{1}{2}(\frac{q_1^2}{\sigma_1^2} + \frac{q_2^2}{\sigma_2^2} + \sigma_m(\frac{q_1}{\sigma_1^2} + \frac{q_2}{\sigma_2^2})^2)}$$

where:

$$\frac{1}{\sigma_m^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \frac{1}{\sigma_p^2} = \frac{(\sigma_1^2 + \sigma_p^2)(\sigma_2^2 + \sigma_p^2) - \sigma_p^4}{(\sigma_1 \sigma_2 \sigma_p)^2}$$

Note that the density cannot be cast in the form of a product of two functions, one function of q_1 and one of q_2 only, the variables q_1 and q_2 are correlated.

The previous expression can be recast in a more compact form using the expression of ρ defined above.

$$f(q_1, q_2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_p^2}\sqrt{2\pi}\sqrt{\sigma_2^2 + \sigma_p^2}\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{q_1^2}{\sigma_1^2+\sigma_p^2} + \frac{q_2^2}{\sigma_2^2+\sigma_p^2} - 2\rho \frac{q_1 q_2}{\sqrt{\sigma_1^2+\sigma_p^2}\sqrt{\sigma_2^2+\sigma_p^2}}\right)}$$

The previous expression can be written in an even more compact form using the weight matrix.

$$f(q_1, q_2) = \frac{1}{(\sqrt{2\pi})^2} \cdot \frac{1}{\sqrt{DetV_y}} \cdot e^{-\frac{1}{2}\vec{Y}^T V_Y^{-1} \vec{Y}}$$

6.4 A Physical Correlation

Consider charged particles which pass through a medium. Their angular deflection and the lateral displacement are correlated variables (Fermi 1941). We will derive the distribution in the simple case of small deflections. Let x be the axis of motion of a particle. (The lengths are measured in g/cm^2)

Let $\xi(\theta_y)$ be the projected scattering probability such that the average scattering angle (in space) is:

$$\frac{\Theta_s^2}{2} = \int \theta_y^2 \, \xi(\theta) d\theta$$

The function ξ is symmetrical:

$$\int \theta_y \xi(\theta_y) d\theta_y = 0$$

and Θ_s is the usual average scattering angle

$$\Theta_s^2 = (\frac{E_s}{\beta \; p \; c})^2 \cdot \frac{1}{X_0}$$

and

$$E_s = \sqrt{\frac{4\pi}{\alpha}} \cdot m_e c^2 = 21 MeV$$

Let $P(x, y, \theta_y)$ be the probability that a particle at depth x in the material is in the cell $(y, y + dy) \cdot (\theta_y, \theta_y + d\theta_y)$. We want to compute the change in the probability $P(x, y, \theta_y)$ when the particle passes from x to $x + dx$. At the depth x the number of particles in the "volume" $dy \; d\theta_y$ at θ_y and y is: $P(x, y, \theta_y) dy d\theta_y$. On traversing the additional thickness dx some particles undergo a collision and are removed from the interval $d\theta_y$; their number is:

$$d\theta_y \cdot dx \cdot dy \cdot P(x, y, \theta_y) \cdot \int \xi(\theta_y') d\theta_y'$$

Other particles which were at y within dy but not in $d\theta_y$ can suffer a scattering and will be brought into $d\theta_y$; their number is:

$$d\theta_y \cdot dx \cdot dy \cdot \int \xi(\theta_y') \cdot P(x, y, \theta_y + \theta_y') d\theta_y'$$

The net change in the interval $d\theta_y$ is:

$$d\theta_y \cdot dx \cdot dy \cdot \int \xi(\theta'_y) \cdot (P(x,y,\theta_y+\theta'_y) - P(x,y,\theta_y))d\theta'_y$$

The spatial density is also changed since in traveling from x to $x+dx$ the particles with an angle θ_y will move from y to $y+dx\cdot\theta_y$. The movement from x to $x+dx$ produces the net change:

$$\begin{aligned}(P(x,y-dx\cdot\theta_y,\theta_y) - P(x,y,\theta_y)) \cdot d\theta_y \cdot dx \cdot dy &= \\ &= -\theta_y \cdot \frac{\partial P}{\partial y} \cdot \theta_y \cdot dx \cdot dy\end{aligned}$$

By adding the effects together we get:

$$\frac{\partial P}{\partial x} = -\theta_y \cdot \frac{\partial P}{\partial y} + \int (P(x,y,\theta_y+\theta'_y) - P(x,y,\theta_y))\xi(\theta'_y)d\theta'_y$$

If we assume that $P(x,y,\theta_y+\theta'_y)$ can be developed in power series of θ_y up to the second order (no hard scattering, this is sometimes called the *diffusive* or Fokker-Plank approximation), we get:

$$\frac{\partial P}{\partial x} = -\theta_y \frac{\partial P}{\partial y} + \frac{\Theta_s^2}{4} \cdot \frac{\partial^2 P}{\partial \theta_y^2}$$

The solution which corresponds to a single particle is:

$$P(x,y,\theta_y) = \frac{2\sqrt{3}}{\pi} \cdot \frac{1}{\Theta_s^2 \cdot x^2} \cdot e^{-\frac{4}{\Theta_s^2}\cdot(\frac{\theta_y^2}{x} - \frac{3y\theta_y}{x^2} + \frac{3y^2}{x^3})}$$

The mechanism by which the variables y and θ_y become correlated is clear. Mathematically this is made more clear by rearranging the variables. Let:

$$\begin{aligned}\sigma_\theta^2 &= \frac{\Theta_s^2 x}{2} \\ \sigma_y^2 &= \frac{\Theta_s^2 x^3}{6} \\ u &= \frac{\theta_y}{\sigma_\theta} \\ v &= \frac{y}{\sigma_y} \\ \rho^2 &= \frac{3}{4}\end{aligned}$$

The *p.d.f.* becomes:

$$P(x,u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot e^{-\frac{u^2-2\rho uv+v^2}{2\cdot(1-\rho^2)}}$$

which is the *p.d.f.* of bivariate normal distribution with correlation ρ. The marginal probability of u (integrating away v) is:

$$P(u) = \int_{-\infty}^{\infty} P(x,u,v)dv = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$$

The variable u is $N(0,1)$. Symmetrically θ_y is $N(0,\sigma_{\theta_y}^2)$.

The conditional probability $P(u \mid v)$ i.e. the *p.d.f.* of u given the value v of the other variable, is given by the standard rules:

$$P(u \mid v) = \frac{P(U \cap V)}{P(V)} = \frac{P(u,v)}{P(v)}$$

which, after some algebra becomes:

$$P(u \mid v) = \frac{1}{\sqrt{2\pi} \cdot \sqrt{1-\rho^2}} \cdot e^{-\frac{(u-\rho v)^2}{2(1-\rho^2)}}$$

The variable u is normal, but centered at ρv. Its variance is reduced from 1 to $1-\rho^2$ (Fig.6.3). This is the consequence of having measured the correlated quantity v. These facts are general of normal correlated variables. The line:

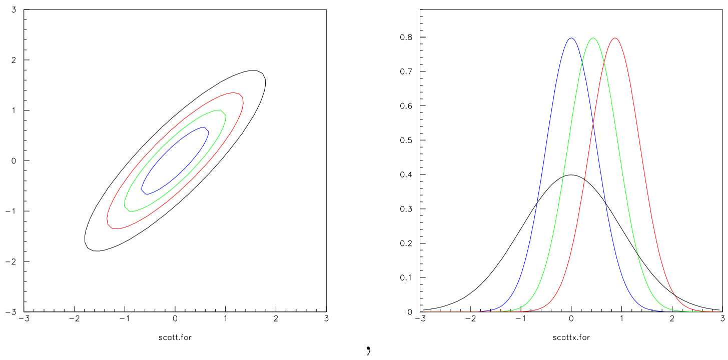


Figure 6.3: *Contour level of $P(u,v)$ and conditional probability distribution.*

$$\begin{aligned} u &= \rho v \\ \theta_y &= y \cdot \frac{3}{2x} \end{aligned}$$

is called regression line [1].

Exercise 6.2 *Prove that if a particle passes through two points A and B separated by x, the p.d.f. of the position halfway between A and B is normal with* $< y^2_{hw} >= \frac{\Theta^2_S \cdot x^3}{96}$.

Exercise 6.3 *Let us measure the track position with drift chambers placed at distance D from each other. The angles between two successive chords are:*

$$\omega_i = \frac{y_{i+1} - y_i}{D}$$

prove that $\sigma^2 \omega = \frac{2}{3}\sigma^2_\theta$ *and that* $< \omega_i \omega_{wi+1} >= \frac{\sigma^2_\theta}{6}$ *so that the correlation coefficient* $\rho_\omega = \frac{1}{4}$.

Example 6.3 *We measure the position and the angle of a particle after a slab of material. The true angle of production θ_0 is changed by the scattering. We can improve our knowledge of the scattering angle θ_0 by measuring y in addition to θ_1. We can write:*

$$\theta_0 = f_1 y + f_2 \theta_1$$

and choose f_1 and f_2 to minimize the variance of θ_0 and compute the variance of θ_0.

[1]The origin of the name is due to a biometric study by Dalton who investigated the correlation of the height of the pair father-son. If the height of the father (x) is higher of the average of the population, the same is for the height of the son (y(x)), but in lesser proportion. The height of the son regresses toward the average of the population.

Chapter 7

Sample Variables

7.1 The Sum of Random Variables

The sum of random variables is often met when we analyze data. For instance the average of repeated measurements etc. We will see that when the number of random variables is very large the random variable defined as the sum of all of them ($S = \sum x_i$), has *p.d.f.* which is independent of the *p.d.f.* of the individual variables. We can use the *m.g.f.* to find the distribution of the sum. We will now use the *m.g.f.* technique to compute the mean and the variance of the sum.

A property of the *m.g.f.* is that the *m.g.f.* of the sum of variables is the product of the individual *m.g.f.*:

$$\begin{aligned} \Phi_{\sum x_i}(t) &= E\big(e^{t\sum x_i}\big) = \\ &= \prod \Phi_x(t) = \big(\Phi_x(t)\big)^n \end{aligned}$$

Let the moments of $\sum x_i$ about the origin be m'_k and the moments of x_i be μ'_k. We have:

$$\begin{aligned} \Phi \sum x_i(t) &= 1 + m'_1 t + m'_2 \frac{t^2}{2!} + \cdots = \\ &= \big(1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \cdots\big)^n = \\ &= 1 + n\ \mu'_1\ t + \frac{n(n-1)}{2}\mu'^2_1\ t^2 + n\mu'_2 \frac{t^2}{2} + \cdots \end{aligned}$$

Equating the coefficients of t we get:

$$m'_1 = n\mu'_1$$

The expected value of the sum of the variables is the sum of their expected values $E(\Sigma) = n \cdot E(X)$.

Equating the coefficients of $\frac{t^2}{2!}$:

$$m'_2 = n \cdot \mu'_2 + (n^2 - n)\mu'^2_1$$

and the variance:

$$\begin{aligned} Var\left(\sum x_i\right) &= m'_2 - m'^2_1 = \\ &= n \cdot (\mu'_2 - \mu'^2_1) = nVar(x_i) \end{aligned}$$

These are the addictive properties of the mean and of the variance.

7.2 The Sample Mean

We consider the sample mean:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

We see easily that:

$$E(\overline{x}) = E\left(\sum x_i\right)/n = E(x_i) = \mu$$

The sample mean is an *unbiased estimate* of the population mean μ. Also:

$$\begin{aligned} Var(\overline{x}) &= Var\left(\frac{\sum x_i}{n}\right) = \frac{Var\left(\sum x_i\right)}{n^2} \\ &= \frac{Var(x_i)}{n} \end{aligned}$$

The sample mean is a more accurate estimate of μ than the single values x_i.

These results are independent of the *p.d.f.* of x.

7.3 The Law of Large Numbers

We have seen that the sample mean is a good estimator of the population mean. We have to prove that when the sample size increases, the sample mean tends to μ. First we prove the Chebicheff's inequality:

Theorem 7.1 *If μ and σ are the mean and the standard deviation of the p.d.f. of x, then, for any k we have:*

$$P(\mid x - \mu \mid \geq k \cdot \sigma) \leq \frac{1}{k^2}$$

The inequality is valid for all p.d.f. with finite mean and variance.

Proof 7.1 *We divide the following integral in three parts: from $-\infty$ to $\mu - k \cdot \sigma$, from $\mu - k \cdot \sigma$ to $\mu + k \cdot \sigma$ and from $\mu + k \cdot \sigma$ to $+\infty$; then we drop the middle term.*

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = \\ &= \int_{-\infty}^{\mu-k\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} \int_{\mu+k\sigma}^{\infty} (\mu - x)^2 \cdot f(x)dx \\ &\geq \int_{-\infty}^{\mu-k\sigma} \int_{\mu+k\sigma}^{\infty} (\mu - x)^2 \cdot f(x)dx \end{aligned}$$

For all x within the range of the integral we have:

$$(x-\mu)^2 \geq (k\sigma)^2$$

Then we obtain:

$$\sigma^2 \geq (k\sigma)^2 \int_{-\infty}^{\mu-k\sigma} \int_{\mu+k\sigma}^{\infty} f(x)dx = k^2\sigma^2 \; P(\mid x - \mu \mid \geq k\sigma)$$

The integral is the probability that $\mid x - \mu \mid \geq k\sigma$, therefore we have proved the theorem. ∎

This inequality is generally rather weak. For instance in the case of Normal distribution with $k = 1$, we get from the statistics tables $P(\mid \mu - x \mid \geq 1) = 0.32$ while the inequality predicts a upper limit of 1. The importance of the Chebicheff inequality consists in the fact that it holds for a very large class of *p.d.f.*.

Let us apply this inequality to the sample mean $\overline{x}$ whose variance is $Var(\overline{x}) = \frac{\sigma^2}{N}$. We get:

$$P\left(\mid \mu - \overline{x} \mid \geq \frac{k\sigma}{\sqrt{N}}\right) \leq \frac{1}{k^2}$$

Calling:

$$\epsilon = \frac{k\sigma}{\sqrt{N}} \qquad\qquad k = \epsilon\frac{\sqrt{N}}{\sigma}$$

We get:

$$P(\mid \overline{x} - \mu \mid) \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$$

If we take the opposite proposition:

$$P(\mid \overline{x} - \mu \mid < \epsilon) > 1 - \frac{\sigma^2}{N\epsilon^2}$$

Or:

$$P(-\epsilon < (\overline{x} - \mu) < \epsilon) > 1 - \frac{\sigma^2}{N\epsilon^2}$$

This is *the weak law of large numbers.* The probability that $x \to \mu$ tends to 1 as $N \to \infty$. There exists also a strong law of large numbers. The difference between the two is similar to the difference between the point wise convergence and uniform convergence in function theory.

Important points are: the theorem holds for a large class of *p.d.f.* and it quantifies the way the sample mean approaches μ.

7.4 The Central Limit Theorem

We have just shown that the accuracy with which the sample mean estimates μ increases as $\sqrt{N}$.

The central limit theorem gives a stronger statement on $\overline{x}$ and its distribution:

Theorem 7.2 *Consider a variable X whose mean is μ and variance is σ^2. If σ^2 is finite, then the distribution of the sample mean $(\overline{x})$ approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{N}$ as $N \to \infty$.*

Proof 7.2 *The proof will be given by proving that the m.g.f. of the variable $z = \sum w_i$ $\left(w_i = \frac{x_i - \mu}{\sigma \cdot \sqrt{N}}\right)$ is the same as the one of a Normal distribution of unit variance and zero mean in the limit $N \to \infty$.*

$$\begin{aligned}
\Phi_{x-\mu}(t) &= E\left(e^{(x-\mu)t}\right) = \\
&= 1 + m_1\,t + m_2\,\frac{t^2}{2!} + m_3\,\frac{t^3}{3!} + \cdots = \\
&= 1 + 0 + \sigma^2\,\frac{t^2}{2!} + m_3\,\frac{t^3}{3!} + \cdots
\end{aligned}$$

Let $w = \frac{x-\mu}{\sigma \cdot \sqrt{N}}$. The m.g.f. of w is:

$$\begin{aligned}\Phi_w(t) &= E\big(e^{\frac{(x-\mu)t}{\sigma\sqrt{N}}}\big) = \\ &= \Phi_x\big(\frac{t}{\sigma\sqrt{N}}\big) = \\ &= 1 + \frac{\sigma^2 t^2}{2!\sigma^2 N} + \cdots = \\ &= 1 + \frac{t^2}{2N} + \cdots\end{aligned}$$

Let's define a new variable z:

$$z = \sum_{i=1}^{N} w_i = \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma\sqrt{N}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{N}}$$

From a previous theorem we know that the m.g.f. of the sum of N identical distributed variables is the Nth power of the individual m.g.f.

$$\Phi_z(t) = \big(\Phi_w(t)\big)^N = \big(1 + \frac{t^2}{2N} + \cdots\big)^N \approx 1 + \frac{t^2}{2} + \cdots \to e^{\frac{t^2}{2}}$$

This density tends to $e^{t^2/2}$ as N tends to infinity which is the m.g.f. of a normal distribution with zero mean and variance 1.

In the limit of N going to infinity z has a normal distribution with zero mean and unit variance. Since $\overline{x} = \frac{\sigma z}{\sqrt{N}} + \mu$ and since a linear function of a Normal variable is also a normal variable, $\overline{x}$ in the limit is normally distributed with mean μ and variance $\frac{\sigma^2}{N}$. ∎

The only condition required by the theorem is that the variance is finite. No other condition on the *p.d.f.* of x is required. Fig. 7.1 shows how variables with different *p.d.f.* have a sample mean which approaches N(0,1),

The theorem has also important practical applications:

Example 7.1 *A die is thrown 300 times. We call success if the die lands 1 or 2 ($x = 1$, $x = 0$ if failure). Hence $p = 1/3$. The total number of successes is $r = \sum x_i$ and the mean value of x is $r/300$ with $E(\overline{x}) = 1/3$. Which is the probability that the number of successes is comprised between 85 and 115? The probability for r is (from the binomial distribution)*

$$P(r) = \left(\begin{array}{c} 300 \\ r \end{array}\right) (\frac{1}{3})^r (\frac{2}{3})^{300-r}$$

Then we should sum the probabilities for $r = 85, 86, \cdots 115$. The computation is rather tedious and lengthy. An approximate answer can be obtained by changing the formulation of the problem:

$$P(85 < r < 115)$$

to the equivalent one:

$$P(\frac{85}{300} < \overline{x} < \frac{115}{300})$$

and as soon as we deal with averages, we can apply the central limit theorem. The average has normal distribution with mean $1/3$ and variance $p(1-p)/N = 2/2700$. Therefore:

$$P(\frac{85}{300} < \overline{x} < \frac{115}{300}) = \int_{85/300}^{115/300} N(\frac{1}{3}, \frac{2}{2700}) dx$$

which can be read from the statistical tables.

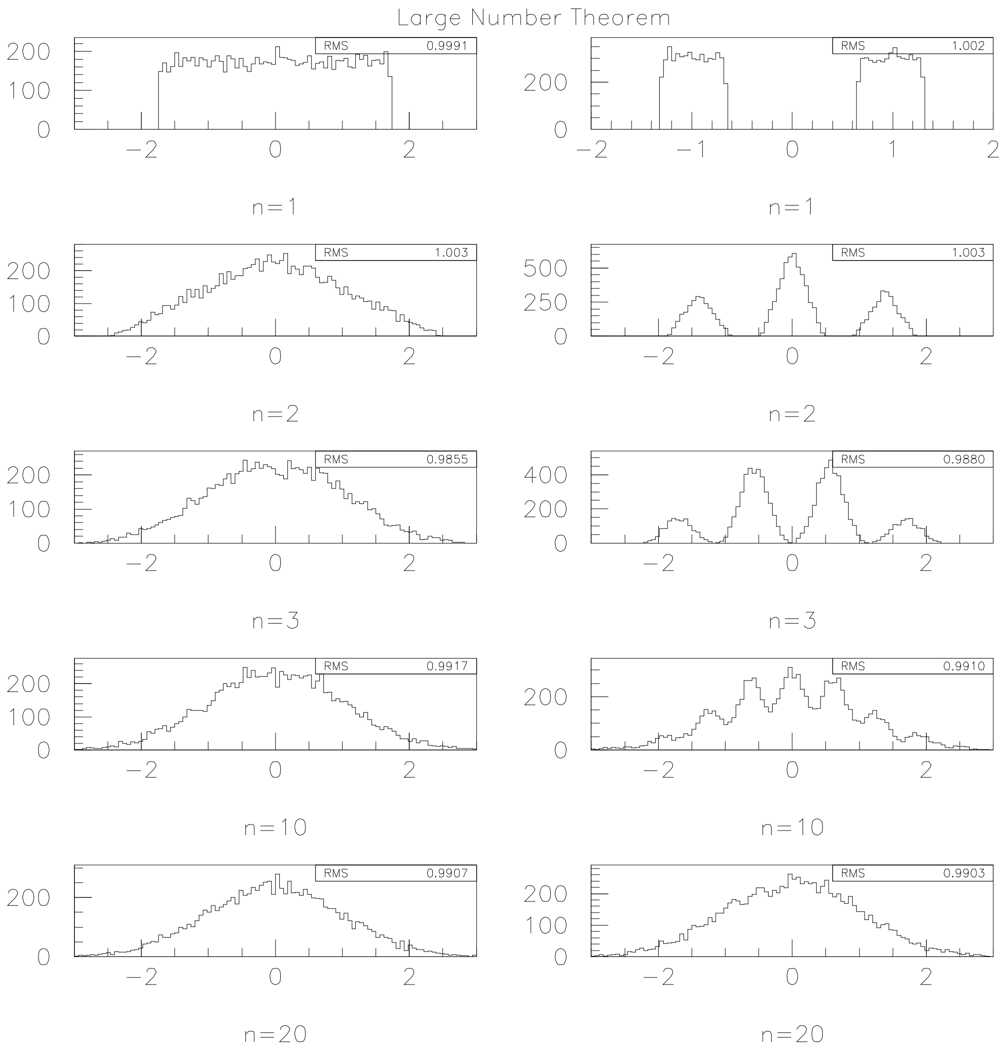


Figure 7.1: *Top figures shows the p.d.f. of two variables. On each row we plot the p.d.f. of the average of a sample of size 2,3,5,10,20.*

7.5 Random Sample

In practical instances we do not have the full "population" to determine its properties, but only a sample of it. We have a few events out of all possible events that can occur in a high energy interaction. Or we can measure the resolution of a few calorimeter modules out of the many employed in an experiment. To study the properties of the interaction or the overall resolution of the complete calorimeter we need to have an unbiased sample.

The word *random* means that every possible samples have the same chance to be selected (simple random sample) or at least a calculable chance.

In the real word we know that very often our measurements are subject to personal taste or prejudices and the very basic notion of *random sample* becomes questionable. In High Energy Physics we talk about *biases* or *systematic effects*. The methods of statistics that we are discussing *do not apply* to these type of effects, which will need special cares and will be discussed later in this course.

There are classical examples of personal biases. Table 7.1 shows the frequency of final digit in a scale reading made by four observers (A, B, C and D) (Yule, 1927):

Final Digit	A	B	C	D
0	158	122	251	358
1	97	98	37	49
2	125	98	80	90
3	73	90	72	63
4	76	100	55	37
5	71	112	222	211
6	90	98	71	62
7	56	99	75	70
8	126	101	72	44
9	129	81	65	16

Table 7.1: *Example of biases. The table shows the frequency of occurrence of the last digit in the scale reading by four different observers.*

The experiment consisted in measuring the length of bins. The frequency of the last digit is expected to be flat. Observer B is the most accurate, but his frequency would rate pretty low in a statistical test for constant probability. The digits 0 occur more often at the expense of 9. Observer A has a preference on 0, 2, 8 and 9. C and even more D are rather poor: 0 and 5 (the easiest digits) occurs very often.

Even if one has no interest in the measurements, inevitably he will tend to influence the outcome of an experiment. A physicist is an observer very interested in the outcome of the measurement and will tend to bias even more the results. In bubble chamber experiments the phase of scanning (the search for events) was always performed by non physicists to whom the topology of the events was described in such a way not to create prejudices or let them know what physicists were looking for. Another example from High Energy Physics are the trigger biases in modern electronic experiments.

Chapter 8

The Method of MonteCarlo

Many physical processes can be idealized as a sequence of choices. For instance an unstable, long lived particle can or cannot decay in the sensitive volume of a detector. If it has decayed inside the detector its detection depends on the way it has decayed. If its decay products are $\mu + \nu_\mu$ or $e + \nu_e$ the signature in the detector is different. If it happened to be the electron mode, there is a finite probability of producing a bremstrahalung photon... The example can be continued further till the description of the energy release into the detector's sensitive elements. If the probability function of each choice is known, then it is possible to make a mathematical description of the physical process by simulating each choice. The analytical description is usually complex, but not impossible, and will aim to compute physically detectable quantities, like the energy distribution of electrons in our example or the efficiency to detect the process. In most of the cases we have to deal with rather complex integrals of probability distributions over the geometry of the experimental apparatus.

The MonteCarlo techniques are often the only practical methods to perform such numerical calculation of the complicate integrals. This is especially true with the modern fast digital computers.

Before entering in the more technical discussion we will discuss two issues: how the MonteCarlo methods originated and how it is possible to reconcile an intrinsically random process with the rigid sequence of operations performed by a digital computers.

8.1 A brief History of MonteCarlo

The method is called after the city in the Monaco Principality famous also for its Casinos and that magic (for somebody!) random number generator that is the roulette. Though the systematic development of MonteCarlo dates from the beginning of the era of digital computers, there have been previous examples of its use.

As early as in 1768 Buffon "experimentally" determined the value of π by casting a needle on a ruled grid. In a sense the experiment was used in a way opposite to its modern usage: it was a validation of the statistical methods, instead of the use of statistical sampling methods to determine answer to a problem.

In 1899 Lord Rayleigh showed that a one dimensional random walk without absorbing barriers could provide an approximate solution to parabolic differential equation. And Kolmogorov, in 1931 showed the deep relationship between Markov stochastic processes and a class of integro-differential equations.

In the early years of the twentieth century the British statistical school made some work in MonteCarlo methods, mostly of didactic character. But in some rare cases the method brought to real discoveries. W.S. Gosset (Student) used experimental sampling to help him towards his discovery of the distribution of correlation coefficient (1908). In the same years Student used statistical sampling to bolster his faith on the so called *t-distribution* of which he had given a somehow incomplete analysis.

According to Emilio Segre, Enrico Fermi in 1930 invented a form of MonteCarlo method when he was studying the moderation of neutrons in Rome. Fermi did not publish anything on the subject, but amazed his colleagues with his predictions of experimental results. After indulging himself, he would reveal that his guesses were derived from the statistical sampling techniques that he performed over night, when he could not sleep. Later, in 1947, while in Los Alamos, Fermi invented a mechanical device called FERMIAC to trace neutron through fissionable materials by the MonteCarlo method.

The real and systematic use of MonteCarlo methods as a research tool starts with the work on the atomic bomb (the Manhattan project). J. von Neumann and S. Ulam defined the mathematical basis for probability densities, inverse distributions and random number generators. The computing project was directed by R. Feynman[1]and it is interesting to remind how primitive was the computing environment at that time and the challenges that the project presented to the researchers.

The work of Harris and Khan in 1948 and immediately after the approximate solution of a Schroedinger equation by Fermi, Metropolis and Ulam opened the field of the modern use of MonteCarlo methods.

8.2 The Pseudo-Random Number Generators

Anyone who considers arithmetical methods of
producing random digits is, of course, in a state of sin.
J. von Neumann (1951)

In the following we will often speak of random numbers. This is not exact, if referred to "random numbers" generated by a digital computer.

- Truly random numbers are those produced in real random processes, like the time between two "clicks" from a Geiger counter exposed to a radioactive source. Truly random numbers exists and are stored in large data banks in some Laboratory. Such truly random numbers are cumbersome and lengthy to use and very rarely they are needed. Before the advent of digital computers and the modern techniques for pseudo-random number generators, there existed also books of random numbers.

- Pseudo-random numbers have only the appearance of randomness. In reality they exhibit a specific, repeatable pattern. These sequences can be produced by methods on digital computers. These methods produce a sequence of numbers which, finally, after a (long) cycle will return to the first one and repeat the sequence.

- Quasi-random is a sequence which fills the solution space. For instance in the integer space $[0, 100]$ a sequence: $\{0, 1, 2, \cdots 99, 100\}$ or $\{100, 99, \cdots 1, 0\}$ are quasi-random sequences.

Since the method consists in simulating with a random sampling the real processes, we should need the first category of events. As a matter of facts the pseudo-random are used and care is placed in choosing an algorithm that:

- Produces almost uncorrelated sequences of numbers.

[1]A fascinating exposition of this first effort in really large scale computation given in Feynman's *Surely you are joking, Mr Feynman.*

- The sequence of pseudo-random number should be very long to avoid that random sampling fails because of the repetition of the numbers.
- Uniform and unbiased sequence. Many generators produce pseudo-random sequences in the interval $[0-1]$. The interval should be evenly populated, as a truly random number should do.
- Efficient algorithms should be employed to avoid that too much overhead computing is spent in the random number generation.

8.2.1 The Linear Congruent Method

One of the most popular method to generate on a computer pseudo-random number is the Linear Congruent method. The sequence is defined by:

$$X_{n+1} = \lambda X_n + \mu \quad Mod\, m$$

By $r = x\ Mod\ m$ we mean that r is the remainder when x is divided by m; x, r and m are integer numbers.

This sequence has often good properties, in the sense specified above, if λ and μ are carefully chosen, for a given m. Of course $0 \leq r \leq m-1$ therefore the cycle of the sequence has a maximum length of m (a poor choice of λ and μ could make the cycle much shorter). This makes clear that it is important that m is large and since in a digital computer the maximum integer that can be contained in a memory of d bits is 2^d, the architecture of the computer itself is relevant.

The series above is of pseudo-random integers that range from zero to m. We can easily convert the series to floats (computer jargon which means fractional numbers) by dividing by the maximum number of the series:

$$Z_n = \frac{X_n}{m}$$

Let us try to generate pseudo-random numbers on an hypothetical 5-bits computer. At most we would generate 32 random numbers before the cycle repeats itself, since the computer has only $2^5 = 32$ stable states. We chose: $X_0 = 1$, $\lambda = 5$ and $\mu = 7$. A possible code (in FORTRAN) is shown in Figure 8.1. The list of the 32 pseudo-random numbers is given in Figure 8.2. It is clear from this example that these numbers fill the interval $[0-1]$ evenly and not in sequence, but that they are not quite random: the last digit repeat the sequence: 0,5,0....

8.2.2 The basic Equations for MonteCarlo Method

Let us assume that we have a good source of random variables R, as discussed in the previous section, which generates random numbers $\{r\}$ evenly distributed in the interval $[0-1]$; that is:

$$P(R \leq r) = r \qquad (0 \leq r \leq 1) \tag{8.1}$$

We require to generate values of a random variable X with a given cumulative distribution $F(x)$:

$$P(X \leq x) = F(x) \qquad (0 \leq F \leq 1) \tag{8.2}$$

Let us consider first the case of discrete variables. Figure 8.3 shows an example. We require a transformation from R to F such that X has the distribution 8.2. The algorithm is the following:

Theorem 8.1 *Find x_i such that:*

$$F(x_{i-1}) < r \leq F(x_i) \tag{8.3}$$

then put: $X = x(r) = x_i$..

```
Mar 04, 02 14:43                                   random8.f
      Program Random
      Real Z(32)
*
* Set parameters
*
      Lambda = 5
      Mu = 7
      Modulus = 32
*
* Set the seed
*
      Ix = 1
*
      Do N = 1,32
         Ix = mod(Lambda*Ix+Mu, Modulus)
         Z(N) = Float(Ix)/Float(Modulus)
      Enddo
*
      Write(15,1000) (Z(1),1=1,32)
 1000 Format(4F10.5/)
      Stop
      End
```

Figure 8.1: Five bits computer code to generate pseudo random numbers.

Mar 04, 02 14:43 **fort.15**

0.37500	0.09375	0.68750	0.65625
0.50000	0.71875	0.81250	0.28125
0.62500	0.34375	0.93750	0.90625
0.75000	0.96875	0.06250	0.53125
0.87500	0.59375	0.18750	0.15625
0.00000	0.21875	0.31250	0.78125
0.12500	0.84375	0.43750	0.40625
0.25000	0.46875	0.56250	0.03125

Figure 8.2: The 32 random numbers produced by the program of Figure 8.1. The next random number is equal to the first.

Proof 8.1 *From 8.3 we see that $X \leq x_i$ implies $R \leq F(x_i))$, hence $P(X \leq x_i) = P(R \leq F(x_i) = F(x_i))$.* ∎

In the continuous case the procedure is similar:

Theorem 8.2 *Find $x(r)$ such that:*

$$F(x(r)) = r \tag{8.4}$$

and put $X = x(r)$.

Proof 8.2 *The equation 8.4 implies a transformation $X = X(R)$ in the implicit form:*

$$F(X) = R \tag{8.5}$$

. Now equation 8.2 implies that $F(X)$ is a strictly increasing function of X, hence:

$$\begin{aligned} P(X \leq x) &= P(F(X) \leq F(x)) \\ &= P(R \leq r) \\ &= r \\ &= F(x) \end{aligned}$$

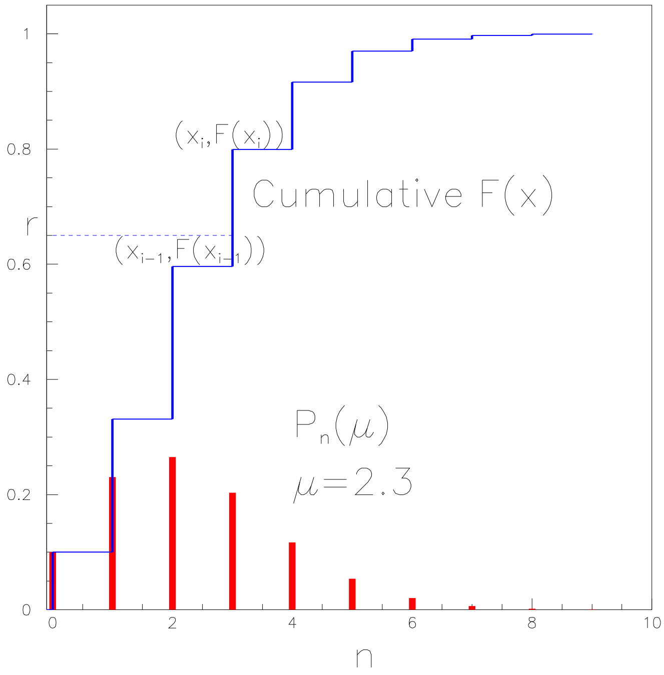


Figure 8.3: Probability function (thick lines) and its distribution.

The second line is a consequence of 8.4 and8.5, the third line follows from 8.1 and the last from 8.4. Again we see that X follows the required distribution. ∎

In the continuous case the distribution function is:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

where $f(x)$ is the probability density function. The basic equation 8.4 can be written in the following way:

$$r = \int_{-\infty}^{x(r)} f(t)\, dt$$

This equation defines $x = x(r)$. In general a solution in closed forms does not exists and a numerical iterative procedure has to be followed.

Example 8.1 *Let us consider a case where an analytical solution exists. Assume that* $f(x) = \lambda e^{-\lambda x}$. *The previous equation leads to the relation:*

$$r = \int_{0}^{x} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \rightarrow x = -\frac{1}{\lambda}\lg(1-r)$$

An example on the implementation in FORTRAN can be seen in Figure 8.4. The output of the program is shown in Figure 8.5

Sometimes is very simple to transform the uniform distribution to the needed one. An example:

```
Mar 04, 02 17:42                randomexp.f
      Program RandomExp
*-----------------------------------------------------------------
* How to generate exponential density
*-----------------------------------------------------------------

      Common/pawc/h(100000)
      Call Hlimit(100000)

*            Open output file

      Open(unit=10,file='rndmexp.eps',form='formatted',
     $                            status='unknown')

      Call HBOOK1(101,'Exponential',50,0.,10.,0.)
      Call HPLint(1)
      Call IGmeta(10,-113)

      Ntot = 5000
      Xlambda = 0.5
      Do I = 1, Ntot
         rr = -log(RNDM(x))/Xlambda
         call HFILL(101,rr,0.,1.)
      EndDo

      Call HPLopt('GRID',101)
      Call HPLopt('NBOX',101)
      Call HPLopt('STAT',101)
      Call HPLset('DMOD',1.)

      Call HPLot(101, ' ',' ',0)
      Call HPLsof(7.5, 14., 'Exponential random number',0.5,0.,99.,-1)

      Call HPLend
      Call Hindex

      END
```

Figure 8.4: The generation of an exponential density.

Example 8.2 *The generation of isotropic distribution of direction in space is very simple. Isotropy means that the density is proportional to the solid angle:* $d\Omega = d(\cos\theta)d\phi$*, hence* $\cos\theta$ *must be uniform in the interval* $[-1, 1]$ *and* ϕ *in the interval* $[0, 2\pi]$*:*

$$\begin{aligned} \cos\theta &= 1 - 2\cdot r \\ \phi &= 2\cdot\pi\cdot r \end{aligned}$$

(limits can be easily implemented).

8.2.3 von Neumann Method

A method different from the transformation of a variable is due to von Neumann. It is used in those cases in which the probability density is known numerically and not analytically. The method implements a simple Acceptance-Reject strategy: assume that you can envelope the density in a box of dimension A and B, then generate two uniform random numbers in A and B (r_x and r_y) respectively. Accept the number r_x if r_y is below the density at $x = r_x$.

Example 8.3 *Figure 8.6 shows a distribution which is confined in a box:* $0 \le x \le 1$ *and* $0 \le y \le 8$*. In the case of this example we have parametrized the density with a simple function and use the code shown in Figure 8.7 The output of the program is shown in Figure 8.8.*

8.3 MonteCarlo Simulation

Historically, the first application of MonteCarlo method were studies of neutron scattering and absorption by dense materials. These are random process and it appeared quite natural to employ random numbers to describe the real processes. These calculation are known as MonteCarlo simulation since we attempt to describe a complex physical process by direct construction, i.e.

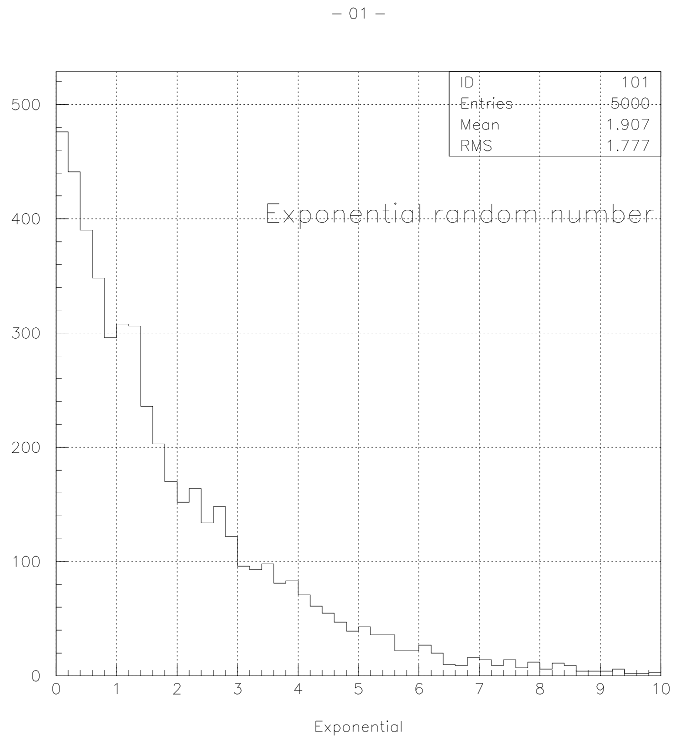


Figure 8.5: An exponential density generated from the program of Figure 8.4.

by mimiking the elementary random processes by suitable distributed pseudo random numbers Though the processes were random, the physical problem is deterministic and could be solved by analytical methods. A MonteCarlo simulation avoids complicate multidimensional integration and, sometimes a more intuitive solution to the physical problem.

8.4 Montecarlo as Integration

Let us consider a simple one dimensional integration:

$$I = \int_a^b f(x)dx$$

This can be easily transformed into:

$$I = (b-a) \cdot \int_a^b f(x) \cdot U(x)dx \qquad U(x) = \frac{1}{b-a}$$

We have formally multiplied the function by the uniform density (over the interval (a, b). This way of writing makes clear that the integral is the expectation value of f if the variable x is uniformly distributed in the interval (a, b).

$$I = (b-a) \cdot E(f)$$

Suppose that we have a random number $x_i \in (a, b)$ the value $f_i = f(x_i)$ is a random value, whose expectation value is $E(f_i) = I/(b-a)$. f_i is a estimation of $I/(b-a)$, albeit not very precise. The precision can be improved if, instead of f_i we take the arithmentic mean over many random points x, uniformly distributed in the interval (a, b).

$$I_n = (b-a) \cdot \frac{\sum_{i=1,n} f(x_i)}{n}$$

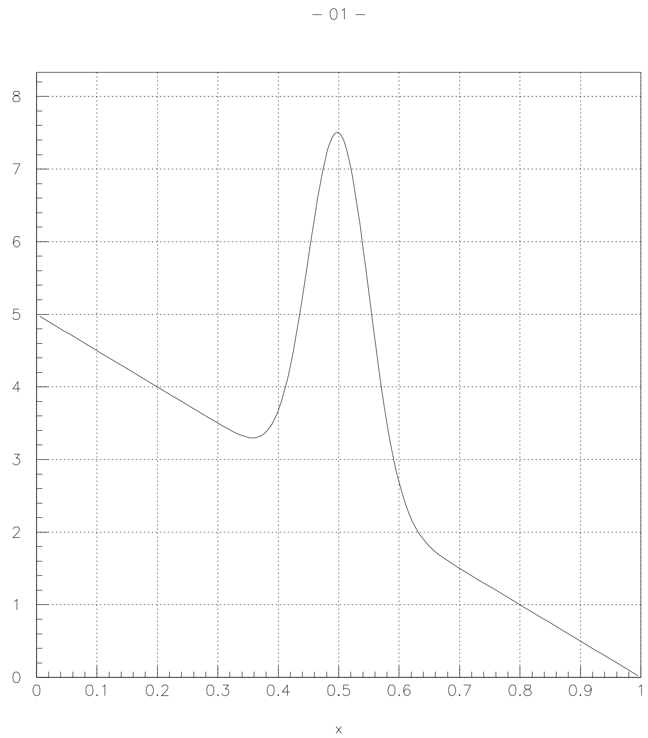


Figure 8.6: Probability density from which we want to extract random numbers.

Mar 04, 02 18:29 **random_neumann.f**

```
      Program vonNeumann
      Common/pawc/h(100000)
*-------------------------------------------------------------
*   The method of von Neumann
*-------------------------------------------------------------

      External Fun
      Call Hlimit(100000)
*
*             Open output file
*
      Open(unit=10,file='rndm1.eps',form='formatted',
     $                              status='unknown')

      Call Hbfun1(100,'x',100,0.,1.,Fun)
      Call HBOOK1(101,'Random secondo 100',100,-0.1,1.1,0.)

      Call HPLint(0)

      Do I = 1,10000
         rr = RNDM(x)
         rs = RNDM(x)*8
         If(fun(rr).GT.rs) call HFILL(101,rr,0.,1.)
      enddo

      call hplcap(-10)
      call HPLopt('GRID',100)
      call HPLopt('NBOX',100)
      call HPLopt('GRID',101)
      call HPLopt('NBOX',101)
      call HPLset('DMOD',1.)

      call Hplot(101,' ',' ',0)

      Call HPLend
      Call Hindex

      END

      Function Fun(x)
      Fun = 5 - 5 * x + 5*Exp(-((x-0.5)**2)/0.005)
      return
      end
```

Figure 8.7: The code to generate random numbers with the Hit-Miss method of von Neumann.

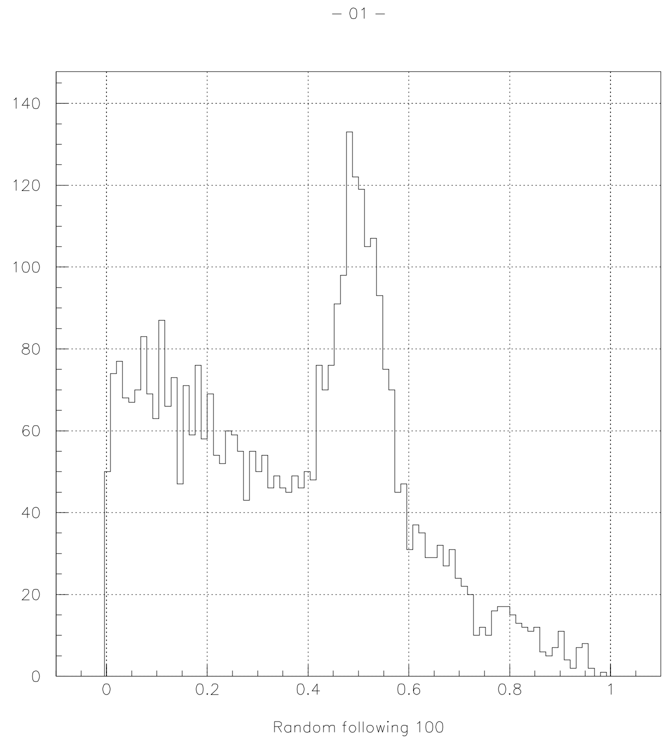


Figure 8.8: Random number generated by the code shown in the previous figure

In fact we can also exploit the Central Limit Theorem: the distribution of the sample mean, when the sample size is large, is a Normal density:

$$I_n \sim N(I, (b-a)\frac{\sigma_f^2}{n})$$

where σ_f^2 is the variance of the function f if x is distrubuted uniformly over the interval (a, b):

$$\sigma_f^2 = \int_a^b (f - E(f)) \cdot U(x)dx$$

To compute the variance of f we would have to perform another integration. We can estimate the variance by the sample itself by:

$$s^2 = \frac{\sum_{i=1,n}(f_i - \overline{f})^2}{n-1} \qquad \overline{f} = \frac{\sum_{i=1,n} f_i}{n}$$

In short we might write:

$$I \approx (b-a) \cdot \frac{\sum_{i=1,n} f(x_i)}{n} \pm \frac{s}{\sqrt{n}}$$

Chapter 9

The Goodness of Fit Test

It happens frequently to compare experimental data with the prediction of a model. Is the model good enough to describe the data? We will not be content of a qualitative answer, rather we want to pose the problem in a quantitative way: we will look for the probability that the model fits the data.

The first method is probably the most general and was designed by K. Pearson [50]. It uses the X^2 or the sum of the squares of the residuals between data and the predictions of a model whose density is known. But this is not the only method. We will describe those most frequently used in High Energy Physics.

9.1 The Pearson Test

Let us consider a single random variable which has discrete values: $x_1, x_2 \cdots x_k$. In case the distribution is continuous we will group data in a finite number K of classes, as shown in figure 9.1:

$$C_1 : x \in (-\infty, x_1) \quad C_2 : x \in (x_1, x_2), \cdots C_K : x \in (x_k, \infty)$$

Let us call H_0 the hypothesis: $X \sim f(x \mid \theta)$ and we will assume that the *p.d.f.* of the random variables is completely determined (simple hypothesis, the parameter θ is known). We can compute the probabilities that an observation falls in the class C_i by integrating the probability density over the sample space of C_i. Let us call these probabilities $p_i = Prob(x \in c_i \mid H_0)$.

We repeat the experiment n times obtaining:

$$\{x_i, i = 1 \cdots n\} \rightarrow \{r_i, i = 1 \cdots K\}$$

Each class has an occupancy: $r_i,\ i = 1, K$. Our problem is: Do the measurements follows the *p.d.f.* $f(x)$?

The following scheme summarizes the problem:

$$\begin{cases} r_i & \textit{the observations} \\ m_i = p_i\ n & \textit{the predictions if } H_0 \textit{ is true} \\ n & \textit{the number of observations} \\ k & \textit{the number of classes} \end{cases}$$

We will compute the quantity:

$$u = \sum_{i=1}^{k} \frac{(r_i - n \cdot p_i)^2}{n \cdot p_i}$$

This is the sum of the squares of reduced variables whose distribution can be approximated by normal density. In fact the probability distribution of r_i is binomial, whatever the *p.d.f.* of x_i is, hence $E(r_i) = n\ p_i$ and $Var(r_i) = n\ p_i\ (1 - p_i) \approx n\ p_i$.

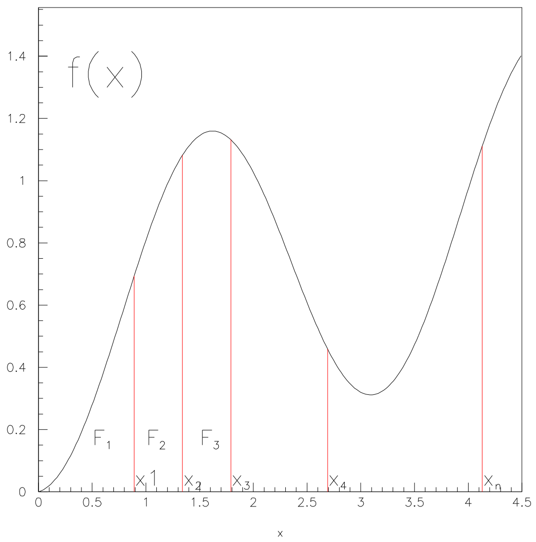


Figure 9.1: *Binning a continuous p.d.f.*

In the limit $n \to \infty$ the Binomial distribution becomes Normal with mean $n\ p_i$ and variance $n\ p_i$.

(Conventionally it is assumed that the Binomial distribution approximates well enough the Normal distribution if $n \cdot p_i > 5$, for all the classes. Cochran has shown that we can relax this requirement. It is assumed that the normality approximation holds if more than 20% of classes have expectations between 1 and 5.)

Assume that the approximation holds, then the variable:

$$z_i = \frac{r_i - n \cdot p_i}{\sqrt{n \cdot p_i}} \sim N(0,1)$$

Hence

$$u = \sum_{i=1,k} z_i^2 = \sum_{i=1,k} \frac{(r_i - m_i)^2}{m_i} \sim \chi^2_{k-1}$$

(The reduction of the number of degrees of freedom by one follows from the constraint $\sum r_i = n$; only $k-1$ variables are independent.)

This result is *independent* of the *p.d.f.* $f(x)$ of the random variable x. The only important condition is the normality in each class, i.e. enough events in each class. Before proceeding, we will play a bit with this quadratic form and understand how this method works.

The Average Value of u

The previous quadratic form can be rewritten as:

$$u = \sum_{i=1}^{k} \frac{r_i^2 - 2n \cdot p_i r_i + n^2 p_i^2}{n \cdot p_i} = \frac{1}{n} \sum_{i=1}^{k} \frac{r_i^2}{p_i} - n$$

Let us compute the expectation value:

$$E(u \mid H_0) = E\left(\frac{1}{n} \sum_{i=1}^{k} \frac{r_i^2}{p_i} - n \right)$$

For that we have to determine the distribution of the random variables r_i. The probability distribution of each r_i individually is binomial. In fact it is the probability that in n independent measurements we have r_i successes, given the probability p_i of a success (a success is the event in which the outcome of the observation falls in the class i) [1]. The previous expectation value reduces to the calculation of $E(r_i^2)$ for a binomial. This is easily done remembering that:

$$\begin{cases} E(r_i) &= n \cdot p_i \\ Var(r_i) &= n \cdot p_i(1-p_i) \\ E(r_i^2) &= Var(r_i) + E^2(r_i) \\ &= np_i(1-p_i) + n^2 p_i^2 \end{cases}$$

In general we do not know if the probabilities p_i given by H_0 are the true probabilities. We will call the true probabilities p_i^0 which will coincide with the probabilities p_i if H_0 is true. The expected value of u can now be computed:

$$\begin{aligned} E(u) &= \frac{1}{n}\sum_{i=1}^{k}\frac{E(r_i^2)}{p_i} - n \\ &= \frac{1}{n}\sum_{i=1}^{k}\frac{np_i^0(1-p_i^0)+n^2(p_i^0)^2}{p_i} - n \\ &= \sum_{i=1}^{k}\frac{p_i^0(1-p_i^0)}{p_i} + n\sum_{i=1}^{k}\left(\frac{(p_i^0)^2}{p_i} - 1\right) \end{aligned}$$

In the equations we have expressed the fact that the expectation value of r_i^2 is determined by the true values p_i^0. If H_0 is true then $p_i^0 = p_i$, $i = 1, k$ hence $E(u \mid H_0) = k - 1$, independently of the number of events. Since we know already that $E(\chi_m^2) = m$, this confirms that the variable u that we know already to be distributed as a χ^2 , has really $k - 1$ degrees of freedom. One degree of freedom is, of course eaten up by the condition that the total number of events is n : $\sum r_i = n$.

The Minimum of u

We will now compute for which set of probabilities p_i the variable u attains the minimum value. We have to minimize

$$u = \frac{1}{n}\sum_{i=1}^{k}\frac{r_i^2}{p_i} - n$$

with respect to p_i with the only condition that $\sum p_i = 1$. This is easily done using Lagrange multipliers method, i.e. by minimizing, with respect to p_i and λ the form:

$$\frac{1}{n}\sum_{i=1}^{k}\frac{r_i^2}{p_i} - n + \lambda(\sum p_i - 1) = Minimum$$

This leads to the equations:

$$\begin{cases} -\frac{1}{n}\frac{r_i^2}{p_i^2} + \lambda &= 0 \\ i = 1 \cdots n & \\ \sum p_i &= 1 \end{cases}$$

[1]Remember that the joint probability of ensemble of all r_i is distributed as a multinomial distribution:

$$P(r_1, r_2, \cdots r_k) = \frac{n!}{r_1! r_2! \cdots r_k!} p_i^{r_1} p_2^{r_2} \cdots p_k^{r_k}$$

This distribution reduces to binomial in the case of only two classes (or, equivalently, if we consider only the class i).

which give

$$\begin{cases} p_i &= \frac{r_i}{\sqrt{\lambda \cdot n}} \\ \sum p_i &= 1 \\ \lambda &= n \end{cases}$$

hence $p_i = r_i/n$. The minimum is attained when the probability is equal to the frequency of the events. In the limits of large number of events, in the frequency theory of statistics, the frequency tends to the probability:

$$p_i = \frac{r_i}{n} \rightarrow \; p_i^0$$

Asymptotically the variable u reaches its minimum when the hypothesis H_0 is true. We have already computed the minimum value:

$$E(u)|_{Min} = k - 1$$

The Value of u if H_0 is not true

It is not easy to compute analytically, in the general case, the expectation value of u if H_0 is not true. There is however a simple case which illustrates how the minimum is obtained. Let us assume that we have built the classes in such a way to have them equally populated (under H_0, all the expected probabilities are equal $p_i = const = 1/k$. If p_i^0 are the true values of the probability in each class, we have:

$$\begin{aligned} E(u) &= \sum_{i=1}^{k} \frac{p_i^0(1-p_i^0)}{p_i} + n \sum_{i=1}^{k} \left(\frac{(p_i^0)^2}{p_i} - 1 \right) \\ &= k \sum_{i=1}^{k} p_i^0(1-p_i^0) + n \sum_{i=1}^{k} \left(k(p_i^0)^2 - 1 \right) \\ &= k - 1 + (n-1) \cdot \left(k \sum (p_i^0)^2 - 1 \right) \end{aligned}$$

It is easy to show that $1/k \leq k \cdot \sum_{i=1}^{k} p_i^2 \leq 1$. In fact:

$$1 = (p_1 + p_2 + \cdots p_k)^2 = \sum_{i=1}^{k} p_i^2 + \sum_{i \neq j} p_i p_j$$

The last term is zero only if all, but one, of the p_i are zero. The minimum value is reached in a particular situation which can be found minimizing, with respect to p_i and λ, the form:

$$f = \sum_{i=0}^{k} p_i^2 + \lambda \left(\sum_{i=1,k} p_i - 1 \right)$$

The minimum value is reached when $p_i = 1/k$. In this case we would have:

$$\sum_{i=1,k} p_i^2 = \frac{1}{k}$$

Hence $E(u) \geq (k-1)$. The equality holds if $p_i = 1/k$.

9.2 How to assign the Probability to u

The previous discussion has shown that the variable u has a probability distribution of χ^2_{k-1} independently of the distribution of the random variable x. The only requirement is to have a sufficiently large number of events in each class, in order to use the Normal approximation to the Binomial.

In the frequency school of Statistics this means that if we repeat the experiment many times (measure many times the ensemble of variables $\{x_i,\ i = 1, n\}$) and the hypothesis H_0 is true, the values of u would be distributed according to the asymmetric, bell shaped, χ^2 distribution.

