The Basic Principles of Statistics

Tarcisio Del Prete
delprete@pi.infn.it

INFN Sezione di Pisa

6^0 Seminario Nazionale sul Software,
Alghero, 1-5 Giugno 2009

Outline

Part 1

- Probability defined by the frequency of repeatable experiments.
- Probability as the measure of the degree of belief an individual has in uncertain proposition.
- Example (and Exercises!)

Probability and Statistics

- The theory of Probability is a branch of pure mathematics. It is based on axioms and definitions. Propositions are then obtained deductively. The neatest approach is based on set theory, measure theory and Lebesgue integration.
- The theory of Statistics is essentially inductive and empirical, since, from the observation of events infers the value of the unknown parameters and of hypothesis. This process is similar to what a physicist does when building a theory from measurements.

Example: a problem of Probability

Find the probability of observing n heads when tossing a coin N times, knowing the probability p of landing heads.
The solution is given by the binomial distribution and is:

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

The result is obtained by a pure deductive method, once the axioms of probability have been stated.

Example: a problem of Statistics

A coin is flipped N times and it falls heads n times. What can we say on the unknown probability p of landing heads?
This is a problem of inference and the answer cannot be as unambiguous as in the previous example. The main questions that we shall try to answer are:

- The *best* estimate of p: $\hat{p} = n/N$.
- The precision of the estimate.
  - the standard deviation of $\hat{p}$, but more precisely
  - the *credible or confidence interval* $[p_-, p_+]$ which contains the true value of p with some *confidence level*.
- The data are compatible with the hypothesis that $P = \alpha$?

Our main tool will be the theory of the Probability.

What is the Probability?

Probability theory is nothing but common sense reduced to calculation.

(Laplace 1818)

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*. Even the previous definition would not be accepted by all Statisticians. Many would prefer to say: "... quantifies the status of knowledge of a proposition whose truth we are not sure".

The two main Schools I

Frequency theory (R. Von Mieses) It applies to repeatable events.
We observe an event A occurring n times in N trials, the probability is defined as the limit:

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

The two main Schools II

Subjective or Bayesian (F.P. Ramsey, B. De Finetti). It is more general than the previous definition since it applies also to non-repeatable experiments.
The probability of an event *A* to occur (or an uncertain proposition to be true) is the measure of one's belief in its occurring. In this definition the probability loses any objective content.
However this definition is more general than the previous and applies also to *propositions* whose truth we are uncertain.
This is the so called *Modern* definition. The power and limitations of this approach will be discusses at length later.

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.