Many experiments would give a value of u around $k-1$ and only few on the tails. The fraction of experiments having $u > u_M$, u_M being the value that we have actually measured, can be computed as the integral of the χ^2_{k-1} from u_M to ∞. (This is a basic property of Frequency theory, where the frequency of an event is the measure of the probability of that event.)

The larger is u_M, the smaller is the probability. Conventionally we take as the probability for the unique measurement that we have made (u_M), the integral:

$$P(u_M \mid H_0) = \int_{u_M}^{\infty} \chi^2_{k-1}(u)\ du$$

Abnormally low values of u_M are also suspect, in fact the probability of having $u \le u_M$ is:

$$P_L(u_M) = \int_0^{u_M} \chi^2_{k-1}(u)\ du = 1 - P_H(u_M)$$

and this value also has to be reasonable.

The integral above can be computed numerically and is in fact tabulated in the tables appended at the end of these notes. Tab. 9.1 shows the values of u at a fixed probability levels CL as a function of the number of d.o.f.. The values of u_{low} is such that $\alpha = Prob(u \le u_{low} \mid k-1)$. Similarly for u_{high}, $\alpha = Prob(u \ge u_{high} \mid k-1)$.

The situation in which the *p.d.f.* is totally specified is called *simple* hypothesis and occurs seldom in practice. In most of the cases some parameters have to be estimated from the sample. Each parameter that has to be evaluated from the data is a constraint which reduces the number of independent variables and consequently reduces the number of d.o.f. by one .

It happen often that the number of events in each bin (or class) is very small and the normal approximation of the Poisson distribution is not valid. In these cases the X^2 is not distributed as a χ^2_{k-1} but has to be computed numerically with Monte Carlo methods [3]. In this case the distribution depends not only on the number of classes but also on the model; the test is not any longer *distribution free.*

9.2.1 Tossing a Coin

We throw 4040 times a coin landing heads 2048 times (h) and tails (t) 1992 times. We want to test the hypothesis that the probability is 0.5. In this case we have only two possible outcomes and we can use a binomial distribution with $p = 0.5$. Numerically:

$$U = \frac{(h - p\cdot n)^2}{p\cdot n} + \frac{(t-(1-p)\cdot n)^2}{n\cdot(1-p)} = \frac{(h-p\cdot n)^2}{n\cdot p\cdot(1-p)} = 0.776$$

The number of d.o.f. is 1 ($k = 2$ and the hypothesis is totally specified). We compute easily the probability $P(\chi^2_1(u) \ge 0.776) = 0.38$. The hypothesis is accepted with a confidence level of 38%.

$\alpha=$	0.05		0.01		0.005	
N_{dof}	u_{low}	u_{high}	u_{low}	u_{high}	u_{low}	u_{high}
1	.393E-02	3.84	.157E-03	6.63	.393E-04	7.88
2	.103	5.99	.201E-01	9.21	.100E-01	10.6
3	.352	7.82	.115	11.3	.716E-01	12.8
4	.711	9.49	.297	13.3	.207	14.9
5	1.15	11.1	.554	15.1	.412	16.7
6	1.64	12.6	.872	16.8	.676	18.5
7	2.17	14.1	1.24	18.5	.989	20.3
8	2.73	15.5	1.65	20.1	1.34	22.0
9	3.33	16.9	2.09	21.7	1.73	23.6
10	3.94	18.3	2.56	23.2	2.16	25.2
15	7.26	25.0	5.23	30.6	4.60	32.8
20	10.9	31.4	8.26	37.6	7.43	40.0
25	14.6	37.7	11.5	44.3	10.5	46.9
30	18.5	43.8	15.0	50.9	13.8	53.7
40	26.5	55.8	22.2	63.7	20.7	66.8
50	34.8	67.5	29.7	76.2	28.0	79.5
75	56.1	96.2	49.5	106.	47.2	110.
100	77.9	124.	70.1	136.	67.3	140.
150	123.	180.	113.	193.	109.	198.
200	168.	234.	156.	249.	152.	255.

Table 9.1: *Values of u for probability content of CL in the low and high tails of a χ^2_n p.d.f. as function of the number of degrees of freedom.*

9.2.2 Mendel's Peas

In a famous experiment, Mendel counted the number of times he got green and yellow peas. He expected $p = 0.75$. The data collected in many experiments are summarized in table 9.2.

Experiment	N.O tests	yellow	u	$P(u)$
1	36	25	.593	.44
2	39	32	1.034	.31
3	19	14	0.018	.89
4	97	70	.416	.52
5	37	24	2.027	.15
6	26	20	0.051	.82
7	45	32	0.363	.55
8	53	44	1.818	.18
9	64	50	0.333	.56
10	62	44	0.538	.46
Total	478	355	0.137	.71

Table 9.2: *Mendel experiment on yellow and green peas.*

In the last raw we have summed the results of all the experiments. What conclusion can we draw from this analysis?

- The probabilities of each experiment is large. If we take an acceptance criterion $P(u) \geq 0.05$, all experiments would be accepted.
- The same is true for the sum of the results of all experiments (last raw).
- If we take the sum all u_i we get the value of $U = 7.191$. Since the measurements are independent U should follow a χ^2_{10} distribution. The probability of U is:

$$P(U) = \int_{7.191}^{\infty} \chi^2_{10dof}(u)\ du \ = \ 0.71$$

- The test of the distribution of u_i is possible since we have many experiments. u_i should follow a χ^2_1 distribution To make the check we group the 10 experiments in 3 classes, as shown in Table 9.3. The binning of the ten experiments in three classes is suggested by

Limits	Nexp	p_i
$u < 0.148$	2	0.3
$0.148 < u < 1.074$	6	0.4
$1.074 < u$	2	0.4

Table 9.3: *The distribution of the values of u in Mendel's peas experiment.*

the poor statistics.

We obtain:

$$u = \frac{(2-3)^2}{3} + \frac{(6-4)^2}{4} + \frac{(2-4)^2}{4} = 1.67$$

The 3 classes have now one constraint (the sum of the experiments is 10) and the probability is:

$$P(u = 1.67) = \int_{1.67}^{\infty} \chi^2_2(u)\ du = .43$$

The hypothesis can be safely accepted.

9.2.3 Rutherford's α

This is an example from the origin of nuclear physics. Rutherford (Phil. Mag. 1912) was investigating whether the decay process of a radioactive atom is independent from the status of all the other atoms. If such is the case, the number of events in a time interval follows the Poisson statistics. However the mean is not known and has to be determined from the data itself. The hypothesis is not *simple*. The experiment consisted in measuring the number of alpha particles emitted from a radioactive source in an interval 7.5 *sec* long. The number of intervals that were measured is $n = 2608$. The statistical problem is to prove that the population is Poisson distributed with (unknown) average μ (see Tab. 9.4).

In the table we report the number of times (N_i, the second column) in which N (the first column) α particles were detected in the time intervals. The third column is computed from the Poisson statistics and is the product of the probability of having N events from an average rate μ and the total number of events $n = 2608$. This is the expected number of events in each bin N. The fourth column is the value of u computed from the second and the third column. The last three classes have been summed to improve the statistics.

The data analysis is rather simple. The average rate of the source (μ) had to be measured from these observations.

N	N_i	$n \cdot p_i$	u
0	57	54.399	0.124
1	203	210.523	0.2688
2	383	407.371	1.4568
3	525	525.496	0.0005
4	532	508.418	1.0938
5	408	393.515	0.5332
6	408	393.515	0.5332
7	139	140.325	0.0125
8	45	67.882	7.7132
9	27	29.189	0.1642
10 11 12	10 4 2	17.075	0.0677
Total	2608	2608	12.8849

Table 9.4: *Number of α particles emitted by a source in 2608 time intervals of* 7.5 *secs with unknown but constant rate.*

The arithmetic average from columns 1 and 2

$$\overline{x} = \frac{\sum_i i \cdot N_i}{n}$$

is the best estimate of μ and yields $\overline{x} = 3.870$.

The number of classes is 11 and we have two constraints (the total number of events and the average), hence the number of d.o.f. is 9. The sum of individual u is 12.9 and this corresponds to a probability of

$$P(u = 12.885) = \int_{12.885}^{\infty} \chi^2_9(u)\ du = 0.17$$

which is quite acceptable. Hence the data are well represented by a model in which the individual radioactive decays happen independently from each other.

9.2.4 Weldon's Dices

As another example of goodness of fit we will consider an experiment done by Weldon who throw 26306 times a set of 12 dice and accounted for success the case where a die landed 5 or 6. The number of dice with face 5 or 6 is shown in Tab.9.5, first column.

A perfect dice would have a probability $p = 1/3$ and the frequency would be given by the binomial.

In the column headed u_1 we have computed, under the assumption $p = 1/3$, the quantity: $u_1 = \frac{(obs-exp)^2}{obs}$ We see that $u = 31.9$ which, for 10 d.o.f gives:

$$P(u = 31.9) = 4 \cdot 10^{-4}$$

the hypothesis $p = 1/3$ is unlikely.

We can compute the probability p from the sample: the average number of successes is $\overline{n} = 4.0522$ which gives

$$p = \frac{\overline{n}}{12} = 0.3377$$

N	N_i	$H_0(p = 1/3)$	u_1	$H_1(p = 0.3377)$	u_2
0	185	202	1.56	187	0.022
1	1148	1216	4.03	1146.7	.0
2	3265	3345	1.96	3215.6	0.74
3	5475	5576	1.86	5465	0.02
4	6114	6273	4.13	6269	3.92
5	5194	5018	3.30	5115	1.20
6	3067	2927	6.40	3043	0.19
7	1331	1255	4.34	1330	0.
8	403	392	0.30	424	1.09
9	105	87	3.09	96	0.77
> 9	18	14	0.90	16	0.22
Total	26306	26306	31.9	26306	8.19
Probability			4.210-4		0.515

Table 9.5: *Number of dices landed 5 or 6 in 26306 throw of 12.*

very near to the theoretical value $p = 1/3$. The last column shows the results for this test (u_2). We have computed one parameter from the data, therefore the d.o.f. is now 9.

In conclusion the goodness of fit gives now a quite acceptable probability; the data are well represented by a Bernoulli model where, however the probability is not what would have been expected by a perfect die ($p = 1/3$) but has a slightly larger value.

Exercise 9.1 *Take a random number in the range 1-999 and compute the frequency of occurrence of the last digit ($P_i, i = 0, 1, \cdots 9$). Make a small Monte Carlo generating 100 numbers in the range* $(1, 999)$ *and record the frequency (r_i) for the occurrence of the last digit. The probability for the occurrence of digit i should be* $1/10$ *independent of the digit, if the random number generator works well. Then compute the quantity $X_i^2 = (r_i - 0.1)^2/0.01$ for the 100 random digits and verify that the distribution of X_i^2 is χ_1^2. Compute the quantity $X^2 = \sum_{i=0,9} X_i^2$ and plot it in a series of 1000 similar experiments. Verify that X^2 follows a χ_{10}^2. (Why 10 and not 9?) (Figs.9.2).*

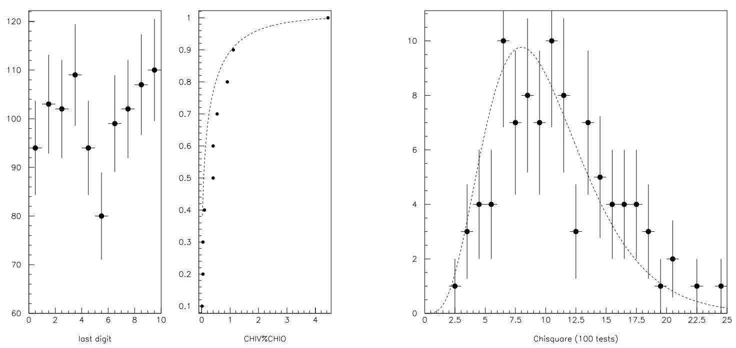


Figure 9.2: *Left: in one experiment (100 random digits) we record the distribution of the last digit. In center is shown the cumulative distribution in one experiment. At right we plot the distribution of the variable X^2 in many experiments(see text)with, overlaid, its theoretical p.d.f..*

9.3 Smirnov-Cramer-Von Mieses ω^2 Test

The X^2 test is not the only possible way to test a hypothesis, many other exists. In this section we discuss the ω^2 test.

Assume we have a sample of data $(x_1, \cdots i = 1, n)$ from a population. We want to test whether the population has *p.d.f.* $f(x)$.

Example 9.1 *We have made n measurements of a variable, like the polar angle of a decay product of a polarized resonance and we would like to know if the measurement is in agreement with a specific prediction, like the distribution expected from a spin 1 resonance.*

Let us assume that we have ordered the sample x_i in increasing order. We will built the *empirical distribution* $F_n(x)$:

$$\begin{cases} F_n(x) &= 0 \quad (x < x_1) \\ F_n(x) &= \frac{i}{n} \quad (x_i < x < x_{i+1}) \\ F_n(x) &= 1 \quad (x > x_n) \end{cases}$$

where i is the number of x_i's less than x. $F_n(x)$ is a function constant between successive x_i and jumps up by $1/n$ at each x_i.

The ω^2 statistics is the expectation value:

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF = \frac{1}{12 \cdot n^2} + \frac{1}{n} \cdot \sum_{i=1,n} (F(x_i) - \frac{2 \cdot i - 1}{n})^2$$

$F(x)$ is the cumulative function of $f(x)$. The terms $S(x) = \frac{m_i}{n}$ are the measured frequencies of a binomial process with probability $p = F(x)$:

$$\begin{aligned} E(S(x)) &= F(x) \\ Var(S(x)) &= \frac{F(x) \cdot (1 - F(x))}{n} \end{aligned}$$

It is now possible to compute the mean and the variance of ω^2

$$E(\omega^2) = \frac{1}{n} \cdot \int_0^1 F(1 - F) \, dF = \frac{1}{6n}$$

and

$$Var(\omega^2) = E(\omega^4) - E(\omega^2)^2 = \frac{4n - 3}{180n^3}$$

The *p.d.f.* of ω^2 is independent of the distribution of x. Its characteristic function is:

$$\lim_{n \to \infty} E(e^{i \cdot tn\omega^2}) = \sqrt{\frac{\sqrt{it}}{sin\sqrt{it}}}$$

The critical values have been computed (Smirnov, 1936 and Andersen, 1952) Tab. 9.6:

9.4 Kolmogorov-Smirnov Test

This test is based on a different approach. Assume that we have measured n values of the quantity X: $x_i, i = 1, 2, \cdots n$, which are then ordered in increasing order. We want to test the hypothesis that the distribution of X is $F(x) = \int_{-\infty}^{x} f(t) \, dt$. We will use the empirical

Probability	$n\omega^2$
0.1	0.347
.05	.461
.01	.743
.001	1.168

Table 9.6: *Probability for some values of* $n \cdot \omega^2$.

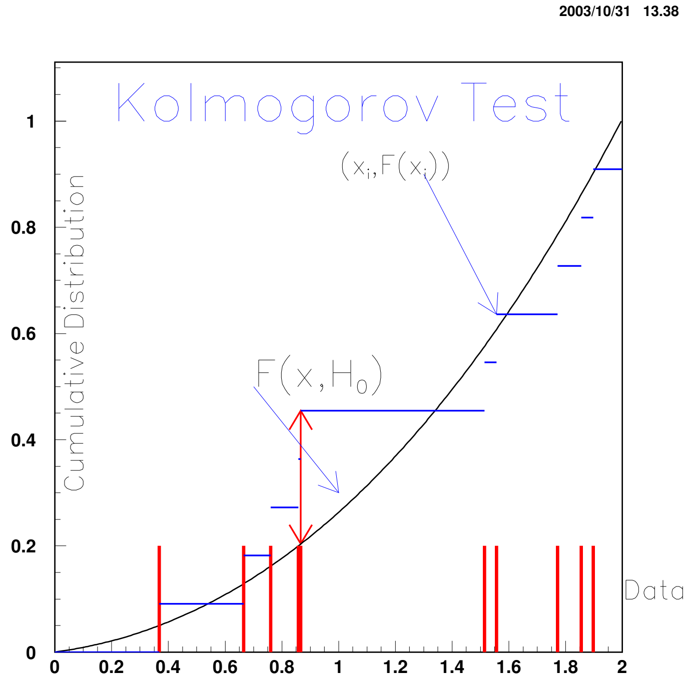


Figure 9.3: The empirical distribution function and the hypothetical distribution.

distribution ($F_n(x)$) defined in the previous section. The plot of F_n it together with $F(x)$ is shown in figure 9.3. The quantity $nF_n(x_i)$ is distributed as a Binomial:

$$nF_n(x_i) \sim B(F(x_i), n)$$

(nF_i is the number of events below x_i. If H_0 is true, the probability for this Bernoulli trial is $F(x_i)$). In formulas we expect:

$$\begin{cases} E(F_n(x_i)) & = & F(x_i) \\ Var(F_n(x_i)) & = & \frac{F(x_i)(1-F(x_i))}{n} \end{cases}$$

From the Chebicheff inequality we have:

$$P\left(|F_n x_i - F| \leq t\sqrt{\frac{F(1-F)}{n}}\right) \geq 1 - \frac{1}{t^2}$$

hence, for $n \to \infty$ we expect $F_n \to F$, in words the empirical distribution tends to the true distribution. It seem reasonable to use as a statistical test a measure of the "*distance*" between the two. Kolmogorov used, as test the quantity:

$$d(F_n, F) = Max(F_n(x_i) - F(x_i)) \quad i = 1, 2, \cdots n$$

Since $F_n(x)$ is constant between successive x_i, The distance $d(F_n, F)$ can be computed only at the x_i.

We shall not go into the study of the distribution of d, we shall only show the following property:

Theorem 9.1 *The distribution of $d(F_n, F)$ is independent of F if F is the true distribution of x_i.*

Proof 9.1 *This can be easily seen in this way: from the measurements $x_1, \cdots x_n$ let us construct the ordered sample: $y_1, \cdots y_n$, defined as: $y_i = F(x_i) = \int_{-\infty}^{x_i} f(t)\, dt$. The variabley $= \int_{-\infty}^{x} f(t)\, dt$ is distributed uniformly in the interval $[0, 1]$ [2]. Let us now construct the empirical distribution of y_i (U_n) and let us call U the uniform distribution $U(t) = t$. We have to prove that $d(F_n, F) = d(U_n, U)$. Since $y = F(x)$ then $y_i \leq y$ if and only if $x_i \leq x$ hence $U_n(y_i) = F_n(x_i)$ and since $U(y) = y = F(x)$, we have:*

$$U_n(y_i) - U(y) = F_n(x_i) - F(x)$$

The distances are the same and the same their distributions.

Hence we can compute the distribution of $d(F_n, F)$ in the particular case of the uniform distribution only, the result will have a general validity.

Kolmogorov was able to find the asymptotic distribution for d and also recurrence relations to compute the distribution also for finite values of n. Figure 9.4 shows the value of the probability $P(\sqrt{n} \cdot d(F_n, F) \geq x)$, as a function of x. The same plot can be used to compare two sets of data of size n and m, and test, the hypothesis that the two samples have the same distribution. In this case the statistics is $x = d(F_n, F) \cdot \sqrt{\frac{m \cdot n}{m+n}}$.

One advantage of Kolmogorov-Smirnov test is that we do not have to discretize the data and loose so any eventual fine structure. A drawback is that the test is valid only for simple hypothesis: no parameter can be computed from the data to determine $F(x)$. It is not easy to modify the Kolmogoroff-Smirnow test to handle composite hypothesis.

9.5 The Run Test

Assume we have two series of observations each arranged in increasing order:

$$(y_i, i = 1, \cdots n) \qquad\qquad (x_i, i = 1, \cdots m) \qquad (n \leq m)$$

Let combine the $(n + m)$ observations and arrange them in increasing order, obtaining a new series, for instance:

$$x_1, x_2, y_1, x_3, y_2, y_3, y_4, x_4, \cdots$$

If the two samples originate from the same population, the combined sample would be "well mixed". We call *run* a sequence of symbols of the same type. In the example above we see first a run of two x, then a run of one y, followed by a run of three y etc.. We can use the number of

[2] In fact $t = F(u)$ is the probability that x is measured less than u ($P(X \leq u)$). But the probability that y is less then t is: $P(y \leq t) = P(F(x) \leq t) = P(x \leq u) = F(u) = t$. Hence P(y$\leq t) = t$ or the distribution of y is uniform. The second equality follows from the monotonic behavior of $F(x)$ as a function of x.

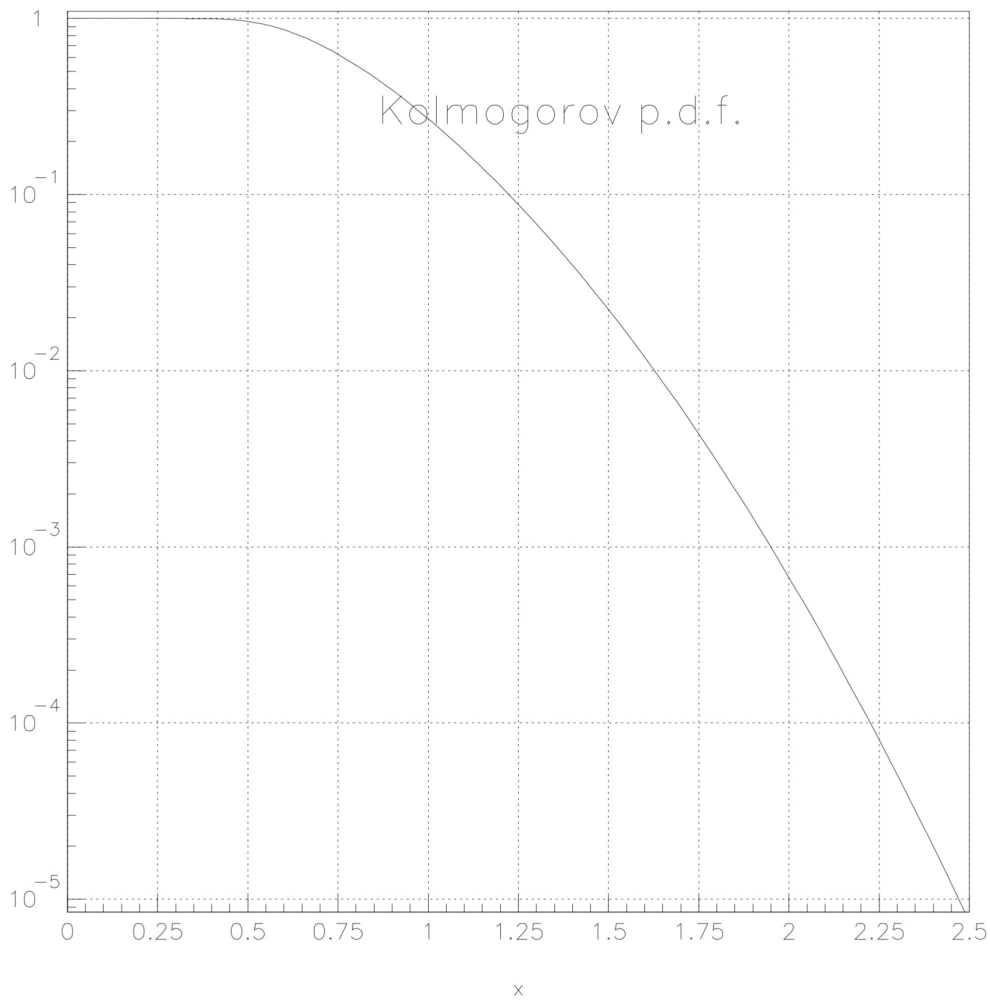


Figure 9.4: Plot of the probability $P(\sqrt{n} \cdot d(F_n, F) \geq x)$.

runs r as a test statistics for the hypothesis H_0 (the two samples are extracted from the same population) to be true.

The number of ways in which $(n + m)$ objects can be arranged is $(n + m)!$. However the order within the x's and the y's is fixed which makes the total number of possible permutations $(n+m)!/(n!m!)$. Each of these permutation has the same probability of occurring, hence to find the probability of r runs we have to count all the permutations that give rise to exactly r runs.

The combinatorial problem can be carried out in a straightforward way (Wilks 1962) yielding:

$$\begin{aligned} P(r=2k) &= 2 \cdot \frac{\binom{n-1}{k-1} \cdot \binom{m-1}{k-1}}{\binom{n+m}{n}} \\ P(r=2k-1) &= \frac{\binom{n-1}{k-2}\binom{m-1}{k-1} + \binom{m-1}{k-2}\binom{n-1}{k-1}}{\binom{n+m}{n}} \end{aligned}$$

It can be shown that the distribution has mean and variance:

$$\begin{aligned} E(r) &= \frac{2\ n\ m}{n+m} + 1 \\ Var(r) &= \frac{2\ n\ m\ (2\ n\ m - n - m)}{(n+m)^2 \cdot (n+m-1)} \end{aligned}$$

For large samples ($m \cdot n > 10$) the probability distribution is very close to normal and the run test statistics is:

$$z = \frac{r - E(r)}{\sqrt{Var(r)}} \sim N(0,1)$$

Example 9.2 *Two experiments have measured the effective mass of gamma pairs (π^0 or background). We want to know if the two spectra are consistent. The samples are small and all the information has to be used. Data are shown in the Fig 9.5.*

The two samples have size 28 and 32 and the number of runs computed after mixing the two samples is 24.

$$E(r) = \frac{2 \cdot 28 \cdot 32}{28+32} + 1 = 30.9 \qquad\qquad V(r) = 14.62$$

The value obtained from the data is $z_{obs} = -1.80$. If we use the large sample approximation we can read from the probability tables of the normal distribution:

$$P(z < -1.80) = 0.035$$

A two sided test gives instead:

$$P(|z| > 1.80) = 0.070$$

This test complements the Pearson X^2 test. In fact the Pearson test does not make distinction among positive and negative fluctuations and would give the same probability even for ordered data. It does not see the eventual *trend* of the data. On the contrary the run test can be used to test the pattern of signs and can be a supplement to the X^2 test.

9.6 Test for Normal distribution

Many important distributions (χ^2, t, F) rely on the assumption that the distribution of the parent population is Normal. It is therefore important to test this assumption.

The simplest test is the one based on the skewness [13] (however one of the most popular end efficient is due to Shapiro, Wilks and Chen [43]).

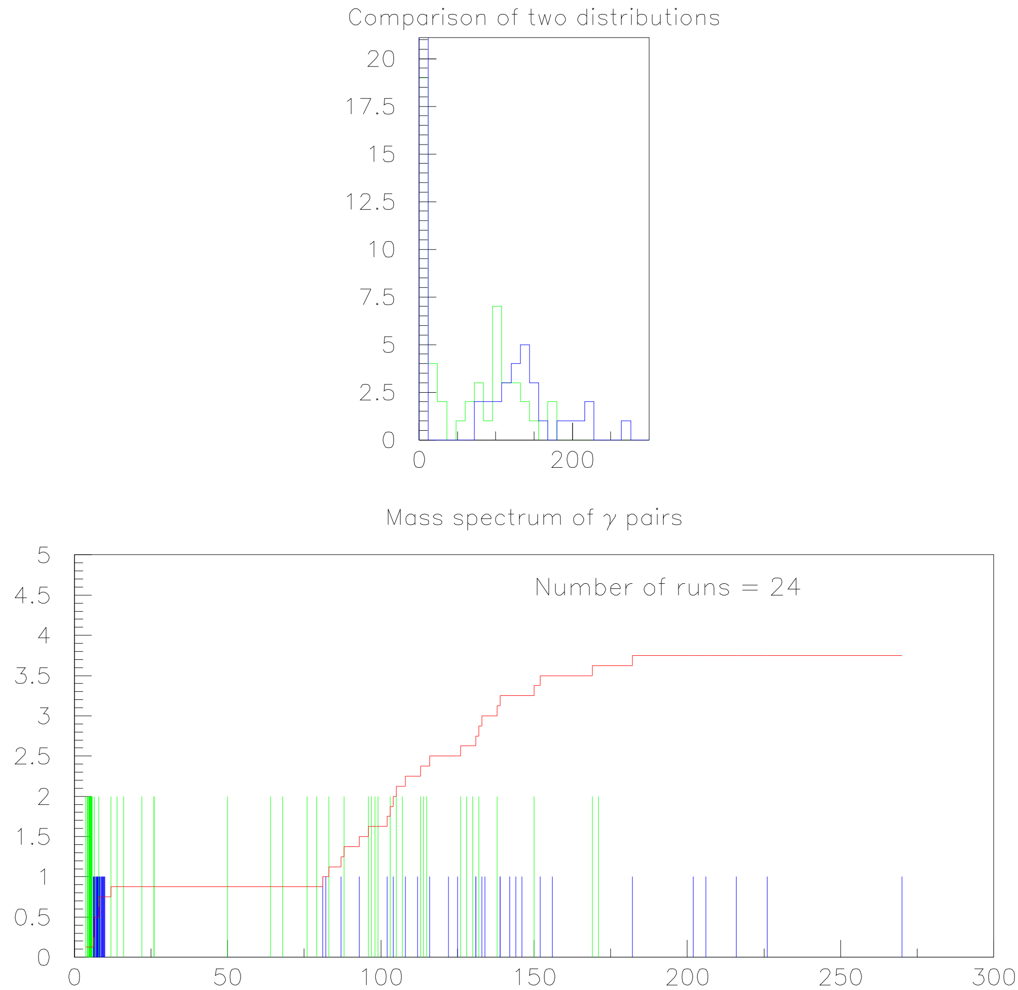


Figure 9.5: *Two experiments measure events containing π^0. The mass spectrum is plot in left high figure. Next to it is the plot of cumulative distributions for the two sets of data. Below the data are shown unbinned with the run counting.*

In a sample of size n: $(x_i, i = 1, n)$, the skewness is defined as [49]:

$$sk = \frac{n}{(n-1)(n-2)} \sum_i \frac{(x_i - \overline{x})^3}{s^3}$$

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \overline{x})^2$$

If $X \sim N(\mu, \sigma^2)$ then it is possible to prove [13] that:

$$\begin{aligned} sk &\sim N(\mu, \sigma^2) \\ \mu &= 0 \\ \sigma^2 &= \frac{6\ n\ (n-1)}{(n-2)(n+1)(n+3)} \end{aligned}$$

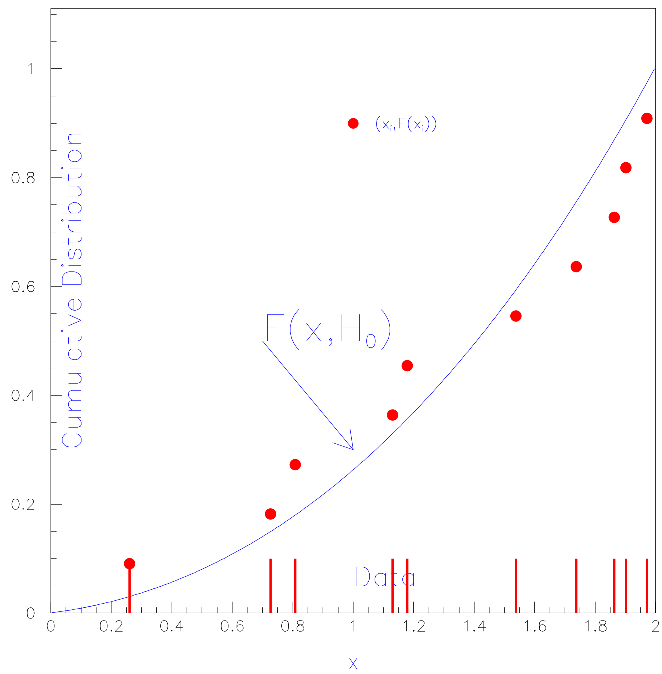


Figure 9.6: Construction of the distribution function.

From the value of the skewness computed in the sample, it is then possible to test the normality of the sample.

9.7 The empirical Distribution

The method that we will discuss makes use of the "empirical" densities. Let us assume that the sample is ordered: $x_1 \leq x_2 \cdots \leq x_n$ (if it is not, then we will order it).

The n observations define $n+1$ intervals:

$$(-\infty, x_1), (x_1, x_2) \cdots (x_n, \infty)$$

Conventionally we say that $\frac{1}{n+1}$ of the sample is contained in each interval:
$\frac{1}{n+1}$ of the sample is $\leq x_1$,
$\frac{2}{n+1}$ of the sample is $\leq x_2$, etc.

Recalling the definition of distribution: $F(x) = \int_{-\infty}^{x} f(t)dt$ is the probability to find one observation in the interval $(-\infty, x)$ we can say that:
$\frac{1}{n+1}$ of the sample estimates $F(x_1)$,
$\frac{2}{n+1}$ of the sample estimates $F(x_2)$, etc.

The set of points:

$$\left\{ x_i, \ \frac{i}{n+1} \right\} \ i = 1 \cdots n$$

is our best estimate of the curve $F(x)$. Figure 9.6 shows an example. The short lines at the bottom are the x values measured in the experiment, the points is the estimation of $F(x_i)$ as described above. Overlaid is the distribution function from which the data have been extracted.

This simple test is already good and useful, even if the fluctuations in the data sometimes mask the similarity between the empirical distribution and the real distribution function.

9.8 Normal Plots

This test that we will discuss is only qualitative, but is also very intuitive and simple. Assume we have a random sample $(x_i, i = 1, \cdots n)$ arranged in increasing order $(x_1 \leq x_2 \leq \cdots \leq x_n)$. If the *p.d.f.* of the parent population is normal, the distribution of the cumulative distribution is:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, dt$$

This means that the probability to have one event with a value $x \leq a$ is $F(a)$. From our sample $1/(n+1)$ estimates $F(x_1)$ (being the fraction of points less or equal to x_1). In the same way $2/(n+1)$ of the sample is less or equal to than x_2 and estimates $F(x_2)$ etc.

In general it can be proved that for all random variables having continuous distribution:

$$E\left(F(x_i)\right) = \frac{i}{n+1}$$

This type of properties which is independent of the shape of the distribution functions is called in statistics *distribution free* or *non-parametric* [3].

Coming back to our problem, it is difficult, in case the number of measurements is small, to compare their distribution to the Gaussian curve and also the cumulative distribution to the sigmoid curve. It is useful to plot the points:

$$(x(F), F) = \left(x_i, \frac{i}{n+1}\right), \qquad i = 1, \cdots n$$

(Remember that $F = i/(n+1)$ is fixed for each point while x_i is an estimate of $x(F)$.) The practical problem of comparing the observations to the sigmoid curve is made easier by linearizing the plots. We change the vertical scale by replacing the variable $(F = i/(i+1))$ to a new variable y such that:

$$F = \int_{-\infty}^{y(F)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, dt \qquad y = F^{-1}\left(F_n(x_i)\right)$$

The plot of the points $(x_i(F), y(F))i = 1, \cdots n$ is called probability plot. If x has a normal distribution $N(\mu, \sigma^2)$, then the points line up along the line:

$$y = (x - \mu)/\sigma$$

An example of probability plot for a normal distributed sample of size 10 is given in the Fig. 9.7 A fit drawn by eye will find easily the mean and sigma (6 and 2).

Example 9.3 *The following data represent the measurement of the attenuation length (in cm) of a set of optical fibers used in ATLAS experiment.*

$$\begin{aligned} \lambda_i \quad = \quad & 303.83,\ 281.30,\ 259.57,\ 296.60,\ 267.29,\ 286.44, \\ & 280.01,\ 204.47,\ 326.39,\ 282.80,\ 207.68,\ 307.14, \\ & 262.93,\ 314.80,\ 295.50,\ 305.46,\ 316.37,\ 225.95, \\ & 324.41,\ 334.49 \end{aligned}$$

[3] *Non-parametric* in statistics means than no assumption is needed concerning the *p.d.f.* of the underlying population. The tests that we have described are all non parametric, with the exception of the Pearson test for few events. We have been able to compute the distribution of the statistics without making use of the distribution of the population.

Parametric methods concern tests or inference on a specific distribution (or class of distributions). In this case the tests concern a specific value of a parameter.

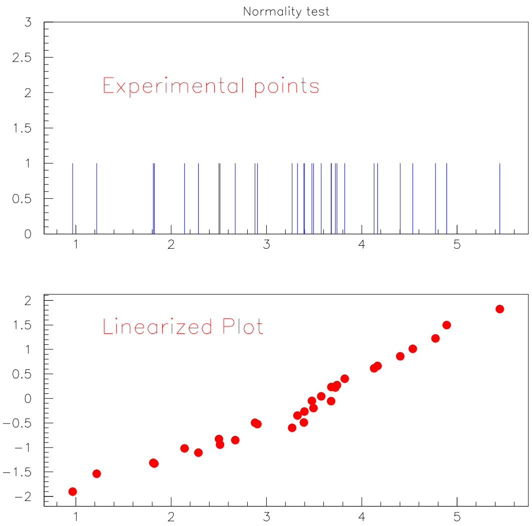


Figure 9.7: *Probability plot. Top plot is the raw measurements. Bottom is the probability plot. An eyeball fit easily finds the mean and standard deviation.*

The average of the distribution is: $\overline{x} = 284.17$ *and the sample standard deviation is:* $s = 37.36$. *In doing the measurement we are interested to verify that the production of the fibers is homogeneous and compatible with a Gaussian shape. Any departure from Gaussian shape is a signal of potential mistakes in the production process of the fiber. In particular the lowest attenuation length (204.5) is more than two standard deviation from the average, a rather rare event.*

We compute the skewness: $s_k = -0.917$. *If the parent population is normal this should be distributed as* $N(0, \sigma = 0.51)$. *The* $P - value$ *for our value of the skewness is:*

$$P - value \;=\; 2(1 - \int_{s_k}^{\infty} N(0,\sigma)\; dx) = 0.073$$

We conclude that, at the level of 5% *the parent population is normal.*

As final check, we show the normal probability plot (Fig. 9.8). We see that, apart the two lowest points, the rest of measurements lines up nicely. The two lowest points are outliers. Outliers are atypical and rare observations. These data points do not appear to follow the characteristic distributions of the expected p.d.f.. Outliers could be produced by measurement errors or other anomalies in the data. These points, since they lie at the edge of the probability distribution, are able to modify considerably any inference from the data. Probability plot are an easy way to identify these anomalous points.

What to do in our case? The probability that the data follow a normal distribution is 7%, *not small, but also not large. Two points are a bit anomalous. Probably the most sensible thing to do is to check again the measurements. If the anomaly persists make more measurements and check if the problem is still there. At this point we will have to judge if we can claim that we have detected an evidence of mistakes in the fiber production.*

9.9 Combining different Tests

It can be shown that in the case of simple hypothesis H_0 (no parameter estimated from the data) the run test and the Pearson test are asymptotically independent, hence the two tests can

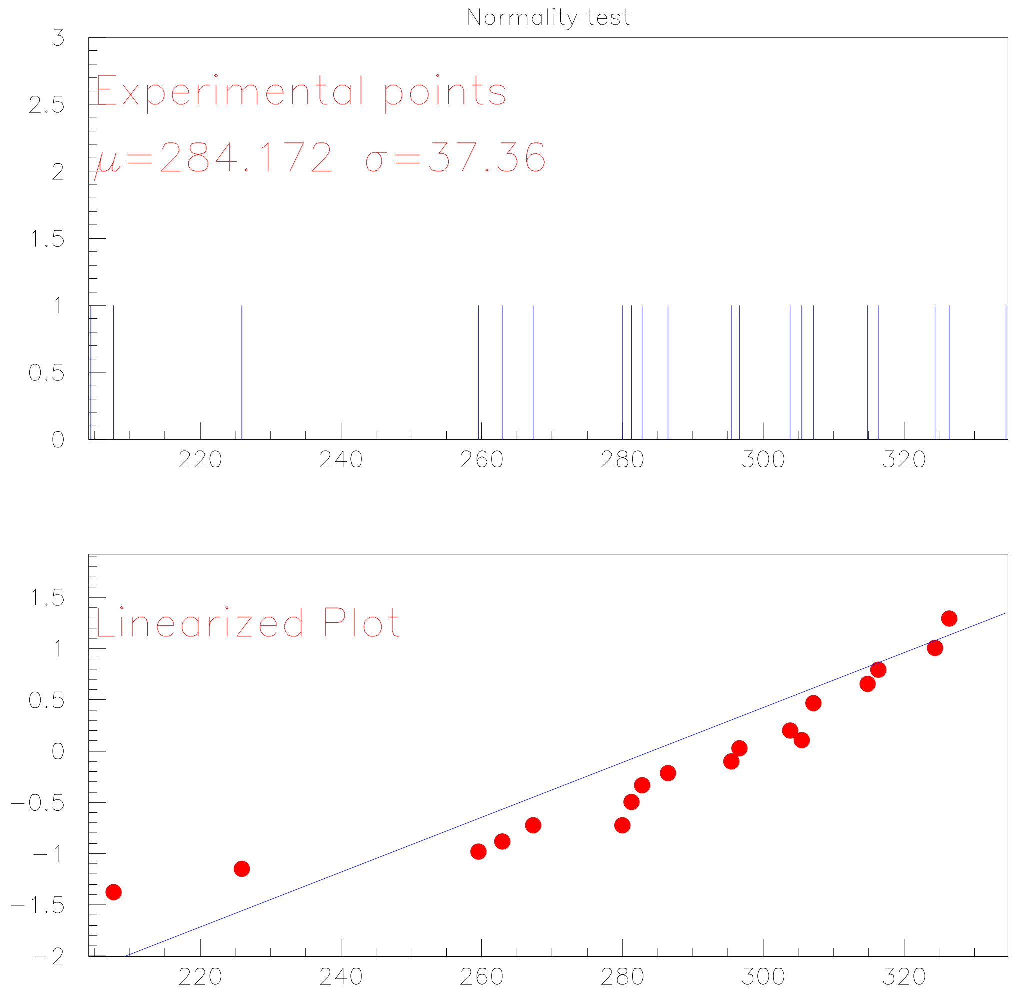


Figure 9.8: *Normal probability plot for ATLAS fibers.*

be combined. The two independent tests have yield the significance level:

$$p_1 = \alpha_1 \qquad\qquad p_2 = \alpha_2$$

In the hypothesis H_0 the probability of $p_{1,2}$ are distributed uniformly:

$$Prob\,(p_{1,2} \leq \alpha_{1,2}) = \alpha_{1,2}$$

The argument that, since the probability of each test is α_1 and α_2, then the probability of the combined test is $\alpha_1 \cdot \alpha_2$ is true only for independent single events. In this case, the repetition of the experiment would have yield different values of p_1, p_2 which populate uniformly the plane (p_1, p_2). In a number of cases the joint probability $p_1 p_2$ would be less then $\alpha_1\alpha_2$. From this it would be clear that we want to compute the probability:

$$Prob\,(p_1 p_2 \leq \alpha_1\alpha_2)$$

Since the space (p_1, p_2) is uniformly populated the probability is obtained is obtained by integrating under the curve:

$$p_1 = \frac{\alpha_1 \cdot \alpha_2}{p_2} \qquad\qquad p_1, p_2 \leq 1.$$

yielding:

$$\alpha = \int_\Gamma \frac{dp_2}{p_1} = \alpha_1 \cdot \alpha_2 \cdot (1 - ln(\alpha_1 \cdot \alpha_2))$$

which is larger than $\alpha_1 \cdot \alpha_2$.

9.10 P-value

There is a way to quantify how well a hypothesis is compatible with measurements. This is done by quoting the *P-value*. The P-value is the probability to obtain a result as compatible or less with the actual observation, if the hypothesis is true.

Example 9.4 *Let us consider the experiment that we have analyzed above consisting of tossing a coin. The actual experiment was done tossing 4040 times a coin and observing 2048 times heads. The estimated probability of getting heads is a binomial random variable whose vales and variance are found to be:*

$$\begin{aligned} p_{exp} &= \frac{2048}{4040} = 0.50693 \\ \sigma_p^2 &= \frac{p(1-p)}{4040} = 6.2 \cdot 10^{-5} \end{aligned}$$

We can use the normal approximation to the binomial distribution:

$$p \sim N(0.5, 6.2 \cdot 10^{-5})$$

and the P-value is:

$$P - value = 2 \cdot \int_{p_{exp}}^{\infty} N\, dp = 0.378$$

This is the fraction of times we would obtain a result equal or "worse" than the one we have obtained, by repeating many times the experiment in similar conditions (toss 4040 times a coin equal to the one we have used).

Chapter 10

Hypothesis Testing

10.1 Introduction

In the previous sections we have considered the problem of measuring the probability that data follows from a given probability distribution. The probability distribution could be completely specified (simple hypothesis) or depend on parameters which are to be determined from data (composite hypothesis). In the latter case the problem is related to the one of parameter estimation. We have to test data against two hypothesis: H_0 and H_1. The case H_1 could be $NOT\ H_0$ is a special case.

We have built a function of the data, a *test statistics* (as the variable u in the previous section), compute its distribution under the hypothesis that H_0 is true and define a *critical region* w and an *acceptance region* W. If the measurement of the test statistics falls in w we will assume that the hypothesis is NOT true.

In the previous examples of Pearson test, large values of u would be *significant* since they would correspond to small probabilities ($P(u|H_0)$). An accumulation of chance effects would also produce large values of u, but it is not likely.

H_1 is an alternative hypothesis: x is distributed as $g(x) \neq f(x)$. If H_1 would be true, then $P(u|H_0)$ would have a different shape (H_1 in Fig. 10.1). A large value of u would then be likely if H_1 would be true.

We have to adjust the critical region w in order to obtain a desired *level of significance* that is the probability that H_0 is rejected even if it is true:

$$P(X \in w \mid H_0) = \alpha$$

The test is useful if it is capable of rejecting H_1 with high probability, what is called *power of the test* $(1 - \beta)$:

$$P(X \in w \mid H_1) = 1 - \beta$$

Since the test statistics is a random variable an erroneous conclusion is always possible. We distinguish two types of errors:

- Type 1 error when H_0 is told false while it is true (probability α).
- Type 2 error when H_0 is told true while H_1 is true (probability β).

Let us consider the example of Mendel experiment. Mendel was obviously concerned with the result of the test that could confirm his theory. The result of the test is open to errors: the test could be favorable but in reality the probability is not 0.75 or vice versa.

Let us consider the two hypothesis:

- H_1: p = 0.75

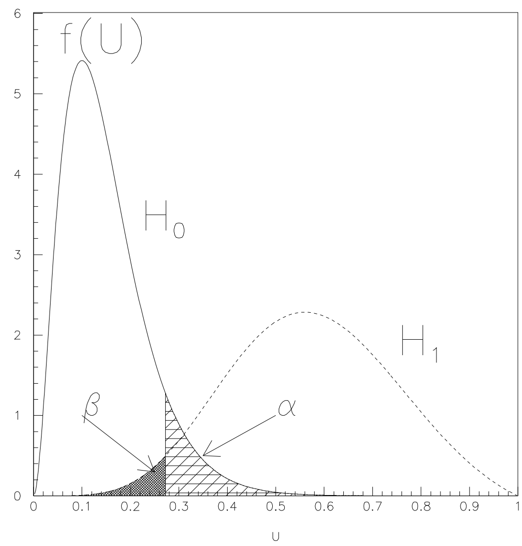


Figure 10.1: *Distribution of u with two different hypothesis setting a proper cut on the critical value of u, we can optimize the choice.*

- H_0: NOT H_1, the null hypothesis.

This is the *Reject/Support or RS* scheme, the null hypothesis is always against to what the research believes. This will set the asymmetry of the two type of errors. We have the scheme shown in Table 10.1.

	H_0 is true	H_1 is true
H_0 is accepted	correct	Type II error (β)
H_1 is accepted	Type I error (α)	correct

Table 10.1: *Possible types of errors in hypothesis test.*

- α is the fraction of cases you accept H_1 (what the researcher believes) while H_0 was true. The scientific community will require that α is small (usually $\leq$ 0.05) since a false theory should not be supported, funded etc.
- β is the fraction of cases you reject H_1, while it was true. Of course the researcher wants β to be as small as possible. A good theory could be abandoned because of an unfavorable fluctuation in the test.

  The quantity
  $$P = 1 - \beta$$
  is the *Power* of the test.

Example 10.1 *A politician makes an opinion poll to test if the majority of the population will support him at the next elections.*

- *π is the true fraction of favorable voters.*

- *p is the measured fraction of favorable voters in a opinion poll of N voters.*
- *H_0 is the hypothesis: $\pi \leq 0.5$*

The variable $p = \frac{Number\ of\ favorables}{N}$ is a binomial variable. Let us assume $N = 100$ (a large polling is very expensive) and $\pi = 0.5$. The variance of the distribution of p is:

$$\sigma_p^2 = \frac{\pi(1-\pi)}{N}$$

One possible strategy to reject H_0 is: $p \geq p^* = 0.58$. We can compute the value of α using the normal approximation to the binomial:

$$\alpha = \int_{p^*}^{\infty} N(\pi = 0.5, \sigma_p^2)\, dp = 0.044$$

This test keeps the Type I error small. In fact it would be very bad for the politician to believe that he could be elected, while the public opinion was in fact against him.

However the power of the test is too small. Assume, for instance that $\pi = 0.55$ (the voters will support him, H_0 is false). The probability to have the wrong result from the test is:

$$Power = 1 - \beta = 1 - \int_{p^*}^{\infty} N(\pi = 0.55, \sigma_p^2)\, dp = 0.27$$

Only in 0.27 of cases a repeated polling test (in the same conditions) would have given the correct positive result. One should instead a power as large as 0.8. In this case the power of the test can be increased by increasing the size of the opinion polling and tuning the value of N and p^* in order to get, at the same time $\alpha \approx 0.05$ and $\beta \approx 0.8$ for any value of π.

This type of test is frequently used in medicine, biology, social sciences etc. In high Energy Physics the requests on α and β are often dictated by consideration specific of the experiment.

Example 10.2 *A particle crossing an experimental apparatus could be an electron or a pion. The response of the apparatus is different in the two cases (shower size etc. if the apparatus is a calorimeter). A function of the measurements is constructed: t is our test statistics. The p.d.f. of t will be different if the particle is an electron or a pion and can be studied experimentally in dedicated Test Beams runs. Assume we are interested in electron initiated events. We will set up the hypothesis test putting: $H_0 = \pi$ and $H_1 = e$. The Type I errors are now called beam contamination of π in our electron sample. Type II errors will be lost electrons. Also in this case we will require that the probability α to be small, since we do not want the data to be flooded with background. On the contrary if β is large (or, in other words, the power is small) we will have to run longer to get the same statistics. A large β is not so deadly as a large α.*

10.2 The Likelihood ratio Test

The likelihood ratio test minimizes the probability of type II errors among all tests based on the observations $\{x_1 \cdots x_n\}$ and for which the probability of type I error does not exceed α (Neyman-Person lemma) [47].

The method is the following:

- Assume that we have two *simple* hypothesis:
  - H_0 for which the true density of x is f_0;

  - H_1 for which the true density of x is f_1;

- Construct the likelihood ratio:
$$t = \frac{f_1(x_1 \cdots x_n)}{f_0(x_1 \cdots x_n)}$$

- Decide a maximum probability of type I error α (say $\alpha = 0.05$) and from this value compute t_α such that:
$$P_0(t \geq t_\alpha) = \alpha$$
i.e. the probability that, under the hypothesis H_0, $t \geq t_\alpha$ is α.

- then, if $t \geq t_\alpha$ accept H_1

In fact we want to find a region S_0 such that:

$$\int_{S_0} f_0(x)\ dx = \alpha$$

and which maximizes the power:

$$1 - \beta \quad = \quad power \quad = \quad \int_{S_0} f_1(x)\ dx = \int_{S_0} \frac{f_1}{f_0} f_0\ dx = E_{S_0}(\frac{f_1}{f_0} \mid f_0)$$

The power is maximized if in all S_0 we maximize $\frac{f_1}{f_0}$. Therefore we have to choose S_0 such that:

$$t = \frac{f_1}{f_0} \geq t_\alpha$$

in all S_0. Of course t_α defines S_0 and must be the largest possible compatible with the condition:

$$\int_{S_0} f_0(x)\ dx = \alpha$$

Example 10.3 *We have to make a decision on the averages of a normal distribution:*

- $x \sim N(\mu_0, \sigma^2)$
- $x \sim N(\mu_1, \sigma^2)$

We make n measurements. The LR test is:

$$t = \frac{f_1}{f_0} = \frac{exp(-\frac{1}{2\sigma^2}\sum(x_1 - \mu_1)^2)}{exp(-\frac{1}{2\sigma^2}\sum(x_1 - \mu_0)^2)} = exp(\frac{n\ (\mu_1 - \mu_0)(2(\overline{x} - \mu_0) + (\mu_0 - \mu_1)}{2\sigma^2}) \geq t_\alpha$$

Rather than computing t_α by:
$$P_0(t \geq t_\alpha) = \alpha$$
we will rewrite it in the form:
$$P_0(ln(t) \geq ln(t_\alpha)) = \alpha$$
$$ln(t) = \frac{1}{2\sigma}\ 2\sqrt{n}\ (\mu_1 - \mu_0)\ (\frac{\overline{x} - \mu_0}{\sigma})\sqrt{n} - \frac{(\mu_1 - \mu_0)^2 n}{2\sigma^2}$$
The variable $z = \frac{(\overline{x} - \mu_0)^2}{\sigma/\sqrt{n}} \sim N(0,1)$ if H_0 is true. Since $ln(t)$ is a linear function of z, the statement:
$$P_0(ln(t) \geq ln(t_\alpha)) = \alpha$$

is equivalent to:

$$P_0(z \geq \lambda_\alpha) = \alpha$$

where λ_α *is the* α*-point of the standard normal distribution* ($\int_{-\infty}^{\lambda_\alpha} N(0,1)\ dx = \alpha$)*. The LR test is equivalent to a cut on* z*, the critical region is* $z \geq \lambda_\alpha$

The power of the test is a function n *and of the difference of the two means.*

$$power \ = \ 1 - \beta = P_1(z \geq \lambda_\alpha) = \int_{\lambda_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\ exp(-\frac{(z - (\mu_1 - \mu_0)\sigma/\sqrt{n})^2}{2})\ dz$$

Upon the change of variable: $t \ = \ z - (\mu_1 - \mu_0)\sqrt{n}/\sigma$*, we get:*

$$power \ = \ \int_{\lambda_\alpha - (\mu_1 - \mu_0)\sqrt{n}/\sigma}^{\infty} N(0,1)\ dx = \Phi((\mu_1 - \mu_0)\sqrt{n}/\sigma - \lambda_\alpha)$$

which is a function of $\mu_1 - \mu_0$ *and* n*.* $\Phi(x)$ *is the cumulative function of the Standard Normal distribution. The power becomes 1 if* $\mu_1 \to \infty$ *or* $n \to \infty$*. If instead* $\mu_1 = \mu_0$ *the power becomes* α *as expected.*

The likelihood ratio is a reasonable test statistics. The difficulty is to determine the critical region: $P_{H_0}(t \geq t_\alpha) = \alpha$ and therefore t_α. We need the $p.d.f.$ of t whose calculation is usually difficult and lengthy. One can use Monte Carlo simulations or use the asymptotic distribution of t.

Another possibility is indicated by Wilks [52] who proved that, for $n \to \infty$ the limiting distribution of $-2\ ln(t)$ is, under H_0:

$$-2\ ln(t) \sim \chi_r^2$$

where r is the number of parameters fixed by H_0. This property can be used to compute t_α.

Chapter 11

Distributions related to the Normal

11.1 The χ^2 Distribution

Let x_i be independent measurements of a normal standard variable ($N(0,1)$) and consider the variable u:

$$u = \sum_{i=1,k} x_i^2$$

u is called a χ^2 variable with k degrees of freedom (d.o.f.). We want to find the *p.d.f.* of u.

11.1.1 The Distribution of x^2

First we find the distribution of a single normal variable x^2.

The inverse function $x = \sqrt{h}$ is a two valued function:

$$x_1(h) = \sqrt{h} \qquad\qquad x_2(h) = -\sqrt{h}$$

The derivatives are

$$\left| \frac{\partial x_{1,2}}{\partial h} \right| = \frac{1}{2\sqrt{h}}$$

The *p.d.f.* of h is obtained by summing the two values:

$$g(h) = N(x_1(h); 0, 1) \left| \frac{\partial x_1}{\partial h} \right| + N(x_2(h); 0, 1) \left| \frac{\partial x_2}{\partial h} \right|$$

and we get:

$$g(h) = \frac{1}{\sqrt{2\pi}}(e^{-\frac{h}{2}} + e^{-\frac{h}{2}})\frac{1}{2\sqrt{h}}$$

We can recast the expression in the following way:

$$g(h) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot h^{\alpha-1} \cdot e^{-\frac{h}{\beta}}$$

($\Gamma(1/2) = \sqrt{(\pi)}$), with $\alpha = 1/2$ and $\beta = 2$ ($0 \le h \le \infty$).

Hence $g(h)$ has a *p.d.f.* of the family of the gamma distribution.

11.1.2 The χ^2 Distribution

To obtain the *p.d.f.* of u we use the *m.g.f.* techniques. In fact the *m.g.f.* of h has already been computed:

$$\Phi_h(t) = (1 - 2 \cdot t)^{-\frac{1}{2}}$$

The sum of independent variables all having the same distribution is the product of the individual *m.g.f.*:

$$\Phi_u(t) = (\Phi_h(t))^k = (1-2t)^{-\frac{k}{2}}$$

This is again of the *m.g.f.* of the gamma function with: $\alpha = k/2$ and $\beta = 2$. Therefore u has a gamma distribution with the above parameters:

$$h(u) = \frac{1}{\Gamma(k/2) \cdot 2^{k/2}} \cdot u^{(\frac{k}{2}-1)} \cdot e^{-\frac{u}{2}}$$

Plots of the distributions are given in the Fig.11.1. The average and variance can be found from

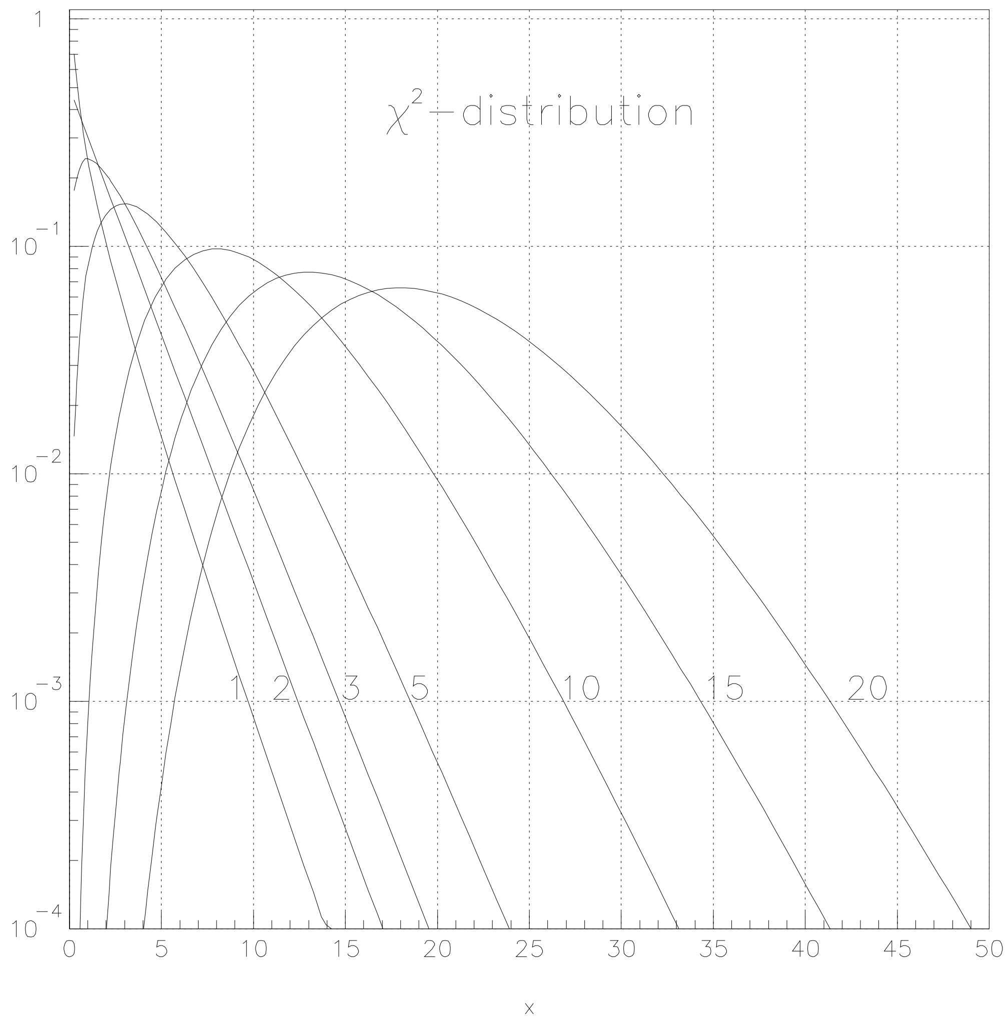


Figure 11.1: *The chisquare distribution for a few values of k*

the *m.g.f.*

$$\begin{aligned}\mu_1 = E(u) &= \frac{\partial M'(t)}{\partial t} \mid_{t=0}= \\ &= k(1-2t)^{-(1-k/2)} \mid_{t=0}= \\ &= k\end{aligned}$$

Exercise 11.1 *Prove that $Var(u) = 2k$.*

In case the variance is not 1 but σ^2, it is sufficient to consider the reduced variable x_i/σ which is again a standard normal variable.

$$u = \sum(\frac{x_i}{\sigma})^2 = \frac{1}{\sigma^2}\sum_{i=1,k} x_i^2 \sim \chi^2_k$$

Therefore that the variable $\sum x_i^2$ has a $\sigma^2\chi^2$ *p.d.f.*

Fig.11.2 shows the probability content of the χ^2 distribution. i.e. the area of the curve, for fixed n, above u. In fact from the definition of *p.d.f.* this area is the probability that a measurement of the variable has a result larger than u. Values of u much different from its average (for that n) are unlikely. The numerical values of u for a fixed probability content and for several degrees of freedom are shown in Tab 11.1.2.

N_{dof}	0.05		0.01		0.005	
	low	high	low	high	low	high
1	.393E-02	3.84	.157E-03	6.63	.393E-04	7.88
2	.103	5.99	.201E-01	9.21	.100E-01	10.6
3	.352	7.82	.115	11.3	.716E-01	12.8
4	.711	9.49	.297	13.3	.207	14.9
5	1.15	11.1	.554	15.1	.412	16.7
6	1.64	12.6	.872	16.8	.676	18.5
7	2.17	14.1	1.24	18.5	.989	20.3
8	2.73	15.5	1.65	20.1	1.34	22.0
9	3.33	16.9	2.09	21.7	1.73	23.6
10	3.94	18.3	2.56	23.2	2.16	25.2
15	7.26	25.0	5.23	30.6	4.60	32.8
20	10.9	31.4	8.26	37.6	7.43	40.0
25	14.6	37.7	11.5	44.3	10.5	46.9
30	18.5	43.8	15.0	50.9	13.8	53.7
40	26.5	55.8	22.2	63.7	20.7	66.8
50	34.8	67.5	29.7	76.2	28.0	79.5
75	56.1	96.2	49.5	106.	47.2	110.
100	77.9	124.	70.1	136.	67.3	140.
150	123.	180.	113.	193.	109.	198.
200	168.	234.	156.	249.	152.	255.

Table 11.1: *Minimum and maximum value of u for probability content of 0.05, 0.01 and 0.005 as function of the number of degrees of freedom for a χ^2_n p.d.f..*

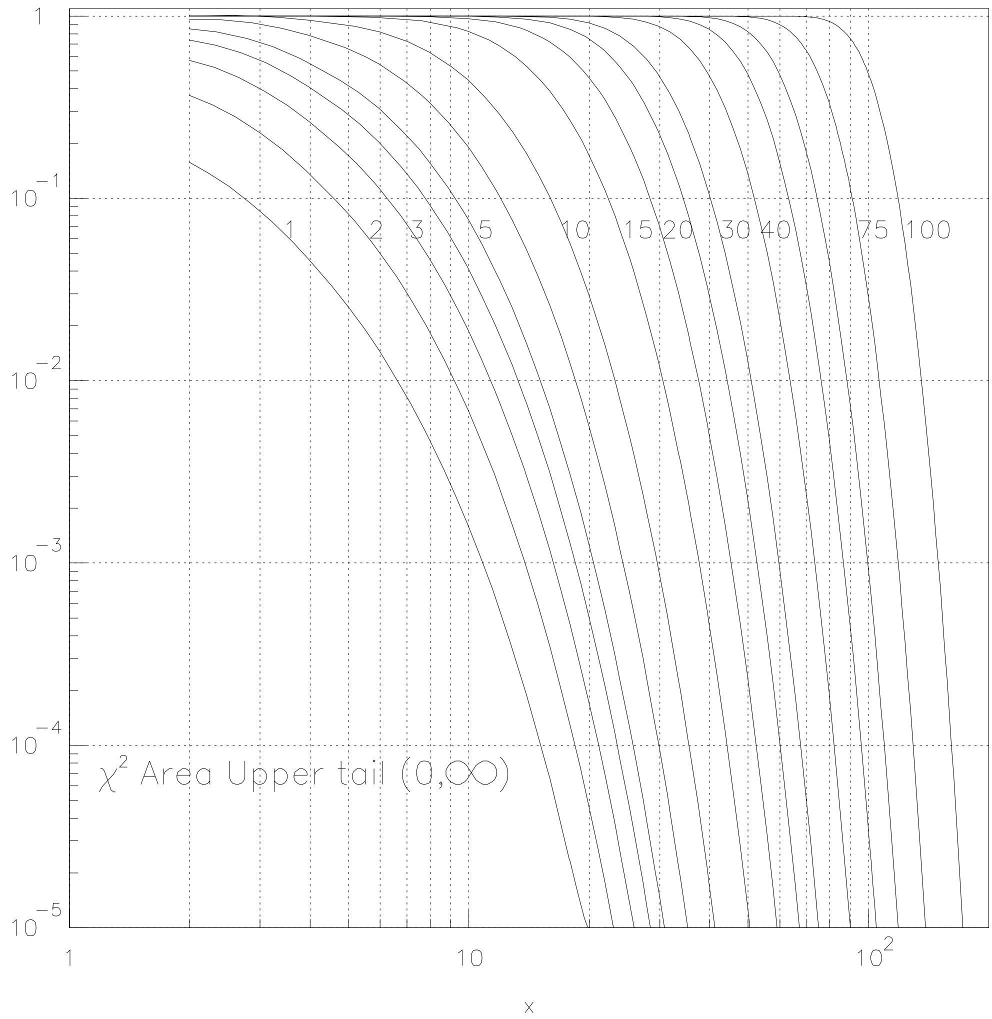


Figure 11.2: *The probability content of the chisquare distribution*

11.1.3 The Sum of Squares about the Mean

We have a random sample x_i from a population $N(\mu, \sigma^2)$, with mean sample $\overline{x}$. Let:

$$w = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n \cdot \overline{x}^2$$

To compute the *p.d.f.* of w we make an *orthogonal transformation*. A linear orthogonal transformation is a linear transformation:

$$h_i = \sum_j c_{i,j} x_j, \qquad i = 1, 2, \cdots n$$

such that:

$$\sum h_i^2 = \sum x_i^2$$

The Jacobian of this transformation is unity. In matrix notation we have:

$$\underline{h} = \underline{C} \cdot \underline{x}$$

with the condition:

$$\underline{h}^T \cdot \underline{h} = \underline{x}^T \underline{C}^T \underline{Cx} = \underline{x}^T \cdot \underline{x}$$

hence $C^T C = I$ .This means that the the transformation is equivalent to a rotation in a n-dimensional Euclidean space.

Let (this is also called Helmert's transformation)

$$h_1 = (x_1 + x_2 + \cdots + x_n)/\sqrt{n} = \overline{x} \cdot \sqrt{n}$$

This choice specifies somehow also the remaining variables such that:

$$E(h_i) = E(\sum_j c_{i,j} \cdot x_j) = \sum_j c_{i,j} \cdot E(x_j) = \mu \sum_j c_{i,j} = 0 \qquad (i > 1)$$

because of the orthogonality condition.

Exercise 11.2 *Prove it. (Hint: use $C^T C = I$)*

Also since:

$$\sum_{i=1,n} h_i^2 = \sum_{j=1,n} x_j^2$$

we have that the variable w transforms into:

$$w = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n \cdot \overline{x}^2 = \sum h_i^2 - h_1^2 = \sum_{j=2,n} h_j^2$$

And, since each h_i has a *p.d.f.* $N(0, \sigma^2)$ we have that w has a *p.d.f.* $\sigma^2 \cdot \chi^2_{n-1}$.

The sum of squares about the mean is an χ^2 variate with $n-1$ degrees of freedom. The loss of one d.o.f. arises from the constraint $\sum (x_i - \overline{x}) = 0$.

11.1.4 The Sample Variance

The variance of a sample is defined as:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1,n} (x_i - \overline{x})^2 = \frac{1}{n-1} \sum_{i=2,n} h_i^2$$

The average value of s^2 is equal to σ^2, independently of the of the *p.d.f.* of x_i. In fact, since $E(h_i) = 0$, we have that $E(h_i^2) = \sigma^2$. Therefore:

$$E(s^2) = \frac{1}{n-1} \sum_{i=2,n} \sigma^2 = \sigma^2$$

whatever the distribution of x. In particular if x_i have a *p.d.f.* $N(\mu, \sigma^2)$, then

$$\frac{n-1}{\sigma^2} \cdot s^2 \sim \chi^2_{n-1}$$

The variance of s^2 is obtained from the variance of the χ^2_{n-1} distribution:

$$Var(s^2) = \sigma^4 \, \frac{2}{n-1}$$

For large n:

$$Var(s^2) \approx \frac{2 \cdot \sigma^4}{n}$$

In summary: from a sample $(x_i, i = 1 \cdots n)$ the mean:

$$\overline{x} = \frac{\sum x_i}{n} \;\sim\; N(\mu, \frac{\sigma^2}{n})$$

The sample variance is:

$$s^2 = Var(x) = \frac{\sum(x_i - \overline{x})^2}{n-1} \;\sim\; \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$$

Since s^2 is an unbiased estimator of σ^2 we get:

$$Var(\overline{x}) = \frac{\sigma^2}{n} \approx \frac{s^2}{n} = \frac{\sum(x_i - \overline{x})^2}{n \cdot (n-1)}$$

$$Var(s^2) = \frac{2\sigma^4}{n-1} \approx \frac{2s^4}{n-1} \approx \frac{2\ s^4}{n}$$

Exercise 11.3 *Prove that in the case of a Normal sample the random variables $\overline{x}$ and s^2 are independent variables.*

An intuitive argument to understand why we normalize the sum of squares to $n-1$ instead of n is the following. The variance of the sample is made by two terms: the dispersion of the events around the sample mean and the variance of the mean:

$$s^2 = \frac{\sum_{i=1,n}(x_i - \overline{x})^2}{n} + \frac{s^2}{n}$$

so that the sample estimate of the population variance is:

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

The determination of the mean from the sample itself has eaten up one d.o.f..

11.1.5 Quantiles and Median

For a continuous distribution the median (μ_M) is defined as:

$$\int_{-\infty}^{\mu_M} f(x)dx = \int_{\mu_M}^{\infty} f(x)dx = \frac{1}{2}$$

This concept can be easily extended to different fractions of the *p.d.f.*: *quantiles* that divide the total frequency into n equal parts.

In the case of a discrete distribution the median is the value of the abscissa that divides the samples into two equal parts. If the sample size is odd $(2N+1)$ and the measurements are ordered in increasing order, then $\mu_M = x_N$. If the sample size is even $(2N)$, conventionally we take $\mu_M = (x_{N-1} + x_N)/2$.

We will now compute the variance of quantiles, in particular of the median. Let us now consider a sample of n measurements from a distribution:

$$dF = f(x)dx$$

and let us compute the probability that, out of the n measurements $l-1$ fall below a value x_q, one in the interval $(x_q - dx_q, x_q + dx_q)$ and $n-l$ fall above x_q. Apart from normalization we have:

$$P(x_q)\ dx_q = (F(x_q))^{l-1} \cdot f(x_q)\ (1 - F(x_q))^{n-l}\ dx_q = F_q^{l-1}(1 - F_q)^{n-l} dF_q$$

Let us put $l = nq$ and $n - l = np$ $(p + q = 1)$.

To compute the maximum of the distribution (the modal point) we first take the logarithm and then differentiate. We get:

$$(l-1)\frac{f_q}{F_q} - (n-l)\frac{f_q}{1-F_q} + \frac{f'_q}{f_q} = 0$$

Since n and l are large, we neglect the last term. To order $1/n$ we get:

$$\frac{q}{F_q} - \frac{p}{1-F_q} = 0$$

This equation is solved for a value x^* such that $F(x^*) = q$.

We can now study the distribution of $P(x_q)$ near its maximum. Let's put $F_q = q + \eta$, we have:

$$P \approx (q+\eta)^{nq}(p-\eta)^{np}$$

We take the logarithm and expand around $\eta = 0$:

$$\begin{aligned}
\log P &= nq(\log q + \log{(1+\frac{\eta}{q})}) + np(\log p + \log{(1+\frac{\eta}{p})}) \\
&= nq(\frac{\eta}{q} - \frac{1}{2}\frac{\eta^2}{q^2} + \cdots) + np(-\frac{\eta}{p} - \frac{1}{2}\frac{\eta^2}{p^2} + \cdots) + const \\
&= -\frac{n\eta^2}{2pq} + const + O(\eta^3)
\end{aligned}$$

We neglect the high order terms and finally get:

$$P(\eta)d\eta \approx e^{-\frac{n\eta^2}{2pq}}$$

In the limit η is distribute normally. The variance of η is easily obtained:

$$Var(\eta) = \frac{pq}{n}$$

In fact we are interested in the variance of x_q that can be obtained since:

$$dF_q = d\eta = f_q dx_q$$

so that:

$$Var(x_q) = \frac{Var(\eta)}{f_q^2}$$

In the case of the median: $p = q = 1/2$ and $f_{\frac{1}{2}}$ is the value of the median ordinate $f_{1/2} = \frac{1}{\sqrt{2\pi}\sigma}$:

$$Var(\mu_M) = \frac{\pi}{2}\frac{\sigma^2}{n}$$

Though the median characterizes the mid point of the distribution as well as the mean, the median μ_M has a variance $\pi/2$ larger than the mean. The mean is more "precise" than the median, its variance is smaller. Among the *estimates* of the population mean value, the sample mean is more *efficient*.

11.2 The t Distribution (W.S. Gosset)

The variable t is used as a test statistics for the null hypothesis in the case the sample is extracted from a normal population whose mean is known, but whose variance is unknown.

Example 11.1 *Let us assume that the mass of a resonance has been measured in an experiment. Only 10 events have been recorded yielding the values:*

$$m_i = 1.60, 1.60, 1.67, 1.70, 1.73, 1.76, 1.78, 1.78, 1.81, 1.81 \qquad GeV/c^2$$

Are the measurements compatible with the known mass of the resonance $M = 1.67\ GeV/c^2$? We assume also that the p.d.f. of the measurements is normal but with unknown variance σ^2.

We find that the average $\overline{m} = 1.724$ and we know that the distribution is: $\overline{m} \sim N(\mu, \sigma^2/n)$. Could the difference with respect to the value of M be due to chance effect and with what probability? If we knew σ^2 we could built the reduced variable:

$$z = \frac{\overline{m} - \mu}{\sigma/\sqrt{n}}$$

which is $N(0,1)$, independent of μ and σ^2 and we can compute $P(|m| > z)$.

In our case we can estimate σ^2 from the sample by the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1,n} (m_i - \overline{m})^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$$

We have:

$$s^2 = 62.9 \cdot 10^{-4} \qquad s = 7.93 \cdot 10^{-2} \qquad s/\sqrt{n} = 2.51 \cdot 10^{-2}$$

If now we use s^2 in the place of σ^2, the probability distribution of z is not any longer $N(0,1)$. To proceed with our test we have to compute the distribution of this variable.

By analogy with z we built the variable:

$$t = \frac{\overline{m} - m}{\frac{s}{\sqrt{n}}} = \frac{\frac{\overline{m}-m}{\frac{\sigma}{\sqrt{n}}}}{\frac{s}{\sqrt{\sigma^2}}} = \frac{N}{\sqrt{U/(n-1)}}$$

The unknown variance σ^2 disappears from t and therefore it can be used to test the difference $\overline{x} - m$. In the last passage we have stressed the fact that the variable t is made of two pieces: the numerator is a $N(0,1)$ variable, while U is a χ^2_{n-1} variable.

The *p.d.f.* of t can now be computed in the standard way. Since $\overline{x}$ and s^2 are independent variables, the joint density of N and U is the product of the two *p.d.f.*:

$$\Theta(N, U) = \frac{e^{-\frac{N^2}{2}}}{\sqrt{2\pi}} \cdot \frac{U^{\frac{n-1}{2}-1} \cdot e^{-\frac{U}{2}}}{\Gamma(\frac{n-1}{2}) \cdot 2^{\frac{n-1}{2}}}$$

Then we make the transformation:

$$h_1 = t = \frac{N}{\sqrt{U/(n-1)}} \qquad\qquad h_2 = U$$

and, inverting:

$$U = h_2 \qquad\qquad N = h_1 \sqrt{\frac{h_2}{n-1}}$$

The Jacobian of the transformation is:

$$J = \begin{pmatrix} \frac{\partial N}{\partial h_1} & \frac{\partial N}{\partial h_2} \\ \frac{\partial U}{\partial h_1} & \frac{\partial U}{\partial h_2} \end{pmatrix} = \begin{pmatrix} \sqrt{h_2/(n-1)} & \frac{h_1}{2}\sqrt{\frac{1}{h_2(n-1)}} \\ 0 & 1 \end{pmatrix} = \sqrt{\frac{h_2}{n-1}}$$

Then we express Θ as a function of h_1 and h_2, then integrate away h_2 to get the marginal probability of t:

$$f_{n-1}(t) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)} \cdot \Gamma\left((n-1)/2\right)} \cdot \frac{1}{(1+\frac{t^2}{n-1})^{\frac{n}{2}}}$$

This is the probability distribution of t with $n-1$ d.o.f.. When $n = 2$ the distribution reduces to Cauchy. For $n = \infty$ the distribution becomes N(0,1). The shape of the t-distribution is shown in the Fig. 11.3 for a few values of n. The probability content of the t-Student distribution

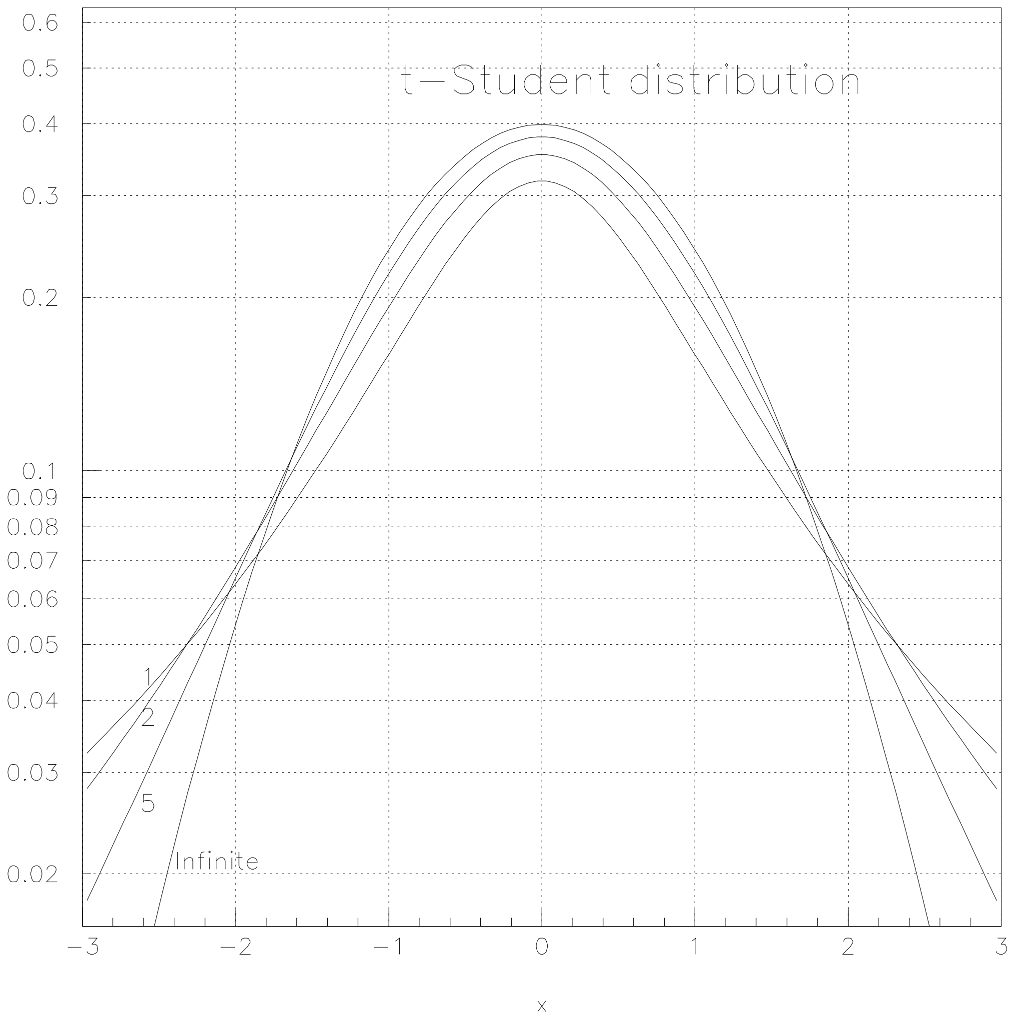


Figure 11.3: *The probability distribution of the t-Student variable as a function of t, for a few values values of ν.*

is shown in Tab 11.2 where, for some N_{dof}, we quote the value of t for a fixed value of the probability.

N_{dof}	$\alpha = .05000$	.01000	.00500
2	2.91999	6.96456	9.92485
3	2.35338	4.54070	5.84091
4	2.13185	3.74696	4.60410
5	2.01505	3.36493	4.03216
6	1.94318	3.14267	3.70743
7	1.89458	2.99795	3.49948
8	1.85955	2.89646	3.35539
9	1.83311	2.82144	3.24984
10	1.81246	2.76377	3.16927
15	1.75305	2.60248	2.94671
20	1.72472	2.52798	2.84534
25	1.70814	2.48511	2.78744
30	1.69726	2.45726	2.75000
40	1.68385	2.42326	2.70446
50	1.67590	2.40327	2.67779
75	1.66543	2.37710	2.64298
100	1.66024	2.36422	2.62589
150	1.65507	2.35147	2.60900
200	1.65251	2.34514	2.60063
250	1.65098	2.34136	2.59564

Table 11.2: *Values of t_α such that $\alpha = \int_{t_\alpha}^{\infty} f(t)\, dt$ for some values of α and for some DoF.*

Exercise 11.4 *Evaluate the integral and complete the derivation of f_ν.*

Exercise 11.5 *Prove that the average value and the variance of t are:*

$$E(t) = 0 \qquad\qquad Var(t) = \frac{n-1}{(n-1)-2} \quad (n-1 > 2)$$

In the case of the example above we get:

$$t_m = 2.15$$

In this case we were concerned with the difference in absolute value of the average of measurements with respect to a known value. The probability is the sum:

$$P(t_m) = 2 \cdot \int_{|t_m|}^{\infty} f(t)\, dt$$

From the tables we find: $P = 6\%$. We accept the hypothesis that the mass of the resonance is $M = 1.67\ GeV/c^2$ at 5% confidence level. If we would have used (erroneously) the normal distribution, we would have found $P = 3.1\%$ and have rejected the hypothesis at 5% confidence level.

Example 11.2 *This case is met very often in the practice. Suppose we have a sample n_1 of observations of the normal variable x and n_2 observations of the normal variable y*

$$x \sim N(\mu, \sigma^2) \qquad\qquad y \sim N(\mu_2, \sigma^2)$$

We assume that the two samples have the same variance. We want to test the null hypothesis that the two mean are equal. We assume that the variance is the same, but unknown. Since we do not know the true variance, we use instead the sample variance and our test variable is:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\overline{x} - \overline{y}}{\sqrt{Var(\overline{x} - \overline{y})}}$$

s^2 *is the estimate derived from both samples:*

$$s^2 = \frac{\sum_{i=1,n_1}(x_i - \overline{x})^2 + \sum_{i=1,n_2}(y_i - \overline{y})^2}{n_1 + n_2 - 2}$$

Once we have the data we can compute the quantity t and get from the probability tables the confidence limit.

The tables usually quote the value:

$$P(t) = \int_t^{\infty} f(t)\, dt$$

which represents the chance that another experiment would yield a larger value of t. This is called *one sided test*. If we are interested to compute the probability that the difference of the means is different from zero either positive or negative, then we have to make a *two sided test*, i.e. to add to the previous integral the symmetric one:

$$P(t) = \int_{-\infty}^{-t} f(t)\, dt$$

A technical point: is it justified to use the Student distribution for the previous expression of t? The numerator $(\overline{x} - \overline{y})$ in the hypothesis of equal means is normal and zero mean. In addition we have already shown that $\overline{x}$ is independent of s_x^2 and the same is true for y. Hence s^2 is independent of $\overline{x} - \overline{y}$. Also:

$$\sum_{i=1,n} (x_i - \overline{x})^2 \sim \sigma^2 \chi^2_{n_1-1}$$

the same is true for the symmetric term in y but with $n_2 - 1$ d.o.f. To complete the proof we have to show that:

$$\chi^2_{n_1-1} + \chi^2_{n_2-1} \sim \chi^2_{n_1+n_2-2}$$

This can be easily proved remembering that the *m.g.f.* of the sum of two independent random variables is the product of the two *m.g.f.*:

$$\begin{aligned} m.g.f.(\chi^2_{n_1-1} + \chi^2_{n_2-1}) &= (1-2s)^{-\frac{n_1-1}{2}} \cdot (1-2s)^{-\frac{n_2-1}{2}} \\ &= (1-2s)^{-\frac{n_1+n_2-2}{2}} = m.g.f.(\chi^2_{n_1+n_2-2}) \end{aligned}$$

Example 11.3 *This is from the experiment NOMAD. The response the e.m. Calorimeter to muon signal can be used to monitor its stability. During the years 1995 and 1996 the average response to muon has been monitored. Fig.11.4 shows the distribution since beginning of 1995 of the average response.*

x_i, i=1... 52 points (average and S.D.) for 1995 and

y_i, i=1... 44 for 1996.

Out of each set (1995 and 1996) we built the averages and sample variances and we get: $(\overline{x}, s_x^2)$ and $(\overline{y}, s_y^2)$.

$$s_x^2 = \frac{\sum_{i=1,n}(x_i - \overline{x})^2}{n-1} \qquad s_y^2 = \frac{\sum_{i=1,m}(y_i - \overline{y})^2}{m-1} \qquad s^2 = \frac{s_x^2\ (n-1) + s_y^2\ (m-1)}{m+n-2}$$

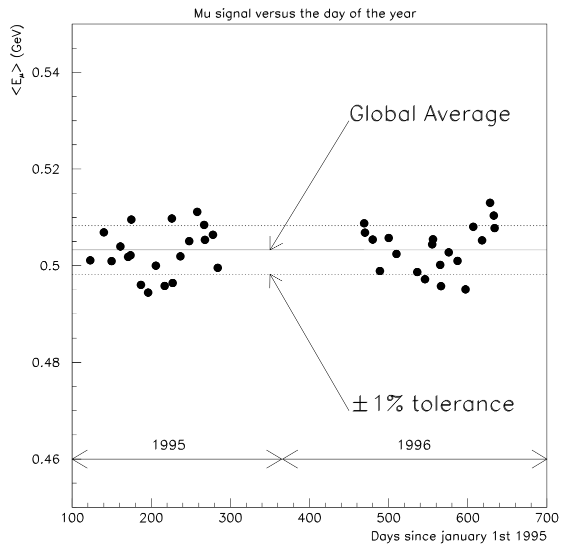


Figure 11.4: *Time dependence of the average Cherenkoff signal over two years of running time in the NOMAD e.m. calorimeter.*

The variable:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s^2}{n} + \frac{s^2}{m}}} \sim f(t)_{n+m-2}$$

has a t-distribution with $n + m - 2$ d.o.f. (n and m are the number of points in the two years).

In numbers:

$$\begin{aligned} \overline{x} &= 0.50391 \qquad && s_x = 4.899 \cdot 10^{-4} \\ \overline{y} &= 0.50277 \qquad && s_y = 3.872 \cdot 10^{-4} \end{aligned}$$

$t = 1.25$ *with 94 d.o.f. yielding a probability* $p(|t| > 1.25) = 0.21$ *(two sided test). The hypothesis that the means are equal is accepted at the confidence level of 20%.*

11.3 The Fisher Snedecor Distribution

This test is used to establish if two variances are significantly different. Assume s_1^2 is an estimate of σ_1^2 from the random sample of size $n + 1$. Its probability distribution will be:

$$s_1^2 \sim \frac{\sigma^2 \cdot \chi_n^2}{n}$$

And assume that from an independent sample of size $m + 1$ we have obtained another sample variance:

$$s_2^2 \sim \frac{\sigma^2 \cdot \chi_m^2}{m}$$

We want to compare s_2^2 with s_1^2. More precisely the hypothesis to be tested (null hypothesis) is:

$$s_1^2 = s_2^2 = s^2$$

In this case it is convenient to use as test statistics the ratio of sample variances:

$$F_{n,m} = \frac{s_1^2}{s_2^2}$$

The distribution of the variable F can be calculated because:

$$F_{n,m} = \frac{s_1^2}{s_2^2} = \frac{\frac{\sigma^2 \cdot X_n^2}{n}}{\frac{\sigma^2 \cdot X_m^2}{m}} = \frac{\frac{X_n^2}{n}}{\frac{X_m^2}{m}}$$

is independent of σ^2. The variables X_ν^2 are distributed as χ_ν^2, hence the ratio $F_{n,m}$ is independent of σ^2 in our hypothesis.

We have to find the probability density of the function F of the two χ^2 variables that we shall call V and W with m and n d.o.f. respectively.

The joint probability of V and W is the product of the individual *p.d.f.*:

$$P(V,W)dUdW = \frac{1}{\Gamma(\frac{m}{2}) \cdot 2^{m/2}} \cdot V^{\frac{m}{2}-1} \cdot e^{-\frac{V}{2}} \cdot \frac{1}{\Gamma(\frac{n}{2}) \cdot 2^{n/2}} \cdot W^{\frac{n}{2}-1} \cdot e^{-\frac{W}{2}} dVdW$$

We change the variables:

$$\begin{aligned} F &= \frac{\frac{V}{m}}{\frac{W}{n}} \\ V &= V \end{aligned}$$

then we solve for the old variables in terms of the new ones:

$$\begin{aligned} V &= V \\ W &= \frac{n}{m} \cdot \frac{V}{F} \end{aligned}$$

The Jacobian of the transformation can easily be found:

$$\frac{\partial(V,W)}{\partial(F,V)} = \begin{bmatrix} \frac{\partial V}{\partial F} & \frac{\partial V}{\partial V} \\ \frac{\partial W}{\partial F} & \frac{\partial W}{\partial V} \end{bmatrix} = \frac{n}{m} \cdot \frac{V}{F^2}$$

and now we can replace V and W in the expression of P(V,W) with the variables V and F:

$$q(F,V)dFdV = p(V, W(F,V)) \cdot \mid \frac{\partial(V,W)}{\partial(F,V)} \mid dFdVb$$

getting:

$$q(F,V) = \frac{(\frac{n}{m})^{\frac{n}{2}}}{(\frac{m}{2}-1)! \cdot (\frac{n}{2}-1)!} \cdot 2^{-\frac{n+m}{2}} \cdot \frac{V^{\frac{m+n}{2}-1}}{F^{\frac{n}{2}+1}} \cdot e^{-(\frac{V}{2} - n\frac{V}{2mF})}$$

The marginal distribution for F is obtained by integrating away V. [1] The resulting distribution is:

$$f(F)_{m,n} = C \cdot \frac{F^{\frac{m}{2}-1}}{(n \cdot F + m)^{\frac{m+n}{2}}}$$

where:

$$C = \frac{\Gamma(\frac{m}{2} + \frac{n}{2})}{\Gamma(\frac{m}{2}) \cdot \Gamma(\frac{n}{2})} \cdot n^{\frac{m}{2}} \cdot m^{\frac{n}{2}}$$

The shape of the *p.d.f* is shown in Fig. 11.5 for a few values of n and m. Given the estimation (pooled) of s_1^2 and s_2^2 we can read from the probability table the significance of the difference.

[1] The most remarkable part of the integration is:

$$\int_0^\infty V^\alpha \cdot e^{-\frac{V}{\beta}}\, dV = \alpha!\beta^{\alpha-1}$$

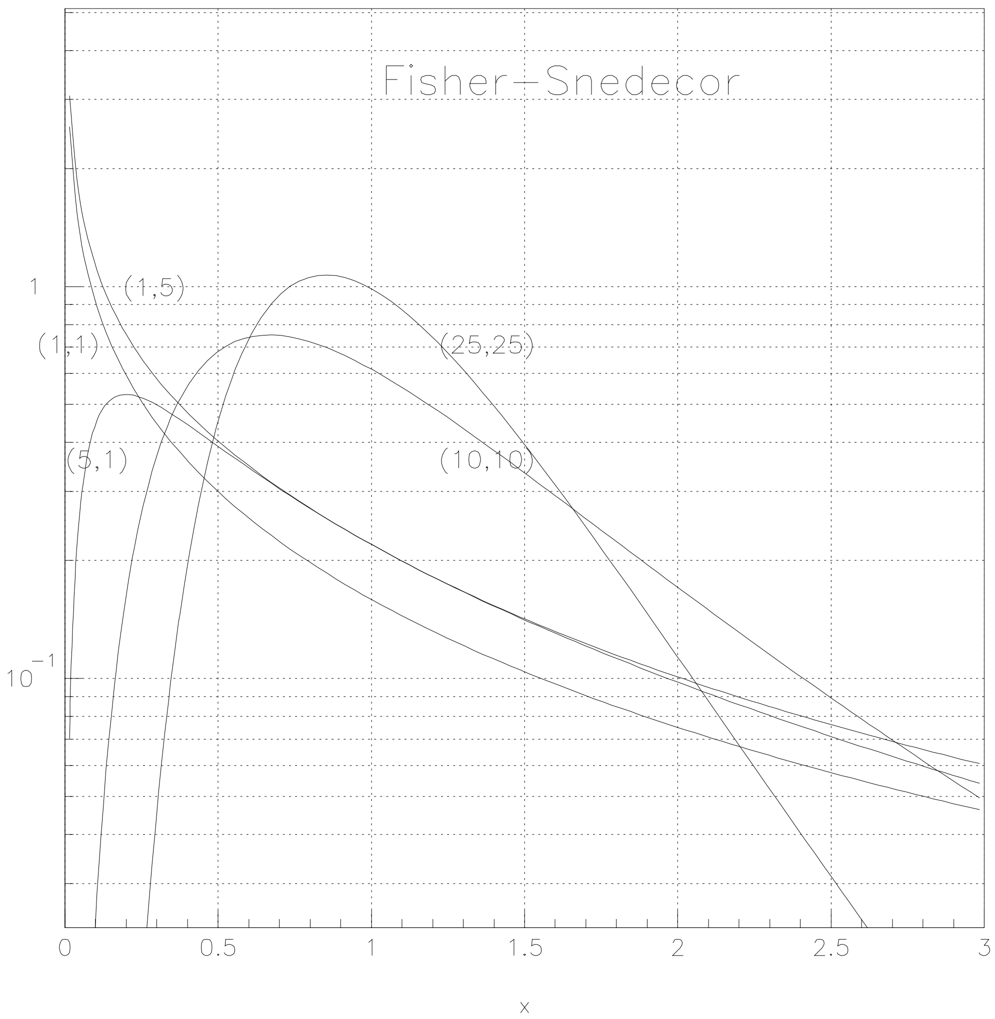


Figure 11.5: *Probability distribution of Fisher-Snedecor variable F for a few values of n and m.*

Exercise 11.6 *Show that:*

$$E(F_{n,m}) = \frac{m}{m-2} \qquad m \geq 3$$

$$Var(F_{n,m}) = \frac{2\ m^2\ (n+m-2)}{n(m-2)^2(m-4)} \qquad m \geq 5$$

Exercise 11.7 *Show that if we interchange the two samples, and therefore the ratio of the sample variances becomes:*

$$y = \frac{1}{F}$$

then the density function is the same with m and n interchanged and also the distribution function $G_{m,n}(t)$ which gives the probability that $F < t$ is related with $G_{n,m}(1/t)$ by:

$$G_{m,n}(F) = 1 - G_{n,m}(\frac{1}{F})$$

The Fisher-Snedecor distribution, used to test the equality of variances, is in fact very important to establish if the means of two samples are equal. This is obtained with the Analysis of Variances, which however is seldom used in High Energy Physics and which we only outline here. As an example of the application of the *Fisher* test in High Energy Physics we will study its application in curve fitting, when describing the X^2 method.

11.4 Analysis of Variances in short

This analysis is better described with an example.

Example 11.4 *Assume that we have two groups of measurements G_1 and G_2 (Tab. 11.4). In the Table the symbol SS means Sum of the Squares about the mean.*

	G_1	G_2
Obs 1	2	6
Obs 2	3	7
Obs 3	1	5
Mean	2	6
SS	2	2
$Mean_{1+2}$	4	
SS_{1+2}	28	

Table 11.3: *Two groups of observations are analyzed within groups and globally.*

We want to test if the two groups are homogeneous, i.e. they have the same mean and variance. In each group we have computed the average and the sum of squares (SS) about the average. If now we mix the two groups we get: $Mean_{1+2} = 4$ and $SS_{1+2} = 28$. SS_{1+2} is much larger than $SS_1 + SS_2$: the two groups have rather different means. The contribution to SS_{1+2} is due only in part to $SS_1 + SS_2$, the rest is due to the difference between the means of the two groups. The test of homogeneity of the two groups can be reduced to a test of compatibility of the two sample variances. Before going to it, we have to construct independent sample variances

Assume that we have a set of measurements divided in p families:

$$\begin{cases} x_{1,1} & \cdots & x_{1,n_1} & \overline{x_1} \\ x_{2,1} & \cdots & x_{2,n_2} & \overline{x_2} \\ \cdots & & & \\ x_{p,1} & \cdots & x_{p,n_p} & \overline{x_p} \end{cases}$$

The following relations hold:

$$\begin{aligned} \sum_{i,j}(x_{i,j} - \overline{x})^2 &= \sum_{i,j}(x_{i,j} - \overline{x_j} + \overline{x_j} - \overline{x})^2 = \sum_{i,j}(x_{i,j} - \overline{x_j})^2 + \sum_{i,j}(\overline{x_j} - \overline{x})^2 \\ &= \sum_{i,j}(x_{i,j} - \overline{x_j})^2 + \sum_j n_j(\overline{x_j} - \overline{x})^2 \end{aligned}$$

If the observations are normally distributed and homogeneous then ($n = \sum n_i$):

$$\frac{1}{\sigma^2}\sum_{i,j}(x_{i,j} - \overline{x})^2 \quad \sim \quad \chi^2_{n-1}$$

$$u_j = \frac{1}{\sigma^2}\sum_{i}(x_{i,j} - \overline{x_j})^2 \quad \sim \quad \chi^2_{n_j-1} \qquad j = 1, \cdots, p$$

$$\sum_j u_j = \frac{1}{\sigma^2}\sum_{i,j}(x_{i,j} - \overline{x_j})^2 \quad \sim \quad \chi^2_{n-p}$$

$$\frac{1}{\sigma^2}\sum_{j}(\overline{x_j} - \overline{x})^2 \quad \sim \quad \chi^2_{p-1}$$

In summary (Tab. 11.4):

Sum of Squares		$D.o.F$
of families means about global mean	$\sum_j n_j(\overline{x_j} - \overline{x})^2$	$p-1$
of individuals about families mean	$\sum_{i,j}(x_{i,j} - \overline{x_j})^2$	$n-p$
of individuals about global mean	$\sum_{i,j}(x_{i,j} - \overline{x})^2$	$n-1$

Table 11.4: *Construction of the analysis of variance.*

The last row in Tab. 11.4 is the sum of the first two rows. The ratios of the second column to the third one are unbiased estimates of the variance of the population from which the p samples were drawn. Only the first two (first and second rows) are independent. We can use the Fisher-Snedecor test for equal variances to test the assumption that the p samples are homogeneous.

$$t = \frac{s_1}{s_2} \sim F_{p-1,n-p}$$

We can now finish the previous example.

Example 11.5 *In the previous example we had two experiments: G_1 and G_2 each comprising 3 measurements, hence $p = 2$ and $n = 6$. The two pieces that give independent estimates of the population variance are:*

$$\sum_{i,j}(x_{i,j} - \overline{x_j})^2 = 4$$

$$\sum_j n_j(\overline{x_j} - \overline{x})^2 = 24$$

The sum of the two terms gives correctly $\sum_{i,j}(x_{i,j} - \overline{x})^2 = 28$. The ratio of the two is $F = 6$ which, if the hypothesis that the two samples originate from the same population is distributed at $F_{1,4}$. From the probability tables we can determine the probability: $P(F > 6 \mid F_{1,4}) = 0.07$. Note that we might have computed $y = 1/F = 4/24 = 1/6 = 0.167$ and the probability: $P(y > 0.167 \mid F_{4,1}) = 0.93$. Of course this is the same as the previous result, since now we have obtained the probability $P(y = 1/F > 1/6) = 1 - P(F > 6)$.

Example 11.6 *A nuclear resonance has been measured in four experiment, leading the results shown in Tab. 11.6. We want to test if the various experiments are consistent. We will use the ANOVA setup to measure the variances within each experiment, the overall variance, mixing all experiments and variance among the mean values of each experiment. The analysis is performed as shown in Table 11.6. The ANOVA analysis shows that there is no significant differences (large probability of the Fisher test) between the observations of the resonance mass in the various experiments.*

Exp.	Measurements	d.o.f	$\sum_j x_{i,j}$	$\overline{x_j}$	$\sum_j x_{i,j}^2$
E_1	1.60, 1.61, 1.65, 1.68, 1.70, 1.72, 1.80	7	11.76	1.68	19.7854
E_2	1.58, 1.64, 1.64, 1.70, 1.75,	5	8.31	1.662	13.8281
E_3	1.46, 1.55, 1.60, 1.62, 1.64, 1.66, 1.74, 1.82	8	13.09	1.63625	21.5037
E_4	1.51, 1.52, 1.53, 1.57, 1.60, 1.68	6	9.41	1.56833	14.7787
All experiments				1.6373077	69.8959
$\sum(x_{i,j}-\overline{x})^2 = \sum\ x_{i,j}^2 - n\ \overline{x}^2 =$			0.195711		
$\sum\ n_j\ (\overline{x_j}-\overline{x})^2 = \sum n_j \overline{x_j}^2 - n\overline{x}^2 =$			0.044297	d.o.f = 3	0.014667
Residual			0.15141	d.o.f = 22	0.0068824
$F_{3,22}$=			2.1454		
Probability =			0.2468		

Table 11.5: *Anova analysis for consistency of 4 sets of measurements.*

Chapter 12

General Properties of Estimators

We have already discussed the Goodness of Fit: how to test if a sample has a given *p.d.f.*. Assuming now that the *p.d.f.* is correct (the model is good) we want to specify fully the *p.d.f.*.

We start from a random sample (a limited number of observations). The *p.d.f.* of the parent population depends on unknown parameters. For instance a random process is known to be Poisson distributed, but we ignore its mean.

In data analysis is very common to have a model which depends on unknown parameters that we want to fit to data. We extract from the data the values of the parameters which fits at best the model. The aim of next chapters is to describe the methods to estimate the parameters. This is called in statistics *point estimate*.

The point estimates (the best values of the unknown parameters) is strictly related with the *interval estimation* (the interval which comprises the true value of the parameter with a given probability, also called *Confidence Interval*). We will describe how to compute *Confidence Intervals* for the two main methods that we will use for the point estimation. We will take again the issue at the end of this course to describe the general construction od Confidence Intervals.

Statistics has its own jargon. Here we explain a few important terms.

- Estimator: a function of observation (or a method) used to find the unknown parameter.
- Estimate: the numerical value of the parameters obtained with the estimator.
- Statistics: a function of the random variables (observations) that does not depend on the parameters. The statistics $A = A(x_1, \cdots, x_n)$ is an estimator of an unknown parameter θ. For instance the sample mean:
$$\overline{x} = \frac{\sum x_i}{n}$$
is a statistics and an estimator of the mean μ.

Important properties that estimators should have are:

- Consistency: the estimates converge to the true parameters as the number of observations increases.
- Unbiasedness: for finite random samples the expected value of an estimate is equal to the true parameter:
$$E(t) = \theta$$
if instead $E(t) = \theta + \theta/n$ then the estimator would be consistent but biased.

- Sufficiency: an estimator is said sufficient when it exhausts all the information of the observations regarding the parameter (in the case of one parameter only). For instance in the case of Normal distribution the sample mean is an estimator of the population mean μ. No other knowledge on μ is obtained from other functions of the observations like $\sum x_i^2$ or $\sum |x_i|$ etc. The sample mean is called *sufficient statistics* for μ.

- Minimum variance: it is possible to show that both the sample mean and the sample median are unbiased estimators of the central value of a normal population with known variance. However the variance of the mean is smaller than the variance of the median. The sample mean is considered a better estimator of the central value of the normal distribution. There is a lower limit to the variance of an estimator given by the Cramer-Rao inequality. An estimator attaining this limit is called *efficient estimator*.

12.1 Sufficient Statistics

We have already seen examples of sufficient statistics. In Normal distribution the estimates of the average and of the variance ($\hat{\mu}$ and $\hat{\sigma}^2$) are functions *only* of $\sum x_i$ and of $\sum x_i^2$. Any other function of the measurements like:

$$\sum \mid x_i \mid \qquad \sum x_i^3 \qquad \mid Max(x_i - Min(x_i) \mid$$

do not add any information to th average and variance. We will say that $\sum x_i$ and $\sum x_i^2$ are sufficient statistics. Also any estimator of the parameters which are functions of $\sum x_i$ and of $\sum x_i^2$ are sufficient estimators.

12.2 Recognition of Sufficient Statistics

Assume that $(x_i, i = 1, n)$ are measurements of a variable whose *p.d.f.* is $f(x|\theta)$. θ is a parameter or a set of parameters that we have to estimate.

Let $\hat{\theta} = \hat{\theta}(x)$ be the function of the measurements (x) that we use as estimator of θ. Suppose that the joint probability of the measurement can be factorized as follows:

$$P(x \mid \theta) = \prod f(x_i \mid \theta) = g(x \mid \hat{\theta}) \cdot h(\hat{\theta} \mid \theta) \tag{12.1}$$

g is the conditional probability of x, given $\hat{\theta}$. θ does not appear in g, hence g cannot be used to provide information on θ. The only reference to x is contained in $h(\hat{\theta}, \theta)$ via $\hat{\theta}$. Hence $\hat{\theta}$ must contain all the information on θ. $\hat{\theta}$ is sufficient statistics [46].

Example 12.1 *Consider a sample of measurements of a random variable distributed normally with variance one.*

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

The probability of the measurements $(\vec{x} : x_1, x_2 \cdots x_n)$*:*

$$P(\vec{x} \mid \mu) = \frac{1}{(\sqrt{2\pi})^n}\, e^{-\frac{\sum(x_i-\mu)^2}{2}}$$

Let us call $\hat{\theta} = \sum x_i/n$. *By using the equality:*

$$\sum_i (x_i - \mu)^2 = \sum_i x_i^2 - 2\mu \sum_i x_i + n\mu^2 = \sum_i x_i^2 + n(\mu^2 - 2\mu\hat{\theta})$$

we can rearrange the probability in the form:

$$P(\vec{x} \mid \mu) = \left[\frac{e^{-\sum x_i^2}}{(2\pi)^{n/2}}\right] \cdot \left[e^{-\frac{\mu^2 - 2\mu\hat{\theta}}{2/n}}\right]$$

The two functions in square brackets are such that the first function depends on x_i and the second on the parameter μ and on the statistics $\hat{\theta}$. Hence $\hat{\theta} = \sum x_i/n$ is a sufficient statistics for μ.

Example 12.2 *Let us consider a sample of size n from a Poisson variate:*

$$\{r_1, r_2, \cdots r_n\} \sim \frac{e^{-\mu}\mu^r}{r!}$$

The joint probability of $\{r_i\}$ is:

$$P(r_1, r_2, \cdots r_n) = \prod_i \frac{e^{-\mu}\mu^{r_i}}{r_i!} = \left[\frac{1}{r_1! r_2! \cdots r_n!}\right]\left[e^{-n\mu}\mu^{\sum_i r_i}\right]$$

The statistics $\overline{r} = \sum_i \frac{r_i}{n}$ is sufficient.

Example 12.3 *Let us consider a sample of n from a Cauchy distribution:*

$$\{x_1, x_2 \cdots x_n\} \sim \frac{1}{\pi}\frac{1}{1 + (x - \mu)^2}$$

The joint probability of the $\{x_i\}$ is:

$$P(x_1, x_2 \cdots x_n) = \frac{1}{\pi^n}\prod_i \frac{1}{1 + (x_i - \mu)^2}$$

and there is no way to bring this equation into the form 12.1. The Cauchy p.d.f. does not admit sufficient estimators.

Example 12.4 *The joint distribution of a sample of n from a Binomial distribution is:*

$$P(r_1, r_2 \cdots) = \prod_i \begin{pmatrix} N \\ r_i \end{pmatrix} \cdot p^{r_i} \cdot (1 - p)^{N - r_i} = \left(\prod_i \begin{pmatrix} N \\ r_i \end{pmatrix}\right)\left(p^{\sum_i r_i}(1 - p)^{nN - \sum_i r_i}\right)$$

The statistics $\sum_i r_i$ is a sufficient statistics hence $\hat{p} = \frac{\sum_i r_i}{n}$ is a sufficient estimator of p.

Example 12.5 *Consider a sample of n of a Normal distribution with zero mean and unknown variance: $N(0, \sigma^2)$. The joint distribution of the observation is:*

$$P(x_1, x_2 \cdots x_n) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_1^2}{2\sigma^2}} = \left(\frac{1}{(2\pi)^{n/2}\sigma^n}\right)\left(e^{-\frac{\sum_i x_i^2}{2\sigma^2}}\right)$$

Hence $\sum_i x_i^2$ is a sufficient statistics of σ^2.

It can be shown (*Darmois Theorem*) that a necessary condition for a sufficient statistics to exist is that the *p.d.f.* is of the exponential family:

$$f(x \mid \theta) = e^{B(\theta) \cdot C(x) + D(\theta) + E(x)}$$

The factorization of the likelihood function

$$L(\theta \mid x) = h(\hat{\theta} \mid \theta) \cdot g(x \mid \hat{\theta})$$

means that a sufficient statistics is:

$$t = \sum_i C(x_i)$$

12.3 Lower Bounds on the Variance

Suppose we have n observations of a variable x which has a *p.d.f.* $f(x \mid \theta)$. Assume also that the likelihood $L(\theta \mid x)$ is differentiable at least twice w.r.t. θ and that the range of x is independent of θ. Be $t(\vec{x})$ an estimator of some function of the parameter θ, say $\tau(\theta)$. We have seen that a ML estimator can be biased. Let us define the bias (b) as that function of θ such that:

$$E(t) = \int \cdots \int \; t(x_i) \cdot L(x \mid \theta) \; dx_1 \cdots dx_n = \tau(\theta) + b(\theta)$$

Differentiate w.r.t. θ:

$$\begin{aligned} \int \cdots \int \; t \cdot \frac{\partial}{\partial \theta} L(x \mid \theta) \; dx_1 \cdots dx_n &= \frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta} \\ \int \cdots \int \; t \cdot \frac{\partial}{\partial \theta} \lg(L) \cdot L \; dx_1 \cdots dx_n &= \frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta} \end{aligned}$$

Since L is the joint probability of the n observations:

$$\int \cdots \int \; L(x \mid \theta) \; dx_1 \cdots dx_n = 1$$

Differentiating:

$$\int \cdots \int \frac{\partial}{\partial \theta} \lg(L) \cdot L \; dx_1 \cdots dx_n = 0$$

If we multiply by $\tau(\theta) + b(\theta)$ and subtract to the previous formula we get:

$$\int \cdots \int \; (t - \tau(\theta) - b(\theta)) \cdot \frac{\partial}{\partial \theta} \lg(L) \cdot L \; dx_1 \cdots dx_n = \frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta} \tag{12.2}$$

and, applying the Cauchy-Schwartz[1] inequality:

$$\left(\int \cdots \int \; (t - \tau(\theta) - b(\theta))^2 \cdot L d\vec{x} \right) \cdot \left(\int \cdots \int (\frac{\partial}{\partial \theta} \lg(L))^2 \cdot L \; d\vec{x} \right) \geq (\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta})^2$$

The first integral is the variance and the second is the expectation value of the derivative square of the likelihood:

$$Var(t) \geq \frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta} \right)^2}{E\left(\left(\frac{\partial}{\partial \theta} \lg(L) \right)^2 \right)}$$

This is the Cramer-Rao inequality and provides a lower bound to the variance of an estimator. In the case $t = \theta$ (the function of the estimator is the estimator itself) and for unbiased estimator ($b = 0$) we have:

$$Var(t) \geq \frac{1}{E\left(\left(\frac{\partial}{\partial \theta} \lg(L) \right)^2 \right)}$$

[1]For a and b in a inner product space, we have:

$$\mid (a \cdot b) \mid \leq \mid\mid a \mid\mid \cdot \mid\mid b \mid\mid$$

the equality holds only if and only if a and b are linerly dependent: $a = \alpha x$. Squaring we get:

$$\mid (a \cdot b) \mid^2 \leq \mid (a \cdot a) \mid \cdot \mid (b \cdot b) \mid$$

The term at the denominator is Fisher's information:

$$I_\theta = E\left(\left(\frac{\partial}{\partial\theta}\lg(L)\right)^2\right)$$

To understand its meaning let us consider the case of a Normal variable in a sample of one measurement.

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

For a single measurement

$$\begin{aligned}
I_\theta &= \int f(x \mid \mu)\cdot\left(\frac{\partial}{\partial\mu}\left(-\lg(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}\right)\right)^2 dx \\
&= \int f(x \mid \mu)\frac{(x-\mu)^2}{\sigma^4}\,dx \\
&= \frac{1}{\sigma^2}
\end{aligned}$$

The information is the inverse of the variance. For n measurement the information is n times larger and the Cramer-Rao becomes:

$$Var(\overline{x}) \geq \frac{\sigma^2}{n}$$

This means that the sample average $\overline{x}$ has the lowest possible variance among all the estimators of the mean. In the Statistics jargon $\overline{x}$ is an *efficient estimator*.

Exercise 12.1 *Derive the following relation:*

$$E\left(\left(\frac{\partial\lg(L)}{\partial\theta}\right)^2\right) = E\left(-\frac{\partial^2\lg(L)}{\partial\theta^2}\right)$$

Hint: differentiate w.r.t. θ the identity:

$$\int\cdots\int\frac{\partial L}{\partial\theta}\,d\vec{x} = \int\cdots\int\frac{\partial\lg(L)}{\partial\theta}\,L\,d\vec{x}$$

12.4 Minimum Variance Bound (MVB)

An estimator attaining the Cramer-Rao limit is called MVB estimator or efficient estimator. The condition to attain the MVB is that the Schwartz inequality becomes an equality. The two "vectors" used to built the inequality: $t-\tau(\theta)$ and $\partial(\lg(L))/\partial\theta$, are linearly related:

$$\frac{\partial\lg(L)}{\partial\theta} = A(\theta)\cdot(t-\tau(\theta)-b(\theta)) \qquad (12.3)$$

The minimum variance bound reads:

$$Var(t) = \frac{\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}}{A(\theta)} = MVB$$

and the efficiency of an estimator t is defined as the ratio between the MVB and its variance:

$$Efficiency = \frac{MVB}{Var(t)}$$

In fact from 12.3:

$$\frac{\partial \lg(L)}{\partial \theta} = A(\theta) \cdot (t - \tau(\theta) - b(\theta))$$

Let us multiply by $t - \tau - b$ and take the average on all the observations:

$$\int \cdots \int (t - \tau - b) \frac{\partial \lg L}{\partial \theta} L \prod_i dx_i = Var(t) A(\theta) = \frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}$$

Hence:

$$Var(t) = \frac{\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}}{A(\theta)}$$

Example 12.6 *In a sample of n from a Normal distribution with unit variance and unknown mean the joint probability of the n observations is:*

$$P(x_1, \cdots x_n) = \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2}} =$$

Therefore:

$$\frac{\partial \lg L}{\partial \mu} = \frac{\partial}{\partial \mu} \left[-\frac{1}{2} \sum_i (x_i - \mu)^2 \right] = -n(\overline{x} - \mu)$$

Hence $\overline{x}$ is a MVB of μ with variance $\frac{1}{n}$.

Example 12.7 *In a sample of n from a Poisson process the joint probability of the observations is:*

$$P(r_1, \cdots x_n) = \prod_i \frac{e^{-\mu} \mu^{r_i}}{r_i!}$$

And the the derivative with respect to μ of the logarithm is:

$$\frac{\partial \lg L}{\partial \mu} = \frac{n}{\mu} (\overline{r} - \mu)$$

And we have proved that $\overline{r}$ is a MVB estimator of μ with variance $\frac{\mu}{n}$.

Example 12.8 *In a sample of n from a Cauchy distribution we have:*

$$P(x_1, \cdots x_n) = \frac{1}{\mu} \frac{1}{1 + (x - \mu)^2}$$

We have:

$$\frac{\partial \lg L}{\partial \mu} = \frac{\partial}{\partial \mu} \lg(\sum_i \frac{1}{1 + (x_1) - \mu)^2}) = \sum_i \frac{x_i - \mu}{1 + (x_i - \mu)^2}$$

There is no way to recover the form 12.3. This distribution does not have an efficient estimator. We could have anticipated it, since not even sufficient estimators exist for a Cauchy distribution.

Example 12.9 *In a single measurement r of a Binomial distribution, the likelihood is:*

$$L = \left(\begin{array}{c} N \\ r \end{array} \right) p^r (1 - p)^{N - r}$$

Hence:

$$\frac{\partial \lg L}{\partial p} = \frac{r}{p} - \frac{N - r}{1 - p} = \frac{N}{p(1 - p)} (\frac{r}{N} - p)$$

We have proved that the statistics $t = \frac{r}{N}$ is a MVB of p with variance $\frac{p(1-p)}{N}$.

Example 12.10 *In a sample of n from a Normal density with zero mean and unknown variance, the likelihood is:*

$$L = \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{-\frac{\sum_i x_i^2}{2\sigma^2}}$$

The derivative with respect to σ of the logarithm is:

$$\frac{\partial \lg L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_i x_i^2}{\sigma^3} = \frac{n}{\sigma^3}(\frac{\sum_i x_i^2}{n} - \sigma^2)$$

This shows that $\frac{\sum_i x_i^2}{n}$ is a MVB estimator of σ^2 with variance $\frac{\sigma^3}{n}\frac{\partial \sigma^2}{\partial \sigma} = \frac{2\sigma^4}{n}$. There is no MVB estimator of σ.

Exercise 12.2 *Compute in the limit of large n the efficiency of the median.*

Exercise 12.3 *Is the weighted mean efficient?*

Exercise 12.4 *Compute the variance of the weighted mean.*

Exercise 12.5 *Show that in the case of the exponential distribution ($f(t \mid \tau) = \frac{1}{\tau}e^{t/\tau}$), the ML estimate of τ is efficient while the ML estimate of $\lambda = \frac{1}{\tau}$ is not efficient.*

12.5 An Efficient Estimator is always Sufficient

The condition for sufficient statistics is:

$$L(\theta \mid x) = g(x \mid \hat{\theta}) \cdot h(\hat{\theta} \mid \theta)$$

Take the logarithm and derive with respect to θ:

$$\frac{\partial}{\partial \theta} \lg(L) = \frac{\partial}{\partial \theta} \lg(h(\hat{\theta} \mid \theta))$$

We recognize that 12.3 is a particular instance of the more general previous equation: an efficent estimator is always sufficient. The converse is not true: a *p.d.f.* can admit several sufficient estimators of the same parameter but only one is efficient.

Chapter 13

The Maximum of Likelihood Method

Let us use an example to illustrate the meaning of this method which is more modern than the Least Square one and is due to R. Fisher.

13.1 Discrete Distributions

Two urns have four balls with different percentage of black and white balls:

- Urn A $\frac{NumberW}{NumberB} = \frac{1}{3}$
- Urn B $\frac{NumberW}{NumberB} = \frac{3}{1}$

One urn is chosen at random, then 3 balls are drawn from the urns (with replacement). The probability to get r black balls follows the binomial law:

$$P(r) = \left(\begin{array}{c} 3 \\ r \end{array} \right) p^r \, (1-p)^{3-r}$$

Where $p = 3/4$ in case of urn A and $p = 1/4$ for urn B.

We intend to estimate p on the basis of the drawn balls. The probability, as a function of r for each urn can be easily computed and the results are shown in Tab.13.1 What we have

	$r = 0$	1	2	3
P(r; p=3/4)	1/64	9/64	27/64	27/64
P(r; p=1/4)	27/64	27/64	9/64	1/64

Table 13.1: *Probability of drawing r black balls, under the two different hypothesis (see the text).*

computed are *a-priory* probabilities i.e. before making any observation. After we have made the observation and measured r black balls, we have to argue on which is the value of p.

Suppose we do not observe any black ball ($r = 0$), then we can argue that the results are 27 times more likely to have occurred if $p = 1/4$ than in the other case. The opposite is true if we would have drawn $r = 3$ black balls. The argument is less compelling in the case we have drawn r=1 or 2 black balls.

Maximum Likelihood Principle: *we are to choose that hypothesis which gives the greatest probability of the observed events.*

Assume an urn contains a completely unknown mixture of white and black balls. The probability p can have any value between 0 and 1. Once we have made the observations and measured r black balls, the function:

$$L(p \mid r) = \begin{pmatrix} 3 \\ r \end{pmatrix} p^r \, (1-p)^{3-r}$$

is not any longer the *p.d.f.* of r, since now r has been measured, it is a function of p called the Likelihood of p.

If we draw one black ball ($r = 1$) then: $L(p \mid 1) = 3p(1-p)^2$. The value of p which is the most likely to have given rise to $r = 1$ can be found by maximizing $L(p \mid 1)$:

$$\frac{\partial L}{\partial p} = 3\,(1-p)^2 - 6\,p\,(1-p) = 0 \qquad\qquad \hat{p} = \frac{1}{3}$$

Comments:

- Any constant factor is irrelevant.
- An alternative form which is easier to handle and is frequently used in the practice is *Log L*.
- We have written $L(p \mid r)$ to stress that it is the conditional probability of p *given* r. In the case of *p.d.f.* we have written $P(r \mid p)$, the probability of having r as the result of a random process whose probability of success is p.

The previous example can be generalized to the case we draw r black balls out of n. The best estimate of p becomes:

$$\hat{p} = \frac{r}{n}$$

13.2 Continuous Distributions

If a variable x has a continuous *p.d.f.* $f(x|\theta)$ where θ is a parameter, then the distribution of a sample of size n (n measurements are distributed as) is:

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1,n} f(x_i \mid \theta)$$

When the measurement is made and the x_i have been observed (θ are unknown) the likelihood of θ is:

$$\begin{aligned} L(\theta \mid x) &= \prod_{i=1,n} f(x_i \mid \theta) \\ Log\ L &= \sum_{i=1,n} Log\ f(x_i \mid \theta) \end{aligned}$$

13.3 ML in the Case of Normal Variables

13.3.1 ML estimate of the mean

We have made n measurements of a variable x. The instrument that we have used has a precision (variance) known to be 1 and the measurements are distributed normally ($x \sim N(\mu, 1)$).

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

Omitting irrelevant factors:

$$L(\mu \mid x) = e^{-\sum \frac{(x_i-\mu)^2}{2}}$$

Notice again the form $f(x \mid \mu)$ and $l(\mu \mid x)$ to make it clear that in the *p.d.f.*, x is the result of our assumption on the value of the average, while μ is the unknown in l.

The logarithm of l:

$$\lg(L(\mu \mid x)) = -\sum_{i=1,n} \frac{(x_i-\mu)^2}{2}$$

L is maximized w.r.t. μ when:

$$\hat{\mu} = \frac{\sum_{i=1,n} x_i}{n} = \overline{x}$$

The likelihood estimate of the mean in a Normal distribution is the arithmetic average of the measurements.

13.3.2 ML Estimate of the Mean and of the Variance in the case of equal errors

Assume now that both the variance and the average are unknown:

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

The likelihood is (omit irrelevant factors):

$$L(\mu, \sigma^2 \mid x) = \frac{e^{-\frac{\sum_{i=1,n}(x_i-\mu)^2}{2\sigma^2}}}{\sigma^n}$$

and its log becomes:

$$\lg(L) = -\frac{n\ log\sigma^2}{2} - \frac{\sum (x_i-\mu)^2}{2\ \sigma^2}$$

Minimizing with respect to μ and σ^2 we get:

$$\begin{cases} \frac{\partial \lg L}{\partial \mu} &= \frac{\sum(x_i-\mu)}{\sigma^2} = 0 \\ \frac{\partial \lg L}{\partial \sigma^2} &= -\frac{n}{2\ \sigma^2} + \frac{\sum(x_i-\mu)^2}{2\ \sigma^4} = 0 \end{cases} \qquad \begin{cases} \hat{\mu} &= \frac{\sum x_i}{n} \\ \widehat{\sigma^2} &= \frac{\sum_{i=1,n}(x_i-\hat{\mu})^2}{n} \end{cases}$$

The maximum likelihood best estimate for the variance is biased. The correct normalization would be $1/(n-1)$ instead of $1/n$, since, as we already have shown, one D.o.F. is lost in the calculation of the unknown average.

13.3.3 ML estimate of mean in case of different errors

If the n measurements are Normally distributed but with different variance (different measurement errors):

$$x_i \sim \frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{-\frac{x_i-\mu}{2\sigma_i^2}}$$

then the likelihood function becomes:

$$L = \frac{1}{(\sqrt{2\pi})^n \prod \sigma_i} \cdot e^{-\frac{1}{2}\sum \frac{(x_i-\mu)^2}{\sigma_i^2}}$$

Its log is:

$$\lg L = -\sum \lg(\sigma_i) - \frac{n}{2}\lg 2\pi - \frac{1}{2}\sum \frac{(x_i-\mu)^2}{\sigma_i^2}$$

The ML estimate of μ is obtained by equating to zero the derivative of the previous expression.

$$\hat{\mu} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

This is called the *weighted mean.* The observation are weighted by the inverse of their variance.

Exercise 13.1 *Shows that the variance of the weigthed mean is:* $Var(\hat{\mu}) = \frac{1}{\sum \frac{1}{\sigma_i^2}}$ *(Hint: use the expression for the variance:* $Var(\mu) = S\ V_x\ S^T$ *with* S *a vector whose components are* $S_i = \frac{\partial \mu}{\partial x_i}$*)*

13.4 ML estimate of the mean in case of exponential distribution

Consider the problem of estimating the mean lifetime of an unstable particle. We have measured n decay times t_i and we know that the *p.d.f.* is

$$f(t_|\tau) = \frac{1}{\tau}e^{-\frac{t_i}{\tau}}$$

The likelihood is:

$$l(\tau \mid \vec{t}) = \prod_i \frac{1}{\tau}\, e^{-\frac{t_i}{\tau}} = \frac{1}{\tau^n}e^{-\frac{\sum t_i}{\tau}}$$

The ML estimate ($\hat{\tau}$) is found by solving the equation

$$\frac{\partial}{\partial \tau}\log l = \frac{\partial}{\partial \tau}(\sum_i(-\log\tau - \frac{t_i}{\tau}))$$

The solution is:

$$\hat{\tau} = \frac{\sum_i t_i}{n} = \bar{t}$$

Exercise 13.2 *Use, instead of lifetime* τ*, its inverse* $\lambda = 1/\tau$*. Use the p.d.f.* $f(t|\lambda)$ *and compute the maximum likelihood estimate of* λ*. Verify that* $\hat{\lambda} = 1/\hat{\tau}$*. This property, the invariance, is of general validity for Maximum Likelihood estimators.*

The variance of $\hat{\tau}$ can be computed directly by:

$$Var(\hat{\tau}) = \int_0^\infty \cdots \int_0^\infty (\hat{\tau}-\tau)^2 \prod_i f(t_i \mid \tau)dt_1\cdots dt_n$$

Note that the term $\prod f(t_i \mid \tau)$ is the joint probability of the n measurements and not the Likelihood.

$$\begin{aligned} Var(\hat{\tau}) &= \int_0^\infty \cdots \int_0^\infty (\hat{\tau}^2 + \tau^2 - 2\hat{\tau}\tau) \prod_{i=1,n} \frac{e^{-\frac{t_i}{\tau}}}{\tau} dt_1\cdots dt_n \\ &= \int \cdots \int [(\sum_i \frac{t_i}{n})(\sum_j \frac{t_j}{n}) - 2\tau\sum_i \frac{t_i}{n} + \tau^2] \prod_i \frac{e^{-t_i/\tau}}{\tau} dt_1 \cdots dt_n \end{aligned}$$

The first integral is made of $n(n-1)$ terms of the form:

$$\int \cdots \int \frac{t_i t_j}{n^2} e^{-t_i/\tau} e^{-t_j/\tau} \frac{1}{\tau^2} dt_i dt_j = \tau^2 \frac{1}{n^2}$$

and of n terms:

$$\int \frac{t_i^2}{n^2} \frac{e^{-t_i/\tau}}{\tau} dt_i = \frac{2\tau^2}{n^2}$$

The second integral has n terms:

$$-2\frac{\tau}{n} \int_i \frac{e^{-t_i/\tau}}{\tau} t_i \; dt_i = -2\frac{\tau^2}{n}$$

Finally:

$$Var(\hat{\tau}) = (\frac{2}{n}\tau^2 + \frac{n-1}{n}\tau^2) - 2\tau^2 + \tau^2 = \frac{\tau^2}{n}$$

13.5 ML Estimate in the Case of an Uniform Distribution

This example illustrates another interesting concept. Assume we have a sample of measurements $(x_i, i = 1, n)$ of a variable which is known to have a uniform *p.d.f.*:

$$f(x|a,b) = \frac{1}{b-a} \qquad a \leq x \leq b$$

We arrange the sample in ascending order:

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$$

We must have $a \leq x_1$ and $b \geq x_n$. This sets a boundary on the region of the (a, b) space allowed to the solutions of the problem. The allowed region is hashed in Figure 13.1 The likelihood function is written:

$$L(a, b \mid x) = (\frac{1}{b-a})^n$$

In this case the Max Likelihood solution cannot be found by differentiation. The lines of constant l are straight lines $b = a+k$ in the plane (a, b), as shown in Fig.13.1. The value $k = 0$ corresponds an infinite value of the likelihood. The minimum value of L, compatible with the boundaries $a \leq x_1$ and $b \geq x_n$ is given by the line that passes through the lower right angle of the allowed region.

$$\hat{a} = x_1 \qquad \hat{b} = x_n$$

All the other measurements *do not* contribute further information to the estimate of a and b. x_1 and x_n are said to be *sufficient* for a and b.

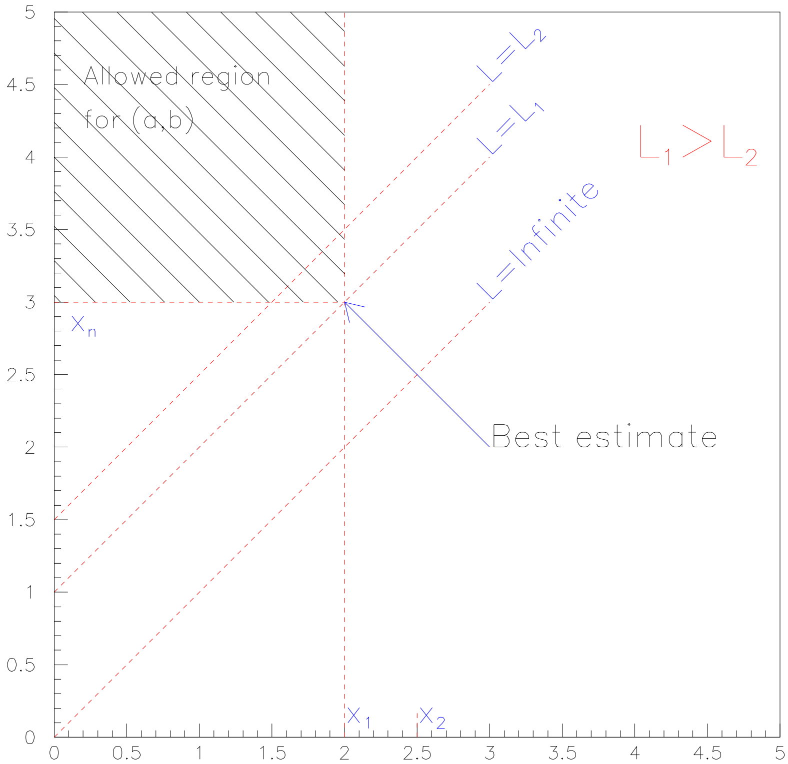


Figure 13.1: *ML estimate of a and b in case of uniform distribution.*

Chapter 14

Properties of ML Estimators

We will discuss, in this chapter, the properties of Maximum Likelihood estimators, not only the sufficiency and the efficiency of the estimator found by the Likelihood method, but also another property which makes this method particularly useful: the *invariance*.

The following section will shortly summarize the properties of ML estimators. We distinguish between *small sample*, i.e. properties which are valid even for a small sample and *large sample*, i.e. properties only asymptotically.

14.1 Sufficiency (small sample)

Not all the *p.d.f.* admit a sufficient estimator, as we have discussed above. If a sufficient estimator exists, say t, then we can factorize the likelihood function as:

$$L(\theta \mid x) = g(x \mid t) \cdot h(t \mid \theta)$$

In the ML method we take the logarithm of L and derive with respect to θ.

$$\frac{\partial \lg L}{\partial \theta} = \frac{\partial \lg h(t \mid \theta)}{\partial \theta} = 0$$

From the previous formula we see that the ML estimation of θ depends on the data through the sufficient statistics t only, if a sufficient estimator exists: if a sufficient statistics exists then the ML method will produce it by solving the previous equation.

14.2 Efficiency (small sample)

Theorem 14.1 *If an efficient estimator exists, the the ML method will produce it.*

Proof 14.1 *The necessary and sufficient condition for an efficient estimator to exist is:*

$$\frac{\partial}{\partial \theta} \lg(L) = A(\theta)(t - \theta - b(\theta))$$

The ML estimator is obtained maximizing the logarithm of the likelihood, i.e. the previous expression is equal to zero when $\theta = \hat{\theta}$. Hence $t = \hat{\theta}$: if there is an efficient estimator, the ML method will produce it. ∎

14.3 Efficiency (asymptotic)

The ML estimators are asymptotically efficient (given without proof). This and the previous property makes the ML method very powerful.

14.4 Uniqueness

Theorem 14.2 *If an efficient estimator exists for some function $\tau(\theta)$, then the ML solution is unique*

Proof 14.2 *If t is an efficient estimator, then:*

$$\frac{\partial \lg L}{\partial \theta} = A(\theta)(t - \tau(\theta) - b(\theta))$$

the second derivative at $\theta = \hat{\theta}$ is:

$$\frac{\partial^2}{\partial \theta^2} \lg(L)\bigg|_{\theta=\hat{\theta}} = \frac{\partial A}{\partial \theta} \cdot (t - \tau(\theta) - b(\theta))\bigg|_{\theta=\hat{\theta}} - A(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta})\bigg|_{\theta=\hat{\theta}}$$

The first term is zero, and:

$$\frac{\partial^2}{\partial \theta^2} \lg(L) \mid_{\theta=\hat{\theta}} = -A^2(\hat{\theta}) \cdot Var(t) \leq 0$$

Every ML solution ($\partial \lg L/\partial \theta = 0$) corresponds to a maximum. Since a regular function must have minima between maxima, the solution is unique. ∎

Unfortunately this property does not hold if the estimator is not efficient; we will describe two simple cases.

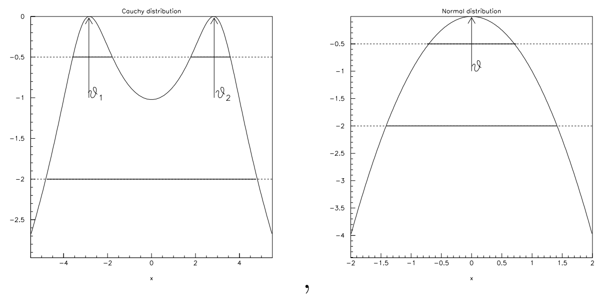


Figure 14.1: *Shape of* $\lg L$ *function, see the text.*

Example 14.1 *Let us assume that we have made two observations of a variable which is distributed as Cauchy (this p.d.f. does not have sufficient statistics for μ:*

$$x \sim \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}$$

The two measurements have yielded the results: $x = -3.\ 3.$. The values of the ML estimator for μ are found maximizing:

$$\lg L = \lg \frac{1}{1+(-3-\mu)^2} + \lg \frac{1}{1+(3-\mu)^2}$$

The plot of the function is shown in Figure 14.1 (left). The maximum is not unique, there are two maxima, symmetric about zero. As we will discuss later, also the 1σ intervals are disconnected. In cases like these, it is appropriate to publish the shape of the likelihood function more than the maximum itself, even if one of the two maxima would be larger that all the others.

Example 14.2 *Assume now that the same two observations would be made from a Normal population with unit variance.*

$$x \sim \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The $\lg L$ *becomes:*

$$\begin{aligned} \lg L &= -\frac{\sum_i (x_i - \mu)^2}{2} \\ &= -\frac{1}{2}\left(n(\overline{x} - \mu)^2 + \sum (x_i - \overline{x})^2\right) \end{aligned}$$

The plot of the function is shown in Figure 14.1 (right). The shape of the function is parabolic and the maximum is unique.

14.5 Normality (asymptotic)

We will prove that a ML estimator will be distributed asymptotically as $N(\theta_0, \sigma_t^2)$ (θ_0 is the true value of the parameter) where:

$$\sigma_t^2 = \frac{1}{E(-\frac{\partial^2}{\partial\theta^2}\lg(L))}$$

Assume we have n measurements of a variable with *p.d.f.* $f(x \mid \theta)$: $(x_i, i = 1 \cdots n)$. Let us Taylor expand about the minimum:

$$0 = \left.\frac{\partial \lg(L)}{\partial\theta}\right|_{\hat\theta} = \left.\frac{\partial}{\partial\theta}\lg(L)\right|_{\theta_0} + (\hat\theta - \theta_0)\cdot\left.\frac{\partial^2}{\partial\theta^2}\lg(L)\right|_{\theta'}$$

Where $\hat\theta$ is the ML estimate, θ_0 the true value and θ' a value $\hat\theta \le \theta' \le \theta_0$. From:

$$\int\cdots\int L\, d\vec{x} = 1$$

differentiating:

$$\int\cdots\int \frac{\partial L}{\partial\theta}\, d\vec{x} = 0 = \int\cdots\int \frac{\partial}{\partial\theta}\lg(L) L d\vec{x}$$

Hence:

$$E\left(\frac{\partial}{\partial\theta}\lg(L)\right) = 0$$

Differentiating again:

$$0 = \int\cdots\int \frac{\partial^2 L}{\partial\theta^2} d\vec{x} = \int\cdots\int \frac{\partial^2}{\partial\theta^2}\lg(L)\ L\ d\vec{x} + \int\cdots\int \left(\frac{\partial}{\partial\theta}\lg(L)\right)^2\ L\ d\vec{x}$$

Hence:

$$\int\ldots\int L\frac{\partial^2}{\partial\theta^2}\lg(L) d\vec{x} = -\int\ldots\int \left(\frac{\partial}{\partial\theta}\lg(L)\right)^2\ L\ d\vec{x}$$

We can now compute the variance:

$$Var\left(\frac{\partial}{\partial\theta}\lg(L)\right) = E\left((\frac{\partial}{\partial\theta}\lg(L))^2\right) - \left(E\left(\frac{\partial}{\partial\theta}\lg(L)\right)\right)^2$$

or:

$$Var\left(\frac{\partial}{\partial\theta}\lg(L)\right) = E\left(-\frac{\partial^2}{\partial\theta^2}\lg(L)\right)$$

If now we write

$$\left.\frac{\partial}{\partial\theta}\lg(L)\right|_{\theta_0} = \sum_i \left.\frac{\partial}{\partial\theta}\lg(f(x_i \mid \theta)\right|_{\theta_0}$$

the r.h.s. is a sum of *independent* random variables which have 0 mean and variance given by the previous expression. Hence from the central limit theorem the quantity:

$$u = \left.\frac{\frac{\partial}{\partial\theta}\lg(L)}{\sqrt{E(-\frac{\partial^2}{\partial\theta^2}\lg(L))}}\right|_{\theta_0} \sim N(0,1)$$

will be asymptotically normal.

Let now go back to the Taylor expansion. Because of the consistency of the ML estimators, $\hat{\theta}$ converges to θ_0 and so does θ' which is bracketed between the two. Because of the law of Large Numbers:

$$\left.\frac{\partial^2}{\partial\theta^2}\lg(L)\right|_{\theta'} = \sum_i \left.\left(\frac{\partial^2}{\partial\theta^2}\lg(f(x_i \mid \theta))\right)\right|_{\theta'} \rightarrow_{n\rightarrow\infty} E\left(\frac{\partial^2}{\partial\theta^2}\lg(L)\right)$$

Replacing the asymptotic values in the Taylor expansion we get:

$$(\hat{\theta}-\theta_0)\cdot\sqrt{E\left(-\frac{\partial^2}{\partial\theta^2}\lg(L)\right)} = \left.\frac{\frac{\partial}{\partial\theta}\lg(L)}{\sqrt{E\left(-\frac{\partial^2}{\partial\theta^2}\lg(L)\right)}}\right|_{\theta_0} = u \sim N(0,1)$$

This proves that the ML estimators are asymptotically Normally distributed:

$$\hat{\theta} \sim N(\theta_0, \sigma^2_{\hat{\theta}}) \qquad \sigma^2_{\hat{\theta}} = \frac{1}{E(-\frac{\partial^2}{\partial\theta^2}\lg(L))}$$

14.6 Invariance

Let us first prove a lemma.

Lemma 14.1 *Let $g(\mu)$ be a one-to-one function of μ. The likelihood of*

$$g_0 = g(\mu_0)$$

is the same as the likelihood of μ.

Proof 14.3 *The proof is based on the very definition of the likelihood: the probability of the observations $\vec{x}$ is completely defined once we specify the value of the parameter μ_0. It is irrelevant whether we specify μ_0 or $g_0 = g(\mu_0)$ since one can be obtained from the other without ambiguity. Hence:*

$$P(\vec{x} \mid \mu_0) = P(\vec{x} \mid g_0)$$

But, once we have made the measurements:

$$L(\mu_0 \mid \vec{x}) = P(\vec{x} \mid \mu_0)$$

Hence:

$$L(g_0 \mid \vec{x}) = L(\mu_0 \mid \vec{x})$$

∎

As a consequence we have the invariance Theorem:

Theorem 14.3 *if $\hat{\mu}$ is the ML estimate of μ, then the ML estimate of a function of μ ($g(\mu)$) is:*

$$\hat{g} = g(\hat{\mu})$$

Proof 14.4 *The proof follows from the previous lemma: Be $g' = g(\hat{\mu})$ then, from the previous lemma it follows that:*

$$L(g') = L(\hat{\mu}) \geq L(\mu)$$

for all μ by definition of ML estimate.

But since $L(\mu) = L(g)$ for all μ it follows that

$$L(g') \geq L(g)$$

for all μ and g. Therefore $g' = \hat{g}$. ∎

Example 14.3 *Let x be a variable $\sim N(0,1)$ and $g(\mu) = \mu^3$. Then:*

$$f(x \mid \mu_0) = C \cdot e^{-\frac{(x-\mu_0)^2}{2}} = L(\mu_0 \mid x)$$

and:

$$f'(x \mid g_0) = C \cdot e^{-\frac{(x-g_0^{1/3})}{2}} = f(x \mid \mu_0) = L(g_0 \mid x) = L(\mu_0 \mid x)$$

The reason for such a simple form is the relation between the likelihood and probability density. The probability of x is the same whether the parameter is given directly as μ_0 or in the implicit form $g(\mu_0)$.

Example 14.4 *Maximum likelihood estimators are invariant under the transformation of the parameter. If we make the transformation:*

$$g = g(\theta)$$

then the max-likelihood estimate of g is also the transformed of the max-likelihood estimate of θ:

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

We will show with an example how this property is related to the biasness of ML estimators. The ML estimator of the variance is:

$$\hat{s}^2 = \frac{\sum_{i=1,n}(x_i - \overline{x})^2}{n}$$

and we know from the previous theorems that the ML estimator of s is $\sqrt{\hat{s}^2}$. Both are biased, in fact the unbiased estimators are:

$$\begin{aligned} \hat{s^2} &= \frac{\sum_{i=1,n}(x_i - \overline{x})^2}{n-1} \\ \hat{s} &= \frac{(\frac{n-3}{2})!}{\sqrt{2} \cdot (\frac{n-2}{2})!} \cdot \sqrt{\sum_{i=1,n}(x_i - \overline{x})^2} \end{aligned}$$

Exercise 14.1 *Prove the previous relation. First transform the variable $u \sim \chi^2_n$ to $v = \sqrt{u}$ and verify that:*

$$v \sim \frac{1}{2^{\frac{n}{2}-1} \cdot \Gamma(\frac{n}{2})} \cdot v^{n-1} \cdot e^{-\frac{v^2}{2}}$$

Then show that the moments are:

$$\mu'_k = 2^{\frac{k}{2}} \cdot \frac{\Gamma(\frac{n+k}{2})}{\Gamma(\frac{n}{2})}$$

Find the distribution of the sample variable

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

(x is Normal variable with mean μ and variance σ^2) and verify that

$$s = \frac{(\frac{n-3}{2})!}{\sqrt{2} \cdot (\frac{n-2}{2})!} \cdot \sqrt{\sum_{i=1,n}(x_i - \overline{x})^2}$$

is an unbiased estimator of the standard deviation.

14.7 Variance in a Large Sample

A ML estimator is asymptotically efficient. In this case (efficient estimator) the expression for the variance can be simplified. In fact the efficiency condition

$$\frac{\partial\ \lg(L)}{\partial\ \theta} = A(\theta)\ (t - \theta - b(\theta))$$

can be used to compute the expectation value:

$$E(-\frac{\partial^2\ \lg(L)}{\partial\ \theta^2}) = -\int \frac{\partial}{\partial\ \theta}(\frac{\partial\ \lg(L)}{\partial\ \theta})\ L\ d\vec{x} = -\int L(\frac{\partial\ A}{\partial\ \theta}(t - \theta - b(\theta)) - A(\theta)(1 + \frac{\partial\ b}{\partial\ \theta}))d\vec{x}$$

Now the first part of the integral vanishes since $E(t) = \theta + b(\theta)$ by definition. Therefore:

$$E(-\frac{\partial^2\ \lg(L)}{\partial\ \theta^2}) = A(\theta)(1 + \frac{\partial\ b}{\partial\ \theta}) = -\frac{\partial^2\ \lg(L)}{\partial\ \theta^2}\ |_{\theta=\hat{\theta}}$$

for an efficient estimator. Clearly if the likelihood function is Normal, the second derivative is constant and equal to $1/Var(\theta)$.

The expression of the variance can be approximated in a convenient way in the limit of large number of observations. If f is the $p.d.f.$ of x, then:

$$-\frac{\partial^2\ \lg(L)}{\partial\ \theta^2} = -\sum_{i=1,n} \frac{\partial^2\ \lg(f(x_i;\theta)}{\partial\ \theta^2}$$

The expectation value of this expression is:

$$E(-\frac{\partial^2\ \lg(L)}{\partial\ \theta^2}) = -\sum_{i=1,n} \int \frac{\partial^2\ \lg(f)}{\partial\ \theta^2}\ f\ dx \approx -n\ \int \frac{\partial^2\ \lg(f)}{\partial\ \theta^2}\ f\ dx$$

If the integration limits are independent of θ we can rearrange the expression as:

$$\frac{\partial^2\ \lg(f)}{\partial\ \theta^2} = \frac{\partial}{\partial\theta}(\frac{1}{f}\frac{\partial\ f}{\partial\ \theta}) = -\frac{1}{f^2}(\frac{\partial\ f}{\partial\ \theta})^2 + \frac{1}{f}\frac{\partial^2\ f}{\partial\ \theta^2}$$

Substituting in the integral, the last term vanishes:

$$E(-\frac{\partial^2\ \lg(L)}{\partial\ \theta^2}) = n\ \int \frac{1}{f}(\frac{\partial\ f}{\partial\ \theta})^2\ dx - n\ \frac{\partial^2}{\partial\ \theta^2} \int\ f\ dx = n\ \int \frac{1}{f}(\frac{\partial\ f}{\partial\ \theta})^2\ dx$$

In conclusion:

$$\frac{1}{Var(\hat{\theta})} = n\ \int \frac{1}{f}(\frac{\partial\ f}{\partial\ \theta})^2\ dx$$

This expression for the variance can be used to evaluate the precision of a measurement prior to the measurement, i.e. in the planning phase of the experiment.

Chapter 15

Confidence Intervals of ML Estimates

In the previous chapter we have shown a general method to compute a Maximum Likelihood estimator of a parameter. The estimator is a random variable, since it is a function of random variables. Its value, given the experiment that we have done does not coincide with the parameter but we assume that it will be *near* the value of the parameter. If we could repeat the experiment several times we would remark that the estimates cluster around the value of the parameter (in absence of bias). The spread of this cluster is related to the precision of our measurement of the parameter.

In this chapter we will study how we can determine this precision. The method consists in defining a *Confidence Level (CL)* and then define a *confidence interval (CI)*. The interval is itself a function of the observations hence a random variable and it will vary from experiment to experiment. This interval must be constructed in such a way that, in repeated experiments, the true value of the parameter is included in the interval a fraction CL of times. Of course the determination of the interval must be based on the observations only, since the parameter has an unknown value.

What we have outlined above is the general scheme that we will elaborate in the following chapters. In this chapter we want to study how we can construct an interval for the specific case of Likelihood estimators, exploiting some of the properties that we have studied. We will see that there is a simple and elegant, albeit often approximate, method to construct such an interval.

15.1 Likelihood interval for Normal variables

Let us consider again the case of a Normal distribution with known variance (one) and unknown mean. In a sample of size n:

$$L(\mu \mid x) \propto e^{-\frac{\sum(x_i-\mu)^2}{2}}$$

Its Log becomes:

$$\lg(L(\mu \mid x)) = -\frac{\sum(x_i-\mu)^2}{2}$$

We have seen that it is possible to add any constant (w.r.t. μ) without changing its statistical properties. If we add: $\frac{\sum x_i^2 - n\overline{x}}{2}$ to the left side, we get:

$$\lg(L(\mu \mid x)) = -n \cdot \frac{(\overline{x}-\mu)^2}{2}$$

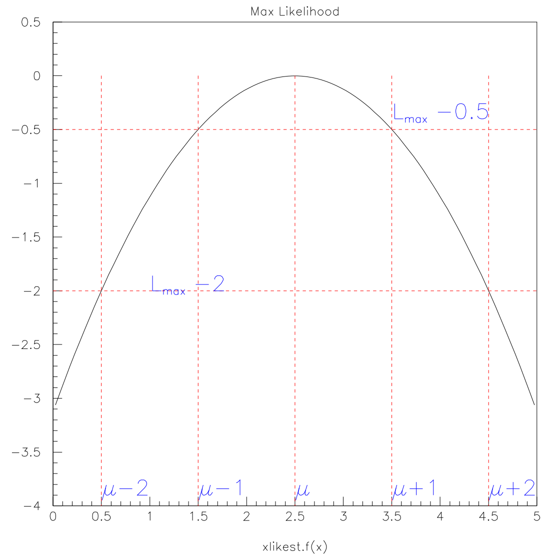


Figure 15.1: *Dependence of L on μ.*

Fig.15.1 shows L as a function of μ. The likelihood is continuous as a function of μ and has a single local maximum at $\mu = \overline{x}$. A line drawn at $L = -1/2$ encloses an interval:

$$\{\overline{x} - 1/\sqrt{n}, \quad \overline{x} + 1/\sqrt{n}\}$$

(because of $(\overline{x} - \mu)^2 = 1/2$). The probability content is known from the normal distribution:

$$P(\mid \overline{x} - \mu \mid \leq \frac{1}{\sqrt{n}}) = P(\overline{x} - \frac{1}{\sqrt{n}} \leq \mu \leq \overline{x} - \frac{1}{\sqrt{n}}) \approx 0.683$$

(Of course the probability statement concerns only the interval and not the value of μ.)

In the same way a line drawn at $L = -2$ provides a ≈ 0.95 CI. (The probability statement concerns $\overline{x}$ and not μ which is a constant.)

15.2 Confidence Intervals in the general case

Assume that $l(\theta \mid y)$ is a continuous function of θ with only one maximum in the allowed range. We can obtain the confidence interval for θ by copying the procedure used for the normal distribution:

- Add a constant to $\log Log(l(\theta \mid y))$ $(= L_\theta(\theta \mid y))$ such that $Max(\log L_\theta(\theta \mid y)) = 0$.
- Draw a line across the plot at $\log L_\theta(\theta \mid y) = -1/2$ and read from the plot the 68% confidence interval. ($L = -2$ for 95% confidence interval).

Let us assume that a one-to-one change of variable exists ($g = g(\theta)$) which transforms the curve $\log L_\theta(\theta \mid y)$ into a parabola:

$$\log L_g(g(\theta) \mid y) = -(g - y')^2/2$$

According to the likelihood principle we must make the same inference about g as we made about θ. The point estimate is $\hat{g} = y'$. We draw a line at $\log L_g = -2$ to obtain the 95% confidence interval:

$$\hat{g} - 2 < g < \hat{g} + 2$$

Then we transform back to obtain the 95% confidence interval:

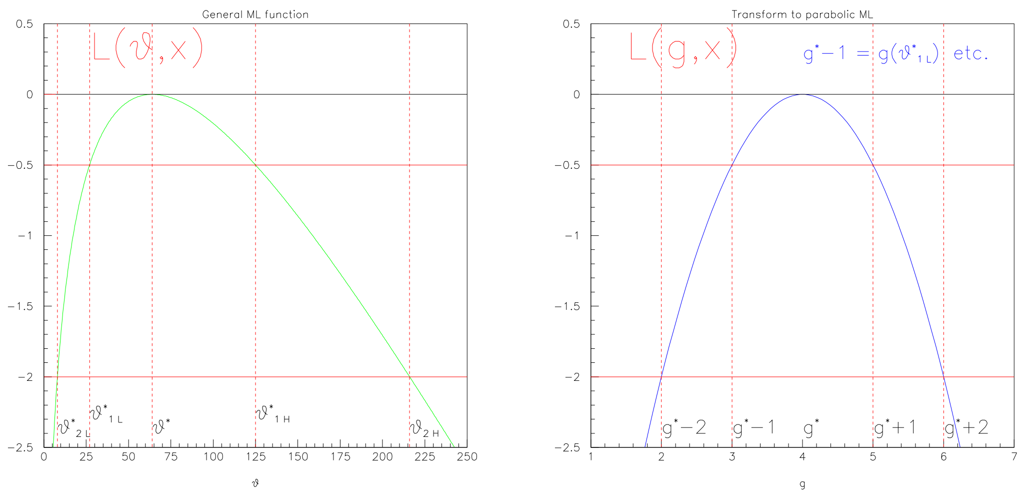
$$\theta_1 = g^{-1}(\hat{g} - 2) \qquad \theta_2 = g^{-1}(\hat{g} + 2)$$

All this is unnecessary since we can read the values of θ_1 and θ_2 directly off the plot of $\log L_\theta$. In fact the values g_1 and g_2 are such that $\log L_g(g_{1,2}) = -2$. Now the values θ_1 and θ_2 are those for which $g_1 = g(\theta_1)$ and $g_2 = g(\theta_2)$. But

$$L_\theta(\theta_1) = L_g(g_1) = -2$$

from the invariance principle.

Example 15.1



,

Figure 15.2: *Invariance in ML interval limit.*

Assume we have made a single measurement of a $N(\mu,1)$ variable and have obtained x. We want to get 68% limits on the parameter $\theta = m^3$. We have the probability density:

$$f(x \mid \mu_0) = k \cdot e^{-\frac{(x-\mu_0)^2}{2}}$$

hence the likelihood function is

$$l(\mu_0 \mid x) = k \cdot e^{-\frac{(x-\mu_0)^2}{2}}$$

On the other hand the probability can also be expressed in terms of θ:

$$f(x \mid \theta_0) = k \cdot e^{-\frac{(x-(\theta_0)^{1/3})^2}{2}}$$

and the likelihood is:

$$l(\theta_0 \mid x) = k \cdot e^{-\frac{(x-(\theta_0)^{1/3})^2}{2}}$$

See again Fig.15.2 for a plot of the $\log L$ for the two functions and for $\mu_0 = 4$.

A line is drawn at -0.5 and we can read the 68% limits for μ and for θ.

We can find the confidence intervals of a non-normal *p.d.f.* variable without actually finding explicitly the transformation $g(\mu)$ which makes g normal. This method is very useful and elegant but has some limitations.

- The transformation $g(\theta)$ which makes the $\log L$ Normal may not exist. What we have done was to use the data themselves to transform the experimental $\log L$ to approximately Normal. This is not a change of variables for which the Invariance property holds. We have not determined a transformation which would have made the theoretical distribution normal.
- The method is exact only in the limit $N \to \infty$. In this limit the likelihood estimates are normal and the method coincide with the exact one described in the previous paragraph.
- If the experimental $\log L$ is far from being normal, the central intervals will be rather asymmetric. In such a case also the procedure might lead to a poor approximation.

Example 15.2 *In an experiment to analyze the angular distribution of the decay products of a resonance, ten events have been recorded. The cosine of the decay angle for each event is:*

$$\begin{aligned} cos(\theta_i) &= 0.244,\ 0.949,\ 0.907,\ 0.045,\ -0.722, \\ &\quad -0.152,\ -0.619,\ -0.139,\ -0.655,\ 0.283 \end{aligned}$$

A theory predicts that the angular distribution should be:

$$\frac{d\ N}{d\ cos(\theta)} = K\ (1 + a\ cos^2(\theta))$$

Determine:

- *the ML value of a,*
- *its error at* 0.68 *Confidence Level,*
- *how well the theory represents the data*

First of all we have to normalize the probability distribution:

$$\begin{aligned} f(x = cos(\theta)) &= K\ (1 + a\ x^2) \\ 1 &= \int_{-1}^{1} f(x)dx = K\ (a+3)\ \frac{2}{3} \\ K &= \frac{3}{2}\frac{1}{3+a} \\ f(cos(\theta)) &= \frac{3}{2}\frac{1}{3+a}\ (1 + a\ cos^2(\theta)) \end{aligned}$$

The Likelihood method is superior to the Least Square method: the number of events is small and to use the LS method we would have to group data in bins, loosing some information.

$$\begin{aligned} L &= (\frac{3}{2})^n (\frac{1}{3+a})^n \prod (1 + acos^2(\theta_i)) \\ log(L) &= -n\ log(3+a) + \sum_i log(1 + a\ cos(\theta_i)) \end{aligned}$$

There is no simple way to solve this equation and it is simpler to use numerical methods. Before doing it let us see what we should expect from the data. Fig. 15.3 shows the individual data and

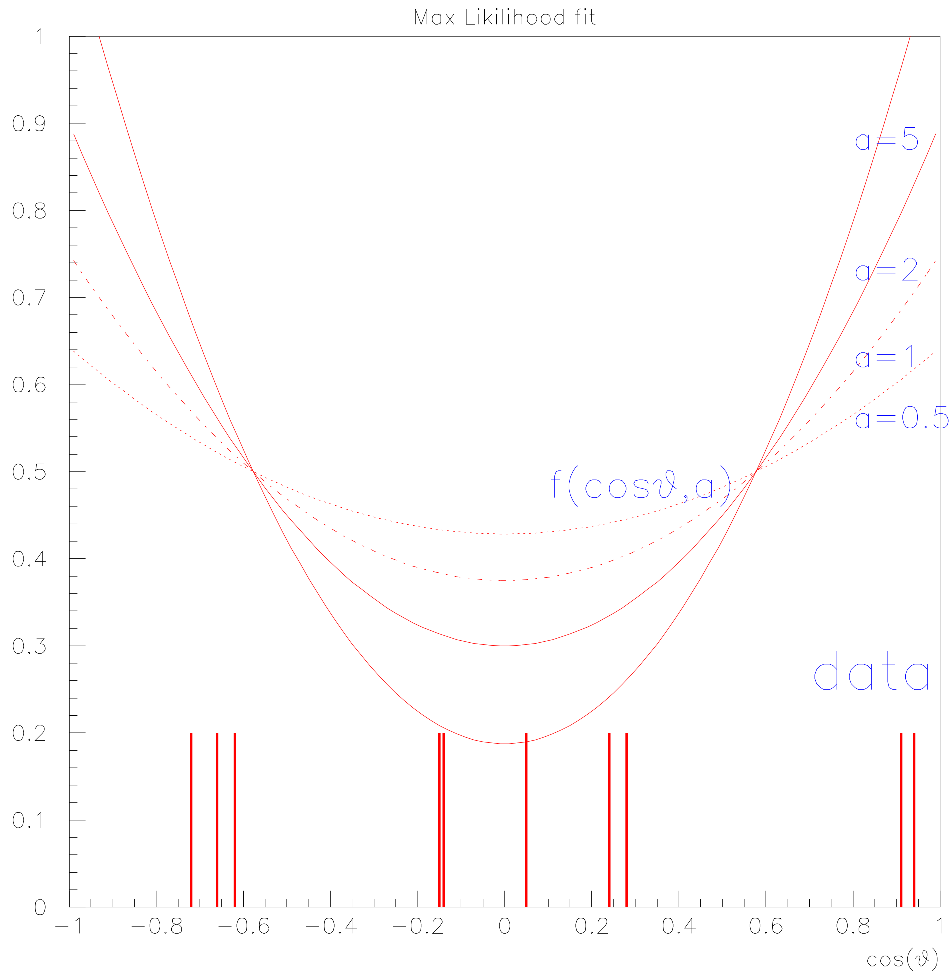


Figure 15.3: *The data and the shape of the function for a few values of a.*

the plot of the theoretical distribution for a few values of a. We see that the data are evenly distributed, as a function of $cos(\theta)$. No dependence on the angle is visible. On the contrary the theoretical function peaks at $\pm$ 1. It is reasonable to expect that the analysis will give small values of a.

The shape of $log(L)$ is shown in Fig. 15.4, as a function of a. A constant has been added to the value of $log(L)$ such that $max(log(L)) = 0$. The value of a at which $log(L)$ is maximum is: $\hat{a} = -0.096$ the intercepts of the horizontal line $log(L) = -0.5$ with the plot give the (approximate) limits at 0.68 CL. Finally: $\hat{a} = -0.096$ and $(-0.73, \ +1.27)$ is the .68 CL interval. The interval is asymmetric about $\hat{a}$.

The value of the variance can be estimated from the expression:

$$\sigma_{\hat{a}} \geq \sqrt{\frac{1}{-\frac{\partial^2 log(L)}{\partial a^2}}} \qquad a = \hat{a}$$

and the numerical value is $\sigma_{\hat{a}} \geq 0.41$. (The estimator is not efficient)

Concerning the goodness of fit we cannot use the Smirnov-Kolmogorov test, since the value

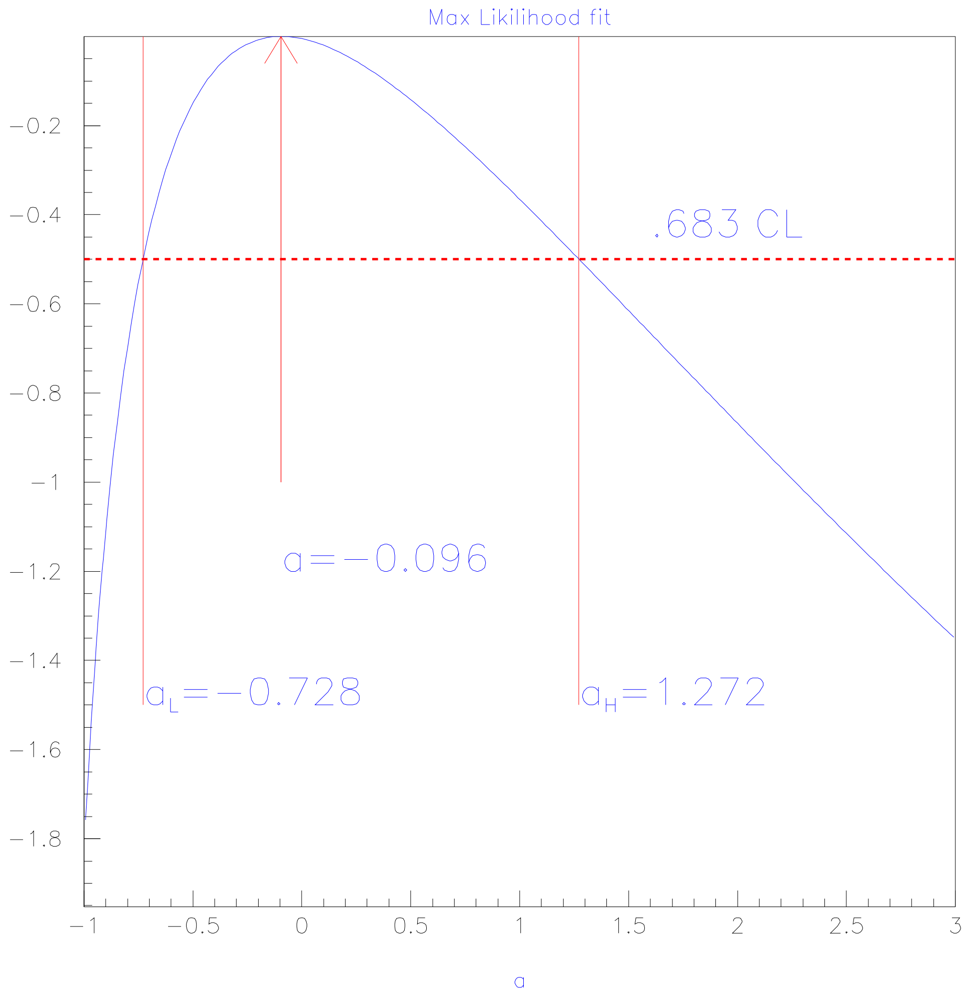


Figure 15.4: *Dependence of L on a.*

of a is obtained from the data and the probability of D_{max} is no longer distribution-free.

We can use, instead the Pearson test with $n-2$ d.o.f. (one is the total number of events one is the value of a). We have to group the data in classes. We can, for instance, use:

$$0 \leq \mid cos(\theta) \mid \leq 0.5 \qquad 0.5 \leq \mid cos(\theta) \mid \leq 1.$$

The theoretical probabilities for the two classes are:

$$\begin{aligned}
P_1 &= \int_0^{0.5} \frac{3}{2} \frac{1}{3+\hat{a}}(1+\hat{a}\ cos^2(\theta))d\ cos(\theta) = 0.5124 \\
P_2 &= 0.4876 \\
X^2 &= (\frac{n_1 - P_1\ N}{N\ P_1})^2 + (\frac{n_2 - P_2\ N}{N\ P_2})^2 = 0.019 \\
P(X^2) &= 0.89 \qquad (d.o.f.\ =\ 1)
\end{aligned}$$

Chapter 16

The Least Square Method

Des tous les principles qu'on peut proposer pou cet objet,
je pense qu'il n'en est pas de plus general, de plus exact, ni d'une
application plus facile que celui qui consiste a' rendre minimum
la somme des carrés des erreurs.
Adrien Marie Legendre (1805)

The method consists in minimizing the sum of the squares of the residuals between data and prediction.

16.1 A Short History

The least square method has a long history. It was used for the first time by Legendre in 1805 (" Nouvelle methode pour la determination des orbite des cometes") and the first sound mathematical basis were provided in 1808 by R. Adrain under the condition that the residuals have a normal distribution. Gauss in 1809 gave another derivation of the normal law and also claimed to have been using the least square method since 1795. In those days other methods were in use. In 1792 Laplace used the criterion to minimize the sum of the absolute value of the residues. In 1831 Cauchy introduced the method to minimize the maximum residuum (the min-max method which is occasionally still used). In 1812 Laplace proved that the least square method produces unbiased estimates, irrespective of the parent distribution. Gauss in 1821 showed that the least square method provided estimates having the least possible variance, among all unbiased, linear estimators, irrespective of the parent distribution. Markoff generalized the theorem in 1912.

Today is still one of the most used estimator method and is implemented in many tools used in the practice of the experimental physics.

16.2 The Linear Model

We will start from the most common problem that arises when we analyze data. We have made measurements of a quantity (say a cross section) at different angles and we want to know the components in a spherical harmonics analysis. The cosine of the angle are the independent quantities that are assumed to be free of any uncertainty. The cross section is subject to statistical errors which are assumed to be distributed normally. Let us translate this language in statistical jargon.

Assume that at the points $x_1, \cdots x_n$ we have measured independent normal variables $y_1, \cdots y_n$ with variances σ_i^2. The $\vec{x}$ are the independent variables while the $\vec{y}$ is the random sample used to estimate a functional dependence of $E(y) = f(x)$.

Assume also that a theoretical model predicts the dependence $f(x \mid \theta)$ where θ is a set of L parameters to be estimated from the sample (data). We minimize the sum of the squares of the residual (theory-experiment):

$$X^2 = \sum_{i=1,n} w_i (y_i - f(x_i \mid \theta))^2 = Min$$

The minimum is attained, by varying the parameters θ, at $\theta = \hat{\theta}$.

The weights w_i are taken to be: $w_i = \frac{1}{\sigma_i^2}$ for consistency with the maximum likelihood method. Very often it is used: $w_i = 1/y_i$ if the y_i come from a counting experiment (they are Poisson distributed), instead of the variance of the distribution. This is empirically justified by the great simplification of the mathematical problem. In some cases $w_i = Const$ (mainly when the variance is constant and unknown) and we have an unweighted Least Square.

If the measurements are correlated the previous quadratic form is written as:

$$X^2 = \sum_i \sum_j ((y_i - f(x_i, \theta)) \cdot V_{i,j}^{-1} \cdot ((y_j - f(x_j, \theta))$$

The weight matrix $w = V^{-1}$ is the inverse of the covariance matrix.

The formulas are more compact in matrix notation:

$$\left\{ \begin{array}{lcl} \vec{y} & = & \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad V_y = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2}^2 & \cdots & \sigma_{2,n} \\ & \cdots & \cdots & \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n}^2 \end{bmatrix} \\ \vec{f} & = & \begin{bmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{bmatrix} \\ \vec{\theta} & = & \begin{bmatrix} \theta_1 \\ \theta_2 \\ \cdots \\ \theta_L \end{bmatrix} \end{array} \right.$$

With this formalism the expression of X^2 becomes:

$$X^2 = (\vec{y} - \vec{f})^T V_y^{-1} (\vec{y} - \vec{f})$$

The value of $\vec{\theta}$ that minimize the value of X^2 are the estimated parameters and will be denoted, as usual with a hat: $\hat{\theta}$

16.2.1 An Elementary Example

Before doing the general case we will work in some detail the case of a straight line fit to a set of uncorrelated points. With a familiar notation:

$$(x_i,\ y_i \pm \sigma_i) \qquad i = 1, n$$

The X^2 form to be minimized is:

$$X^2 = \sum_{i=1,n} (\frac{y_i - (a + b \cdot x_i)}{\sigma_i})^2$$

The parameters to be determined are a and b. To find the minimum of X^2 we derive X^2 with respect to a and b and make the derivatives equal to zero.

$$\begin{aligned}\frac{\partial X^2}{\partial a} &= \sum_{i=1,n}\left(\frac{y_i-(a+b\cdot x_i)}{\sigma_i^2}\right)=0\\ \frac{\partial X^2}{\partial b} &= \sum_{i=1,n}\left(\frac{y_i-(a+b\cdot x_i)\cdot x_i}{\sigma_i^2}\right)=0\end{aligned}$$

If we indicate:

$$\begin{aligned} S_1 &= \sum_i \frac{1}{\sigma_i^2} \qquad S_x=\sum_i\frac{x_i}{\sigma_i^2}\\ S_{xx} &= \sum_i\frac{x_i\cdot x_i}{\sigma_i^2} \qquad S_y=\sum_i\frac{y_i}{\sigma_i^2}\\ S_{xy} &= \sum_i\frac{x_i\cdot y_i}{\sigma_i^2} \qquad D=S_1\cdot S_{xx}-S_x^2\end{aligned}$$

the LS system becomes:

$$\begin{cases} S_1\,a+S_x\,b &= S_y\\ S_x\,a+S_{xx}\,b &= S_{xy}\end{cases}$$

whose solution is:

$$\begin{cases} \hat{a} &= \frac{S_y\cdot S_{xx}-S_x\cdot S_{xy}}{D}\\ \hat{b} &= \frac{S_1\cdot S_{xy}-S_x\cdot S_y}{D}\end{cases}$$

The variance of the parameters can be found with the usual formulas for the *propagation of the errors*. It is convenient to use matrix notation calling:

$$\vec{\theta}=\begin{pmatrix} a\\ b\end{pmatrix}$$

The vector $\vec{\theta}$ is a linear function of $\vec{y}$ and the variance of θ is given by (see Chapter 5):

$$Var(\hat{\vec{\theta}})=S\;Var(\vec{y})\;S^T$$

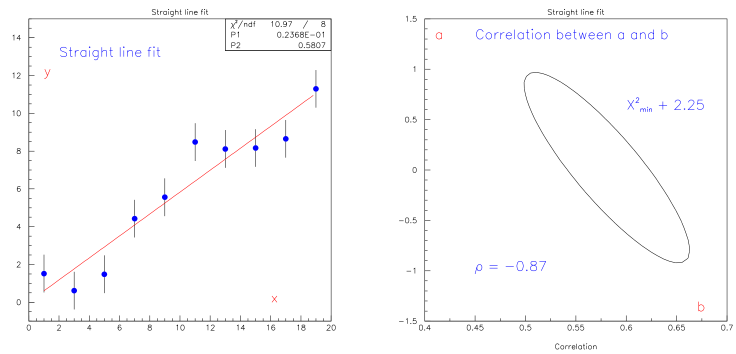
S id the matrix of the derivatives of a and b with respect to y_i:

$$S=\begin{pmatrix}\frac{\partial a}{\partial y_1} & \cdots & \frac{\partial a}{\partial y_n}\\ \frac{\partial b}{\partial y_1} & \cdots & \frac{\partial b}{\partial y_n}\end{pmatrix}$$

$$\begin{aligned} Var(\hat{\vec{\theta}}) &= \begin{bmatrix}\sigma_{a,a} & \sigma_{a,b}\\ \sigma_{b,a} & \sigma_{b,b}\end{bmatrix}=\\ &= \begin{pmatrix}\frac{\partial \hat{a}}{\partial y_1} & \cdots & \frac{\partial \hat{a}}{\partial y_n}\\ \frac{\partial \hat{b}}{\partial y_1} & \cdots & \frac{\partial \hat{b}}{\partial y_n}\end{pmatrix}\cdot\begin{pmatrix}\sigma_1^2 & \cdots & 0\\ 0 & \sigma_2^2 & 0\\ 0 & \cdots & \sigma_n^2\end{pmatrix}\cdot\begin{pmatrix}\frac{\partial \hat{a}}{\partial y_1} & \frac{\partial \hat{b}}{\partial y_1}\\ \cdots & \cdots\\ \frac{\partial \hat{a}}{\partial y_n} & \frac{\partial \hat{b}}{\partial y_n}\end{pmatrix}\\ &= \begin{pmatrix}\sum(\sigma_i\cdot\frac{\partial a}{\partial y_1})^2 & \sum\sigma_i^2\frac{\partial a}{\partial y_i}\frac{\partial b}{\partial y_i}\\ \sum\sigma_i^2\frac{\partial a}{\partial y_i}\frac{\partial b}{\partial y_i} & \sum(\sigma_i\cdot\frac{\partial b}{\partial y_1})^2\end{pmatrix}=\frac{1}{D}\begin{bmatrix}S_{xx} & -S_x\\ -S_x & S_1\end{bmatrix}\end{aligned}$$

The correlation coefficient is:

$$\rho=-\frac{S_x}{\sqrt{S_{xx}S_1}}$$



,

Figure 16.1: *Fit of uncorrelated points with a straight line..*

Fig.16.1 shows an example of a linear fit, the correlation of the two parameters.

The fitted parameters are correlated, even if the measurements are not. The correlation coefficient is negative. The fact that the variables are correlated means that if we change a bit a (the intercept) also the value of b will change in the fit, in order to keep a (relative) minimum to the X^2. Notice that the correlation of the two parameters vanishes if $S_x = 0$, i.e. if we perform a transformation of the coordinates $x \to x - w$ with:

$$w = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

The correlation of the parameters can be avoided with a wise choice of the origin of the $x-axis$. This fact will be discussed again when considering the orthogonal polynomials.

It is interesting to perform the calculation in a direct way:

$$\Delta\hat{a} = \hat{a} - \overline{\hat{a}} = \frac{1}{D}(S_y \cdot S_{xx} - S_x \cdot S_{xy} - \overline{S_y} \cdot S_{xx} + S_x \cdot \overline{S_{xy}}) =$$

The average is made on the observations (y_i).

$$= \frac{1}{D}\sum_i \frac{(S_{xx} - S_x x_i)(y_i - \overline{y_i})}{\sigma_i^2}$$

The variance is:

$$\overline{\Delta\hat{a}^2} = \frac{1}{D^2}\sum_{i,j} \frac{S_{xx} - S_x \cdot x_i}{\sigma_i^2}\frac{S_{xx} - S_x \cdot x_j}{\sigma_j^2}\overline{(y_i - \overline{y_i})(y_j - \overline{y_j})}$$

If the measurements are uncorrelated we have:

$$\overline{(y_i - \overline{y_i})(y_j - \overline{y_j})} = \sigma_i^2 \delta_{i,j}$$

Hence:

$$\overline{\Delta\hat{a}^2} = \frac{S_{xx}}{D}$$

In a similar way we obtain:

$$\overline{\Delta\hat{b}^2} = \frac{S_1}{D} \qquad \overline{\Delta\hat{a}\Delta\hat{b}} = -\frac{S_x}{D}$$

The parameters $\hat{a}$ and $\hat{b}$ are linear functions of y_i which are normally distributed. Hence $\hat{a}$ and $\hat{b}$ are also normally distributed with the variances and correlation coefficient that we have computed. The joint density function of $\hat{a}$ and $\hat{b}$ is:

$$f(\hat{a},\hat{b}) = \frac{1}{2\pi}\frac{1}{1-\rho^2}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(\hat{a}-\bar{\hat{a}})^2}{\sigma_{\hat{a}}^2}+\frac{(\hat{b}-\bar{\hat{b}})^2}{\sigma_{\hat{b}}^2}-2\rho\frac{(\hat{a}-\bar{\hat{a}})(\hat{b}-\bar{\hat{b}})}{\sigma_{\hat{a}}\sigma_{\hat{b}}}\right)}$$

The variance of the fitted function can be computed in a similar way:

$$\begin{aligned}
Y_0 &= \hat{b}x_0 + \hat{a} \\
\Delta y_0 &= x_0\Delta\hat{b} + \Delta\hat{a} \\
\overline{\Delta y_0^2} &= \overline{x_0^2\Delta\hat{b}^2 + \Delta\hat{a}^2 + 2x_0\Delta\hat{b}\Delta\hat{a}} = \\
&= x_0^2\frac{S_1}{D} + \frac{S_{xx}}{D} - 2x_0\frac{S_x}{D} = \\
&= \frac{S_1}{D}(x_0 - \frac{S_x}{S_1})^2 + \frac{1}{S_1}
\end{aligned}$$

The variance of the fit is minimum for a particular value of $x = S_x/S_1 = w$ and then grows quadratically with $x - w$.

16.3 The Linear Model in the General Case

Here we will show the fit machinery in the case of a linear model. In general the linear model (L parameters and N data) is such that:

$$\begin{aligned}
\vec{f} &= A\cdot\vec{\theta} \\
A &= \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,L} \\ a_{2,1} & a_{2,2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ a_{N,1} & \cdots & \cdots & a_{N,L} \end{bmatrix}_{(N\times L)}
\end{aligned}$$

The coefficients $a_{i,j}$ are functions of x. In the polynomial fit they are powers of x but in general they can be any function of x like $a_{i,j} = Sin(i\cdot x) * Cos(j\cdot x)$ etc. The term *linear* refers to the dependence of f on the parameters $\vec{\theta}$, not on $\vec{x}$; A can be any function of $\vec{x}$. The X^2 form becomes:

$$X^2 = (\vec{y} - A\cdot\vec{\theta})^T\cdot V_y^{-1}\cdot(\vec{y} - A\cdot\vec{\theta})$$

The quadratic form is minimized by taking the derivatives to zero. This is more easily done working with the components:

$$X^2 = \sum_{i,j,k,l}(y_i - A_{i,j}\vec{\theta}_j)V_{i,k}^{-1}(y_k - A_{k,l}\vec{\theta}_l)$$

differentiating with respect to θ_m we have:

$$\begin{cases} \frac{\partial X^2}{\partial\theta_m} &= -2\sum_{i,k,l}A_{i,m}V_{i,k}^{-1}(y_k - A_{k,l}\theta_l) \\ m &= 1\cdots L \end{cases}$$

or, in matrix notation:

$$A^TV^{-1}\vec{y} - A^TV^{-1}A\vec{\theta} = 0$$

and the solution is, provided the matrices are not singular:

$$\widehat{\vec{\theta}} = (A^T \cdot V_y^{-1} \cdot A)^{-1} \cdot A^T \cdot V_y^{-1} \cdot \vec{y}$$

The expression shows the linear relation between $\vec{\theta}$ and $\vec{y}$. The variance of the parameters is estimated from the usual formula:

$$Var(\widehat{\vec{\theta}}) = S\, V_y\, S^T$$

S is the matrix of the derivatives of $\hat{\theta}$ with respect to $\vec{y}$. This relation was derived in the case of a general dependence $\theta(y)$ by linearizing the function obtaining an approximation of the variance. In this case $\vec{\theta}$ is linear in $\vec{y}$ and the expression is exact. We get: [1]

$$\begin{aligned} Var(\widehat{\vec{\theta}}) &= [(A^T \cdot V_y^{-1} \cdot A)^{-1} \cdot A^T \cdot V_y^{-1}] \cdot V_y \cdot [(A^T \cdot V_y^{-1} \cdot A)^{-1} \cdot A^T \cdot V_y^{-1}]^T = \\ &= (A^T \cdot V_y^{-1} \cdot A)^{-1} \end{aligned}$$

The Linear Least Square estimate has three important properties which are independent of the distribution of the residual $\vec{y} - A\vec{\theta}$ and will be discussed in the next three sections.

16.3.1 $\hat{\theta}$ is unbiased Estimates of θ

It is not difficult to show that the estimates are unbiased:

$$\begin{aligned} E(\widehat{\vec{\theta}}) &= (A^T \cdot V_y^{-1} \cdot A)^{-1} \cdot A^T \cdot V_y^{-1} \cdot E(\vec{y}) = \\ &= (A^T \cdot V_y^{-1} \cdot A)^{-1} \cdot A^T \cdot V_y^{-1} \cdot A \cdot \vec{\theta} = \\ &= \vec{\theta} \end{aligned}$$

since $E(\vec{y}) = A\,\vec{\theta}$.

16.3.2 The Gauss-Markoff Theorem

We will now prove the Gauss-Markoff theorem that states: *Among all the unbiased estimator which are linear functions of the observations, the Least Square estimate has the smallest variance.*

Consider a set of estimators $\hat{t} = B \cdot y$ linear in the observations. If the new estimate are unbiased we have:

$$E(\hat{t}) = B \cdot E(y) = (B \cdot A) \cdot \theta = \theta \qquad \text{hence} \quad BA = I$$

We want to find when the covariance of t has the minimal diagonal form.

The variance of $\hat{t}$ is:

$$Var(\hat{t}) = \frac{\partial \hat{t}}{\partial \vec{y}} V_y \left(\frac{\partial \hat{t}}{\partial \vec{y}}\right)^T$$

[1] Note that for any matrix B such that $B^T = B$ we have:

$$BB^{-1} = I \quad \left(B^{-1}\right)^T B^T = I \qquad \text{hence} \quad \left(B^{-1}\right)^T = \left(B^T\right)^{-1}$$

The variance matrix V_y self-transposed by construction, hence:

$$\left(A^T V_y^{-1} A\right)^{-1} = \left(\left(A^T V_y^{-1} A\right)^{-1}\right)^T$$

The matrix of the derivatives is:

$$S = \frac{\partial \vec{t}}{\partial \vec{y}} = B$$

A little of matrix algebra shows that the following identity holds:

$$B\ V_y\ B^T = \left((A^T V_y^{-1} A\right)^{-1} + [B - (A^T V_y^{-1} A)^{-1} (A^T V_y^{-1})] V_y [B - (A^T V_y^{-1} A)^{-1} (A^T V_y^{-1})]^T$$

Each of the two terms on the RHS is a quadratic form of the type UVU^T and has non negative diagonal terms. Only the second term is a function of B. The sum of the two diagonal terms will have a minimum value for all elements when the second term has zero diagonal elements. This fixes the matrix B:

$$B = (A^T V_y^{-1} A)^{-1} A^T V_y^{-1}$$

But with this choice:

$$\hat{t} = (A^T V_y^{-1} A)^{-1} A^T V_y^{-1} \vec{y} = \hat{\theta}$$

Hence the Least Square linear estimators have the minimum variance.

16.3.3 Average of the Sum of Residual Square

If n is number of observations and L the number of parameters estimated from the data, we will prove that:

$$E(X^2_{min}) = (n - L)$$

We will prove the theorem in the simpler case of independent observations, all with the same variance:

$$V_y = \sigma^2 I_{n \times n}$$

In this case:

$$\hat{\theta} = (A^T A)^{-1} A^T \vec{y}$$

First, let us prove a general lemma about certain quadratic forms.

Lemma 16.1 *Let*

$$Q = \vec{y}^T U \vec{y}$$

be a quadratic form in n random variables $\{y_i\}$ and U a given $n \times n$ matrix. Let us call

$$\vec{\eta} = E(\vec{y})$$

and

$$Var(\vec{y}) = \sigma^2 I_{n \times n}$$

then

$$E(Q) = \sigma^2 Trace(U) + \vec{\eta}^T U \vec{\eta}$$

Proof 16.1 *Let us call $U_{i,j}$ the element i, j of the matrix U. We have:*

$$\begin{aligned} \vec{y}^T U \vec{y} &= \sum_{i,j} U_{i,j} y_i y_j = \\ &= \sum_{i,j} U_{i,j} (y_i y_j - \eta_i \eta_j) + \sum_{i,j} U_{i,j} \eta_i \eta_j \end{aligned}$$

Let us now take the expectation of this expression:

$$E(\vec{y}^T U \vec{y}) = \sum_{i,j} U_{i,j} Cov(y_i y_j) + \sum_{i,j} U_{i,j} \eta_i \eta_j$$

But, by assumption $Cov(y_i y_j) = \sigma^2 \delta_{i,j}$*, hence:*

$$E(\vec{y}^T U \vec{y}) = \sigma^2 Trace(U) + \eta^T\ U\ \eta =$$

∎

Theorem 16.1 $E(X^2_{Min}) = (n - L)$

Proof 16.2 *The minimum of the square of the residuals is:*

$$X^2_{min} = \frac{1}{\sigma^2}(\vec{y} - A\hat{\theta})^T(\vec{y} - A\hat{\theta})$$

The proof follows easily with a trick (Plackett [20]). Let us consider the following identity:

$$(\vec{y} - A\theta) = (\vec{y} - A\hat{\theta}) + A(\hat{\theta} - \theta)$$

Let us square it:

$$\begin{aligned}
(\vec{y} - A\theta)^T(\vec{y} - A\theta) &= (\vec{y} - A\hat{\theta})^T(\vec{y} - A\hat{\theta}) \\
&+ (\hat{\theta} - \theta)^T A^T A(\hat{\theta} - \theta) + 2(\hat{\theta} - \theta)^T A^T(\vec{y} - A\hat{\theta})
\end{aligned} \tag{16.1}$$

The last term is zero by definition of $\hat{\theta}$*, in fact, after replacing* $\hat{\theta}$*:*

$$A^T(\vec{y} - A(A^T A)^{-1} A^T \vec{y}) = A^T \vec{y} - A^T \vec{y} = 0$$

The second term of 16.1 is a positive definite quadratic form which becomes zero when $\theta = \hat{\theta}$*. This confirms that* $\hat{\theta}$ *is an absolute minimum. The second term of 16.1, after replacing* $\hat{\theta}$ *becomes:*

$$\begin{aligned}
(\hat{\theta} - \theta)^T A^T A(\hat{\theta} - \theta) &= ((A^T A)^{-1} A^T \vec{y} - \theta)^T A^T A((A^T A)^{-1} \vec{y} - \theta) \\
&= (\vec{y} - A\theta)^T\ A(A^T A)^{-1} A^T\ (\vec{y} - A\theta)
\end{aligned}$$

Since $A^T(A(A^T A)^{-1} A^T) = A^T$ *and* $(A(A^T A)^{-1} A^T)A = A$*. Finally we have:*

$$\begin{aligned}
X^2_{Min} &= \frac{1}{\sigma^2}\left[(\vec{y} - A\theta)^T(\vec{y} - A\theta) - (\vec{y} - A\theta)^T\ A(A^T A)^{-1} A^T\ (\vec{y} - A\theta)\right] = \\
&= \frac{1}{\sigma^2}(\vec{y} - A\theta)^T(I - A(A^T A)^{-1} A^T)(\vec{y} - A\theta)
\end{aligned} \tag{16.2}$$

This relation seems to contain again θ *but this is only apparent since the coefficients of* θ *vanish. For instance:*

$$\theta^T A^T(I - A(A^T A)^{-1} A^T)\vec{y} = \theta^T(A^T - A^T)\vec{y} = 0$$

Now we can use the previous Lemma:

$$\begin{aligned}
E(\vec{y} - A\theta) &= 0 \\
Var(\vec{y} - A\theta) &= \sigma^2 I_{n\times n} \\
Trace(I_{n\times n} - (A(A^T A)^{-1} A^T)_{n\times n}) &= n - Trace((A(A^T A)^{-1} A^T)_{n\times n} = \\
&= n - Trace((A^T A)_{L\times L}(A^T A)_{L\times L}) = n - L
\end{aligned}$$

Where we have used the simple relation:

$$Trace(A\ B\ C) = \sum_{l,i,j} A_{l,i} B_{i,j} C_{j,l} = \sum_{l,i,j} C_{j,l} A_{l,i} B_{i,j} = Trace(C\ A\ B) = Trace(B\ C\ A)$$

Then finally:

$$E(X^2_{Min}) = n - L$$

∎

This relation is important since we can estimate σ^2 if not known:

$$\widehat{\sigma^2} = \frac{X^2_{min}}{n-L} \tag{16.3}$$

We can also rewrite X^2_{Min} in more convenient form. From equation (16.2) we know that:

$$X^2_{Min} = \frac{1}{\sigma^2}(\vec{y}-A\theta)^T(I - A(A^TA)^{-1}A^T)(\vec{y}-A\theta)$$

This expression does not depend on θ, as we have discussed above. Let us then put $\theta = 0$:

$$X^2_{Min} = \frac{1}{\sigma^2}(\vec{y}^T(I - A(A^TA)^{-1}A^T)\vec{y} = \frac{1}{\sigma^2}(\vec{y}^T\vec{y} - A\hat{\theta}) \tag{16.4}$$

All the previous properties do not depend on the distribution of $\vec{y}$

16.4 A Technical Point

It happens frequently to use a polynomial expression to fit data:

$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + ... + \theta_r x^r$$

The design matrix A is in this case:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^r \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^r \end{bmatrix}$$

and the minimization methods proceeds as described above.

We want to point out a difficulty in applying blindly the formalism that we have developed in the previous section. Assume, to make the expressions simpler, that the observations are independent and have the same variance ($V = I$). The form we have to invert to get the Least Square estimate is:

$$A^TA = \begin{bmatrix} \sum 1 & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^r \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum x_i^r & \sum x_i^{r+1} & \cdots & \cdots & \sum x_i^{2r} \end{bmatrix}_{(r+1)\times(r+1)}$$

This type of matrices is very ill conditioned and their numerical evaluation can give serious rounding errors even in double precision computer packages. Let us evaluate the previous expression in the simple assumption that the x's lie between 0 and 1 and are evenly spread. Each of the terms in the matrix can be approximated to:

$$\sum x_i^r \to n\int_0^1 x^r\,dx = \frac{n}{r+1}$$

The matrix A^TA becomes:

$$A^TA = \propto \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{r+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \cdots & \frac{1}{r+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{r+1} & \frac{1}{r+2} & \cdots & \cdots & \frac{1}{2(r+1)} \end{bmatrix} = H_{r+1}$$

H_{r+1} is the Hilbert matrix of order $r+1$ which is known to be very ill conditioned for large r.

We have tried to compute how bad is the calculation of such a matrix with a simple FORTRAN program which uses the standard CERNLIB matrix package. We have computed the Hilbert matrix of order n and its inverse and the made the product of the two. In principle the final matrix would be the identity matrix with one's in the diagonal and zero elsewhere. The results (shown in Table 16.1) have been obtained in 64 bit precision arithmetics. For each $r = 3$ to 14 we have computed the average diagonal element of the matrix $A^T A$, the average non diagonal element and the maximum non diagonal element. Severe round off problems when the order of the matrix is larger than 11. This means that the algorithm that we have used, nice and very compact formally, is not the best to make numerical calculations. With this algorithm we have to perform differences between very large and almost equal numbers, therefore we must make arithmetics with numbers with very many digits. This is just what a computer (human or mechanical) likes the least! To solve the problem we have to use polynomial orthogonal on

n	Average diag	Average non diag	Max non diag
3	1	$3\ 10^{-15}$	$3\ 10^{-14}$
4	1	10^{-14}	10^{-13}
5	1	$3\ 10^{-13}$	$3\ 10^{-12}$
6	1	$6\ 10^{-11}$	10^{-11}
7	1	$4\ 10^{-10}$	$7\ 10^{-9}$
8	1	$2\ 10^{-8}$	$3\ 10^{-7}$
9	1	$6\ 10^{-7}$	$2\ 10^{-5}$
10	1	$3\ 10^{-6}$	$2\ 10^{-4}$
11	1	$4\ 10^{-4}$	$2\ 10^{-2}$
12	0.995	0.02	0.6
13	0.96	1	51
14	1.13	12	780

Table 16.1: *Difficulties in computing the Least Square solutions. The table shows the rounding errors in computing a Hilbert matrix.*

the observations.

16.5 Orthogonal Polynomials, a Simple Case

Consider the straight line fit problem to experimental points that we have discussed above and assume to simplify the notations that the variance for all the points is the same: $V_y = \sigma^2\ I_{n\times n}$. Instead of using the model:

$$f(x,\theta) = \theta_0 + \theta_1 \cdot x$$

let us use:

$$\begin{aligned} g(x,\phi) &= \phi_0 + \phi_1 \cdot (x - \overline{x}) \\ \overline{x} &= \frac{1}{n} \sum_{i=1,n} x_i \end{aligned}$$

The matrix A becomes:

$$A = \begin{bmatrix} 1 & x_1 - \overline{x} \\ 1 & x_2 - \overline{x} \\ & \cdots \\ 1 & x_n - \overline{x} \end{bmatrix}$$

The expression $(A^T A)$ can be easily computed yielding:

$$(A^T A) = \begin{bmatrix} n & \sum_{i=1,n}(x_i - \overline{x}) \\ \sum_{i=1,n}(x_i - \overline{x}) & \sum_{i=1,n}(x_i - \overline{x})^2 \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \overline{x^2} - \overline{x}^2 \end{bmatrix}$$

which is diagonal and easily inverted. Since its inverse is also diagonal, the variance of the parameters is diagonal (if the observations are uncorrelated) therefore the parameters ϕ_0 and ϕ_1 are uncorrelated.

This is he simplest example of an orthogonal polynomial. Since it depends from the observations it is called orthogonal on the data. In case of unequal variances we have to use $w = S_x/S_1$ instead of the arithmetic mean, as we have discussed above.

16.5.1 Orthogonal Polynomial (Same Variance)

Fitting the data with a polynomial model

$$f(x) = \theta_0 + \theta_1 x + \cdots + \theta_r x^r$$

leads to practical difficulties. The idea is to rearrange the coefficients such that the mathematical form of the model is now:

$$f(x) = \phi_0 + \phi_1 P_1(x) + \cdots + \phi_r P_r(x)$$

Where the polynomial $P_j(x)$, of order J, has the form:

$$P_j(x) = k_{j,0} + k_{j,1}x + \cdots + k_{j,j}x^j \qquad (j = 0, \cdots r)$$

In this way we have the freedom to compute the coefficients $k_{j,k}$ such that:

$$\sum_{i=1,n} P_m(x_i) \cdot P_l(x_i) = 0 \qquad m \neq l$$

The previous relation is a set of r equations that can be solved to obtain the coefficients $k_{j,l}$. There exist methods to compute the polynomials (Gram-Schmidt or Forsythe). The simplest method is due to Forsythe (1957) [51] who found a recurrence relations whose proof will be given in the appendix.

Assume that the polynomial normalization is:

$$\sum_{i=1,n} P_j^2(x_i) = 1$$

then:

$$\begin{cases} \lambda \cdot P_j(x) &= xP_{j-1}(x) - \alpha P_{j-1}(x) - \beta P_{j-2}(x) \\ \alpha &= \sum_{i=1,n} x_i P_{j-1}^2 \\ \beta &= \sum_{i=1,n} x_i P_{j-1} P_{j-2} \end{cases}$$

λ is a normalization factor chosen such that $\sum P_j^2 = 1$. With this recurrence formula, knowing the first polynomial, it is possible to compute all the others.

With this choice the matrix $A^T A$ becomes:

$$(A^T A) = \begin{bmatrix} \sum_{i=1,n} P_0^2 & 0 & \cdots \\ 0 & \sum_{i=1,n} P_1^2 & \cdots \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sum_{i=1,n} P_r^2 \end{bmatrix} = I$$

All the off-diagonal elements are zero *by construction*. Algebraically the method is equivalent to the standard method, but leads to fewer rounding errors.

Another advantage can be seen when we increase of one unit the degree of the polynomial. We simply have to add a new orthogonal polynomial (ϕ_{r+1}) since the lower degree polynomial are *unchanged*. For this reason the new matrix to be inverted has only one element anew with respect to the old one:

$$(A^T A)_{r+2,r+2} = \begin{bmatrix} (A^T A)_{r+1,r+1} & 0 \\ 0 & \sum_{i=1,n} P^2_{r+1} \end{bmatrix} = I_{(r+1)\times(r+1)}$$

The solution (coefficients ϕ_j) is obtained with the same machinery that we have discussed above. In fact:

$$f = A \cdot \phi \qquad\qquad A = \begin{bmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_r(x_1) \\ P_0(x_2) & P_1(x_2) & \cdots & P_r(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ P_0(x_n) & P_1(x_n) & \cdots & P_r(x_n) \end{bmatrix}$$

And, if for simplicity we assume $V_y = \sigma^2 I$, we get:

$$\widehat{\vec{\phi}} = A^T \vec{y}$$

or:

$$\widehat{\phi_j} = \sum_{i=1,n} y_i P_j(x_i)$$

And this shows clearly a nice property of the coefficients $\hat{\phi}$:

$$\begin{cases} Var(\widehat{\phi_j}) & = & \sum_{i=1,n} P_j^2(x_i) Var(y_i) = \sigma^2 \\ Var(\widehat{\phi_j}, \widehat{\phi_k}) & = & \sum_{i=1,n} P_j(x_i) \cdot P_k(x_i) Var(y_i) = 0 \end{cases}$$

All the $\widehat{\phi_j}$ are uncorrelated.

The value of the function $f(x)$ at the points x_i is sometimes called *improved measurements* since the value of the function at x_i is obtained from the whole set of measurements:

$$f(x_i) = y_i^* = \sum_{j=0,r} \phi_j P_j(x_i)$$

All these results can be converted back in the original polynomial form:

$$\phi_j P_j(x) = \phi_j \cdot (q_{j,0} + q_{j,1}x + \cdots + q_{j,j}x^j)$$

and rearranging the coefficients:

$$\theta_m = \phi_r q_{r,m} + \phi_{r-1} q_{r-1,m} + \cdots + \phi_m q_{m,m} = \sum_{i=m,r} \phi_i q_{i,m}$$

(the coefficient of x_m comes from P_m as $\phi_m q_{m,m}$, from P_{m-1} as $\phi_{m-1}\ q_{m-1,m}$ and so on.)

The residual sum of squares (the X^2 of the fit) is (16.4):

$$\sigma^2\ X^2_{Min} = (\vec{y} - A\hat{\theta})^T \cdot (\vec{y} - A\hat{\theta}) = \vec{y}^T \vec{y} - \vec{y}^T A \hat{\theta}$$

(The raw sum of squares is reduced by the sum of squares due to fitting.)

In our case:

$$X^2_{Min} = y^T y - y^T A\phi = y^T y - (A^T y)^T \phi = y^T y - \hat{\phi}^T \hat{\phi}$$

If the variance of y_i, σ^2 is unknown, it can be estimated by (16.3):

$$s^2 = \frac{S_r}{n-(r+1)}$$

Finally the variance of the "improved measurements" becomes:

$$Var(y^*) = \sum_{j=0,r} P_j^2(x) Var(\phi_j) = \sigma^2 \sum_{j=0,r} P_j^2(x)$$

Let now compute how the variance of the improved measurements changes when we increase the order of the polynomial by one.

We now change a bit the notations to enhance the dependence of X^2_{Min} on the degree of the polynomial. Let us call $S_{(r)} = X^2_{Min}$ the minimum of the sum of squared residuals for a polynomial of degree r. The increase the degree of the polynomial will make $S_{(r)}$ to decrease, (when $r = n-1$ the polynomial will pass through all the measurements and $S_{(n-1)} = 0$) therefore the fit "improves" by increasing r.

The design matrix A_{r+1} is:

$$A_{r+1} = \begin{bmatrix} & P_{r+1}(x_1) \\ [A_r] & \cdots \\ & P_{r+1}(x_n) \end{bmatrix}$$

$$\overrightarrow{\phi_{r+1}} = A^T_{r+1}\vec{y} = \begin{bmatrix} & [A_r] & \\ P_{r+1}(x_1) & \cdots & P_{r+1}(x_n) \end{bmatrix} \cdot [\vec{y}] = \begin{bmatrix} \vec{\phi_r} \\ \phi_{r+1} \end{bmatrix}$$

$\vec{\phi_r}$ are the r parameters obtained by fitting the polynomial of degree r and are unchanged when fitting with the polynomial of degree $r+1$.

$$\phi_{r+1} = \sum_i y_i P_{r+1}(x_i)$$

is the new coefficient.

$$\begin{cases} y^*_{(r+1)}(x_i) &= y^*_r + \phi_{r+1}P_{r+1}(x_i) \\ S_{(r+1)} &= S_{(r)} - \phi^2_{r+1} \\ s^2 &= \frac{S_{r+1}}{n-1-(r+1)} \\ Var(y^*_{(r+1)}) &= Var(y^*_{(r)}) + \sigma^2 P^2_{r+1}(x) \end{cases}$$

Therefore the sum of residual squares decreases but at the expense of increasing the variance of the fitted curve. There will be a critical value of r above which the reduction of the sum of residual squares does not correspond to an increase of information on the fitted curve. In more rigorous terms the increase of the degree of the polynomial is justified as long as the fitted parameter is significantly different from zero.

All this part is "distribution free". We shall specialize the argument considering the case where the measurement have a Normal distribution.

16.5.2 Normal Regression

Assume now that the residual are distributed as normal with mean zero and the same variance.

$$\epsilon_i = y_i - f_j \sim N(0, \sigma^2)$$

The residual are assumed independent and with identical Normal *p.d.f.* Also since y_i are normal variables and since a linear combination of Normal variables is also a Normal variable, the fitted parameters:

$$\widehat{\phi}_j = \sum_i P_j(x_i) y_i$$

are also Normally distributed and are unbiased since the Linear Least Square estimates are unbiased:

$$E(\widehat{\phi}_j) = \phi_j \qquad Var(\widehat{\phi}_j) = \sigma^2 \qquad \widehat{\phi}_j \sim N(\phi_j, \sigma^2)$$

(ϕ_j are the true parameters, unknown, that we estimate with $\hat{\phi}_j$.)

It is important to know the distribution of the parameter $\hat{\phi}_r$, since if we could tell that the parameter of order r has a value consistent with zero, we could reduce the order of the polynomial and improve the variance of the fit.

If $\phi_j = 0$, then:

$$\frac{\widehat{\phi}_i^{\,2}}{\sigma^2} \sim \chi_1^2$$

If σ^2 are known then we could use χ_1^2 distribution to test if $\widehat{\phi}_j = 0$.

When σ^2 is unknown we can use as its estimate:

$$\widehat{\sigma^2} = s^2 = \frac{S_{(r)}}{n-(r+1)}$$

To proceed we need to find the distribution of s^2. It is possible to fit a polynomial of degree $n-1$ to n points $(x_i, y_i, i = 1 \cdot n)$ such that the fitted line passes in each point and $S_{(n-1)} = 0$. Therefore:

$$S_{(n-1)} = 0 = \sum_{i=1,n} y_i^2 - \sum_{j=0,n-1} \phi_j^2$$

Hence:

$$S_{(r)} = \sum_{i=1,n} y_i^2 - \sum_{j=0,r} \phi_j^2 = S_{(n-1)} + \sum_{j=r+1,n-1} \phi_j^2$$

Hence, if all $\phi_i = 0, i = r+1 \cdots n-1$:

$$\frac{S_{(r)}}{\sigma^2} \sim \chi_\nu^2 \qquad \nu = (n-1) - (r+1) + 1 = n - r - 1$$

Since all ϕ_i are independent. Then we have:

$$\frac{s^2}{\sigma^2} = \frac{S_{(r)}/\sigma^2}{n-r-1} \sim \chi_{n-r-1}^2$$

We can now build a statistics to test if $\widehat{\phi_r}$ is consistent with zero. In fact $\widehat{\phi_r^2}/\sigma^2 \sim \chi_1^2$ and s^2 has the chisquare distribution written above. Their ratio is a variable which follows the Fisher-Snedecor distribution.

The variable $\widehat{\phi_r}^2$ depends on $\widehat{\phi_r}^2$ alone, while $S_{(r)}$ depends on the variables $\widehat{\phi_{r+1}}^2 \cdots \widehat{\phi_{n-1}}^2$, therefore $\widehat{\phi_r}^2$ and $S_{(r)}$ are independent. The ratio:

$$F = \frac{\frac{\widehat{\phi_r}^2}{1}}{\frac{S_{(r)}}{n-r-1}} = \frac{\widehat{\phi_r}^2}{s^2} = \frac{S_{(r)} - S_{(r-1)}}{S_{(r)}}\,(n-r-1) \sim F_{1,n-r-1}$$

The variable F is distributed as a Fisher-Snedecor if $\widehat{\phi_r}^2/\sigma^2 \sim \chi_1^2$ i.e. if the r^{th} degree is NOT justified.

A practical example of how to use this method will be discussed in the next example.

Example 16.1 *Assume we have made* n *measurements at the points* $(x_i, i = 1 \cdots n)$ *obtaining the results* $(y_i, i = 1 \cdots n)$*. The problem is to find the best interpolating polynomial. The mathematical model for the observations is:*

$$f(x) = \sum_{j=1,k} \Psi_j \cdot \xi_j(x) \qquad y_i = f(x_i) + \epsilon_i$$

The observations are a linear sum of the orthogonal polynomials (ξ_j)*. The coefficients* Ψ_j *are uncorrelated and independent of the degree of the polynomial* (j)*.*

If the model is correct, then ϵ *is a normal variable with mean 0. If the degree of the polynomial is too low, then it fits badly the observations, it is of no use and* ϵ_i *are no longer normal variables distributed about 0. If the degree is too high then the* X^2 *is very good (too good!) but the variance of the values predicted by the function increases: we loose precision in the prediction. For these reasons it is important to choose carefully the degree of the interpolating polynomial.*

As an example we have made a small MonteCarlo starting from a polynomial of 3^{rd} *degree:*

$$f(x) = a_0\, x^3 + a_1\, x^2 + a_2\, x + a_3$$

n *points are chosen, evenly spread in* x *and the values of* $f(x_i)$ *has been computed and then "randomized" assuming a p.d.f.* $\sim N(f(x), 1)$*. Table 16.2 shows the data used in the exercise.*

x_i	y_i
0.00000	0.62057
2.00000	2.12587
4.00000	3.84922
6.00000	5.19096
8.00000	7.65269
10.00000	6.63452
12.00000	6.16101
14.00000	8.14717
16.00000	7.49177
18.00000	7.20705
20.00000	6.32668
22.00000	5.27304
24.00000	6.16175
26.00000	7.13883
28.00000	7.04120
30.00000	8.26449
32.00000	7.63721
34.00000	9.23390
36.00000	10.61503
38.00000	11.03811

Table 16.2: *Values of the points used in the example.*

We will consider separately the case where we know the variance of the points (1 in our case) and the case where the variance is unknown.

- *The variance is known. In this case the value of*

$$X^2 = \sum \left(\frac{y_i - f(x_i)}{\sigma} \right)^2$$

has a precise statistical meaning and can be used to test the maximum degree of the interpolating polynomial. We will proceed with two different tests.

– $\frac{X^2}{\sigma^2\ (n-r-1)} \sim \chi^2_{n-r-1}$ *if the fit is correct, i.e. the residuals* $\frac{y_i - f(x_i)}{\sigma}$ *have mean 0 and variance 1. In Table 16.3 we report the analysis performed in this way. Clearly the*

Degree	X^2	*Dof*	*Prob*
0	119.133	19	$1.6\ 10^{-16}$
1	41.44	18	$1.3\ 10^{-3}$
2	40.06	17	$1.3\ 10^{-3}$
3	9.50	16	0.89
4	8.27	15	0.91
5	7.07	14	0.93

Table 16.3: *Analysis of the Chisquare as a function of the maximum degree of the polynomial.*

3^{rd} *polynomial gives a good probability and the forth degree is not justified.*

– *Since* $\hat{\phi}_r$ *are independent from each other and, in case they are compatible with 0, are distributed as a* χ^2_1*, we can test each of them, as shown in Table 16.4. Also with this*

Degree	X^2	$\hat{\phi}_r$	*Prob*
0	119.133		
1	41.44	77.693	$1.2\ 10^{-18}$
2	40.06	1.38	0.24
3	9.50	30.56	$3.2\ 10^{-8}$
4	8.27	1.23	0.27
5	7.07	1.20	0.27

Table 16.4: *Analysis of the Chisquare as a function of the maximum degree of the polynomial.*

type of analysis the fourth and fifth coefficient are compatible with 0.

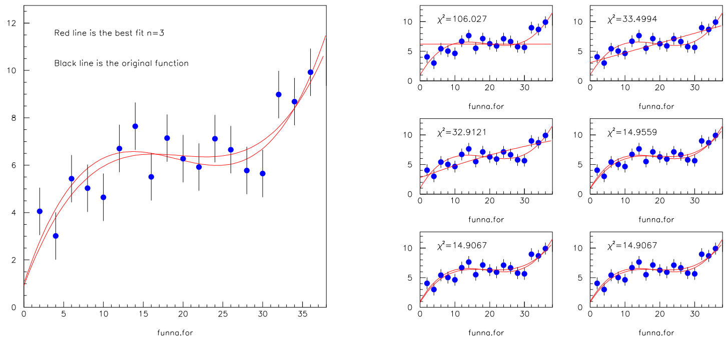
- *If the value of* σ^2 *is not known, as it happens in most of the cases, we can use the Fisher-Snedecor distribution. The analysis is shown in Table 16.5. In this analysis, we see that*

Degree	X^2	$F = \frac{X^2_{r-1} - X^2_r}{X^2_r}\ (n-r-1)$	*Prob*
0	119.133		
1	41.44	11.74	$3\ 10^{-3}$
2	40.06	0.566	.46
3	9.50	12.20	$3\ 10^{-3}$
4	8.27	1.94	0.18
5	7.07	2.03	0.18

Table 16.5: *Analysis of the Chisquare as a function of the maximum degree of the polynomial.*

already with $r = 2$ *the test gives a good probability but then it becomes poor again. This can happen and a safe strategy is to check the probability for a few degrees higher than the minimum suggested by the analysis. Also in this case the analysis indicates that a fourth degree polynomial is not justified.*

In conclusion we can safely use a third degree polynomial as best interpolating polynomial. Fig. 16.2 (right) shows the function, the simulated observations and and a fitted third degree polynomial through the points. The coefficients of the function and the fitted parameters are shown in Tab. 16.6.



,

Figure 16.2: *Fit to points of a polynomial. At left is shown the points, the original curve (continuous line) and the best fit. At right we fit to the same points polynomials of increasing order.*

Degree	True Coefficients	Fitted Coefficients	Error
0	1	0.519	0.545
1	1	1.134	0.081
2	-0.057	-0.0615	0.004
3	0.001	0.00104	0.0008

Table 16.6: *Coefficients (a_i) of a polynomial used to generate a set of points and the fitted parameters (A_i) with their Standard Deviations.*

16.6 Linear Least Square Estimation with Linear Constraints

This happens when the observable (η) are related through algebraic relations. The observations (y_i) are subject to experimental inaccuracies and shall not fulfill the constraints, however the "improved measurements" will do so.

Example 16.2 *Gauss measured the internal angles of a triangle formed by the top of three mountains. The idea was to check if the Euclidean geometry was true. In this case the sum of the internal angles must be π. The test was made by comparing the sum of the angle with π and decide if it is compatible or not. We shall use the measurements in a different way, imposing the constraint and getting an "improved measurement". Let us assume that the three angles were found to be:*

$$y_1 = 63^0, y_2 = 34^0, y_3 = 85^0$$

The variances are the same for all the measurements. We assume that the Euclidean geometry is true and we impose the constraint:

$$\sum_{i=1,3} \eta_i = 180^0$$

(note that the measurements fail the constraint by 2^0.)

The cleanest way to solve the problem is via the Lagrange multipliers. We minimize the form:

$$X^2 = \sum_{i=1,3} (\frac{y_i - \eta_i}{\sigma_i})^2 + 2\lambda(\sum_{i=1,3} \eta_i - 180^0)$$

The problem is a linear one and we have four unknown (η_i and λ). We will assume that all the variances are the same, for simplicity.

We can perform the minimization, as we did in the previous sections, by equating to zero the derivatives of X^2 and solving the system in four equations and four unknowns. The result is:

$$\begin{aligned} \hat{\lambda} &= \frac{1}{3\sigma^2}(\sum_{i=1,3} y_i - 180^0) \\ \hat{\eta_i} &= y_i - \sigma^2 \cdot \hat{\lambda} = y_i - \frac{1}{3}(\sum_{i=1,3} y_i - 180^0) \end{aligned}$$

The excess is equally shared among the "improved measurements" $\hat{\eta_i}$.

This procedure can be easily generalized to a set of linear constraints. Formally we have N observations y_i that we can arrange as a vector $\vec{y}$ of one column, L parameters $\vec{\theta}$ and K linear constraints that we can write as:

$$B\,\vec{\theta} - \vec{b} = 0$$

(B is a matrix $K \times L$ and $\vec{b}$ is a column vector with K components). Call $\vec{\lambda}$ a K-components vector of Lagrange multipliers. We minimize the form:

$$X^2 = (\vec{y} - A\vec{\theta})^T \cdot v^{-1} \cdot (\vec{y} - A\vec{\theta}) + 2\vec{\lambda} \cdot (B\theta - \vec{b})$$

Equalizing to zero the derivatives with respect to the unknown ($\vec{\theta}$) and the Lagrange multipliers ($\vec{\lambda}$) we obtain $K + L$ equations. Calling:

$$C = A^T V_y^{-1} A \qquad\qquad \vec{c} = \vec{A}^T \vec{V}_y^{-1} \vec{y}$$

we have the equations:

$$\begin{aligned} C\vec{\theta} + B^T\vec{\lambda} = \vec{c} \\ B\vec{\theta} = \vec{b} \end{aligned}$$

Provided that the inverse of C exists, we can left multiply the first equation by BC^{-1} and solve for λ:

$$\hat{\vec{\lambda}} = V_B^{-1}(BC^{-1}\vec{c} - \vec{b}) \qquad\qquad (V_B = BC^{-1}B^T)$$

and then replace in the fist equation to get:

$$\hat{\vec{\theta}} = C^{-1}\vec{c} - C^{-1}B^T V_B^{-1}(BC^{-1}\vec{c} - \vec{b})$$

Notice that the term $C^{-1}\vec{c}$ is the unconstrained solution. The term $(BC^{-1}\vec{c} - \vec{b})$ measures of how much the constrains are missed by the observations; the unconstrained solutions are corrected by an amount proportional to it.

The Lagrange multipliers and the parameters are linear combinations of the observations through the vector $\vec{c}$. The expectation value:

$$E(C^{-1}\vec{c}) = C^{-1}E(\vec{c}) = C^{-1}A^T V_y^{-1} E(\vec{y}) = C^{-1}A^T V_y^{-1} A\vec{\theta} = \vec{\theta}$$

$E(BC^{-1}\vec{c} - \vec{b}) = 0$ because of the constraints. Therefore:

$$E(\hat{\lambda}) = 0 \qquad\qquad E(\hat{\theta}) = \theta$$

The parameters are unbiased.

Exercise 16.1 *Verify that:*

$$E(\hat{\lambda}) = 0 \qquad E(\hat{\theta}) = \theta$$

Exercise 16.2 *Verify that the variance of $\hat{\theta}$ is:*

$$Var(\theta) = C^{-1} - (BC^{-1})^T V_B^{-1}(BC^{-1})$$

Exercise 16.3 *Verify that the variance of unconstrained parameters ($C^{-1}\vec{c}$) and the diagonal terms of $(BC^{-1})^T V_B^{-1}(BC^{-1})$ are always non-negative: the constraints reduce the parameters errors.*

16.7 General Least Square Fit

This is the case where even the x variable is subject to uncertainties that cannot be neglected. This means that even in the linear model $\eta = A(x')\theta$ where x' is a value of x around the average quoted in the measure. The Least Square method can be used to minimize the "distance" between the measured values $(x_i, y_i, i = 1, n)$ and the best values $(\xi_i, \eta_i, i = 1, n)$ which must fulfill the functional dependence $\eta = f(\xi)$ (the fit function):

$$\begin{aligned} X^2 &= (\eta - y)^T V_y^{-1}(\eta - y) + (\xi - x)^T V_x^{-1}(\xi - x) \\ f_k(\eta, \xi) &= 0 \qquad k = 1 \cdots n \end{aligned}$$

The problem can be solved with the Lagrange multipliers and is in general intrinsically non linear. We have to start from a guess $(\eta_i^0, \xi_i^0, i = 1, n)$ and solve numerically to get, by iteration an approximate solution.

Chapter 17

Confidence Intervals for the Least Square estimates

We will proceed as in the case of Maximum Likelihood estimates and postpone the general discussion on the confidence intervals. In this case we will exploit the known distribution of the X^2 variable, in the assumption that the observables are Normally distributed.

17.1 The Least Square Intervals

The results of the following sections are exact if the following conditions hold:

- Linear model: $f(\vec{\theta}) = A \cdot \vec{\theta}$
- The measurements $\vec{y}$ are normally distributed.

If such is the case, then the Least Square estimates $\hat{\vec{\theta}}$ are also normally distributed and the X^2 has also a well defined distribution.

If the previous conditions do not hold, then the results are only approximate. We recover the normality of the estimates in the limit of very large number of observations, by exploiting the central limit theorem [8].

17.2 The Normal LS Confidence Intervals when σ^2 is known

In matrix notation the model is:

$$\vec{\eta} = A\,\vec{\theta}$$

which express how the parameters are correlated to the observations.

The form to be minimized is:

$$X^2(\vec{\theta}) = (\vec{y} - A\,\vec{\theta})^T V_y^{-1} (\vec{y} - A\,\vec{\theta})$$

(The matrix V_y^{-1} is completely known.) and the linear solution is:

$$\begin{aligned} \hat{\theta} &= (A^T V_y^{-1} A)^{-1} A^T V_y^{-1} y \\ Var(\hat{\theta}) &= (A^T V_y^{-1} A)^{-1} \end{aligned}$$

We can rearrange the terms in a more useful form:

$$X^2(\theta) = (y - A\theta)^T V_y^{-1} (y - A\theta) + X^2_{min} - X^2_{min}$$

($X^2_{min} = X^2(\hat{\theta})$ i.e. the value of X^2 at the minimum, the Least Square solution).

$$\begin{aligned} X^2(\theta) &= X^2_{min} + y^T V^{-1} y - y^T V^{-1} A\theta - \theta^T A^T V^{-1} y + \theta^T A^T V^{-1} A\theta \\ &- y^T V^{-1} y + y^T V^{-1} A\hat{\theta} + \hat{\theta}^T A^T V^{-1} y + \hat{\theta}^T A^T V^{-1} A\hat{\theta} \end{aligned}$$

Using $(A^T V^{-1} A)\hat{\theta} = A^T V^{-1} y$ and rearranging the terms, we obtain finally (check with the results of the previous chapter):

$$\begin{aligned} X^2(\theta) &= X^2_{min} + (\hat{\theta} - \theta)^T (A^T V^{-1} A)(\hat{\theta} - \theta) \\ X^2(\theta) &= X^2_{min} + (\hat{\theta} - \theta)^T V_\theta^{-1} (\hat{\theta} - \theta) \end{aligned}$$

All this is true independently of the distribution of y.

Let us assume that the observations ($y_i, i = 1, n$) have multi normal *p.d.f.*. We know already that the individual terms have a χ^2 *p.d.f.*. The number of *d.o.f.* is for each term:

- $X^2(\theta) \sim \chi^2_N$ since the observations are N
- $X^2_{min} \sim \chi^2_{N-L}$. The *d.o.f.* are $N - L$ if L is the number of parameters.
- $(\hat{\theta} - \theta)^T V_\theta^{-1} (\hat{\theta} - \theta) \sim \chi^2_L$.

If we have K constraint equations, then we have to replace L with $L-K$ since the constraints reduce the number of independent variables.

17.2.1 One Parameter Case

In this simple case $L = 1$ and the variable $X^2(\theta) - X^2_{min}$ has a distribution of χ^2 with 1 *d.o.f.*.

Let us define a confidence level α such that:

$$Pr(\chi^2 \geq \beta_\alpha) = 1 - \alpha$$

The confidence interval, relative to the confidence limit α is interval of θ for which the previous relation holds.

An important case, often used, is $\alpha = 0.68$ (1σ). This represents the probability that the true value of the parameter θ is outside the confidence interval in 32% of cases. (Also often used is the confidence level $\alpha = 0.5$).

From the Probability tables we read the value of $\beta_{0.68}$ for one *d.o.f.*:

$$P(\chi^2 \geq 1) \approx 0.32$$

Hence:

$$P(X^2(\theta) - X^2_{min} \geq 1) \approx 0.32 \qquad\qquad P(X^2(\theta) \geq X^2_{min} + 1) \approx 0.32$$

This is the origin of the recipe "$X^2_{min} + 1$". Graphically this is illustrated in Fig. 17.1. The X^2 has a parabolic shape as a function of θ. This is exactly true in the linear model.

$$X^2 = X^2(\theta) - X^2_{min} = (\frac{\hat{\theta} - \theta}{\sigma_\theta})^2$$

The 1 σ case ($P(X^2 \geq 1)$ or 68% Confidence Interval) is:

$$(\frac{\hat{\theta} - \theta}{\sigma_\theta})^2 \leq 1 \qquad\qquad \hat{\theta} - \sigma_\theta \leq \theta \leq \hat{\theta} + \sigma_\theta$$

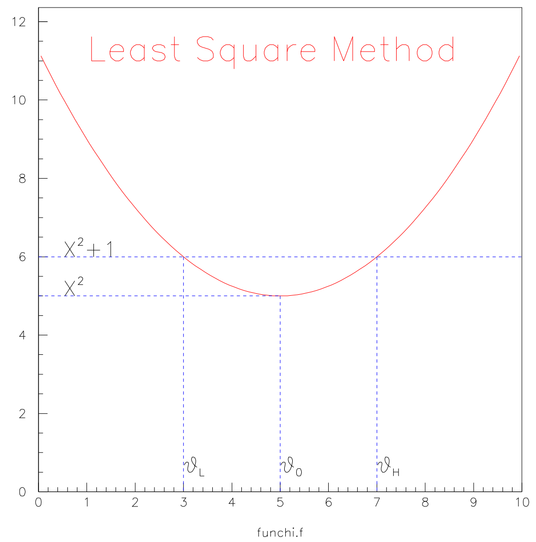


Figure 17.1: *How we define the confidence limits in the LS method.*

Let us expand the X^2 function about its minimum:

$$X^2(\theta) = X^2_{min} + \frac{1}{2}\frac{\partial^2\ X^2}{\partial\ \theta^2}\bigg|_{\theta=\hat{\theta}}\ (\theta - \hat{\theta})^2 + \cdots$$

In case of Linear Least Square the function X^2 is a second order function and there are no other terms beyond the second derivative. From the equation:

$$\begin{aligned} X^2(\theta) &= X^2_{min} + (\theta - \hat{\theta})^T\ V_\theta^{-1}(\theta - \hat{\theta}) \\ &= X^2_{min} + (\theta - \hat{\theta})^2 \cdot \frac{1}{\sigma^2_{\hat{\theta}}} \end{aligned}$$

we obtain:

$$Var(\hat{\theta}) = 2\ \frac{1}{\frac{\partial^2 X^2}{\partial\theta^2}\Big|_{\hat{\theta}}}$$

In the case of Normal distribution we have $X^2 = -2\ L$ and we recover the results of the ML method. This type of relation holds in general; in the case of more than one parameter we have:

$$Var(\theta)_{i,j} = 2\ \frac{1}{\frac{\partial^2 X^2}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\theta}}}$$

All this is exact in case of linear Least Square and will also be numerically correct if the truncation of the Taylor series to second order is acceptable.

17.2.2 Two Parameters Case

This case is analogous to the previous one, but now we have two d.o.f.:

$$X^2(\theta) - X^2_{min} \sim \chi^2_2$$

We cannot use the same interval as before, since the probability content is different, as it can be verified from the probability tables. In fact for two d.o.f.:

$$P(X^2_{2dof} \geq 1) = 0.606 \qquad instead \qquad P(X^2_{2dof} \geq 2.25) = 0.325$$

The value of β for 0.68 confidence interval is:

$$\beta_{0.68} = 2.25 \qquad 2\ d.o.f.$$

We recover the "1 σ" C.I. at larger values of $X^2 - X^2_{min}$. We obtain the α % Confidence Regions by cutting the surface

$$z = X^2(\theta_1, \theta 2)$$

with the plane:

$$z = X^2_{min} + 2.25$$

The Confidence Intervals are ellipses in the plane θ_1, θ_2. Fig. 17.2 illustrate the CI. The probability that the true value of the parameters is inside the ellipsis $(X^2_{min} + 2.25)$ is 68%. Of

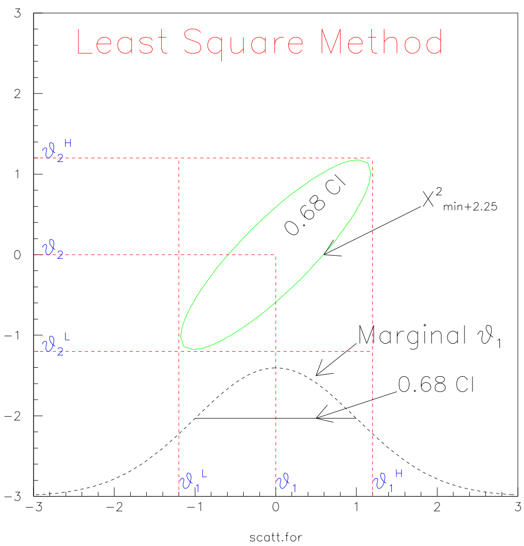


Figure 17.2: *The ellipse of two parameter confidence limit.*

course the "$1 - \sigma$" interval for one parameter is computed by the marginal probability (integrate away the other variable) and we are again in the previous case. Therefore in many dimensional parameter estimation the $\pm$ 1 σ interval has a probability content of 68% only if we ignore the interval of *all* other parameters.

17.3 The Normal LS Confidence Intervals when σ^2 is unknown

This case happens frequently in practice. The measurements $\vec{y}$ are distributed normally with unknown variance. (We have to assume that the variance is the same for all the observations.)

In the quadratic form:

$$X^2 = (\vec{y} - A\ \vec{\theta})^T V_y^{-1} (\vec{y} - A\ \vec{\theta})$$

the matrix V_y^{-1} is not known and must be estimated from the data, i.e. it is a random variable. Hence X^2 has NOT a χ^2_L *p.d.f.* In this case we have to a function of the data which has a known distribution, a distribution independent of the unknown parameters.

17.3.1 The Case of a Single Parameter

We have to find a new variable whose *p.d.f.* does not depend on θ or σ^2.

We have first to estimate the variance of the observations by:

$$\hat{\sigma^2} = s^2 = \frac{X^2_{min}}{N-1}$$

The estimate of the variance of the parameter is then: $\hat{\sigma}_{\hat{\theta}} = s^2(A^TA)^{-1}$ (for independent measurements). The estimate $\hat{\theta}$ is distributed as:

$$\hat{\theta} \sim N(\theta, \sigma^2_{\theta})$$

and the variable:

$$t = \frac{\hat{\theta}}{\hat{\sigma}_{\theta}} \sim S_{N-1}$$

has a Student distribution with $N-1$ *d.o.f.*.

If t_α is the α probability point ($P(|t| \geq t_\alpha) = 1-\alpha$), then the interval $(\hat{\theta} - t_\alpha\hat{\sigma}_\theta,\ \hat{\theta} + t_\alpha\hat{\sigma}_\theta)$ is a α% CI.

17.3.2 The Case of Several Parameters

Also in this case we have to transform the variable, but here the transformation is different. The unknown variance can be estimated in two independent ways:

$$s^2_\theta = \frac{(\hat{\theta}-\theta)^T V_\theta^{-1}(\hat{\theta}-\theta)}{L} = \frac{X^2 - X^2_{min}}{L}$$

(L is the number of parameters)

and from the value of the minimum of X^2:

$$s^2_{X^2} = \frac{X^2_{min}}{N-L}$$

We can prove that the two estimates are independent hence the variable:

$$f = \frac{s^2_\theta}{s^2_{X^2}} = \frac{(\hat{\theta}-\theta)^T V_\theta^{-1}(\hat{\theta}-\theta)}{L} \cdot \frac{N-L}{X^2_{min}} \sim F_{L,N-L}$$

That is f is distributed according to Fisher-Snedecor and does not depend on the parameters or on the unknown variance. The probability content for each parameter can be computed as we have seen in the previous case.

Chapter 18

Confidence Intervals

The methods that we have studied in the previous chapters provide an estimate of the value of the unknown parameters. All the methods construct a function of the measurements, the estimator. This function is optimal in average for all the samples that can be drawn from the population. Once we compute it for the actual observations, we have an estimate. This is does not coincide, in general, with the value of the parameter, since the fluctuations in the sampling introduce an uncertainty in the estimate. This uncertainty is indicated by the variance of the estimator.

Here, however we want to address the problem from a different point of view. We will define a range of values of the parameter in which the value of the true parameter lies with a predefined probability.

18.1 The basic Idea

Before going into the detailed description, let us describe the procedure that we will follow. Assume that we have measured a quantity that we know to be distributed Normally with unit variance: $x \sim N(\mu, 1)$. In the measurement we have obtained the value $x = 3$, as shown in figure 18.1. The maximum likelihood solution of the point estimate leads to the estimate $\hat{\mu} = 3$ (continuous line in the figure).

The estimate does not coincide, in general with the unknown value of μ, the probability density shows that repeated experiments would lead to measurements that cluster around the true value with a variance which is known to be one. Hence the true value of the parameter, though unlikely to coincide with the result of the measurement, should be not too far from it, the scale of the dispersion of repeated measurements being one.

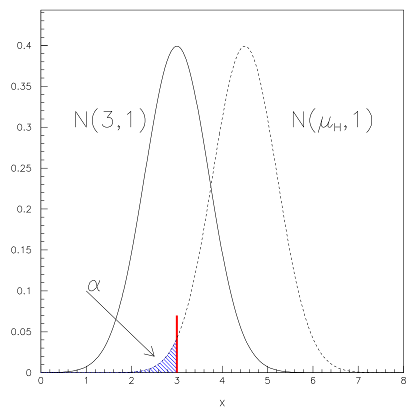


Figure 18.1: *The Upper Confidence Limit.*

The inference of the CI has to be based on: the actual measurement and on the frequency principle of repeatability (at least in principle) of the measurements whose frequency, in the long run, coincide with the probability. If the true value is far too away from the actual measurement (the dashed curve in the figure), then the probability of measurements equal or smaller than what we have measured would be small. Our event would be a rare event, which is quite possible, but unlikely.

The CI are defined in this way: first we decide a confidence level (α) which is a measure of how unlikely our result could be; then we compute the largest value of μ such that the probability of having a result smaller (worse) or equal to the measurement is equal to α (dashed line in the figure). This is the upper limit of μ (μ_H). Then we determine the lower limit (μ_L) in a similar way. The interval: $[\mu_L,\ \mu_H]$ is called a $1-2\alpha$ Confidence Interval.

We have now to discuss carefully the meaning of this procedure.

18.2 The construction of the CI

Let us start from a distribution which depends on a single parameter θ and assume that we have a random sample of n observations $\{x_1, x_2, \cdots, x_n\}$. Let us assume also that we can construct a variable z, function of the observations and of θ whose distribution $F(z)$ is <u>independent of θ</u>. Then, given any probability α, we can find a value of $z = z_1$ such that the following relation holds:

$$\int_{z_1}^{\infty} dF = \alpha$$

or, in the language of the theory of probability,

$$Prob(z \geq z_1) = \alpha$$

If we can transform the relation $z \geq z_1$ in the form $\theta \geq \Theta_L$, where Θ_L is a function of z_1 but not of θ, then we can recast the previous relation in the form:

$$Prob(\theta \geq \Theta_L) = \alpha$$

Even if mathematically this procedure is correct, we have to interpret it. In fact the probability that the unknown constant θ is larger than a constant is either 1 or 0, but we do not know which of the two possibilities is correct. We have, instead, to look at the previous statement as a probability statement on Θ_L. If we repeat the experiment, in uniform conditions, very many times, the statement:

The statistics Θ_L is smaller or equal to θ

is correct in a fraction α of cases, *whatever the value of* θ. Before proceeding, let us make some examples to clarify what we mean.

18.2.1 The Normal CI

A variable X is distributed as a Normal: $X \sim N(\mu, \sigma^2)$. We make n measurements of X: $\{x_1 \cdots x_n\}$. Assume that σ^2 is known, what is the Confidence Interval (CI in the following) for the mean? We can use as statistics the sample mean: $\overline{x} = \sum x_i/n$ whose density is:

$$\overline{x} \sim \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma} \cdot exp\left(-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right)$$

The density depends on the unknown parameter μ, but it is sufficient to change the variable: $z = \overline{x} - \mu$ and the explicit dependence of F on μ disappears. From the Normal distribution tables we know that:

$$Prob\left(z \leq 2\frac{\sigma}{\sqrt{n}}\right) = 0.97725$$

which is the same as:

$$Prob\left(\overline{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu\right) = 0.97725$$

Therefore if we assert that:

$$\mu \geq \overline{x} - 2\frac{\sigma}{\sqrt{n}}$$

we will be correct in 97.725% of cases, in a long run of experiments.

In the same way we can prove that:

$$Prob\left(\mu \leq \overline{x} + 2\frac{\sigma}{\sqrt{n}}\right) = .97725$$

Let us call $A = \left[\mu \geq \overline{x} - 2\frac{\sigma}{\sqrt{n}}\right]$ and $B = \left[\mu \leq \overline{x} + 2\frac{\sigma}{\sqrt{n}}\right]$ the two events. We get easily, by DeMorgan law:

$$\begin{aligned} Prob(A \cap B) &= Prob(\overline{\overline{A} \cup \overline{B}}) \\ &= 1 - Prob(\overline{A} \cup \overline{B}) \\ &= 1 - (Prob(\overline{A}) + Prob(\overline{B})) \qquad \text{(disjoint events)} \\ &= 1 - (1 - Prob(A) + 1 - Prob(B)) \end{aligned}$$

($\overline{A}$ and $\overline{B}$ are disjoint events certainly if $P(\overline{A}),\ P(\overline{B}) < 0.5$.) Hence:

$$Prob\left(\overline{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 2\frac{\sigma}{\sqrt{n}}\right) = 1 - 2(1 - 0.97725) = 0.9545$$

The interval:

$$\left[\overline{x} - 2\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + 2\frac{\sigma}{\sqrt{n}}\right]$$

is a 95.45% interval for the mean μ.

This example can be generalized to any value of probability content α. Let us call $\alpha - Point$, z_α the value of z such that:

$$Prob(z \geq z_\alpha) = \alpha$$

(see figure 18.2). If now $z_{\frac{1-\alpha}{2}}$ is the $(1-\alpha)/2-Point$ of the Standard Normal distribution:

$$\int_{-\infty}^{-z_{\frac{1-\alpha}{2}}} N(0,1)\ dz = \frac{1-\alpha}{2} = \int_{z_{\frac{1-\alpha}{2}}}^{\infty} N(0,1)\ dz$$

then, using the same type of argument as we have done above, it is easy to show that the interval:

$$\left[\overline{x} - z_{\frac{1-\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{1-\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

is a CI with a Confidence Level α.

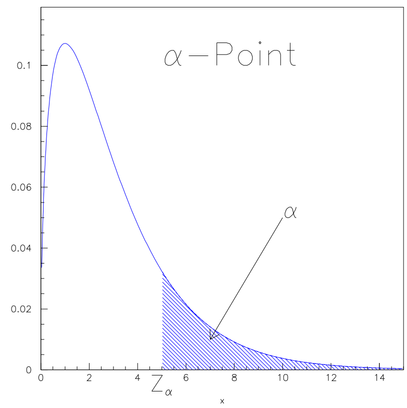


Figure 18.2: *Definition of $\alpha - Point$.*

Other CIs

There is a further complication; as it is easy to understand, there is no unique way of defining a CI at a given confidence level. What we have done is the construction of a *central interval*[1].

[1] The word *central* concerns the probability of the tails of the distribution outside the CI. Unless the distribution of the variable is symmetrical, the limits will not be equidistant from the estimate.

Other intervals exist which have the same probability content. Important intervals often used in practice are the *lower* and *upper* intervals. These intervals are used to answer to the question: How large (or how small) is the parameter likely to be? The interest is shifted from the definition of an interval which is likely to contain the true parameter to an upper or lower bound to the parameter. (If a drug is known to be beneficial to threat some disease but noxious in large quantities, we have to establish the maximum dose of the drug compatible with a small (and known) probability of sideffects.) Mathematically these are simply special cases of the general case.

The variable $z = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}} \sim N(0,1)$ hence, calling $z_{1-\alpha}$ the value of z such that:

$$P(z > z_{1-\alpha}) = 1 - \alpha$$

we have:

$$P(\frac{\overline{x}-\mu}{\sigma_{\overline{x}}} \geq z_{1-\alpha}) = 1 - \alpha$$

or:

$$P(\mu < z_{1-\alpha} \cdot \sigma_{\overline{x}} + \overline{x}) = 1 - \alpha$$

In other words: the interval: $\left[-\infty;\ \overline{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\right]$ is a α lower CI and the value $\mu_H = \overline{x} + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ is the upper limit of the parameter (with CL α).

In the same way we can define a upper CI: $\left[\overline{x} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\ ;\ +\infty\right]$ and the value $\mu_L = \overline{x} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ is the upper limit of the parameter.

This is summarized in table 18.1.

CI in the Normal distribution	
Assumptions	$x \sim N(\mu, \sigma^2)$ $\quad \sigma^2$ *known*
Parameter	μ
Central Interval	$\left[\overline{x} - z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$
Lower Interval	$\left[-\infty,\ \overline{x} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right]$
Upper Interval	$\left[\overline{x} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}},\ \infty\right]$

Table 18.1: *Summary of Normal CI. In the case of Normal density we have:* $z_{1-\alpha} = -z_\alpha$.

18.2.2 The Student CI

If the variance of the parent distribution is not known and it has to be estimated from the sample, we have to proceed differently. Since the sample mean and variance are estimated from the sample:

$$\begin{cases} \overline{x} &= \frac{\sum_i x_i}{n} \\ s^2 &= \frac{\sum_i (x_i - \overline{x})^2}{n-1} \end{cases}$$

the variable:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim S_{n-1}(t)$$

is distributed as a Student with $n-1$ *DoF*. Let us call $t_{\frac{1-\alpha}{2},n-1}$ the value of t such that:

$$\int_{-\infty}^{-t_{\frac{1-\alpha}{2},n-1}} S_{n-1}(t)\ dt = \frac{1-\alpha}{2} = \int_{t_{\frac{1-\alpha}{2},n-1}}^{\infty} S_{n-1}(t)\ dt$$

then, as in the case of a Normal variable:

$$Prob\left(\frac{\overline{x}-\mu}{s/\sqrt{n}} \leq t_{\frac{1-\alpha}{2},n-1}\right) = \frac{1-\alpha}{2}$$

The confidence intervals are summarized in table 18.2.

CI in the Student distribution	
Assumptions	$x \sim N(\mu,\sigma^2) \qquad \sigma^2$ *unknown*
Parameter	μ
Central Interval	$\left[\overline{x} - t_{\frac{1-\alpha}{2},n-1}\frac{s}{\sqrt{n}},\ \overline{x} + t_{\frac{1-\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right]$
Lower Interval	$\left[-\infty,\ \overline{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}}\right]$
Upper Interval	$\left[\overline{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}},\ \infty\right]$

Table 18.2: *Summary of Student CI*

18.2.3 The χ^2 CI

If the parameter that we want to estimate is not μ but σ^2, we have to remember that the variable:

$$X^2 = \frac{s^2 \cdot (n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

is distributed as a χ^2 with $n-1$ DoF. If $X^2_{\beta,n-1}$ is the β point of the χ^2 density:

$$\int_{X^2_{\beta,n-1}}^{+\infty} \chi^2 dX = \beta$$

then it is easy to compute the intervals, with a CL equal to α, summarized in table 18.3.

CI in the χ^2 distribution	
Assumptions	$x \sim N(\mu,\sigma^2) \qquad \sigma^2$ *unknown*
Parameter	σ^2
Central Interval	$\left[\frac{(n-1)s^2}{X^2_{\frac{1-\alpha}{2},n-1}},\ \frac{(n-1)s^2}{X^2_{\frac{1+\alpha}{2},n-1}}\right]$
Lower Interval	$\left[0,\ \frac{(n-1)s^2}{X^2_{\alpha,(n-1)}}\right]$
Upper Interval	$\left[\frac{(n-1)s^2}{X^2_{1-\alpha,(n-1)}},\ \infty\right]$

Table 18.3: *Summary of* χ^2 *CI*

A list of α points for the various distributions and for some confidence levels is presented in table 18.4.

Example 18.1 *Let us consider a sample of* n *measurements of an exponential variable with parameter* τ*:*

$$t \sim \frac{e^{-t/\tau}}{\tau} \qquad t\ >\ 0$$

	The χ^2 distribution						The *Student* distribution		
	β						β		
DoF	0.95	0.90	.8414	0.1586	0.1	0.05	0.05	0.10	0.1586
1	0.004	0.0158	0.0400	1.987	2.706	3.841	6.314	3.078	1.838
2	0.103	0.211	0.345	3.683	4.605	5.991	2.920	1.886	1.322
3	0.352	0.584	0.833	5.187	6.251	7.815	2.353	1.638	1.197
5	1.145	1.610	2.056	7.957	9.236	11.07	2.015	1.476	1.111
10	3.940	4.865	5.680	14.33	15.99	18.31	1.812	1.372	1.053
15	7.261	8.546	9.646	20.36	22.31	25.00	1.753	1.341	1.035
20	10.85	12.44	13.78	26.22	28.41	31.41	1.725	1.325	1.025
30	18.49	20.60	22.34	37.66	40.26	43.77	1.697	1.310	1.017
50	34.76	37.69	40.07	59.94	63.17	76.50	1.676	1.299	1.010

The *Normal* Distribution									
β									
0.0005	0.001	0.005	0.01	0.025	0.05	0.10	0.1586	0.25	0.50
3.291	3.090	2.576	2.326	1.960	1.645	1.282	1.000	0.674	0.000

Table 18.4: *Value of the β point: $X^2_{\beta,DoF}$, the for the $\chi^2_{\beta,Dof}$, $t_{\beta,DoF}$ for the Student and z_β for the and the StandardNormal distribution. Note that for the symmetric distributions $S_n(t)$ and the Normal $N(0,1)$ the symmetric point can be obtained with: $z_{1-\beta,DoF} = -z_{\beta,DoF}$*

We are required to define the CI at a confidence level β. A good statistics is the parameter estimator itself:

$$\bar{t} = \frac{\sum_i t_i}{n}$$

The first step is to find the density of $\bar{t}$. This can be computed with the method of moment generating functions since:

$$\Phi_t = E\left(e^{tx}\right) = \int_0^\infty e^{tx}\frac{e^{-t/\tau}}{\tau}dt = \frac{1}{1-\tau x}$$

Hence:

$$\Phi_{\sum t_i}(x) = \left(\frac{1}{1-\tau x}\right)^n$$

With the change of variable: $z = 2\sum t_i/\tau$ this is transformed into:

$$\Phi_{\sum t_i}(x) = \left(\frac{1}{1-2z}\right)^n$$

This is the MGF of a χ^2_{2n}. Since the MGF is unique, we have:

$$z = 2\frac{\sum t_i}{\tau} \sim \chi^2_{2n}$$

The central β-interval is then:

$$\left[2\frac{\sum t_i}{X^2_{\frac{1-\beta}{2},2n}},\ 2\frac{\sum t_i}{X^2_{\frac{1+\beta}{2},2n}}\right]$$

and the upper and lower intervals are:

$$\left[0, 2\frac{\sum t_i}{X^2_{\beta,2n}}\right] \qquad \left[2\frac{\sum t_i}{X^2_{1-\beta,2n}},\ \infty\right]$$

18.3 CI for large samples

The methods discussed in the previous sections can be applied only in special conditions: if a variable could be identified whose density is independent of the unknown parameter. We will postpone the discussion of the general method to construct CI, now we want to discuss how we can use approximations to the exact CI.

The derivative of the logarithm of the likelihood function is our starting point. Under regularity conditions, for large samples:

$$\frac{\partial \log L}{\partial \theta} = \sum_i \frac{\partial \log f(x_i;\theta)}{\partial \theta}$$

From the Central Limit Theorem the derivative will be, asymptotically, normally distributed.

The expectation value of $\partial_\theta \log L$ is zero:

$$E\left(\frac{\partial \log f(x_i;\theta)}{\partial \theta}\right) = E\left(\frac{1}{f}\frac{\partial f}{\partial \theta}\right) = \int_{-\infty}^{+\infty} \frac{\partial f}{\partial \theta} dx = \frac{\partial}{\partial \theta}\int_{-\infty}^{+\infty} f\ dx = 0$$

The variance is:

$$V\left(\frac{\partial \log L}{\partial \theta}\right) = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)$$

hence the variable:

$$\Psi = \frac{\frac{\partial \log L}{\partial \theta}}{\sqrt{\left(-E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)\right)}}$$

is a standard normal variable and can be used to determine approximate CI.

Example 18.2 *Consider the Poisson distribution.*

$$f(n,\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

It is not possible to find a statistics with the properties discussed in the previous sections, we will see later how is possible to construct exact CI. Now we will discuss approximate CI based on the derivative of $\log L$.

$$\frac{\partial \log L}{\partial \mu} = \frac{n}{\mu}(\overline{n} - \mu)$$

and

$$-\frac{\partial^2 \log L}{\partial \mu^2} = \frac{n\overline{n}}{\mu^2}$$

The expectation value is:

$$-E\left(\frac{\partial^2 \log L}{\partial \mu^2}\right) = \frac{n}{\mu}$$

Hence:

$$\Psi = (\overline{n} - \mu)\sqrt{\frac{n}{\mu}} \sim N(0,1)$$

As an example, if we consider a 95% *confidence interval, corresponding to a normal deviate* ± 1.96, *the approximate CI is given by the solution of the equation:*

$$(\overline{n} - \mu)\sqrt{\frac{n}{\mu}} = \pm 1.96$$

Squaring and solving the second degree equation, we get:

$$\mu_{L,H} = \overline{n} + \frac{1.92}{n} \pm \sqrt{\frac{3.84\overline{n}}{n} + \frac{3.69}{n^2}} \approx \overline{n} \pm \sqrt{\frac{\overline{n}}{n}}$$

18.4 Construction of CI in the general Case

The trick used in the previous section (avoid the explicit dependence on the parameter) works only in special cases. It exists a general method for the construction of CIs, due to Neyman [45], that we will discuss here.

Assume that t is the estimate of a parameter θ, whose value is unknown. t is a function of the measured data x. Assume also that $f(t|\theta)$ is the known probability density of t given the value of θ. (The method works in general for a function of the x whose density can be computed exactly, but taking for that function the estimator of θ helps to grasp better the argument.) Let now consider a range of values (t_1, t_2) of t. We can compute the probability β to obtain a value t comprised in this range for any value of θ:

$$P(t_1 \leq t \leq t_2) = \int_{t_1}^{t_2} f(t|\theta)\, dt = \beta$$

Clearly this probability is a function of the parameter θ whose true value is unknown. It seems that this way takes us nowhere. However now we will go through a mathematical argument: let us compute the integral as a function of θ and make those deductions that will be valid for *any* θ and so also for the unknown true value.

Given β, we can compute the values: t_1 and t_2 that satisfy the above relation. $t_{1,2}$ are functions of θ, once β has been defined. We have seen that there are many ways to construct t_1 and t_2. However t_1 and t_2 can be made unique by requiring, for instance, that the range (t_1, t_2) is central:

$$\int_{-\infty}^{t_1} f(t|\theta)\, dt = \int_{t_2}^{\infty} f(t|\theta)\, dt = \alpha$$

with:

$$\alpha = \frac{1-\beta}{2}$$

If we fix β, we obtain:

$$t_1 = t_1(\theta) \qquad\qquad t_2 = t_2(\theta)$$

which define two lines in the $(t-\theta)$ plane (figure 18.3). By construction in any horizontal line ($\theta = const$) an interval (t_1, t_2) will be cut by the two lines such that the probability of measuring a value t in this interval is exactly β.

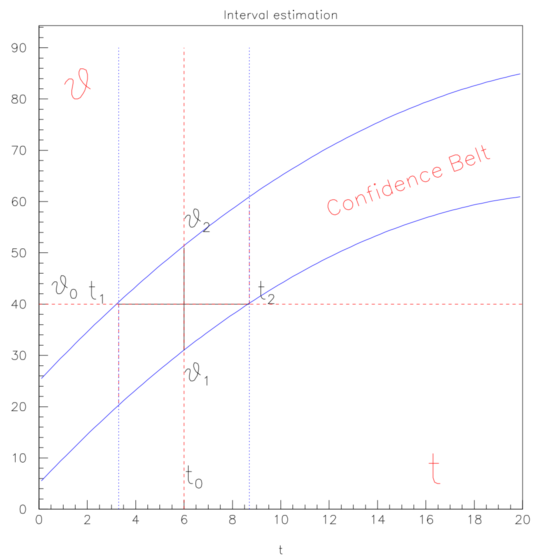


Figure 18.3: *The construction of the classical confidence interval.*

This diagram is useful since the same is true also for a vertical line: for any measured value t_0 the vertical line $(t = t_0)$ will define an interval (θ_1, θ_2) with a coverage probability β for the parameter θ. In a sense the diagram is constructed horizontally, but it is read vertically. The region comprised between $t_1(\theta)$ and $t_2(\theta)$ is called confidence belt.

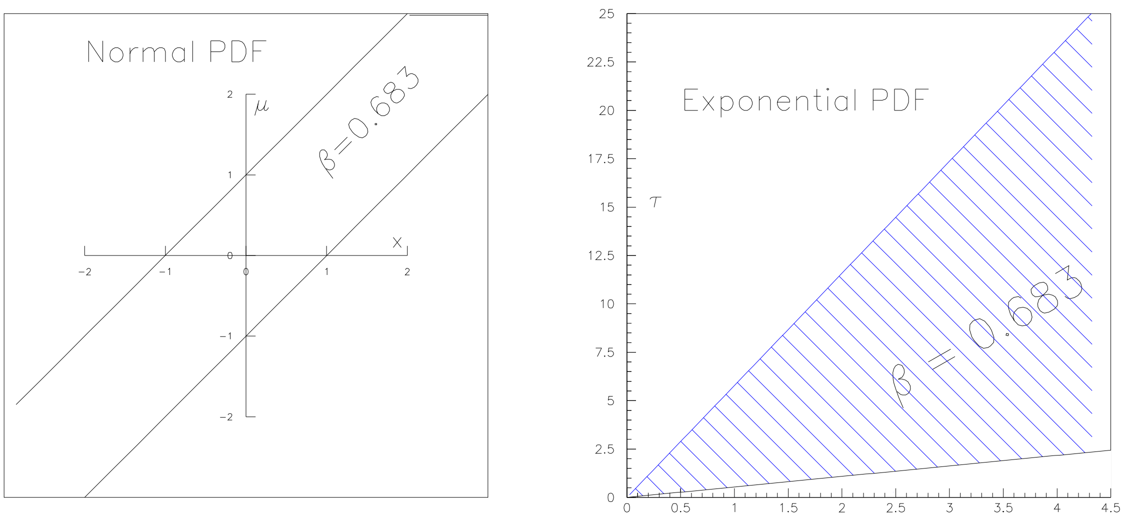
Remember that this is not a probability statement for θ but rather for the interval (θ_1, θ_2) which is a random variable, being a function of the random variable t_0. The interval comprises the true value θ in a fraction β of cases, in a long run of experiments.

We can convince ourself that the interval (θ_1, θ_2) as a coverage probability β in the following way. Assume that the true value of the parameter is θ_0. We make a measurement yielding t_0. By construction the interval (θ_1, θ_2) overlaps θ_0 if the measurement t_0 falls in the interval (t_1, t_2), i.e. in a proportion β of cases, as can be seen from figure 18.3. This is true for any value of θ_0, so it is true also for the real value of the parameter.

Of course, repeating the experiment, we would obtain different t_0 hence different intervals. In a long run the proportion of cases in which we cover the true value of the parameter is β and this is true even if we do not know its value.

The previous discussion shows that the coverage probability is a property of the belt as a whole, it is not a local property of a part of the belt. There is a lot of freedom in the construction of the belt and we have discussed only one method. The only strict requirement is that, for a fixed θ the interval (t_1, t_2) has a probability content at least of β.

Often physicists decide *after* they have made the measurement which confidence belt to use, choosing the one which gives the *best* physical result (central or upper/lower intervals). Of course this is not statistically correct, as described in detail in [48] since the ensemble of many of such results would produce a serious under-coverage in the resulting confidence belt.



,

Figure 18.4: *The construction of the classical confidence interval in the case of Normal (left) and Exponential density.*

Example 18.3 *Let us consider a Normal variable x with variance one:*

$$P(x, \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

The equations that define the central belt with a Confidence Level $=\beta$ are:

$$P(x \leq x_1 \mid \mu) = \frac{1-\beta}{2} \qquad P(x \geq x_2 \mid \mu) = \frac{1-\beta}{2} \tag{18.1}$$

And let us consider the case: $\beta = 0.683$ (one σ). The equations are:

$$x_1 = \mu - 1 \qquad x_2 = \mu + 1$$

that are graphically shown in figure 18.4 (left).

Example 18.4 *Let us now look at an exponential variable:*

$$P(t\tau) = \frac{e^{-\frac{t}{\tau}}}{\tau}$$

The equation of the central belt of the previous example 18.3 still hold and give the following equations for the belt boundaries:

$$t_1 = -\tau \log\left(\frac{1+\beta}{2}\right) \qquad t_2 = -\tau \log\left(\frac{1-\beta}{2}\right)$$

The belt is shown in figure 18.4 (right).

18.4.1 The lower Limit

Let us analyze a bit more closely a particular case of the confidence belt, that which define the lower limit of the parameter. Mathematically this means that the upper arm of the belt is moved to infinity and the belt is the part of the plane above the line $t_2(\theta) \equiv t_L(\theta)$.

The equation is:

$$P(t \leq t_L \mid \theta) = \beta$$

(β is the Confidence Level.) If the statistics t is continuous then the function t_L is regular, a strictly monotonously increasing function of θ (see figure 18.5).

In this case the relation: $t \leq t_L(\theta)$ can be inverted: $\Theta_L = t_L^{-1}(\hat{t}) \leq \theta$. the two expressions:

$$t \leq t_L(\theta) \qquad \Theta_L(t) \leq \theta$$

are equivalent and we can write:

$$P(\theta \geq \Theta_L) = \beta$$

(Once more: Θ_L is the random variable to which the probability relation applies.)

(Don't get trapped by the figure: if $\theta \leq \Theta_L$ the relation $\hat{t} \leq t_2$ would not *always* be true (for instance $\theta = 0$); instead it is clearly always true if $\theta \geq \Theta_L$.)

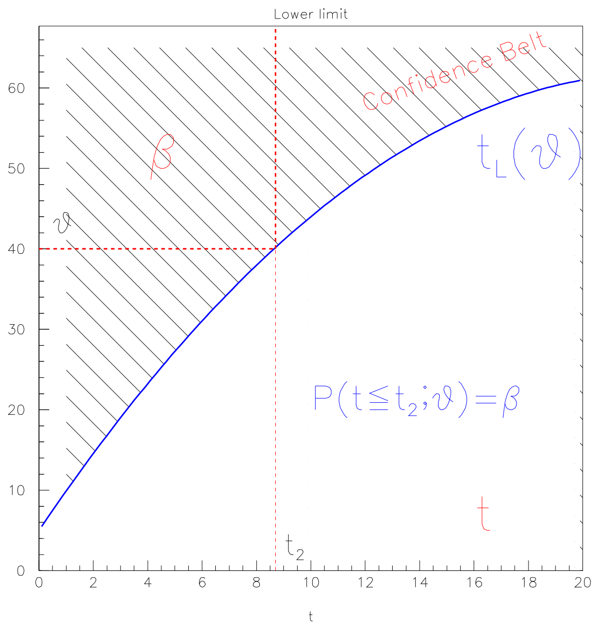


Figure 18.5: *Low limits in case of discrete variables.*

18.4.2 The Lower limit in Case of discrete Variable

If the variable can assume only discrete values, the method is basically the same as in the continuous case, but some more care has to be taken, since the equality:

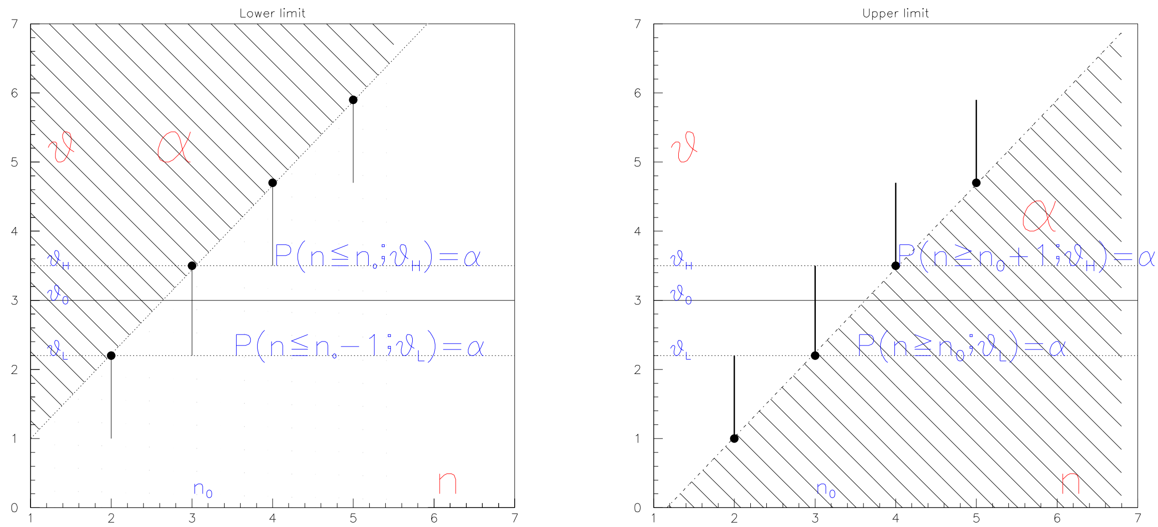
$$P(n \leq n_0 \mid \theta) = \beta$$

cannot always be satisfied. To construct the belt it is preferred to *overcover* i.e. replace the previous relation with:

$$P(n \leq n_0 \mid \theta) \geq \beta \qquad (18.2)$$

and choose the minimum value of n_0 for which the previous equation holds. The previous figure is replaced by a staircase (figure 18.6). For a given value of $\theta = \theta_0$, we have to accept the value n_0 since:

$$P(n \leq n_0 - 1 \mid \theta) < \beta \qquad \text{and} \qquad P(n \leq n_0 \mid \theta) > \beta$$



,

Figure 18.6: *Lower (left) and Upper (right) confidence belts.*

Since the function n_0 is strictly non decreasing, there is a value $\theta_H > \theta_0$ for which the equality holds:

$$P(n \leq n_0 \mid \theta_H) = \beta \qquad \text{(full dots in figure)}$$

We would like to manipulate the expressions $n \leq n_0$ as we have done in the continuous case:

$$[n \leq n_0(\theta)] \qquad \left[\theta \geq n_0^{-1}\right]$$

However in the discrete case the inverse does not exist: for a given n_0 there is a range of $\theta \in [\theta_L,\ \theta_H]$ for which the probability relation holds. We need some more care in the manipulation of these expressions.

The two propositions: $[n \leq n_0]$ and $[\theta > \Theta_L]$ are not equivalent in the discrete case, rather:

$$[n \leq n_0] \subseteq [\theta > \Theta_L]$$

Hence:

$$P(\theta > \Theta_L) \geq P(n \leq n_0) \geq \beta$$

Θ_L has to be the *smallest* θ. This can be seen geometrically from figure 18.6 (left). Also from the figure we can read the lower limit of the parameter.

Once measured n_0 (see figure 18.6) the lower limit on θ at a confidence level β is the smallest of $\theta \in [\theta_H,\ \theta_L]$. Hence it is obtained by the solution of the equation:

$$P(n \leq n_0 - 1 \mid \theta_L) = \beta \qquad \text{or} \qquad P(n\ < n_0 \mid \theta_L) = \beta \qquad \text{or} \qquad P(n \geq n_0) = 1 - \beta \qquad (18.3)$$

18.4.3 The upper Limit

This case is, mathematically, very similar to the lower limit: we move the lower arm of the belt (figure 18.3) to infinity so that the belt is the half space below the line $t_1(\theta)$ $(= t_H(\theta))$. The equation of the line is:

$$P(t \geq t_H \mid \theta) = \beta$$

or

$$\int_{t_H}^{\infty} f(t)dt = \beta$$

which has to be solved for t_H.

18.4.4 The upper Limit (discrete)

The same arguments as for the lower limits holds also in this case. The difference here is that the equation for which we have to find the *largest* θ is:

$$P(n \geq n_0 \mid \theta) \geq \beta$$

This is shown graphically in figure 18.6 (right). The appropriate value for the upper limit Θ_H is the solution of one of the equations:

$$P(n \geq n_0 + 1 \mid \theta) = \beta \qquad \text{or} \qquad P(n > n_0 \mid \theta) = \beta \qquad \text{or} \qquad P(n \leq n_0 \mid \theta) = 1 - \beta \tag{18.4}$$

18.4.5 Lower and Upper Limits in the Poisson and Binomial Distribution

The solution of equations 18.3, 18.4 is facilitated in the case of a Poisson variable since the following relation holds:

$$P(n \geq K \mid \Theta) = P(\chi^2_{2K} \leq 2\Theta)$$

The right hand side is the inverse of the cumulative density of a chi-square which is tabulated (short table is given in the appendix).

A similar relation holds also in the case of a Binomial distribution. In this case:

$$P(n \geq n_0 \mid \theta, M) = P(\beta_{n_0, M-n_0+1} \leq \theta)$$

where β is the incomplete *Beta* function:

$$\beta_{n_0,M-n_0+1}(u) = \frac{\Gamma(M+1)}{\Gamma(n_0)\Gamma(M-n_0+1)} u^{n_0-1}(1-u)^{M-n_0}$$

A summary of the formulas that define the upper and lower limits for Poisson and Binomial distributions is presented in table 18.5 and table 18.6 shows numerical values of the limits in some special cases.

	The *Poisson* Distribution	The *Binomial* Distribution
Θ_H	$1 - \beta = \int_0^{2\Theta_H} \chi^2_{2n_0}(x)dx$	$1 - \beta = \int_0^{\Theta_H} \beta_{n_0,M-n_0+1}(u)du$
Θ_L	$\beta = \int_0^{2\Theta_L} \chi^2_{2(n_0+1)}(x)dx$	$\beta = \int_0^{\Theta_L} \beta_{n_0+1,M-n_0}(u)du$

Table 18.5: *Upper and lower limits for a Poisson and Binomial distributions*

By an appropriate set of the Confidence Level β we can construct the central intervals. This is shown in figure 18.7. The dots at the upper extreme of the segments mark the values of the

n_0	$\beta = 0.8413$		0.90		0.95		0.99	
	Θ_H	Θ_L	Θ_H	Θ_L	Θ_H	Θ_L	Θ_H	Θ_L
0	-	1.84	-	2.30	-	3.0	-	4.60
1	.173	3.30	.105	3.89	.051	4.47	.010	6.64
2	.709	4.64	.532	5.32	.355	6.29	.149	8.41
3	1.38	5.92	1.102	6.68	.818	7.75	.436	10.05
4	2.09	7.16	1.745	7.99	1.366	9.15	.823	11.60
5	2.84	8.38	2.433	9.27	1.970	10.51	1.279	13.11
6	3.62	9.61	3.152	10.53	2.613	11.84	1.785	14.57
7	4142	10.80	3.895	11.77	3.285	13.15	2.330	16.0
8	5.23	11.97	4.656	12.99	3.981	14.43	2.906	17.40
9	6.05	13.14	5.432	14.21	4.695	15.70	3.507	18.78
10	6.89	14.27	6.221	14.40	5.425	16.96	4.130	20.14
15	11.17	20.0	10.300	21.29	9.246	23.01	7.477	26.74
20	15.67	25.54	14.525	27.05	13.255	29.06	11.082	33.10

Table 18.6: *Lower and Higher limits for the mean of a Poisson process as a function of the number of observed events.*

parameter for which the equality in equations **??** holds. Assume that we have observed n events, the confidence intervals for the parameter are obtained reading from the figure the values of μ at the *outer* extremes of the segments which define the two arms of the belt.

Let us look now more closely the Poisson case. Table 18.7 shows, for some values of the outcomes the value of μ_L and the probabilities of a Poisson process for some value of μ. The lowest value of μ_L is equal to 2.30, obtained in case the the outcome of the measurements is zero. However there is a non negligible probability of getting zero even if μ is equal to one or two. Due to the discreteness of the outcomes, if the true value of the parameter is smaller than 2.30, the probability of the proposition: $\theta \leq \theta_L$ is one.

Consider now the column $\mu = 5$. The probability of $n < 2$ is less than 4. In the case of $n = 0, 1$, μ_L is smaller than 5 and is just above 5 for $n = 2$. This means that if μ would have had the value of 5 the coverage would have been larger than the 90%, expected from the construction of μ_L. Our construction cannot give a coverage probability exactly equal to β but *at least* β. This is shown graphically in figure 18.8.

n	μ_U	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\mu = 6$	$\mu = 7$
0	2.30258	0.36788	0.13534	0.04979	0.01832	0.00674	0.00248	0.00091
1	3.89032	0.36788	0.27067	0.14936	0.07326	0.03369	0.01487	0.00638
2	5.32245	0.18394	0.27067	0.22404	0.14653	0.08422	0.04462	0.02234
3	6.68071	0.06131	0.18045	0.22404	0.19537	0.14037	0.08924	0.05213
4	7.99355	0.01533	0.09022	0.16803	0.19537	0.17547	0.13385	0.09123
5	9.27465	0.00307	0.03609	0.10082	0.15629	0.17547	0.16062	0.12772
6	10.53206	0.00051	0.01203	0.05041	0.10420	0.14622	0.16062	0.14900
7	11.77091	0.00007	0.00344	0.02160	0.05954	0.10444	0.13768	0.14900
8	12.99471	0.00001	0.00086	0.00810	0.02977	0.06528	0.10326	0.13038
9	14.20598	0.00000	0.00019	0.00270	0.01323	0.03627	0.06884	0.10140
10	15.40664	0.00000	0.00004	0.00081	0.00529	0.01813	0.04130	0.07098

Table 18.7: *Probability of n events in a Poisson process.*

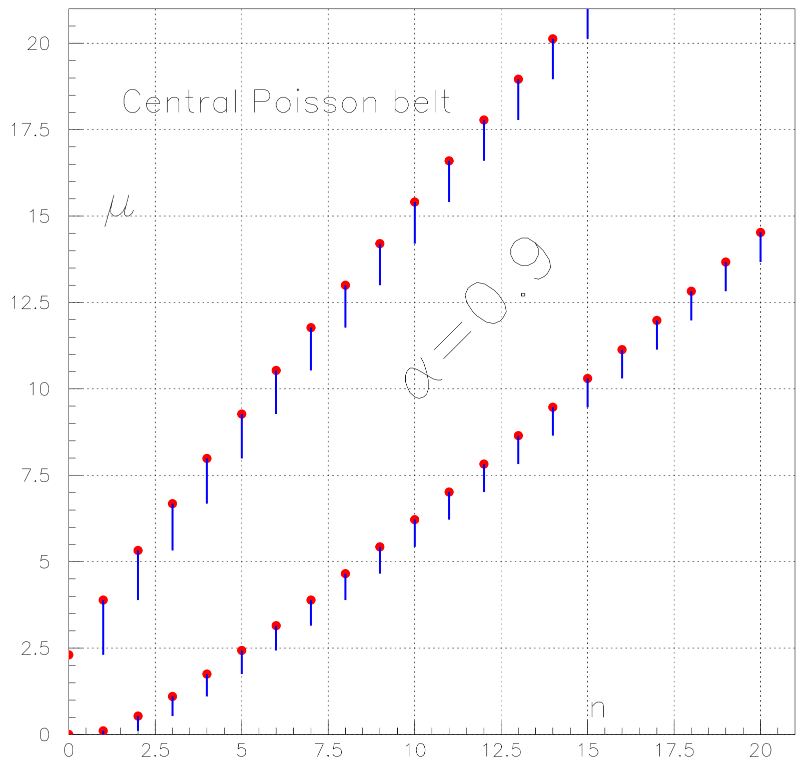


Figure 18.7: *A 90% central confidence belt for a Poisson variable.*

18.5 The Unified Approach

The belt is a powerful tool to construct the exact (whenever possible or at least a conservative) confidence interval in the general case. The only requirement for the construction of the belt is that the probability content in each segment $[t_1,\ t_2]$ is at least the Confidence Level β:

$$P\left(t \in [t_1,\ t_2]\right) \leq \beta$$

We are free to define any interval which has the above property. In the previous paragraphs we have considered three possible belts, leading to central, upper or lower limits in the space of the parameter. In all these cases we have started from the equation:

$$P\left(t \leq t_1 \mid \theta\right) = \alpha \tag{18.5}$$

α has the appropriate value to construct the belt.

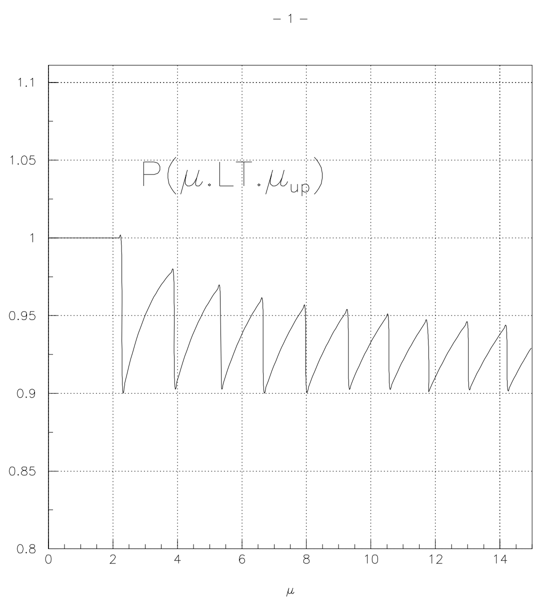


Figure 18.8: *Coverage probability for a Poisson process.*

There are other ways to construct the belt.
Crow and Gardner [53] have considered a Poisson process and defined an interval $[n_1,\ n_2]$ by minimizing its length. Defining a length is always arbitrary, since we have always to specify the metric we intend to use. In this case the discreteness of the variable make the definition of length unambiguous, it is just the difference $n_2 - n_1$. One then order n by decreasing values of $P(n \mid \theta)$ (n_1 corresponds to the largest probability) and takes all the n_i till $\sum_i P(n_i \mid \theta) \geq \beta$. This is a Neyman interval, different from those that we have discussed since equation 18.5 does not hold. The Crow-Gardner intervals are interesting since they introduce the concept of an ordering principle to construct an interval which is shorter than a central one.

Another interesting approach was introduced by Feldman and Cousins [48]. Before discussing the Feldman-Cousins strategy let us describe briefly its motivations with a simple example.

Example 18.5 *Assume that we are measuring the mass of an elementary particle via:* $M^2 = E^2 - p^2$. *The true value of the parameter* $\theta = M^2$ *has to be positive but, being the difference of two numbers, it can happen that it is found to be negative in some cases. In this case the point estimate is rather meaningless and we need to quote a Confidence Interval.*

Assume that we know, by ancillary experiments that $M^2 \sim N(\theta, 1)$ *(the resolution of our apparatus is one and the distribution is Normal). Let us suppose that our measurement has given the value* $\hat{M}^2 = -1.5$, *outside the physical region for* M^2. *It would be meaningless to quote this result of our experiment. But the Confidence Intervals are not any better, since by inspection of figure* **??** *we see that a* 0.683 *CI would read:*

$$P\left(M^2 \in (-2.5, -0.5)\right) = 0.683$$

Which is a wrong statement, since we know that the probability of that interval is zero.

We might argue that we have not implemented the physical requirement that the mass squared has to be positive, in the construction of the belt and we have used blindly the Normal density even for values of the parameter which has no physical meaning. If we take into account the physical boundary, the new belt would be equal to the old one in the upper half plane and non existing in the lower half plane. In such a belt our confidence interval would be empty. This is more comfortable on a conceptual ground but does not solve our problem.

The empty CI are not a pathology of the belt since we know that the coverage probability has a meaning over a long run of repeated measurements and not just the only one that we have made. Our measurement belong to those 15% measurements that fall on one side of the central interval. But we still have to quote a result of the measurement and we may argue that the experiment has indicated that the mass is small (smaller that our resolution) hence we decide to quote an upper limit for the mass of the particle, for instance a 90% limit is $M_U^2 = 0.28$ *(see* **??***). (A 683% limit would again be problematic.) This procedure is mathematically correct but has a serious statistical problem, since we have decided what limit to quote after we have seen the result of the measurement. If we construct the belt choosing the upper or central limit on the basis of the experiment the belt would have serious under-coverage problems, as discussed in [48].*

Feldman and Cousins suggest to construct the belt in a way to correct these problems:

- avoid, in most of practical cases the intervals in the unphysical regions (or the empty CI);
- leave no freedom to choose the intervals (upper, lower, central etc.) by changing from one sided (lower, upper) to a two sided intervals as the measurement become statistically more significant, while keeping the correct coverage.

Let us illustrate how it works with an example.

Let us consider a Poisson process with background:

$$n \sim \frac{\mu^n e^{-\mu}}{n!} \qquad \mu = \mu_s + \mu_b$$

With ancillary experiments we determine accurately the value of the background rate μ_b. The likelihood estimate of μ_s is $\mu_{ML}\hat{\mu_s} = n - \mu_b$, once we impose the physical boundary on the signal rate. We see that this case is similar to the case of the Normal variable discussed above.

To be specific let us assume that $\mu_b = 3$ and construct the belt for $\mu_s = 1$. (see table 18.8).

n	$P(n \mid \mu)$	μ_B	$P(n \mid \mu_B)$	R	Rank
0	0.0183165	0.0	0.0497875	0.3678795	8
1	0.0732635	0.0	0.1493615	0.4905065	6
2	0.1465255	0.0	0.2240425	0.6540085	4
3	0.1953675	0.0	0.2240425	0.8720115	3
4	0.1953675	1.0	0.1953675	1.0000005	1
5	0.1562935	2.0	0.1754675	0.8907275	2
6	0.1041965	3.0	0.1606235	0.6486965	5
7	0.0595405	4.0	0.1490035	0.3995925	7
8	0.0297705	5.0	0.1395875	0.2132745	8
9	0.0132315	6.0	0.1317565	0.1004225	9
10	0.0052925	7.0	0.1251105	0.0423035	10
11	0.0019255	8.0	0.1193785	0.0161215	11
12	0.0006425	9.0	0.1143685	0.0056095	12
13	0.0001975	10.0	0.1099405	0.0017955	13
14	0.0000565	11.0	0.1059895	0.0005325	14

Table 18.8: *The Feldman-Cousins ordering method* ($\mu_s = 1$).

The first column of table 18.8 is the outcomes of the measurement $n = 0, 1 \cdots$, the second is the probability of the outcome ($P(n \mid \mu = \mu_s + \mu_b)$. The third column is the value of μ_{ML} and the forth is $P(n \mid \mu_{ML} + \mu_b)$. The ordering is performed using the ratio $R = P(n \mid \mu_s + \mu_b)/P(n \mid \mu_{ML} + \mu_b)$, that is we measure the probability of n with the highest probability of the process, allowed by the physical boundaries.

We start with the values if n corresponding to the largest value of R and we keep accepting the ns in decreasing order of R till the sum of the probabilities $\sum_i P(n_{|}\mu_s + \mu_b)$ meets or exceeds the confidence level. The column labeled *Rank* indicates the order of ns. This procedure is repeated for all values of μ_s. The result is shown in figure **??**. Once the measurement is done and assume that the value $n = 7$ is found then we can read the limits from the figure as the outer extremes of the two segments limiting the belt (see the figure). If $n < 5$, in this case, only the upper bound is given showing how the belt cope with the transition from central to upper bound.

In this way we avoid, almost always, the construction of empty sets

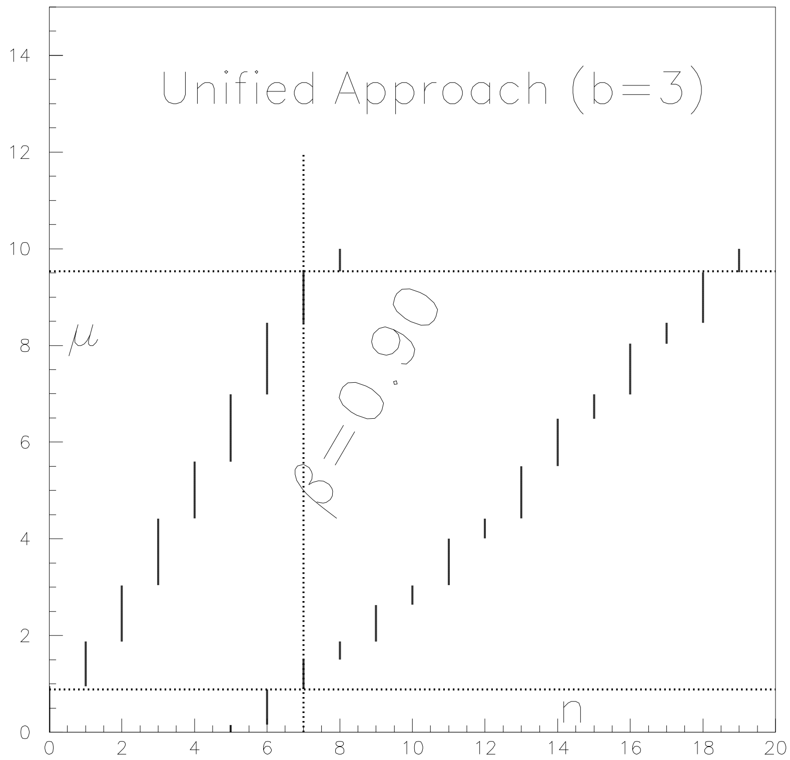


Figure 18.9: *A 90% F-C confidence belt for a Poisson variable.*

Chapter 19

Bayesian Inference

19.1 The Posterior Density

In Bayesian statistics the inference is performed by the posterior density. The posterior *p.d.f.* $P(\theta \mid \vec{x}, I)$ contains all the information about the parameter to be measured. The parameter is unknown hence in Bayes statistics is subject to probability analysis. The Bayes theorem is at the basis of the construction of the posterior distribution:

$$P(\theta \mid D, I) \propto P(D \mid \theta, I) \cdot P(\theta \mid I)$$

The ingredients are:

- the likelihood function $P(D \mid \theta, I)$,
- the prior distribution $P(\theta \mid I)$,
- D is the data: $\vec{x} = \{x_1, x_2 \cdots x_n\}$ and
- I is the relevant *background* information.

There is no such a thing as *absolute probability.* All probability statements are conditional to some other information. Sometimes the background information is not explicitly expressed but logically is part of the problem. In many fields of science the background information is written in all the textbooks that students are required to study and which forms the substrate of all scientific work. In this sense the background information is what makes possible the scientific exchange of opinions and results.

Example 19.1 *The probability of the statement it will rain tomorrow will depend on whether it is winter or summer, or if there are clouds etc.*

Prior probability contains all the previous knowledge on the parameter. This could be either the result of previous experiments or reflect the personal opinion of the experimenter. Via Bayes theorem the information is updated and increased at each new measurement.

This is often the case. However in Physics we often pretend to ignore previous results and insist in analyzing our experiment independently of the results of previous experiments. The task of making an average with world data is left to a later stage of data reduction. This method is important and permits to compare independent measurements and judge their consistency.

For instance the mass of electron neutrino has been measured by many different experiments each being analyzed independently. A compilation of some results of this important quantity is shown in figure 19.1. In fact what is plotted is the mean value of the *square* of the electron mass,

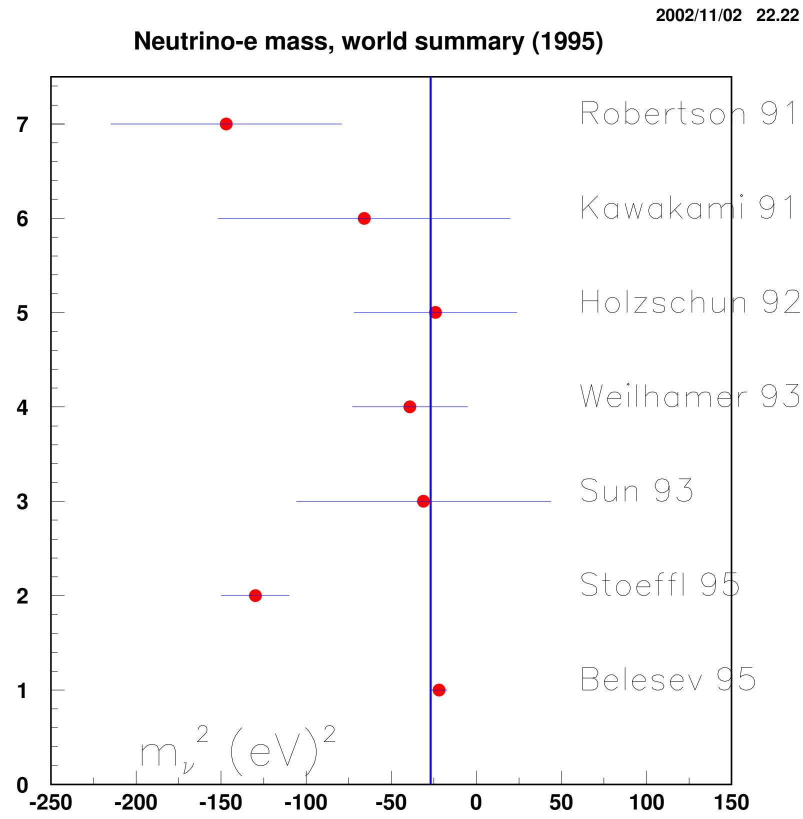


Figure 19.1: *A compilation of results of neutrino mass squared measurements.*

obtained by the end point of Kurie plot of tritium β decay. The world average is the computed only after the different experiments have been checked to be consistent.

The insistence in Physics to ignore previous results raises the problem of which prior $P(\theta \mid I)$ should we use to construct, via Bayes' theorem, the *posterior density*:

$$P(\theta \mid D, I) \propto P(D \mid \theta, I) \cdot P(\theta \mid I)$$

The prior should then convey in a mathematical form our ignorance (true or pretended) on the value of the parameter. The prior should also minimally influence the form of the posterior that will used for the inference. This is the crucial problem of Bayes statistics and needs a more careful discussion.

19.2 Bayes Inference in Binomial Distribution

Let us postpone the discussion of the best prior to represent our ignorance on the parameter. For the time being a flat prior is assumed:

$$P(\theta \mid I) = const$$

In the case of a Binomial experiment if H is the number of *heads* and T the number of *tails* in a binomial trial of $T + H$, the posterior distribution is:

$$P(\theta \mid D, I) = K \cdot \theta^H \cdot (1 - \theta)^T$$

The normalization is computed directly:

$$1 = K \cdot \int_0^1 \theta^H (1-\theta)^T \; d\theta = K \cdot \frac{\Gamma(H+1)\Gamma(T+1)}{\Gamma(H+T+2)} = K \cdot \frac{H!T!}{(H+T+1)!}$$

hence:

$$K = \frac{(H+T+1)!}{H!\; T!}$$

At this point, once we have measured the number of heads and tails, all the information is gathered into the posterior distribution and we can proceed with the point estimate and the determination of the credible intervals. The point estimate can be either the *mode* or the *mean*:

- The mode is the maximum of the density (with this choice of prior, the mode coincide with the Maximum Likelihood point estimate, of course the meaning is different):

$$\theta_{mode} = \frac{H}{H+T}$$

- The mean is:

$$\overline{\theta} = \int_0^1 \; \theta \cdot P(\theta \mid D, I) \; d\theta = \frac{H+1}{H+T+2}$$

  Notice that the mean has no intuitive meaning, from the point of view of frequency theory. The mean, in Bayes theory is computed over the parameter space and not over the sample space, as in the Frequency theory. We cannot expect that it coincides with our intuitive expectation over repeated trials.

The *Credible Intervals* are also computed from the posterior density. Also in this case we have to determine a region of the parameter space that has a given probability content. The equation:

$$\alpha = \int_{\theta_L}^{\theta_H} P(\theta \mid D, I) \; d\theta$$

defines a *credible interval*: (θ_L, θ_H) such that the probability that $\theta \in \{\theta_L, \theta_H\}$ is α.

We will use only *Central Intervals*:

$$(1-\alpha)/2 = \int_{-\infty}^{\theta_L} P(\theta \mid D, I) \; d\theta \qquad (1-\alpha)/2 = \int_{\theta_H}^{\infty} P(\theta \mid D, I) \; d\theta$$

Figure 19.2 shows the posterior distribution after each trial. We have simulated a real toss of coins with even probability of heads and tails. The clear part of the posterior of each curve has a probability content of 68.3%. The curves are modified after each measurement showing how the information is updated. Figure 19.3 shows the posterior after 100 trials.

It is interesting to see how the information is updated. Before the measurements the posterior coincides with the prior:

$$\begin{aligned}
P(\theta \mid noData, I) &= P_0(\theta \mid I) = 1 \\
P(\theta \mid H, I) &\propto \theta \\
P(\theta \mid H, H, I) &\propto \theta^2 \\
P(\theta \mid H, H, , T, I) &\propto \theta^2(1-\theta) \\
&\dots \\
P(\theta \mid nN, mT, I) &\propto \theta^n(1-\theta)^m
\end{aligned}$$

In the previous example it is shown how the posterior, at each trial, becomes the prior of the following one. It is also clear that the posterior obtained from the final score coincide with the one built on each trial.

In addition the posterior does not depend on the order of the outcomes neither on the *stopping rule*, i.e. the rule that has driven the data collection and decided the end of the experiment.

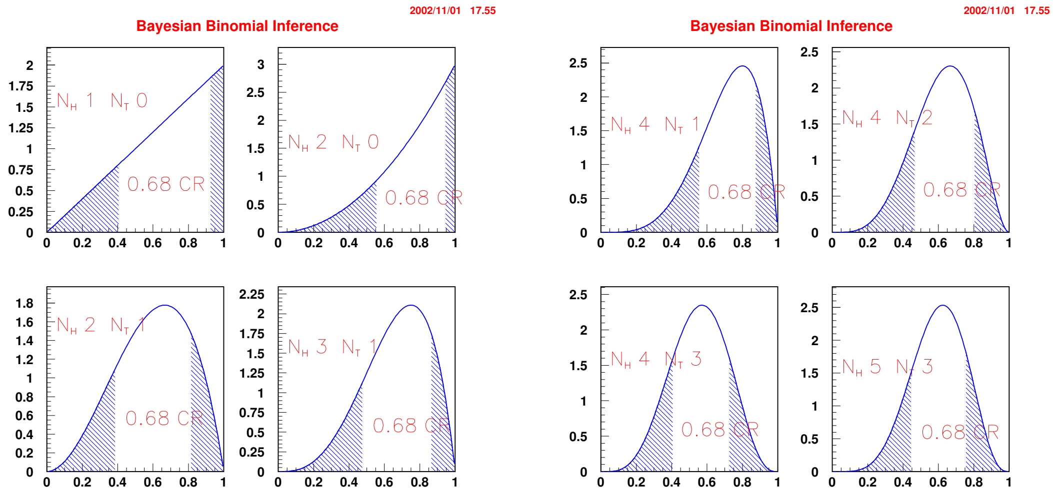


Figure 19.2: *Change in the posterior probability after each measurement (flat prior).*

19.3 Jefreys Prior

Another prior that we will discuss later on is the so called *Jeffreys* prior:

$$P(\theta \mid I) \propto 1/\theta$$

The point estimate is, for this prior:

- The mode is the maximum of the density:

$$\theta_{mode} = \frac{H-1}{H+T-1}$$

- The mean is:

$$\overline{\theta} = \int_0^1 \theta \cdot P(\theta \mid D, I)\ d\theta = \frac{H}{H+T+1}$$

The binomial experiment described in the previous section is analyzed again with the use of this prior. We might suspect that the results will be different, since the two priors are rather different. The plots of the posterior probability with Jeffreys prior is shown in figure 19.4 and a direct comparison of the two priors at each trial is shown in figure 19.5

19.4 The Choice of the Prior

There are many difficulties in defining a *logically consistent and realistic* representation of prior ignorance.

It is not difficult to convince oneself that the uniform density does not represent consistently the ignorance on the parameter with two examples, one from discrete and one from continuous densities.

Example 19.2 *$\theta \in \{\theta_1, \theta_2 \cdots \theta_k\}$ with $k \geq 3$. Nothing is known on θ hence we assume: $P(\theta = \theta_i) = 1/k$. We define now a new variable ϕ such that: $\phi = 1$ if $\theta = \theta_1$, else $\phi = 0$. Being ignorant on θ we are also ignorant on ϕ but the two recipes:*

$$P(\phi = 1) = \frac{1}{2} \qquad \textit{and} \qquad P(\theta = \theta_1) = \frac{1}{k}$$

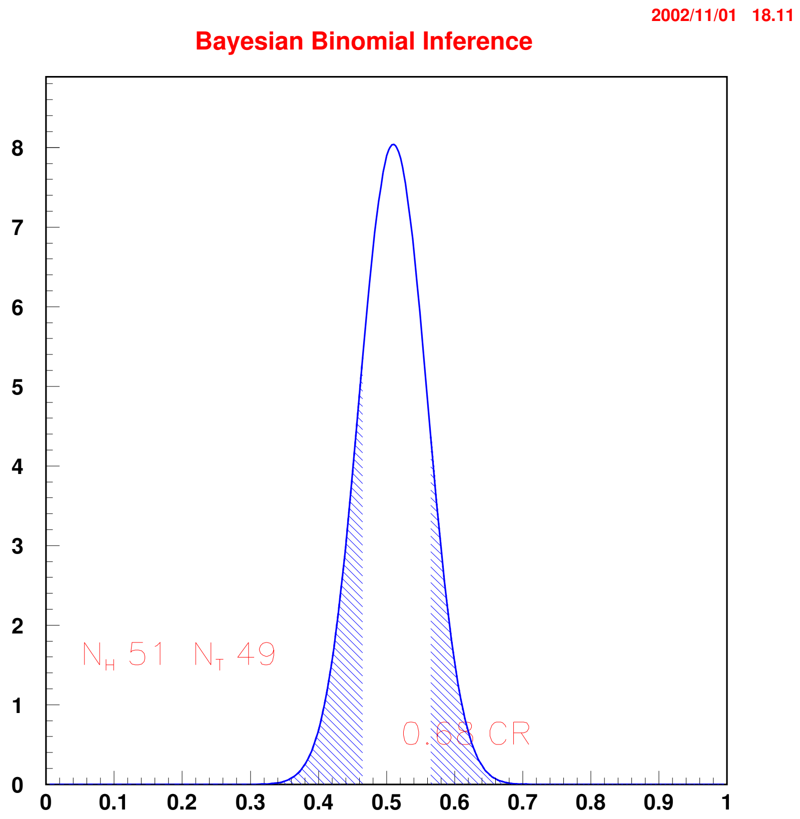


Figure 19.3: *Posterior probability after one hundred trials (flat prior).*

are inconsistent.

Example 19.3 *If θ has a continuous distribution the ignorance on θ implies ignorance on $g(\theta)$. (We assume for simplicity a one-to-one transformation).*

$$\theta \in \{0, 1\}$$

A proper prior is $f(\theta) = 1$. Now let $\phi = \theta^2$. The assertion $f(\phi) = 1$ implies $f(\theta) = 2\theta$: θ and ϕ cannot both have uniform densities.

Assume that the prior on the parameter θ is $f_0(\theta)$, and consider the transformation (one-to-one):

$$\phi = g(\theta)$$

The proper prior for ϕ is:

$$f_0(\phi) = f_0\big(g^{-1}(\phi)\big) \cdot \Big|\frac{\partial g^{-1}(\phi)}{\partial \phi}\Big|$$

Jeffreys (1967) proposes the following solution: if the Fisher *Information* is I_θ

$$I_\theta = -E\Big(\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2}\Big)$$

the one should use the prior:

$$f_0(\theta) \propto \big|I_\theta\big|^2$$

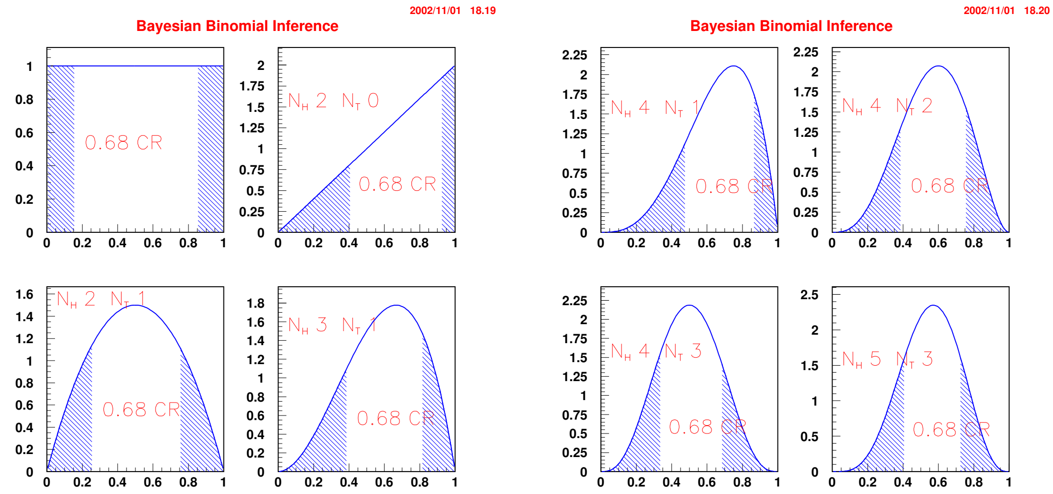


Figure 19.4: *Change in the posterior probability after each measurement. The prior is now the Jeffreys prior.*

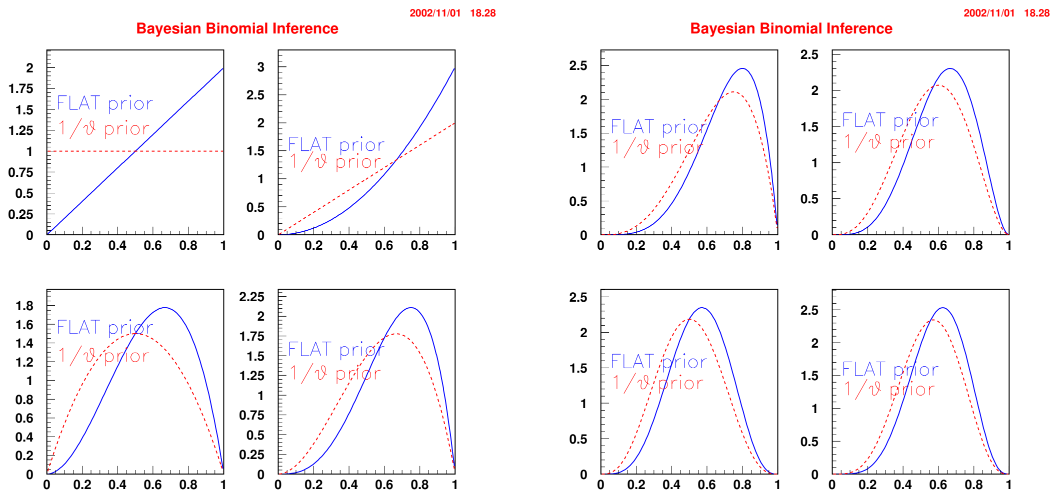


Figure 19.5: *Comparison of posterior probability for the two priors.*

Let us consider a few cases.
If θ is the mean of a normal distribution and the inference is performed from a set of data:

$$\{x_1, x_2 \cdots x_n\} \sim N(\theta, \sigma^2)$$

$$f(\vec{x} \mid \theta) \propto exp(-n(\overline{x} - \theta)^2/\sigma^2)$$

Then the prior suggested by Jeffreys is:

$$\frac{\partial^2 \log f}{\partial \theta^2} = -\frac{n}{\sigma^2} \quad \rightarrow \quad f_0(\theta) \propto \sqrt{\frac{n}{\sigma^2}} \propto 1$$

Therefore in the case of an *unbounded location* parameter the use of a flat prior is justified by Jeffreys argument.

The deep reason for this choice can be better understood from the following, less formal argument. If we do not have any knowledge on θ then the prior should be invariant under the transformation:

$$\theta \to \theta + \theta_0$$

then we have:

$$P(\theta \mid I)\ d\theta = P(\theta + \theta_0 \mid I)\ d(\theta + \theta_0) = P(\theta + \theta_0 \mid I)\ d\theta$$

Hence:

$$P(\theta \mid I) = constant$$

It is essential that the parameter is known to be unbounded.

In the following case the parameter is bounded from below. Let θ be the variance of a Normal density that we infer from the data:

$$\begin{aligned}
\{x_1, x_2 \cdots x_n\} &\sim N(\mu, \theta) \\
f(\vec{x} \mid \theta) &\propto \theta^{n/2} exp(-\sum (x_1 - \mu)/2\theta) \\
\frac{\partial^2 f}{\partial \theta^2} &= \frac{n}{2\theta^2} - \frac{\sum (x_i - \mu)^2}{\theta^3}
\end{aligned}$$

Hence:

$$E(\frac{\partial^2 f}{\partial \theta^2}) = \frac{n}{2\theta^2} - \frac{n\theta}{\theta^3} = \frac{n}{\theta^2} \to f_0(\theta) \propto \frac{1}{\theta}$$

In this case we have the $1/\theta$ prior that we have used in a previous example.

An intuitive argument for this prior is the following.

The problem here is the inference of a bounded parameter whose scale we do not know. (We are called to measure the length of fishes but we ignore if they are whales or shrimps.) The probability density has to be the same in both cases; it means that it has to be invariant under the transformation:

$$\theta \to k \cdot \theta$$

In formulas:

$$P(\theta \mid I) d\theta = P(k \cdot \theta \mid I) d(k \cdot \theta) = k \cdot P(k \cdot \theta \mid I) d\theta$$

Hence:

$$P(\theta \mid I) \propto \frac{1}{\theta}$$

19.5 Frequency and Probability

We will go now in more details in the study of the different concepts at the basis of the two schools of Statistics. It may appear that the greatest difference between Bayes and Frequency statistics is the concept of *p.d.f.* of a parameter. There are cases where a parameter has a frequency distribution, even in the Frequency school of statistics.

Example 19.4 *The mean lifetime of transistors (τ) depends on the production batch. Within a production batch it can be considered constant but varies from batch to batch. It is then reasonable to speak of probability distribution of transistor mean lifetime $P(\tau)$. The measurement of lifetime of individual transistors in each batch has a well defined frequency distribution $f(t \mid \tau)$ which depends on a parameter that has itself a frequency distribution within batches. Both individual transistor lifetimes and mean lifetime are random variables.*

In the previous example there is nothing like absolute quantities, the *transistor mean lifetime* τ depends on the details of production batch. There are cases, instead, where we attempt to estimate a *real constant*, like the mass if a particle, where the frequency distribution is not allowed in Frequency Statistics.

Example 19.5 $\{x_1, x_2 \cdots x_n\}$ *are measurements of the resistivity of a particular alloy whose true resistivity is* θ*. Here the data are repeatable (I can continue my measurements as long as I like) with a well defined frequency probability. On the contrary the alloy is unique and it is quite difficult to conceive a frequency probability to* θ*.*

In *Frequency Statistics* θ is a fixed (though unknown) parameter and NOT a random variable. All probability statements are derived from the data for a given value of the parameter: $f(x \mid \theta)$. Probability is an intrinsic property of the system.

This point of view is considered very distorted by *Bayesian Statistics*. The *parameter* θ *is unknown hence subject to probability distribution.* On the contrary *data, once observed, are not any longer random variables, but fixed* data. (Of course *before* being observed data are random variables.) Bayesian Inference proceeds from the posterior distributions of the random parameters conditioned on the values of the observations.

19.6 The Likelihood Principle

The prior $f(\theta)$ and the likelihood $f(x \mid \theta)$ are both subjective distributions that express the experimenter's beliefs about θ and x prior to observing the data. There is no such a thing as *Absolute Probability*: Probability does not exist, it only reflects our ignorance.

Bayes theorem reads:

$$f(\theta \mid x) = \frac{f(x \mid \theta) \cdot f(\theta)}{\int f(x \mid \theta) \cdot f(\theta) d\theta}$$

The data affect the posterior distribution through the likelihood function $f(x \mid \theta)$. Note that if we modify the likelihood by an arbitrary multiplicative constant or a function of data, the posterior distribution does not change.

Consider two experiments yielding the data x and y. If the two likelihoods: $f(x \mid \theta)$ and $f(y \mid \theta)$ are identical up to a multiplicative factor, function of x (y), they contain identical information on the parameter θ and lead to identical posterior distribution. The two experiments might be very different in other respects but the difference is irrelevant for the inference on θ. This is the *Likelihood Principle*.

The Likelihood Principle represents the key difference between the Frequency and the Bayesian Statistics.

Example 19.6 *We have seen an intriguing example in the analysis of Gamma Ray Bursts data. Two experiments have obtained identical results: in* n *observation of the sky they have detected* r *GRB. But the first experiment had planned in advance* n *observations while the second stopped after having detected* r *GRB.*

$$f_A(x = r \mid \theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r}$$

$$f_B(y = n \mid \theta) = \binom{n-1}{r-1} \theta^r (1-\theta)^{n-r}$$

We have stressed the fact that, independently of the stopping rule, the two experiments had obtained the same result and (up to an arbitrary factor function of the data) the same likelihood *functions. Hence the same inference will result, in Bayesian Statistics. The Bayes Inference does not depend on the stopping rule.*

On the contrary, in Frequency Statistics, we have to know how the experiment was performed, since we must know the sample space. The sample space is required in order to compute the value of the parameter which is optimal (minimum variance, unbiasedness) in repeated samplesand not just in the particular experiments that we have performed. Frequency Inference depends on the stopping rule. The request of being optimal in repeated samples violates the Likelihood Principle since it makes reference to all possible values of the observations. The value of the parameter θ will be different for the two experiments.

Note: $f(x \mid \theta)$ and $f(y \mid \theta)$ are NOT identical for all values of x and y! We have selected $y = n$ and $x = r$ to make then identical, as it happened in this particular case. In general the two experiments will provide different informations; not in this case.

19.7 Discussion on Frequency and Bayes Theories

Bayesian inference always agrees with LP. Frequency minimum variance unbiased estimators of θ are different in the two cases. This is not surprising since unbiasedness itself violates the LP. The bias is defined as:

$$b(\theta) = E(t(x) \mid \theta) - \theta$$

The expectation is made on *all* values of x. This is in disagreement with LP which requires that all inference must be based on the observed values of x.

The Frequency school recommends to use the MVU estimator r/n for a binomial experiment. This is NOT based only on the likelihood $f(x = r \mid \theta) \propto \theta^r(1-\theta)^{n-r}$ but on the whole distribution $f(x \mid \theta)$ at values of x NOT observed. The difference in the analysis of the *Binomial* and the *Negative Binomial* experiment is because the two experiments differs over what *might* have been measured.

The essential of Frequency Inference is an inference rule on the basis of the performances in *repeated sampling*. Bias is one example, others are the variance of an estimator, the probability of first and second error etc. The Frequency inference inevitably depends on what might have been observed, but was not in the actual experiment. LP insists that inference must use only the probability of the actual observations.

This is the deep reason of the conflict the two school of Statistics. Frequency Statistics abandon the LP in favor of the repeated sampling. Frequency Statistics has desirable properties of the estimators *on the average* over the sample of all possible observations.

Example 19.7 *An experiment is designed to measure the lifetime of an unstable particle. The experiment was planned to be sensitive in an interval of times: $t \in (0, T)$. One unstable particle was produced and its decay time was observed: t_m. No other particle was produced that might have been escaped to the measure. Under these circumstances the experimenter asked his favorite Statistician (that happened to be of Frequency school) how to analyze the result. The answer was to use a truncated probability distribution:*

$$f(t \mid \tau) \propto e^{-t/\tau} \quad t \in (0, T) \quad \textit{and} \quad 0 \quad \textit{outside}$$

The experimenter then told the Statistician that in fact his detector had a safety device that would have measured the decay time of a particle that would have decayed at time larger than T. The Statistician promptly replied that in this case the standard exponential distribution had to be used.

After a careful study of the detector, the experimenter discovered that the safety device was off, during the measurement. At that point the Statistician stated that the truncated exponential was the correct distribution.

This is a standard joke among Bayesians: so much fuss over something that does not influence the only measurement that was recorded! This is not correct, according to the Frequency theory. It matters to know how the experiment was done, since on this information I can built an estimator which is unbiased and has the minimum variance. The construction of this estimator proceeds by averaging over the whole space of the measurements, which is different in the two cases (truncated or fully sensitive detector). In a sense the Frequency theory prepares a function of variables that is optimal for the analysis of any future experiment done in those conditions. Of course the repeated samples is the diving concept in the Frequency theory.

Chapter 20

Examples from High Energy Physics

20.1 Measurement of a Normal variable

Assume that we have made a measurement of the variable $X \sim N(\mu, \sigma^2)$ getting the result x and assume also that the average μ is unknown, while the standard deviation σ is known. What can we say of the average?

We take X as an estimator of μ. In classical statistics (frequency theory approach) μ is a parameter and has a defined, though unknown, value. X is a random variable normally distributed, x is the result of a measurement and again is a number. The best value for μ ($\hat{\mu}$) is x but how uncertain is this information? The problem can be stated precisely in the following way: find the interval (x_1, x_2) which is likely to contain the true value of μ with a given probability (coverage interval).

$$P(x_1 < \mu < x_2) = \alpha$$

Note that the previous statement is a probability statement on x_1 and x_2 not on μ. In other words x_1 and x_2 are random variables, while μ is a parameter. X has normal distribution centered around and unknown value μ.

The problem was solved by J.Neyman (1937): x_1 is such that the Gaussian centered at x_1 ($< x$) has a higher tail whose integral is $\frac{1-\alpha}{2}$:

$$\int_{-\infty}^{x} N(x_2, \sigma^2)dx = \frac{1-\alpha}{2} = \int_{x}^{\infty} N(x_1, \sigma^2)dx$$

The values for x_1 and x_2 can be found in the tables of error function (Tab. 4.2). This is shown in Fig 20.1 where we have assumed $\sigma = 1$, $x = 5$ and we have decided for a Confidence Interval $\alpha = 0.863$ (one σ).

This procedure is different from the way we think of. We mentally construct a Gaussian centered at x and compute x_1 and x_2 such that the integral of the tails of this Gaussian is equal to the confidence limits (Fig. 20.2). It is intuitive but not sound and is precisely the Bayesian approach. In this latter case the unknown parameter μ is considered to have a probability density as any random variable.

$$P(\mu \mid x) = \frac{P(x \mid \mu) \cdot P(\mu)}{\int_{-\infty}^{\infty} P(x \mid \mu) \cdot P(\mu) d\mu}$$

Before the measurement NO information is available on μ and a flat distribution is assumed as prior probability distribution ($P(\mu)$). The conditional probability $P(x|\mu)$ represents the probability that given μ we measure x (likelihood). Once the measurement is made the posterior probability ($P(\mu|x)$) becomes equal to the Likelihood function i.e. a Gaussian centered at x.

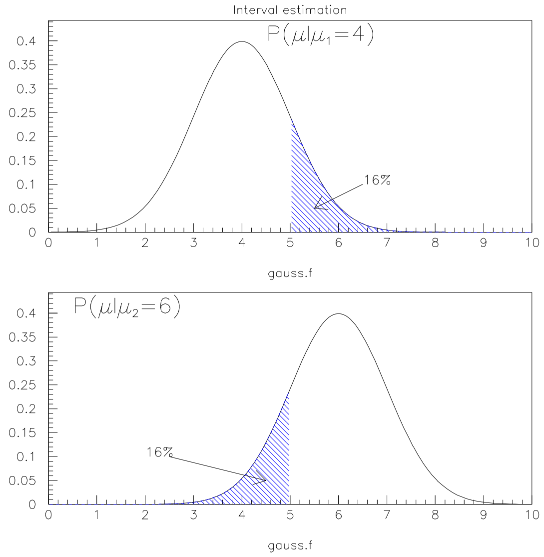


Figure 20.1: *How we set confidence intervals in the frequency theory statistics. Here we assume a normal distribution with $\sigma = 1$. The measurement have yielded the value of $x = 5$. The Confidence Interval is $\alpha = 0.32$.*

This posterior probability can be integrated to get the confidence interval. In the frequency theory approach the likelihood is a number which gives the probability of measuring x given μ and has no probability content on μ. The integration of the likelihood to get confidence intervals is not justified.

In the specific example the confidence intervals obtained by the frequency theory and the Bayesian statistics coincide, as a consequence of the symmetric shape of the Normal *p.d.f.*.

Exercise 20.1 *Prove the previous statement.*

In many cases experimental results are given quoting the average and an "error" given by one standard deviation. From the tables we read that this corresponds to a confidence interval of 0.683. This means that the result in about 1/3 of cases the true value of the parameter will be outside the interval $(\mu - \sigma, \mu + \sigma)$. On the other side the interval $(\mu - 2\sigma, \mu + 2\sigma)$ has a probability only of 1/20 of missing the true value. It is more certain but the interval is wider (If σ is estimated from the data, then the probability content of the interval $(\mu - \sigma, \mu + \sigma)$ is a bit less than 0.683).

Exercise 20.2 *Verify these statements using Tab 4.2.*

The frequency theory and the Bayesian statistics give different coverage in the case of asymmetric distribution.

W have seen in a previous chapter the example of Poisson distribution function. We assumed to have made a measurement and obtained $n = 3$ of a Poisson distributed variable. The 68% confidence interval is computed differently in Bayesian and frequency theory statistics.

In classical statistics we find the values of μ_1 and μ_2 such that:

$$\frac{1-\alpha}{2} = \sum_{i=0}^{3} \frac{{\mu_1}^i}{i!} \cdot e^{-\mu_1}$$

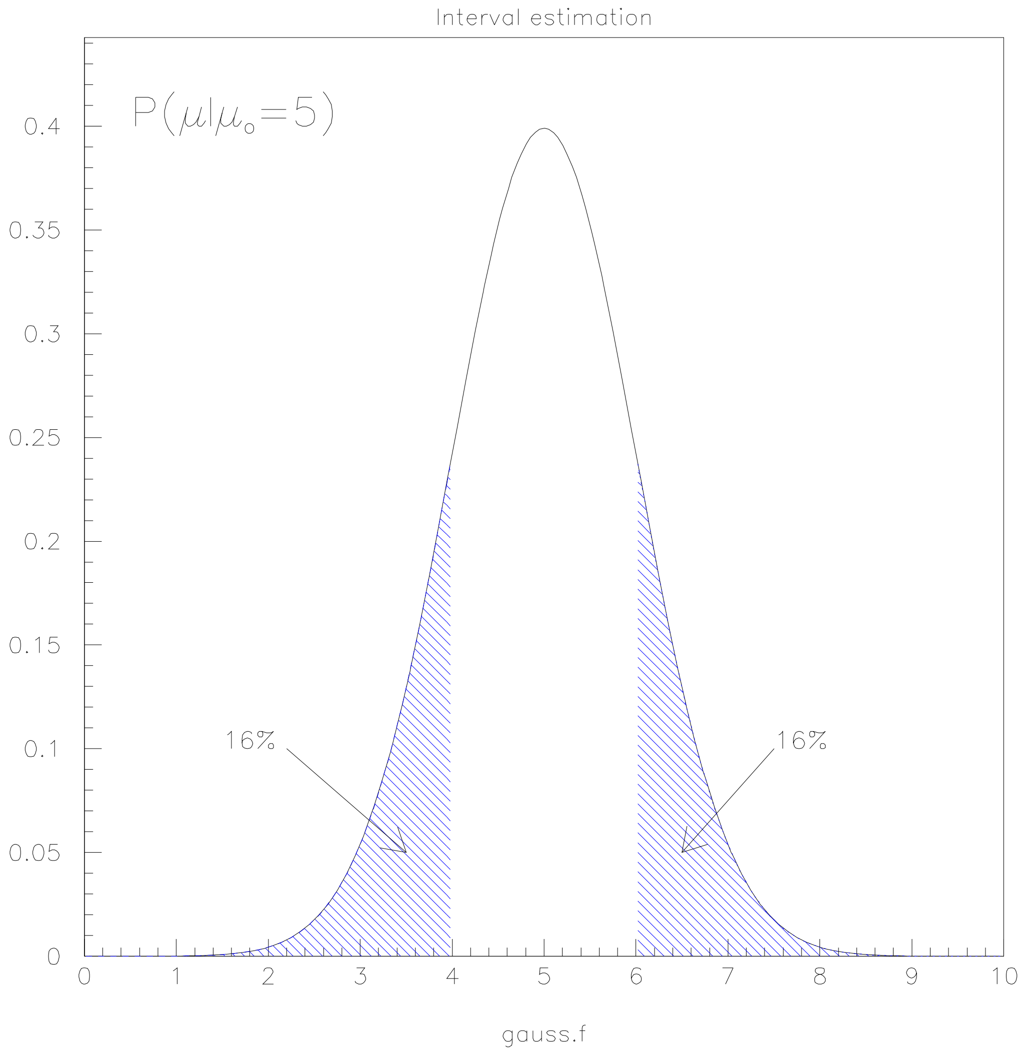


Figure 20.2: *Construction of confidence interval in Bayesian statistics.*

and

$$\frac{1-\alpha}{2} = \sum_{i=3}^{\infty} \frac{{\mu_2}^i}{i!} \cdot e^{-\mu_2}$$

(see Fig.**??**).

In the case of Bayesian statistics we integrate the *p.d.f.* with respect to μ (Fig. **??**).

20.2 Limits on mean lifetimes

Assume that we have measured the decay time of an unstable particle t_1. We know that the *p.d.f.* is:

$$f(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

We have already computed the estimate of τ and its variance:

$$\hat{\tau} = t_1 \qquad\qquad Var(\hat{\tau}) = \hat{\tau}^2$$

The central interval can be obtained in frequency theory statistics with Neyman construction:

$$\int_0^{t_1} \frac{e^{-t/\tau_H}}{\tau_H} dt = \frac{1-\alpha}{2} = \int_{t_1}^{\infty} \frac{e^{-t/\tau_L}}{\tau_L} dt$$

which can be easily solved:

$$\tau_L = -\frac{t_1}{\log\frac{1-\alpha}{2}} \qquad\qquad \tau_H = -\frac{t_1}{\log\frac{1+\alpha}{2}}$$

It can be easily shown that the interval (τ_L, τ_H) has the coverage probability α, by computing it directly. In fact the interval will not cover τ if $\tau_H \leq \tau$ or if $\tau_L \geq \tau$ and the two cases are disjoint and the probabilities add.

$$\begin{aligned} P(\tau_H \leq \tau) &= \int_0^{\tau \log\frac{2}{1+\alpha}} e^{-t/\tau}\frac{1}{\tau}\, dt = \frac{1-\alpha}{2} \\ P(\tau_L \geq \tau) &= \int_{\tau \log\frac{2}{1-\alpha}}^{\infty} e^{-t/\tau}\frac{1}{\tau}\, dt = \frac{1-\alpha}{2} \\ P((\tau_L, \tau_H) \in \tau) &= 1 - P(\tau_H \leq \tau)P(\tau_L \geq \tau) = \alpha \end{aligned}$$

This exact method for constructing interval is shown in Fig. 20.3, where it is also shown the results of the $Log(L)$ method in the case $t_1 = 5$. The intercept of the line $Log(L) = 0.5$ with the curve labelled “One Event” is the interval found with this method. Numerically the two results do not coincide. This is not surprising since we know that the $Log(L)$ method is for low statistics (in this case one event) is approximate.

If we would have measured n events the $\log L(\tau \mid t)$ would have the form:

$$\log L(\tau \mid \bar{t}) = (\frac{\tau - \bar{t}}{\tau}) \cdot n + \log\frac{\bar{t}}{\tau}$$

and with $n = 10$ we can see that the shape is much more symmetric and Gaussian like.

20.3 Difference of Poisson Variables

We have seen that the sum of Poisson variables is again a Poisson variable whose mean is the sum of the means. The inverse problem is the following:

- We measure a process (A) consisting of a signal and background (their means being μ_A and μ_B).
- We also measure a control sample (B) consisting in background only.
- We want to compute the distribution of the signal $S = A - B$ (its mean being μ_S).

The *Gaussian* approximation is often used in practice and gives adequate results if the number of events is large. Only in this case the Poisson distribution can be approximated by a Normal one:

$$P(n, \mu) \rightarrow \frac{1}{\sqrt{2\pi}\sqrt{\mu}} e^{-\frac{(n-\mu)^2}{2\sqrt{\mu}^2}}$$

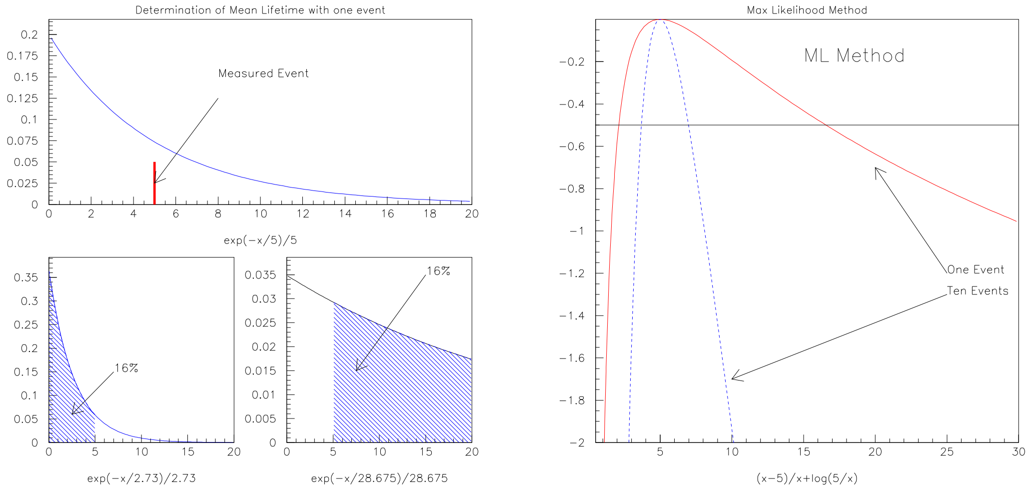


Figure 20.3: *Limits on the lifetime of an unstable particle when 1 event has been measured. At left is the Pearson construction, at right the Log(L) method.*

The difference of Poisson variables is in this approximation a normal variable:

$$P(n_a - n_b) = N(\mu_a - \mu_b \mid \mu_a + \mu_b)$$

If the number of events is small then the exact method has to be used, since the Gaussian approximation underestimates the real Confidence Interval.

Exercise 20.3 *Assume $\mu_B = 50$ events and we measure $n_A = 70$ events. What is $\hat{\mu_S}$? What is the 0.90 level confidence interval?*

20.3.1 The Classical Statistics Solution

We shall follow the analysis of H.B. Prosper [31]. The processes A and B have Poisson distribution with mean μ_A and μ_B respectively:

$$P_A(i_A) = \frac{e^{-\mu_A}\mu_A^{i_A}}{i_A!} \qquad\qquad P_B(i_B) = \frac{e^{-\mu_B}\mu_B^{i_B}}{i_B!}$$

Assuming that $\mu_{A,B}$ are different from zero, the probability of the variable $s = i_A - i_B$ can be found considering that the outcome s can be obtained in an infinite number of ways:

$$\begin{aligned} i_A &= s, \qquad i_B = 0 \\ i_A &= s+1, \qquad i_B = 1 \\ &\ldots \end{aligned}$$

then, when $s > 0$:

$$\begin{aligned} P(s \mid \mu_A, \mu_B) &= P(s \mid \mu_A)P(0 \mid \mu_B) + P(s+1 \mid \mu_A)P(1 \mid \mu_B)\cdots = \\ &= e^{-(\mu_A+\mu_B)}\mu_A^s \sum_{r=0,\infty} \frac{(\mu_A\mu_B)^r}{(r+s)!r!} \end{aligned}$$

and, for $s \leq 0$:

$$\begin{aligned} P(s \mid \mu_A, \mu_B) &= P(0 \mid \mu_A)P(|s| \mid \mu_B) + P(1 \mid \mu_A)P(|s|+1 \mid \mu_B) \cdots = \\ &= e^{-(\mu_A+\mu_B)}\mu_B^{|}s| \sum_{r=0,\infty} \frac{(\mu_A\mu_B)^r}{(r+|s|)!r!} \end{aligned}$$

The two distributions can be combined in a single formula. Define:

$$G_n = (\frac{iy}{2})^{-n} J_n(iy) \qquad\qquad y = 2\sqrt{\mu_A\mu_B}$$

Then:

$$P(s \mid \mu_A, \mu_B) = e^{-(\mu_A+\mu_B)}(\mu_A + \eta(\mu_A - \mu_B))^{|s|}G_{|s|}(y)$$

η is a step variable equal to zero if $s < 0$ and equal to one if s is positive. The probability problem is now solved.

The Statistics problem is the following: we have observed N and B events $s = N - B$. The confidence level associated to s is:

$$\beta = \sum_{i=s+1,\infty} P(i \mid \overline{\mu_A}, \overline{\mu_B})$$

This relation define a curve in the $(\overline{\mu_A}, \overline{\mu_B})$ plane. The part of the plane comprising the origin has a probability β of including the true (unknown) values (μ_A, μ_B).

However this was not our problem, rather we were concerned in the confidence interval of the "signal": $\mu_S = \mu_A - \mu_B$. The parameter μ_B has no interest, is a so called *nuisance parameter* in the statistics jargon, which we would like to "integrate away". In the framework of Classical Statistics there is no procedure to eliminate such parameter. The procedure, which up to now was rigorous, is made approximate, by replacing the parameters μ_B with its estimate.

We shall made the exercise to replace μ_A and μ_B with their maximum likelihood estimate. The likelihood function is:

$$L = \frac{e^{-\mu_B}\mu_B^B}{b!} \cdot \frac{e^{-(\mu_S+\mu_B)}(\mu_s+\mu_B)^N}{n!}$$

and:

$$\begin{aligned} 0 &= \frac{\partial L}{\partial \mu_S} = N - (\mu_B + \mu_S) \\ 0 &= \frac{\partial L}{\partial \mu_B} = N\mu_B + (\mu_s + \mu_B) + 2N\mu_B(\mu_S + \mu_B) \end{aligned}$$

with the result:

$$\begin{aligned} \hat{\mu_S} &= N - B \\ \hat{\mu_B} &= B \end{aligned}$$

We can now compute the probability of the statistics s by replacing μ_B with its estimate $\hat{\mu_B}$:

$$P(s \mid \mu_S, \hat{\mu}_B) = e^{-(\mu_S+2\mu_B)}(\mu_S\eta + \mu_B)^{|s|}G_{|s|}(y)$$

In this case $y = 2\sqrt{\mu_S(\mu_S + \hat{\mu}_B)}$. The β confidence level for μ_S is now obtained by solving:

$$\beta = \sum_{i=s+1,\infty} P(i \mid \overline{\mu_S}, \hat{\mu}_B)$$

N	$B=0$	$B=1$	$B=2$	$B=3$
0	2.30	1.91	1.39	.77
1	3.89	3.34	2.73	2.07
2	5.32	4.70	4.03	3.33
3	6.68	6.01	5.31	4.58
4	7.99	7.29	6.56	5.82

Table 20.1: *90% Classical limits for the difference of Poisson variables.*

The upper limit at 0.90 confidence level for some values of N and of B are reported in Tab. 20.1 The "Gaussian" approximation, often used in practice, for the one sided 0.90 Confidence Interval [1] :

$$\overline{\mu_S} = N - B + 1.28\sqrt{N+B}$$

underestimate the limit for low values of N, B

20.3.2 Bayesian Solution

Bayesian approach has the advantage of offering a general ground to the inference problem, while no general procedure is provided by the Classical approach.

We shall follow the analysis of H. B. Prosper [32]. Assume we have recorded N events from a Poisson process with mean μ_A (signal + background). In a control sample the background is monitored recording B events. The probability to obtain this result is:

$$P(N,B \mid \mu_A,\mu_B) = \frac{{\mu_A}^N}{N!} \cdot e^{-\mu_A} \cdot \frac{{\mu_B}^B}{B!} \cdot e^{-\mu_B}$$

Calling $\mu_S = \mu_A - \mu_B$, the unknown mean of the signal, we get:

$$P(N,B \mid \mu_S,\mu_B) = \frac{(\mu_s+\mu_B)^N \mu_B^B e^{-(\mu_S+2\mu_B)}}{N!B!}$$

We then apply Bayes theorem:

$$P(\mu_S \mid N,B) = \frac{\int_0^\infty \mu_B P(N,B,\mid \mu_S,\mu_B)\cdot P(\mu_S,\mu_B)}{\int_0^\infty d\mu_S \int_0^\infty P(N,B \mid \mu_S,\mu_B)P(\mu_S,\mu_B)}$$

$P(\mu_S,\mu_B)$ is the prior distribution (the personal rational knowledge) on μ_S and μ_B. As Jayes points out : *After nearly two centuries of discussions and debates, we still do not seem to have the principles needed to translate prior information into a definite prior probability assignment.* Prosper assumes a prior distribution corresponding to no prior knowledge as follows:

$$P(\mu_S,\mu_B) = \frac{1}{\mu_S^\nu} \cdot \frac{1}{\mu_B^\nu}$$

- $\nu = 1$ corresponds to the generalization to two dimensions of Jeffreys recipe (uniform in $log(\mu)$).
- $\nu = 0$ corresponds to an intuitive uniform prior, as used by Helene [36]

[1]From the Normal probability tables:

$$0.9 = \int_{-\infty}^{1.28} N(0,1)dx$$

With this choice Prosper obtains:

$$P(\mu_S \mid N, B) = \frac{e^{-\mu_S} \sum_{j=0,N} \mu_S^{n_j} C_j}{\sum_{j=0,N} \Gamma(n_j + 1) C_j}$$

With:

$$\begin{aligned} C_j &= \frac{\begin{pmatrix} N \\ j \end{pmatrix} \Gamma(b_j + 1)}{2^j} \\ n_j &= N - j - \nu \\ b_j &= B + j - \nu \end{aligned}$$

There is a special case in which μ_B is known from an independent measurement:

$$P(\mu_S \mid N, \mu_B) = \frac{e^{-\mu_S} (\mu_S + \mu_B)^N \mu_S^{-\nu}}{\sum_{j=0,N} \begin{pmatrix} N \\ j \end{pmatrix} \Gamma(n_j + 1) \mu_B^j}$$

This coincides (for $\nu = 0$) with Helene's results.

The numerical values for 0.90 confidence limits for this particular case and when $\nu = 0$ are summarized in Tabs. 20.2 and 20.3 for some values of total number of recorded events and for some background rates.

N	$B = 0$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
0	.12	.12	.12	.12	.12
1	.53	.21	.17	.15	.14
2	1.10	.43	.26	.21	.18
3	1.74	.85	.44	.30	.24
4	2.43	1.45	.77	.46	.33
5	3.15	2.15	1.28	.74	.49
6	3.89	2.89	1.93	1.18	.74
7	4.65	3.65	2.66	1.77	1.12
8	5.43	4.43	3.43	2.47	1.65
9	6.22	5.22	4.22	3.23	2.31

Table 20.2: *Bayesian lower limit (0.90 C.L.) for some values of total number of recorded events and average background rate.*

20.3.3 Another *Classical* Solution to the Upper Limit

Zech [38] has tried to interpret the Bayesian formula in a frequency theory way. His starting point is the probability of observing n events from the Poisson processes of signal and background with mean s and b respectively:

$$P(n \mid s, b) = \frac{e^{-(s+b)}\ (s+b)^n}{n!}$$

which is obtained from adding all the possible combination of signal and background yielding the sum n:

$$p(n \mid s, b) = \sum_{n_b=0,n} \sum_{n_s=0,n-n_b} P(n_b \mid b) \cdot P(n_s \mid b)$$

N	$B = 0$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
0	2.46	2.46	2.46	2.46	2.46
1	3.89	3.36	3.12	2.97	2.88
2	5.32	4.48	3.96	3.63	3.41
3	6.68	5.72	4.97	4.44	4.06
4	7.99	7.00	6.11	5.39	4.85
5	9.27	8.27	7.31	6.46	5.77
6	10.53	9.53	8.54	7.61	6.79
7	11.77	10.77	9.77	8.80	7.89
8	12.99	11.99	10.99	10.00	9.05
9	14.20	13.20	12.20	11.20	10.22

Table 20.3: *Bayesian upper limit* (0.90 *C.L.) for some values of total number of recorded events and average background rate.*

as we have seen previously. Zech argues that, after the experiment has been performed and n events have been recorded, $P(n_b \mid b)$ is no longer correctly normalized, since we know that $n_b \leq n$ and cannot extend to infinity as in a Poisson. We have to normalize the distribution of the background to the new range:

$$P(n_b \mid b) \to \frac{P(n_b \mid b)}{\sum_{n_b=0,n} P(n_b \mid b)}$$

Therefore:

$$p(n \mid s, b) \to \frac{p(n \mid s, b)}{\sum_{n_b=0,n} P(n_b \mid b)}$$

The probability of observing n events or less is:

$$\epsilon = \frac{\sum_{n=0,n} p(n \mid s, b)}{\sum_{n_b=0,n} p(n_b \mid b)}$$

and this relation can be used to obtain a upper limit s_u of s with a Confidence Level $1 - \epsilon$ by solving:

$$CL = 1 - \epsilon = \frac{\sum_{n=0,n} P(n \mid s_u, b)}{\sum_{n_b=0,n} P(n_b \mid b)}$$

This formula is mathematical identical to the one obtained with Bayesian method with uniform prior, but is interpreted by Zech in a frequency way: In an infinitely large number of experiments if signal and background are distributed as Poisson variables with means s_u and b, the relative frequency of observing n or less events is ϵ.

Zech also observes that if the background mean b is poorly known, the limits can be obtained by averaging over the probability distribution of b:

$$CL = 1 - \epsilon = \frac{\sum_{n=0,n} \int \; P(n \mid s_u, b) g(b) db}{\sum_{n_b=0,n} \int \; P(n_b \mid b) g(b) db}$$

and the Bayesian flavor is now explicit.

Exercise 20.4 *Prove that Zech formula for the upper limit is identical to the Bayesian one.*

20.4 Ratio of Small Number of Events

Often a quantity R is obtained as the ratio:

$$R = \frac{N \pm \sigma_N}{D \pm \sigma_D}$$

N and D are either normal or Poisson variables. It is well known that R is neither distributed as a Gaussian or Poisson variable and also, since the probability of obtaining $D = 0$ is finite, its variance is infinite.

It is common practice to determine the Confidence Interval using:

$$\sigma_R = \frac{N}{D}\sqrt{(\frac{\sigma_N}{N})^2 + (\frac{\sigma_D}{D})^2}$$

and assuming $R \sim N(\frac{N}{D}, \sigma_r^2)$ The interval $R \pm \sigma_R$ is assumed to cover the true value of R in a 0.683 fraction of cases. This is a good approximation if N and D are large. In this case the distribution of R around the average value is well approximated by a Gaussian. However the tails are larger than Gaussian and this produces shorter confidence interval.

20.4.1 Frequency Theory Solution

James and Roos [29] give the exact frequency theory solution to the problem.

In the definition of R we use events of $D - type$ and of $N - type$. Without loosing generality we can consider the sum $N + D$ as fixed. The problem is to determine the probability p of observing N events out of $N + D$. The distribution is binomial and the observation of N and D sets the limits on p.

This problem was solved in a previous section with the Clopper-Pearson construction. Instead of using the graphical construction it is preferable to use an analytical method based on a transformation of the cumulative Fisher-Snedecor distribution, as it was realized soon after the Clopper-Pearson work.

The Clopper-Pearson confidence limits for p at confidence level $\beta = 1 - \alpha$ (C_-, C_+) are given by:

$$\begin{aligned}
C_-(\alpha, N, N + D) &= \frac{N}{N + (D + 1)\ f_{1-\alpha/2,2(D+1),2N}} \\
C_+(\alpha, N, N + D) &= 1 - C_-(\alpha, D, N + D)
\end{aligned}$$

Where $f_{1-\alpha/2,n,m}$ is the $1 - \alpha/2$ quantile [2] of the Fisher distribution with (n, m) degrees of freedom. The β confidence interval for R is the transformation of the previous interval:

$$(\frac{C_-}{1 - C_-}, \frac{C_+}{1 - C_+})$$

The values of the quantiles can be obtained from the Statistical tables.

These exact results are valid only if there is no background in the data. To take into account the background the authors suggest an approximate method. While we are interested in the ratio:

$$R = \frac{S_N}{S_D}$$

[2] The α quantile (q_α) is defined as the value of x such that

$$\alpha = \int_{-\infty}^{q_\alpha} f(x \mid \theta)\, dx$$

what we measure is:

$$R' = \frac{N}{D} = \frac{S_N + B_N}{S_D + B_D}$$

As above, we will consider $k = N + S$ constant and we will obtain the limits R'. Since:

$$R = \frac{R' - (B_N/k)(1 + R')}{1 - (B_D/k)(1 + R')}$$

we can transform the limits for R' to R. Unfortunately B_N and B_D are not known and we have to use their expectation: $\hat{B_N}$ and $\hat{B_D}$. In this way we neglect the fluctuations of B_N and B_D and the confidence limits will be smaller than the exact limits.

The confidence intervals for 0.683 and 0.90 are reported in Tabs. 20.4 and 20.5.

D	$N = 0$	1	2	3	4	5	6	7	8	9	10
1	0.00	.03	.16	.33	.52	.72	.92	1.12	1.33	1.54	1.74
1	19.00	38.50	57.99	77.48	96.98	116.48	135.98	155.47	174.97	194.47	213.97
2	.00	.02	.11	.23	.37	.52	.67	.82	.97	1.13	1.28
2	3.47	6.38	9.25	12.07	14.92	17.74	20.57	23.40	26.22	29.04	31.86
3	.00	.01	.08	.18	.29	.41	.53	.65	.77	.90	1.02
3	1.71	3.02	4.29	5.54	6.77	8.01	9.23	10.47	11.70	12.93	14.15
4	.00	.01	.07	.15	.24	.34	.44	.54	.64	.75	.85
4	1.12	1.92	2.69	3.44	4.19	4.93	5.66	6.40	7.14	7.87	8.62
5	.00	.01	.06	.12	.20	.29	.37	.46	.55	.64	.73
5	.82	1.39	1.93	2.46	2.98	3.50	4.02	4.52	5.04	5.54	6.07
6	.00	.01	.05	.11	.18	.25	.33	.40	.48	.56	.64
6	.65	1.09	1.50	1.90	2.29	2.69	3.08	3.46	3.85	4.24	4.62
7	.00	.01	.04	.10	.16	.22	.29	.36	.43	.50	.57
7	.53	.89	1.22	1.54	1.86	2.18	2.49	2.80	3.11	3.42	3.72
8	.00	.01	.04	.09	.14	.20	.26	.32	.39	.45	.52
8	.45	.75	1.03	1.30	1.56	1.82	2.08	2.34	2.59	2.84	3.09
9	.00	.01	.03	.08	.13	.18	.24	.29	.35	.41	.47
9	.40	.65	.89	1.12	1.34	1.56	1.78	2.00	2.21	2.43	2.65
10	.00	.00	.03	.07	.12	.16	.22	.27	.32	.38	.43
10	.35	.57	.78	.98	1.18	1.37	1.56	1.75	1.93	2.12	2.31

Table 20.4: *Frequency theory confidence intervals (0.90 C.L.) for ratio $R = N/D$ for a few values of N and of D.*

D	$N=0$	1	2	3	4	5	6	7	8	9	10
1	0.00	.09	.34	.62	.91	1.20	1.50	1.80	2.10	2.40	2.70
1	5.32	11.10	16.89	22.69	28.48	34.28	40.07	45.87	51.66	57.46	63.25
2	.00	.06	.23	.42	.63	.83	1.04	1.25	1.46	1.68	1.89
2	1.51	2.97	4.41	5.83	7.25	8.67	10.09	11.51	12.92	14.34	15.75
3	.00	.04	.17	.32	.48	.64	.80	.97	1.13	1.30	1.47
3	.85	1.62	2.37	3.11	3.86	4.59	5.32	6.07	6.79	7.52	8.26
4	.00	.04	.14	.26	.39	.52	.65	.79	.92	1.06	1.20
4	.58	1.10	1.60	2.09	2.58	3.06	3.55	4.02	4.51	4.98	5.47
5	.00	.03	.12	.22	.33	.44	.55	.67	.78	.90	1.02
5	.45	.83	1.20	1.56	1.92	2.28	2.64	2.99	3.34	3.70	4.05
6	.00	.02	.10	.19	.28	.38	.48	.58	.68	.78	.88
6	.36	.67	.96	1.25	1.53	1.81	2.09	2.38	2.66	2.93	3.21
7	.00	.02	.09	.16	.25	.33	.42	.51	.60	.69	.78
7	.30	.56	.80	1.04	1.27	1.50	1.73	1.96	2.19	2.42	2.65
8	.00	.02	.08	.15	.22	.30	.38	.46	.54	.62	.70
8	.26	.48	.68	.88	1.08	1.28	1.47	1.66	1.86	2.05	2.24
9	.00	.02	.07	.13	.20	.27	.34	.41	.49	.56	.63
9	.23	.42	.60	.77	.94	1.11	1.28	1.45	1.62	1.78	1.95
10	.00	.02	.06	.12	.18	.25	.31	.38	.45	.51	.58
10	.20	.37	.53	.68	.83	.98	1.13	1.28	1.43	1.58	1.72

Table 20.5: *Frequency theory confidence intervals (0.68 C.L.) for ratio $R = N/D$ for a few values of N and of D.*

Appendix A

Sampling Distributions

The knowledge of the density of sampling distribution is important in statistical inference. The exact sampling distribution is not always easy. Here we want to discuss some analytical methods to find such distributions. We will assume that the parent distribution is known: $dF(x)$. In a sample of size n the joint distribution of the n values $x_1, x_2 \cdots x_n$ is

$$dF(x_1)dF(x_2)\cdots dF(x_n)$$

Let

$$z = z(x_1, x_2, \cdots x_n)$$

be the statistics, the distribution of the statistics is given by:

$$F(z_0) = \int \cdots \int_S dF(x_1)dF(x_2)\cdots dF(x_n)$$

The integration domain S is such that $z = z(x_1, x_2, \cdots x_n) \leq z_0$. Thus the solution of our problem is, formally, the evaluation of multiple integrals which can be done in various ways.

Example A.1 *To find the distribution of a linear function of n independent variables $x_1, \cdots x_n$ where each x_j is distributed Normally with zero mean and unit variance:*

$$z = a_1 x_1 + \cdots a_n x_n$$

it is convenient to make a transformation of the type:

$$v_i = \sum_j c_{i,j} x_j \quad (\vec{v} = C\vec{x} \text{ in matrix notation})$$

such that $v_1 = z$ and, moreover, such that the transformation is orthogonal:

$$\sum_i c_{i,j} c_{i,k} = \delta_{l,k} \quad (C^T C = I)$$

This is possible always and in many ways, since we have only $\frac{n(n-1)}{2} + n$ constraints and n^2 constants. The Jacobian of the transformation is one and also:

$$\sum v_i^2 = \sum x_i^2$$

since the transformation is orthogonal. The joint distribution of all the x_i is:

$$dF = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2}\sum x_i^2} \prod dx_i$$

hence the joint distribution of v_i is:

$$dF = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2}\sum v_i^2} \prod dv_i$$

which shows that each v_i is independent of the others. In particular $v_1$is distributed as:

$$dF = \frac{1}{\sqrt{2\pi}} e^{-\frac{v_1^2}{2}} dv_1$$

Thus v_1 is distributed normally with zero mean and unit variance (of course, by symmetry this is true for all v_i).

The calculations yielding the distribution can also be rather complex.

Example A.2 *The sample mean of a Cauchy distribution has interesting properties. We have measured a sample of size n: $x_1, x_2 \cdots x_n$ from a Cauchy distribution:*

$$dF = \frac{1}{\pi}\frac{1}{1+x^2} \qquad -\infty \leq x\infty \leq$$

The joint distribution is:

$$\frac{1}{\pi^n}\prod_{i=1}^{n}\left(\frac{1}{1+x_i^2}\right)$$

the statistics is:

$$nz = \sum x_j$$

We have to integrate the joint distribution subject to $\sum x_j \leq nz$. Let us again make a transformation:

$$\begin{cases} u_1 &= x_1 \\ u_2 &= x_2 \\ \cdots & \\ u_{n-1} &= x_{n-1} \\ u_n &= z = \frac{1}{n}\sum_j x_j \end{cases}$$

The Jacobian of the transformation is equal to n. The new variables $u_1 \cdots u_{n-1}$ can now extend from $-\infty$ to $+\infty$ while u_n varies from $-\infty$ to z. The distribution of z is then:

$$F(z) = \frac{n}{\pi^n}\int_{-\infty}^{z}\int \cdots \int_{-\infty}^{+\infty}\prod_{j=1}^{n-1}\left(\frac{1}{1+u_j^2}\right)\frac{du_j}{1+(nu_n - u_1 - u_2 \cdots - u_{n-1})^2}$$

The integral is made step by step using the elementary integral:

$$\int_{-\infty}^{+\infty}\frac{1}{1+x^2}\frac{1}{r^2+(a-x)^2} = \pi\frac{r+1}{r}\frac{1}{a^2+(r+1)^2}$$

Let us integrate away u_{n-1} :

$$
\begin{aligned}
F(z) &= \frac{n}{\pi^n}\int_{-\infty}^{z}\int\cdots\int_{-\infty}^{+\infty}\prod_{j=1}^{n-2}\left(\frac{du_j}{1+u_j^2}\right)\int_{-\infty}^{+\infty}\frac{1}{1+u_{n-1}^2}\frac{du_{n-1}}{1+(nu_n-u_1-u_2\cdots-u_{n-1})^2}\\
&= \frac{n}{\pi^n}\int_{-\infty}^{z}\int\cdots\int_{-\infty}^{+\infty}\prod_{j=1}^{n-2}\left(\frac{du_j}{1+u_j^2}\right)\left(2\pi\frac{1}{2^2+(nu_n-u_1\cdots-u_{n-2})^2}\right)\\
&= \frac{2n}{\pi^{n-1}}\int_{-\infty}^{z}\int\cdots\int_{-\infty}^{+\infty}\prod_{j=1}^{n-2}\left(\frac{du_j}{1+u_j^2}\right)\left(\frac{1}{2^2+(nu_n-u_1\cdots-u_{n-2})^2}\right)\\
&= \frac{2n}{\pi^{n-1}}\int_{-\infty}^{z}\int\cdots\int_{-\infty}^{+\infty}\prod_{j=1}^{n-3}\left(\frac{du_j}{1+u_j^2}\right)\int_{-\infty}^{+\infty}\left(\frac{1}{1+u_{n-2}^2}\frac{du_{n-2}}{2^2+(nu_n-u_1\cdots-u_{n-2})^2}\right)\\
&= \frac{2n}{\pi^{n-1}}\int_{-\infty}^{z}\int\cdots\int_{-\infty}^{+\infty}\prod_{j=1}^{n-3}\left(\frac{du_j}{1+u_j^2}\right)\left(\frac{3\pi/2}{3^2+(nu_n-u_1\cdots-u_{n-3})^2}\right)\\
&= \frac{1}{\pi}\frac{1}{1+z^2}
\end{aligned}
$$

Thus the sample mean of a Cauchy distribution has the same distribution as the single observation. This is an example of the failure of the Central Limit theorem.

A.1 Order Statistics

Assume that we have a sample of size n: $x_1, x_2 \cdots x_n$ from a population having a distribution $F(x)$ (and a density $f(x)$, it it exists). Put order the observations in ascending order of magnitude:

$$y_1 \leq y_2 \leq \cdots \leq y_n$$

The y_i are called *order statistics* and are the same observations as the x_i but ordered. Note that the x_i are independent, while y_i are not.

We are concerned with the marginal distribution of y_r. The probability that, in a sample of n, $l-1$ falls below a value y_r is $(F(y_r))^{r-1}$. Similarly the probability that $n-r$ falls above y_r is $(1-F(y_r))^{n-r}$ and the probability that a single observation falls in an interval dy_r about y_r is $dF(y_r)$. Therefore the probability of the event in which the observation r^{th} falls at y_r is:

$$dG(y_r) = K \times (F(y_r))^{r-1} dF(y_r)(1-F(y_r))^{n-r}$$

The constant K is the normalization which is readily obtained:

$$dG(y_r) = \frac{n!}{(r-1)!(n-r)!}(F(y_r))^{r-1}(1-F(y_r))^{n-r}f(y_r)dy_r$$

The distribution of y_1, the smallest value of the sample is:

$$P(y_1) = n(1-F(y_1))^{n-1}f(y_1)$$

Example A.3 *In the case of an uniform distribution:*

$$\left\{\begin{array}{lcl} f(x) &=& 1\ (0 \leq x \leq 1)\\ &=& 0\ \textit{elsewhere} \end{array}\right. \qquad \left\{\begin{array}{lcl} F(x) &=& x\ (0 \leq x \leq 1)\\ &=& 0\ (x \leq 0)\\ &=& 1\ (x > 0) \end{array}\right. \tag{A.1}$$

we obtain:

$$P(y_1) = n(1-y_1)^{n-1}$$

The distribution of the largest observation is:

$$P(y_n) = F(y_n)^{n-1} f(y_n)$$

Example A.4 *In the case of an exponential distribution we have:*

$$\begin{cases} f(x) &= \frac{e^{-x/\lambda}}{\lambda}\ (x \geq 0) \\ &= 0\ (x < 0) \end{cases} \qquad F(x) = 1 - e^{-x/\lambda}\ (x \geq 0)$$

The density of the last observation is:

$$P(y_n) = \frac{n}{\lambda} e^{-y_n/\lambda} (1 - e^{-y_n/\lambda})^{n-1}$$

The joint distribution of y_s and y_r $(s > r\ )$ can be easily obtained with a similar argument:

$$d^2G(y_r, y_s) = \frac{F(y_r)^{r-1}(F(y_s) - F(y_r))^{s-r-1}(1 - F(y_s))^{n-s}}{B(r, s-r)B(s, n-s+1)} dF(y_r) dF(y_s)$$

In particular if $s = n$ and $r = 1$ we have:

$$d^2G(y_1, y_n) = n(n-1) \left[F(y_n) - F(y_1)\right]^{n-1} f(y_1) f(y_n)\ dy_1 dy_n$$

From this distribution we can compute the distribution of the *range* $R = y_n - y_1$. We have to make the change of variables:

$$\begin{cases} R &= y_n - y_1 \\ M &= y_n + y_1 \end{cases} \begin{cases} y_n &= R + M \\ y_1 &= R \end{cases}$$

and integrate away M. The result is:

$$dG(R) = n(n-1) \int_{-\infty}^{+\infty} \left[F(x+R) - F(x)\right]^{n-2} f(x+R) f(x)\ dx$$

Example A.5 *In the case of uniform distribution we have:*

$$\begin{aligned} dG(R) &= n(n-1) \int_0^{1-R} \left[F(x+R) - F(x)\right]^{n-2} dx = \\ &= n(n-1) R^{n-2} (1-R) \end{aligned}$$

Appendix B

Polynomial orthogonal to Data

Least square method leads to hill conditioned systems. The problem can be avoided by the use of orthogonal polynomials. We will describe briefly here how these polynomials can be generated. We will expose the method due to G. E. Forsythe [51].

Let us assume that the LSQ problem has n points $\{x_i,\ i = 1 \cdots n\}$ each with weights $\{w_i\}$. The orthogonal property that we require, in order to have a diagonal Least Square matrix is.

$$\sum_{i=0,n} w_i P_l(x_i) P_m(x_i) = 0 \; if \; l \neq m \tag{B.1}$$

where w_i are the weights of the points x_i. We will prove a recurrence relation which makes the construction of a sequence of orthogonal polynomial very simple.

Theorem B.1 *Orthogonal polynomials can be obtained by a recurrence relation of the form:*

$$\left\{ \begin{array}{rcl} P_{j+1}(x) & = & (x - \alpha_{j+1}) P_j(x) - \beta_j P_{j-1}(x) \\ j = 0, 1 \cdots n & & \\ P_{-1}(x) & = & 0 \\ P_0(x) & = & 1 \end{array} \right. \tag{B.2}$$

α_{j+1} and β_j are constants to be determined. The polynomials have the property B.1.

Proof B.1 *We will prove the previous relation by induction. If $j = 0$ we have immediately from B.2 and B.1:*

$$\left\{ \begin{array}{rcl} P_1(x) & = & x - \alpha_1 \\ \alpha_1 & = & \frac{\sum_i w_i x_i}{\sum_i w_i} \end{array} \right.$$

(α_1 is the weighted mean of x_i). Let us now assume that the polynomials $P_j(x)$ ($j = 0 \cdots k$) satisfy B.1 and B.2 we want to show that we can choose α_{k+1} and β_k such that $P_{k+1}(x)$, obtained by B.2 also satisfy the orthogonality conditions:

$$\sum_i w_i P_j(x_i) P_{k+1}(x_i) = 0 \; j = 0 \cdots k$$

Using B.2 with $j = k$ we obtain:

$$\left\{ \begin{array}{rcl} 0 & = & \sum_i w_i P_j(x_i) P_{k+1}(x_i) = \\ \sum_i w_i P_j(x_i) \left[(x - \alpha_{k+1}) P_k(x_i) - \beta_k P_{k-1}(x_i) \right] & = & \underbrace{\sum_i w_i x_i P_j(x_i) P_k(x_i)}_{C} - \alpha_{k+1} \underbrace{\sum_i w_i P_j(x_i) P_k(x_i)}_{A} \\ & & - \beta_k \underbrace{\sum_i w_i P_j(x_i) P_{k-1}(x_i)}_{B} \end{array} \right.$$

We distinguish three cases:

1. For $j = 0 \cdots k-2$ we have $A = B = 0$ by assumption. For the term C we have to consider that the polynomial xP_j is a polynomial of degree $j+1$ and can be expressed as combination of P_m $m = 0 \cdots j+1$ therefore also this term is zero.

2. If $j = k-1$ then the term A is zero by assumption, the others are not zero. This allow us to determine β_k:

$$\beta_k = \frac{\sum_i w_i x_i P_{k-1}(x_i) P_k(x_i)}{\sum_i w_i P_{k-1}^2(x_i)} = \frac{\sum_i w_i P_k^2(x_i)}{\sum_i w_i P_{k-1}^2(x_i)}$$

The last passage comes from the recurrence and orthogonality relations:

$$P_k = (x - \alpha_k)P_{k-1} - \beta_{k-1}P_{k-2} \;\rightarrow\; \sum_i w_i P_k^2(x_i) = \sum_i w_i x_i P_{k-1}(x_i) P_k(x_i)$$

3. If $j = k$ then B vanishes and we can compute α_{k+1} such that the orthogonality relation is satisfied:

$$\alpha_{k+1} = \frac{\sum_i w_i x_i P_k^2(x_i)}{\sum_i w_i P_k^2(x_i)}$$

This completes the proof.

In the proof we have assumed that the denominator does not vanish; this would happen only if

$$P_j(x_i) = 0 \quad i = 1 \cdots n \tag{B.3}$$

for some j. This would also stop the recurrence relation and no other orthogonal polynomial of higher order could be obtained. Since we are given only n points one would expect to generate a maximum of n orthogonal polynomial and that the $n+1^{th}$ polynomial (P_n) would satisfy B.3. This is really so as we will see in a moment.

Let us first see a property of orthogonal polynomials.

Theorem B.2 *The orthogonal polynomial P_j has j distinct zeros in the interval spanned by $\{x_i\}$.*

Proof B.2 *Let us concentrate on odd multiplicity zeros and assume that P_j has m zeros $\{a_1, a_2 \cdots a_m\}$ in the interval $[Min(x_i), Max(x_i)]$. P_j will change sign at each $\{a_i\}$ therefore the polynomial:*

$$(x - a_1) \cdots (x - a_m) P_j(x)$$

does not change sign in the above interval. Using the orthogonality property we can assert that:

$$\sum_{i=1,n} w_i \underbrace{(x_i - a_1) \cdots (x_i - a_m)}_{y^{(m)}} P_j(x) = 0$$

In fact the polynomial in bracket is of order m and can be expressed as a combination of P_k $k = 0 \cdots m$. But the sum cannot vanish because it does not change sign by construction. Hence $m = j$ and we can write

$$P_j(x) = \prod_{i=0,j} (x - a_i)$$

Theorem B.3 *If $P_j(x)$ $j = 0 \cdots n$ are polynomial orthogonal to the points $\{x_i \; i = 1 \cdots n\}$ with $w_i \geq 0$, then $P_n(x_i) = 0$ $i = 1 \cdots n$.*

Proof B.3 *This is true if*

$$P_n(x) = \prod_{i=0,n} (x - x_i) \tag{B.4}$$

This is polynomial has degree n, is orthogonal and has n zeros. We have only to show that it is the only polynomial of degree n. Let us assume that there is another orthogonal polynomial of degree n. $Q_n(x)$. We would have:

$$\begin{cases} \sum_i w_i P_{n-1}(x_i) P_n(x_i) &= 0 \\ \sum_i w_i P_{n-1}(x_i) Q_n(x_i) &= 0 \\ \sum_i w_i P_{n-1}(x_i) \left[P_n(x_i) - Q_n(x_i) \right] &= 0 \end{cases}$$

Now the polynomial in square bracket or is identically zero, in case $P_n \equiv Q_n$ or has degree $n-1$. (This is so because by construction from the recurrence relation the orthogonal polynomial of degree k have the coefficient of x^k equal to 1.) This second case is impossible hence P_n is unique and has the form B.4.

Appendix C

The Combinatorials of the Run Test

C.1 The Run Test

In this test we compare two sets of observations:

$$x_i \; i = 1, \cdots n \quad y_i \; i = 1, \cdots m$$

and test the hypothesis that the two samples have the same distribution.

Let us assume that the sets are ordered in increasing order. Let us now combine the two sets and arrange the observations in increasing order:

$$x_1, x_2, y_1, x_3 \cdots$$

If the two samples originate from the same distribution the combined sample would be "well mixed". The problem that we will solve now is to make quantitative the idea of "well mixed".

One possible measure of mixture of the two samples is the number of *runs*. We define a run a sequence of symbols of the same type. In the sequence above we have a run of two x then a run of one y followed by a run of x etc. If the number of runs is too low (for instance two, all the x precede the y) we have to suspect that H_0 is not true and the clustering of the x's and of the y's hides some dynamics different from H_0. We will study the distribution of the number of runs under the hypothesis that the samples originate from the same population.

We will start from a combinatorial problem: let us consider r indistinguishable balls which have to be placed in n boxes. Any number of balls can be put in each box. Since the balls are indistinguishable a configuration of balls in the box is identified by the occupation number of each box: $r_1, \cdots r_n$.

Theorem C.1 *The number of possible distinct configurations is* $A_{r,n} = \binom{n+r-1}{n-1}$.

Proof C.1 *The argument is easy if we reason in this way. he following figure shows on possible configuration:*

The n boxes define $n + 1$ boundary bars: the first bar is always in position one and the last in position $n + 1$. The r balls fall in between two bars. We have $n - 1$ bars and r balls that we can dispose in any possible ways. A group of r objects (bars or balls) define in an unique way a configuration of balls in the boxes. For instance if we have 3 balls and 2 cells ($n = 2$, $r = 3$, i.e. one bar (B) and three balls (b)) the possible groups are:

Objects	Configuration	Occupation
Bbb	I_IxxxI	(0,3)
bBb	IxIxxI	(1,2)
bbB	IxxIxI	(2,1)
bbb	IxxxI_I	(3,0)

In this case the number of distinguishable configuration is four. In the general case the number of distinguishable configurations is the number of ways we can chose r objects from $n+r-1$ i.e.:

$$A_{r,n} = \left(\begin{array}{c} n-1+r \\ r \end{array} \right) = \left(\begin{array}{c} n+r-1 \\ n-1 \end{array} \right)$$

Another relevant theorem is the following:

Theorem C.2 *The number of distinct configurations which leave NO empty cell is*

$$A_r = \left(\begin{array}{c} r-1 \\ n-1 \end{array} \right)$$

Proof C.2 *There must not be two adjacent bars. Since between r balls there are $r-1$ slots. out of these $n-1$ must be filled. The number of possibilities is the number of ways in which we can take $n-1$ objects out of $r-1$ i.e.* $\left(\begin{array}{c} r-1 \\ n-1 \end{array} \right)$.

The combinatorial part is finished, we can go back to our statistical problem. Assume we have a elements of kind α and b elements of kind β. Once mixed we have, for instance ($a=6$ and $b=8$) :

$$\alpha\alpha\alpha\beta\alpha\beta\beta\beta\alpha\alpha\beta\beta\beta\beta$$

we have six runs of length: $3,1,1,3,2,4$. Of course, by definition, the number of runs is equal to the number of times the kind changes plus one.

Assume that n_α runs of type α are found in a combination of the $a+b$ objects, the number of runs of kind β is not arbitrary, it must be one of the three numbers:

- n_α (if the series begins with a α run and ends with a β run or vice versa)
- $n_\alpha - 1$ (if the series begins and ends with a α run)
- $n_\alpha + 1$ (if the series begins and ends with a β run)

The total number of ways in which $a+b$ objects can be ordered is

$$N_{tot} = \left(\begin{array}{c} a+b \\ a \end{array} \right) = \left(\begin{array}{c} a+b \\ b \end{array} \right)$$

The number of different ways to make n_α runs from a elements is the same as the number of ways in which a balls can be put in n_α cells *without* leaving any cell empty, i.e.:

$$\left(\begin{array}{c} a-1 \\ n_\alpha - 1 \end{array} \right)$$

If, in the same sequence there are n_β runs, the total number of ways is:

$$\left(\begin{array}{c} a-1 \\ n_\alpha - 1 \end{array} \right) \left(\begin{array}{c} b-1 \\ n_\beta - 1 \end{array} \right)$$

The total number of runs is $r = n_\alpha + n_\beta$ and we distinguish two cases:

- $r = even = 2k$. The only possibility is $n_\alpha = n_\beta = k$ which gives a number of combinations:

$$N = 2\begin{pmatrix} a-1 \\ n_\alpha - 1 \end{pmatrix}\begin{pmatrix} b-1 \\ n_\beta - 1 \end{pmatrix}$$

  (the factor 2 comes from $n_\alpha \leftrightarrow n_\beta$). The probability of having $2k$ runs is the ratio of distinct ways of making $2k$ runs and the total number of combinations of $a+b$ objects:

$$P_{2k} = \frac{2\begin{pmatrix} a-1 \\ k-1 \end{pmatrix}\begin{pmatrix} b-1 \\ k-1 \end{pmatrix}}{\begin{pmatrix} a+b \\ a \end{pmatrix}}$$

- $r = odd = 2k+1$. then we have two possibilities: $n_\alpha = k$ and $n_\beta = k+1$ and vice versa. The probability is:

$$P_{2k+1} = \frac{\begin{pmatrix} a-1 \\ k-1 \end{pmatrix}\begin{pmatrix} b-1 \\ k \end{pmatrix} + \begin{pmatrix} a-1 \\ k \end{pmatrix}\begin{pmatrix} b-1 \\ k-1 \end{pmatrix}}{\begin{pmatrix} a+b \\ a \end{pmatrix}}$$

The mean and variance are:

$$\begin{aligned} E(r) &= \frac{2\,a\,b}{a+b} + 1 \\ Var(r) &= \frac{2\,a\,b\,(2\,a\,b - a - b)}{(a+b)^2 \cdot (a+b-1)} \end{aligned}$$

For large samples the distribution of r tends to a Normal and a common test statistics is:

$$z = \frac{r - E(r)}{\sqrt{Var(r)}} \sim N(0,1)$$

Example C.1 *: the arrangements of seats (empty or occupied) along a lounge counter is found to be: EOEEOEEEOEEEOEOE. There are no occupied seats adjacent, is this due to chance? The lounge counter may be occupied by families, hence the tendency for occupants to cluster together and this would produce a small number of runs. On the contrary singles could tend to isolate and the run count would be large. The number of seats is 16 and the number of runs is 11. Empty seats and occupied seats are the two samples. In case of complete random mixture we should have:*

$$\begin{cases} E(r) &= \frac{5\times11\times2}{5+11} + 1 = 7.875 \\ Var(r) &= \frac{5\times11\times2\times(5\times11\times2-11-5)}{(5+11)^2\times(11+5-1)} = 2.69 \end{cases}$$

Then we get $z = 1.91$ *and the probability* $P(z > 1.91) = 0.028$ *(one sided) or* $P(|z| > 1.91) = 0.056$ *(two sided). The number of runs is a bit edgy but not completely unlikely.*

Appendix D

Tables and useful Formulas

We collect here the standard tables of statistical functions. Not all of them, only those that are more commonly used in the statistics practice. More complete set of tables can be founf in specialized literature. The tables collect the values of the density of some important distribution functions, their probability content and the values of the inverse probability:

$$\begin{cases} N(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ P(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ \beta &= \int_{-\infty}^{x(\beta)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \end{cases}$$

and similar tables for χ^2_ν and $t - Student$ functions.

D.1 Formulas

β-Function

$$B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}dx = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \quad (m,n\ > 0)$$

$$\begin{aligned} \int_0^\infty \frac{x^{m-1}}{(1+x)^{m+n}}dx &= B(m,n) \\ \int_0^{\pi/2} \sin^m\theta \cdot \cos^n\theta \, d\theta &= \frac{1}{2}B(\frac{m+1}{2}, \frac{n+1}{2}) \\ \int_0^1 x^m(1-x^2)^{\frac{n-1}{2}} dx &= \frac{1}{2}B(\frac{m+1}{2}, \frac{n+1}{2}) \\ \int_0^1 \frac{x^{m-1}+x^{n-1}}{(1+x)^{m+n}}dx &= B(m,n) \\ \int_0^a x^{m-1}(a-x)^{n-1}dx &= a^{m+n-1}B(m,n) \\ \int_0^\infty \frac{x^{m-1}}{(ax+b)^{m+n}}dx &= \frac{B(m,n)}{a^m b^n} \end{aligned}$$

γ-Function

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx = \int_0^1 (\ln\frac{1}{x})^{n-1}dx$$

$$\begin{aligned}
\Gamma(n+1) &= n\Gamma(n) \\
\Gamma(n)\Gamma(n-1) &= \frac{\pi}{\sin n\pi} \\
\Gamma(n) &= (n-1)! \qquad n\ integer\ > 0 \\
\Gamma(1)=\Gamma(2) &= 1 \\
\Gamma(\frac{1}{2}) &= \sqrt{\pi} \\
\Gamma(n+\frac{1}{2}) &= 1\cdot 3\cdots(2n-3)\cdot(2n-1)\frac{\sqrt{\pi}}{2^n}
\end{aligned}$$

χ^2-Function

$$\chi_k^2(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

$$\left\{ \begin{array}{rcl} x_i & \sim & \chi^2_{k_i} \\ i=1,\cdots n & & \end{array} \right. \Rightarrow \sum x_i \sim \chi^2_{\sum k_i}$$

$$\left\{ \begin{array}{rcl} x_1 & \sim & \chi^2_{k1} \\ x_2 & \sim & \chi^2_{k2} \end{array} \right. \left\{ \begin{array}{rcl} f & = & \frac{x_1}{x_2} \sim \frac{\Gamma(\frac{k_1+k_2}{2})}{\Gamma(\frac{k_1}{2})\Gamma(\frac{k_2}{2})} \frac{f^{\frac{k_1}{2}}}{(1+f)^{\frac{k_1+k_2}{2}}} \\ F & = & \frac{x_1/k_1}{x_2/k_2} \sim \frac{(\frac{k_1}{k_2})^{k_1/2}}{B(\frac{k_1}{2},\frac{k_2}{2})} \frac{F^{\frac{k_1}{2}-1}}{(1+\frac{k_1 F}{k_2})^{\frac{k_1+k_2}{2}}} \end{array} \right.$$

$$z = \frac{1}{2}\lg F \sim 2\frac{(\frac{k_1}{k_2})^{k_1/2}}{B(\frac{k_1}{2},\frac{k_2}{2})} \frac{e^{k_1 z}}{(1+\frac{k_1}{k_2}e^{2z})^{\frac{k_1+k_2}{2}}}$$

Poisson

$$Prob(r \le n_0 \mid \mu) = \sum_{r=0}^{n_0} \frac{e^{-\mu}}{r!}\mu^r = 1 - Prob(\chi^2_{2n_0+2} < 2\mu)$$

$$\int_0^\infty x^n e^{-ax} dx = \frac{n!}{a^{n+1}}$$

$x, \Delta x =$	0.000	0.002	0.004	0.006	0.008	0.010	0.012	0.014	0.016	0.018
0.000	0.5000	0.5004	0.5008	0.5012	0.5016	0.5020	0.5024	0.5028	0.5032	0.5036
0.020	0.5080	0.5084	0.5088	0.5092	0.5096	0.5100	0.5104	0.5108	0.5112	0.5116
0.040	0.5160	0.5164	0.5168	0.5171	0.5175	0.5179	0.5183	0.5187	0.5191	0.5195
0.060	0.5239	0.5243	0.5247	0.5251	0.5255	0.5259	0.5263	0.5267	0.5271	0.5275
0.080	0.5319	0.5323	0.5327	0.5331	0.5335	0.5339	0.5343	0.5347	0.5351	0.5355
0.100	0.5398	0.5402	0.5406	0.5410	0.5414	0.5418	0.5422	0.5426	0.5430	0.5434
0.120	0.5478	0.5482	0.5486	0.5489	0.5493	0.5497	0.5501	0.5505	0.5509	0.5513
0.140	0.5557	0.5561	0.5565	0.5569	0.5572	0.5576	0.5580	0.5584	0.5588	0.5592
0.160	0.5636	0.5640	0.5643	0.5647	0.5651	0.5655	0.5659	0.5663	0.5667	0.5671
0.180	0.5714	0.5718	0.5722	0.5726	0.5730	0.5734	0.5738	0.5742	0.5746	0.5750
0.200	0.5793	0.5797	0.5800	0.5804	0.5808	0.5812	0.5816	0.5820	0.5824	0.5828
0.220	0.5871	0.5875	0.5878	0.5882	0.5886	0.5890	0.5894	0.5898	0.5902	0.5906
0.240	0.5948	0.5952	0.5956	0.5960	0.5964	0.5968	0.5972	0.5975	0.5979	0.5983
0.260	0.6026	0.6030	0.6033	0.6037	0.6041	0.6045	0.6049	0.6053	0.6057	0.6060
0.280	0.6103	0.6106	0.6110	0.6114	0.6118	0.6122	0.6126	0.6129	0.6133	0.6137
0.300	0.6179	0.6183	0.6187	0.6191	0.6194	0.6198	0.6202	0.6206	0.6210	0.6213
0.320	0.6255	0.6259	0.6263	0.6267	0.6270	0.6274	0.6278	0.6282	0.6285	0.6289
0.340	0.6331	0.6334	0.6338	0.6342	0.6346	0.6350	0.6353	0.6357	0.6361	0.6365
0.360	0.6406	0.6409	0.6413	0.6417	0.6421	0.6424	0.6428	0.6432	0.6436	0.6439
0.380	0.6480	0.6484	0.6488	0.6491	0.6495	0.6499	0.6503	0.6506	0.6510	0.6514
0.400	0.6554	0.6558	0.6562	0.6565	0.6569	0.6573	0.6576	0.6580	0.6584	0.6587
0.420	0.6628	0.6631	0.6635	0.6639	0.6642	0.6646	0.6649	0.6653	0.6657	0.6660
0.440	0.6700	0.6704	0.6708	0.6711	0.6715	0.6718	0.6722	0.6726	0.6729	0.6733
0.460	0.6772	0.6776	0.6780	0.6783	0.6787	0.6790	0.6794	0.6797	0.6801	0.6805
0.480	0.6844	0.6847	0.6851	0.6855	0.6858	0.6862	0.6865	0.6869	0.6872	0.6876
0.500	0.6915	0.6918	0.6922	0.6925	0.6929	0.6932	0.6936	0.6939	0.6943	0.6946

Table D.1: Values of a standard normal probability $P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Values for negative x can be obtained by: $P(-x) = 1 - P(x)$.

0.520	0.6985	0.6988	0.6992	0.6995	0.6999	0.7002	0.7006	0.7009	0.7013	0.7016
0.540	0.7054	0.7057	0.7061	0.7064	0.7068	0.7071	0.7075	0.7078	0.7082	0.7085
0.560	0.7123	0.7126	0.7129	0.7133	0.7136	0.7140	0.7143	0.7146	0.7150	0.7153
0.580	0.7190	0.7194	0.7197	0.7201	0.7204	0.7207	0.7211	0.7214	0.7217	0.7221
0.600	0.7257	0.7261	0.7264	0.7267	0.7271	0.7274	0.7277	0.7281	0.7284	0.7287
0.620	0.7324	0.7327	0.7330	0.7334	0.7337	0.7340	0.7343	0.7347	0.7350	0.7353
0.640	0.7389	0.7392	0.7396	0.7399	0.7402	0.7405	0.7409	0.7412	0.7415	0.7418
0.660	0.7454	0.7457	0.7460	0.7463	0.7467	0.7470	0.7473	0.7476	0.7479	0.7483
0.680	0.7517	0.7521	0.7524	0.7527	0.7530	0.7533	0.7536	0.7540	0.7543	0.7546
0.700	0.7580	0.7583	0.7587	0.7590	0.7593	0.7596	0.7599	0.7602	0.7605	0.7608
0.720	0.7642	0.7645	0.7649	0.7652	0.7655	0.7658	0.7661	0.7664	0.7667	0.7670
0.740	0.7703	0.7707	0.7710	0.7713	0.7716	0.7719	0.7722	0.7725	0.7728	0.7731
0.760	0.7764	0.7767	0.7770	0.7773	0.7776	0.7779	0.7782	0.7785	0.7788	0.7791
0.780	0.7823	0.7826	0.7829	0.7832	0.7835	0.7838	0.7841	0.7844	0.7847	0.7849
0.800	0.7881	0.7884	0.7887	0.7890	0.7893	0.7896	0.7899	0.7902	0.7905	0.7907
0.820	0.7939	0.7942	0.7945	0.7947	0.7950	0.7953	0.7956	0.7959	0.7962	0.7964
0.840	0.7995	0.7998	0.8001	0.8004	0.8007	0.8009	0.8012	0.8015	0.8018	0.8021
0.860	0.8051	0.8054	0.8057	0.8059	0.8062	0.8065	0.8068	0.8070	0.8073	0.8076
0.880	0.8106	0.8108	0.8111	0.8114	0.8117	0.8119	0.8122	0.8125	0.8127	0.8130
0.900	0.8159	0.8162	0.8165	0.8167	0.8170	0.8173	0.8175	0.8178	0.8181	0.8183
0.920	0.8212	0.8215	0.8217	0.8220	0.8223	0.8225	0.8228	0.8230	0.8233	0.8236
0.940	0.8264	0.8266	0.8269	0.8272	0.8274	0.8277	0.8279	0.8282	0.8284	0.8287
0.960	0.8315	0.8317	0.8320	0.8322	0.8325	0.8327	0.8330	0.8332	0.8335	0.8337
0.980	0.8365	0.8367	0.8370	0.8372	0.8374	0.8377	0.8379	0.8382	0.8384	0.8387

Table D.2: (Continued) Values of a standard normal probability $P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Values for negative x can be obtained by: $P(-x) = 1 - P(x)$.

$x, \Delta x =$	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.500	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
0.510	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
0.520	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
0.530	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
0.540	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
0.550	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
0.560	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
0.570	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
0.580	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
0.590	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
0.600	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
0.610	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029
0.620	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
0.630	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
0.640	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
0.650	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
0.660	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
0.670	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
0.680	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
0.690	0.4959	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
0.700	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
0.710	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
0.720	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
0.730	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
0.740	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
0.750	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031

Table D.3: Values of $x(\beta)$ (the inverse probability function of a standard normal): $\beta = \int_{-\infty}^{x(\beta)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$.

0.760	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
0.770	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
0.780	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
0.790	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
0.800	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8633	0.8669	0.8705	0.8742
0.810	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
0.820	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
0.830	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
0.840	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
0.850	1.0364	1.0407	1.0450	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
0.860	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
0.870	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
0.880	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
0.890	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
0.900	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.910	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.920	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.930	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.940	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.950	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.960	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.970	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.980	2.0537	2.0749	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.990	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

Table D.4: (Continued) Values of $x(\beta)$ (the inverse probability function of a standard normal): $\beta = \int_{-\infty}^{x(\beta)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$.

x, $DoF =$	1	2	3	4	5	7	10	12	15	20
0.100	1.2000	0.4756	0.1200	0.0238	0.0040	0.0001	0.0000	0.0000	0.0000	0.0000
0.200	0.8072	0.4524	0.1614	0.0452	0.0108	0.0004	0.0000	0.0000	0.0000	0.0000
0.300	0.6269	0.4304	0.1881	0.0646	0.0188	0.0011	0.0000	0.0000	0.0000	0.0000
0.400	0.5164	0.4094	0.2066	0.0819	0.0275	0.0022	0.0000	0.0000	0.0000	0.0000
0.500	0.4394	0.3894	0.2197	0.0974	0.0366	0.0037	0.0001	0.0000	0.0000	0.0000
0.600	0.3815	0.3704	0.2289	0.1111	0.0458	0.0055	0.0001	0.0000	0.0000	0.0000
0.700	0.3360	0.3523	0.2352	0.1233	0.0549	0.0077	0.0002	0.0000	0.0000	0.0000
0.800	0.2990	0.3352	0.2392	0.1341	0.0638	0.0102	0.0004	0.0000	0.0000	0.0000
0.900	0.2681	0.3188	0.2413	0.1435	0.0724	0.0130	0.0005	0.0000	0.0000	0.0000
1.000	0.2420	0.3033	0.2420	0.1516	0.0807	0.0161	0.0008	0.0001	0.0000	0.0000
1.200	0.1999	0.2744	0.2398	0.1646	0.0959	0.0230	0.0015	0.0002	0.0000	0.0000
1.500	0.1539	0.2362	0.2308	0.1771	0.1154	0.0346	0.0031	0.0005	0.0000	0.0000
2.000	0.1038	0.1839	0.2076	0.1839	0.1384	0.0553	0.0077	0.0015	0.0001	0.0000
2.500	0.0723	0.1433	0.1807	0.1791	0.1506	0.0753	0.0146	0.0036	0.0003	0.0000
3.000	0.0514	0.1116	0.1542	0.1673	0.1542	0.0925	0.0235	0.0071	0.0008	0.0000
4.000	0.0270	0.0677	0.1080	0.1353	0.1440	0.1152	0.0451	0.0180	0.0033	0.0001
5.000	0.0146	0.0410	0.0732	0.1026	0.1220	0.1220	0.0668	0.0334	0.0085	0.0004
10.000	0.0009	0.0034	0.0085	0.0168	0.0283	0.0567	0.0877	0.0877	0.0629	0.0181
20.000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0022	0.0095	0.0189	0.0384	0.0626
30.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0010	0.0036	0.0162
40.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0015
50.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

Table D.5: Values of the density of the χ^2_ν distribution.

x, $DoF =$	1	2	3	4	5	7	10	12	15	20
0.100	0.7518	0.9512	0.9918	0.9988	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
0.200	0.6547	0.9048	0.9776	0.9953	0.9991	1.0000	1.0000	1.0000	1.0000	1.0000
0.300	0.5839	0.8607	0.9600	0.9898	0.9976	0.9999	1.0000	1.0000	1.0000	1.0000
0.400	0.5271	0.8187	0.9402	0.9825	0.9953	0.9997	1.0000	1.0000	1.0000	1.0000
0.500	0.4795	0.7788	0.9189	0.9735	0.9921	0.9994	1.0000	1.0000	1.0000	1.0000
0.600	0.4386	0.7408	0.8964	0.9631	0.9880	0.9990	1.0000	1.0000	1.0000	1.0000
0.700	0.4028	0.7047	0.8732	0.9513	0.9830	0.9983	1.0000	1.0000	1.0000	1.0000
0.800	0.3711	0.6703	0.8495	0.9384	0.9770	0.9974	0.9999	1.0000	1.0000	1.0000
0.900	0.3428	0.6376	0.8254	0.9246	0.9702	0.9963	0.9999	1.0000	1.0000	1.0000
1.000	0.3173	0.6065	0.8013	0.9098	0.9626	0.9948	0.9998	1.0000	1.0000	1.0000
1.200	0.2733	0.5488	0.7530	0.8781	0.9449	0.9909	0.9996	1.0000	1.0000	1.0000
1.500	0.2207	0.4724	0.6823	0.8266	0.9131	0.9823	0.9989	0.9999	1.0000	1.0000
2.000	0.1573	0.3679	0.5724	0.7358	0.8491	0.9598	0.9963	0.9994	1.0000	1.0000
2.500	0.1138	0.2865	0.4753	0.6446	0.7765	0.9271	0.9909	0.9982	0.9999	1.0000
3.000	0.0833	0.2231	0.3916	0.5578	0.7000	0.8850	0.9814	0.9955	0.9996	1.0000
4.000	0.0455	0.1353	0.2615	0.4060	0.5494	0.7798	0.9473	0.9834	0.9977	1.0000
5.000	0.0253	0.0821	0.1718	0.2873	0.4159	0.6600	0.8912	0.9580	0.9921	0.9997
10.000	0.0016	0.0067	0.0186	0.0404	0.0752	0.1886	0.4405	0.6160	0.8197	0.9682
20.000	0.0000	0.0000	0.0002	0.0005	0.0012	0.0056	0.0293	0.0671	0.1719	0.4579
30.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0009	0.0028	0.0119	0.0699
40.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0050
50.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002

Table D.6: Values of upper tail χ^2_ν distribution, $P(x) = \int_0^x \chi^2_\nu(t)dt$

Dof,$Prob$	0.001	0.010	0.050	0.100	0.500	0.900	0.950	0.990	0.999
1	0.000	0.000	0.004	0.016	0.455	2.706	3.841	6.635	10.828
2	0.002	0.020	0.103	0.211	1.386	4.605	5.991	9.210	13.816
3	0.024	0.115	0.352	0.584	2.365	6.253	7.815	11.342	16.269
4	0.091	0.297	0.711	1.064	3.356	7.781	9.488	13.276	18.468
5	0.210	0.554	1.146	1.610	4.351	9.237	11.071	15.086	20.516
6	0.381	0.872	1.635	2.204	5.348	10.645	12.592	16.811	22.458
7	0.598	1.239	2.167	2.833	6.346	12.017	14.067	18.475	24.322
8	0.857	1.647	2.733	3.490	7.344	13.361	15.507	20.090	26.125
9	1.152	2.088	3.325	4.168	8.343	14.684	16.919	21.666	27.878
10	1.479	2.558	3.940	4.865	9.342	15.987	18.307	23.210	29.589
11	1.834	3.054	4.575	5.578	10.341	17.275	19.675	24.725	31.265
12	2.214	3.571	5.226	6.304	11.340	18.549	21.026	26.217	32.911
13	2.617	4.107	5.892	7.041	12.340	19.812	22.362	27.688	34.531
14	3.041	4.660	6.571	7.790	13.339	21.064	23.685	29.141	36.124
15	3.483	5.229	7.261	8.547	14.339	22.307	24.996	30.578	37.698
16	3.942	5.812	7.962	9.312	15.338	23.542	26.296	32.000	39.253
17	4.416	6.408	8.672	10.085	16.338	24.769	27.587	33.409	40.790
18	4.905	7.015	9.390	10.865	17.338	25.989	28.869	34.805	42.313
19	5.407	7.633	10.117	11.651	18.338	27.204	30.144	36.191	43.820
20	5.921	8.260	10.851	12.443	19.337	28.412	31.410	37.566	45.315
25	8.649	11.524	14.611	16.473	24.337	34.382	37.652	44.314	52.620
30	11.588	14.953	18.493	20.599	29.336	40.256	43.773	50.892	59.703
35	14.688	18.509	22.465	24.797	34.336	46.059	49.802	57.342	66.619
40	17.916	22.164	26.509	29.051	39.335	51.805	55.758	63.691	73.402
50	24.674	29.707	34.764	37.689	49.335	63.167	67.505	76.154	86.661
60	31.738	37.485	43.188	46.459	59.335	74.397	79.082	88.379	99.607
70	39.036	45.442	51.739	55.329	69.334	85.527	90.531	100.425	112.317
80	46.520	53.540	60.391	64.278	79.334	96.578	101.879	112.329	124.839
90	54.155	61.754	69.126	73.291	89.334	107.565	113.145	124.116	137.208
100	61.918	70.065	77.929	82.358	99.334	118.498	124.342	135.807	149.449

Table D.7: Values of Inverse χ^2_ν probability: $\beta = \int_0^{x(\beta)} \chi^2_\nu(t)dt$.

x, DoF	1	2	3	4	5	7	10	12	15	20
0.000	0.3183	0.3536	0.3676	0.3750	0.3796	0.3850	0.3891	0.3907	0.3924	0.3940
0.100	0.3152	0.3500	0.3639	0.3713	0.3758	0.3812	0.3852	0.3868	0.3885	0.3901
0.200	0.3061	0.3398	0.3532	0.3604	0.3648	0.3699	0.3739	0.3754	0.3770	0.3786
0.300	0.2920	0.3238	0.3364	0.3431	0.3472	0.3521	0.3558	0.3572	0.3587	0.3602
0.400	0.2744	0.3031	0.3145	0.3206	0.3243	0.3287	0.3320	0.3333	0.3346	0.3359
0.500	0.2546	0.2794	0.2891	0.2942	0.2974	0.3011	0.3040	0.3051	0.3062	0.3073
0.600	0.2341	0.2539	0.2616	0.2657	0.2681	0.2710	0.2732	0.2740	0.2749	0.2758
0.700	0.2136	0.2281	0.2335	0.2362	0.2379	0.2398	0.2412	0.2417	0.2423	0.2428
0.800	0.1941	0.2029	0.2058	0.2071	0.2079	0.2087	0.2092	0.2095	0.2096	0.2098
0.900	0.1759	0.1791	0.1794	0.1793	0.1792	0.1789	0.1786	0.1784	0.1783	0.1781
1.000	0.1592	0.1571	0.1551	0.1536	0.1526	0.1512	0.1500	0.1495	0.1490	0.1485
1.100	0.1440	0.1372	0.1330	0.1303	0.1284	0.1261	0.1242	0.1234	0.1225	0.1217
1.200	0.1305	0.1195	0.1134	0.1096	0.1071	0.1039	0.1013	0.1003	0.0992	0.0981
1.300	0.1183	0.1039	0.0962	0.0916	0.0885	0.0847	0.0816	0.0804	0.0791	0.0778
1.400	0.1075	0.0902	0.0813	0.0761	0.0726	0.0684	0.0650	0.0636	0.0622	0.0607
1.500	0.0979	0.0783	0.0686	0.0629	0.0592	0.0547	0.0511	0.0497	0.0482	0.0467
1.600	0.0894	0.0680	0.0577	0.0518	0.0480	0.0434	0.0398	0.0384	0.0369	0.0354
1.700	0.0818	0.0591	0.0486	0.0426	0.0388	0.0343	0.0307	0.0293	0.0279	0.0265
1.800	0.0751	0.0515	0.0408	0.0349	0.0312	0.0269	0.0235	0.0222	0.0209	0.0196
1.900	0.0690	0.0449	0.0344	0.0286	0.0251	0.0209	0.0178	0.0166	0.0154	0.0143
2.000	0.0637	0.0393	0.0289	0.0234	0.0201	0.0163	0.0135	0.0124	0.0113	0.0103
2.500	0.0439	0.0208	0.0125	0.0087	0.0066	0.0044	0.0030	0.0026	0.0021	0.0017
3.000	0.0318	0.0117	0.0057	0.0034	0.0022	0.0012	0.0006	0.0005	0.0003	0.0002
3.500	0.0240	0.0070	0.0028	0.0014	0.0008	0.0003	0.0001	0.0001	0.0001	0.0000
4.000	0.0187	0.0044	0.0014	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
5.000	0.0122	0.0019	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.000	0.0086	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7.000	0.0064	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8.000	0.0049	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9.000	0.0039	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10.000	0.0032	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table D.8: Values of the Student density

x, DoF	1	2	3	4	5	7	10	12	15	20
0.100	0.5317	0.5353	0.5367	0.5374	0.5379	0.5384	0.5388	0.5390	0.5392	0.5393
0.200	0.5628	0.5700	0.5729	0.5744	0.5753	0.5764	0.5773	0.5776	0.5779	0.5782
0.300	0.5928	0.6038	0.6081	0.6104	0.6119	0.6136	0.6148	0.6153	0.6159	0.6164
0.400	0.6211	0.6361	0.6420	0.6452	0.6472	0.6495	0.6512	0.6519	0.6526	0.6533
0.500	0.6476	0.6667	0.6743	0.6783	0.6809	0.6838	0.6861	0.6869	0.6878	0.6887
0.600	0.6720	0.6953	0.7046	0.7096	0.7127	0.7163	0.7191	0.7202	0.7213	0.7224
0.700	0.6944	0.7218	0.7328	0.7387	0.7424	0.7467	0.7501	0.7514	0.7527	0.7540
0.800	0.7148	0.7462	0.7589	0.7657	0.7700	0.7750	0.7788	0.7804	0.7819	0.7834
0.900	0.7333	0.7684	0.7828	0.7905	0.7953	0.8010	0.8054	0.8071	0.8088	0.8106
1.000	0.7500	0.7887	0.8045	0.8130	0.8184	0.8247	0.8296	0.8315	0.8334	0.8354
1.200	0.7789	0.8235	0.8419	0.8518	0.8581	0.8654	0.8711	0.8734	0.8756	0.8779
1.300	0.7913	0.8384	0.8578	0.8683	0.8748	0.8826	0.8886	0.8910	0.8934	0.8958
1.400	0.8026	0.8518	0.8720	0.8829	0.8898	0.8979	0.9041	0.9066	0.9091	0.9116
1.500	0.8128	0.8638	0.8847	0.8960	0.9030	0.9114	0.9177	0.9203	0.9228	0.9254
1.750	0.8348	0.8889	0.9108	0.9225	0.9297	0.9382	0.9447	0.9472	0.9497	0.9523
2.000	0.8524	0.9082	0.9303	0.9419	0.9490	0.9572	0.9633	0.9657	0.9680	0.9704
2.250	0.8669	0.9233	0.9450	0.9562	0.9629	0.9704	0.9759	0.9780	0.9801	0.9821
2.500	0.8789	0.9352	0.9561	0.9666	0.9728	0.9795	0.9843	0.9860	0.9877	0.9894
3.000	0.8976	0.9523	0.9712	0.9800	0.9850	0.9900	0.9933	0.9945	0.9955	0.9965
3.500	0.9114	0.9636	0.9803	0.9876	0.9914	0.9950	0.9971	0.9978	0.9984	0.9989
4.000	0.9220	0.9714	0.9860	0.9919	0.9948	0.9974	0.9987	0.9991	0.9994	0.9996
5.000	0.9372	0.9811	0.9923	0.9963	0.9979	0.9992	0.9997	0.9998	0.9999	1.0000

Table D.9: Values of the Student probability: $\beta = \int_{-\infty}^{x} f_\nu(t)dt$.

$DoF, Prob$	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995	0.999
1	0.325	0.727	1.376	3.078	6.314	12.706	31.820	63.653	317.910
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385

Table D.10: Values of t_β the inverse Student probability:$\beta = \int_{-\infty}^{t_\beta} f_\nu(x)dx$

Bibliography

[1] J. R. Blum and J. I. Rosenblatt *Probability and Statistics* W. B. Saunders, Philadelphia, 1972.

[2] G. Castelnuovo *Calcolo delle Probabilita"* Zanichelli 1961

[3] G. Cowan *Statistical Data Analysis* Clarendon Press - Oxford, 1998

[4] H. Cramer *Mathematical Methods of Statistics* Princeton Univ. Press 1954

[5] A. Dalla Corte *Appunti sul trattamento statistico dei dati sperimentali* Arcetri, Firenze 1957

[6] F. N. David *Games, Gods and Gambling* Griffin, London 1962

[7] R.D. Evans *The Atomic Nucleus* McGraw, New York 1955

[8] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet *Statistical Methods in Experimental Physics* North-Holland, Amsterdam 1971

[9] A.G. Frodesen, O. Skjeggestad and H. Tofte *Probability and statistics in Particle Physics* Universitetsforlaget, Bergen 1979

[10] W. Feller: *An Introduction to Probability Theory and Its Applications* John Wiley & Sons, New York 1968.

[11] R. Fisher *Statistical Methods and scientific inference* Oliver and Boyd 1956

[12] R. Fisher *The Design of Experiments* Oliver and Boyd 1951

[13] M.G. Kendall and A. Stuart *Advanced Theory of Statistics* Vols. 1,2 and 3 Griffin 1958

[14] H. Jeffreys *Advanced Theory of Probability* Clarendon, Oxford, 1939.

[15] D. Hudson *Lectures on Elementary Statistics and Probability* CERN 63-29 1963

[16] D. Hudson *Maximum Likelihood and Least Square Theory* CERN 64-18 1964

[17] A. Kolmogorov *Foundations of the Theory of Probability* , Chelsea, 1956.

[18] I. Hacking *The Emergence of Probability* Cambridge University Press, New York 1975

[19] A. O'Heare *Introduction to the Philosophy of Science* Claredon Press, Oxford 1989

[20] R. L. Plackett *Principles of Regression Analysis* Oxford at the Clarendon Press 1960.

[21] K. R. Popper *Realism and the Aim of Science* Rowman and Littlefield, Totowa, New Jersey 1983

[22] A. Ralston *A First Course in Numerical Analysis* McGraw-Hill, New York 1965.

[23] C. R. Rao *Linear Statistical Inference and its Applications* John Wiley and Sons, New York 1973

[24] E. Severino *Legge e Caso* Adelphi, Milano 1979

[25] S. Ulam *Adventures of a Mathematician* University of California Press, 1991

[26] C. J. Clopper and E. S. Pearson *Biometrika 2 (1934) 404*

[27] C. McCusker and I. Cairns *Phys. Rev. Lett. 23 (1969) 658*

[28] R. Adair and H. Kasha *Phys. Rev. Lett. 23 (1969) 1355*

[29] F. James and M. Roos *Nucl. Phys. B172 (1980) 475*

[30] F. James *Nucl. Instr. and Methods A240 (1985) 203*

[31] H. B. Prosper *Nucl. Instr. and Methods A238 (1985) 500*

[32] H. B. Prosper *Nucl. Instr and Methods A241 (1985) 236*

[33] H. B. Prosper *Phys. Rev. D37 (1988) 1153*

[34] O. Helene *Nucl. Instr. and Methods 212 (1983) 319*

[35] O. Helene *Nucl. Instr. and Methods A300 (1991) 132*

[36] O. Helene *Nucl. Instr. and Methods 228 (1984) 120*

[37] R. D. Cousins and V. L. Highland *Nucl. Instr. and Methods A320 (1992) 331*

[38] G. Zech *Nucl. Instr. and Methods A277 (1989) 608*

[39] J. Orear *Am. J. Phys. 50 (1982) 912*

[40] F. James *Determining the Statistical Significance of Experimental Results. CERN-Data Handling Division DD/81/02 Lectures presented at the 1980 CERN school of computing.*

[41] G. U. Yule *J.R.S.S. 90 (1927) 570.*

[42] W. F. R. Weldon *Cited in G.M. Kendall and A. Stuart.*

[43] Shapiro, Wilks and Chen *Jou. Am. Stat. Ass. 63 (1968) 1343*

[44] B. Rossi *High-Energy Particles, Prentice Hall inc. 1952*

[45] J. Neyman *Phyl. Trans. Roy. Soc. London A236 (1937) 333*

[46] J. Neyman *Giorn. Ist. Ital. Att. 6 (1935) 320*

[47] J. Neyman and E.S. Pearson *Phyl. Trans. Roy. Soc. London A231 (1933) 289*

[48] G.J. Feldman and R. Cousins *Phys. Rev. D57 (1998) 3873*

[49] K. Pearson *Phil. Trans. Poy. Soc. London A186 (1895) 343*

[50] K. Pearson *Phyl. Mag. 50 (1900) 157*

[51] G. E. Forsythe *Generation and Use of Orthogonal Polynomials for Data Fitting on a Digital Computer J. Soc. Indust. Appl. Math 5 (1957) 74*

[52] S.S. Wilks *J. Ann. Math. Stat. 9 (1938) 60*

[53] E.L. Crow and R.S. Gardner *Biometrika 46 (1959) 441*

[54] B.P. Roe and M.B: Woodroofe *Phys. Rev. D60 (1999) 053009-1*

[55] C. Giunti *Phys. Rev D59 (1999) 053001*

[56] G. Punzi *INFN-PI/AE 99/08*

[57] G. D'Agostini *CERN 99-03, 19 July 1999*

[58] R. T. Cox *Probability, frequency and reasonable expectation, Am. J. Phys. 14 (1946) 1*

[59] The Atom *Fermi Invention Rediscovered at LASL*, *Los Alamos Scientific Laboratory*, October 1966

Contents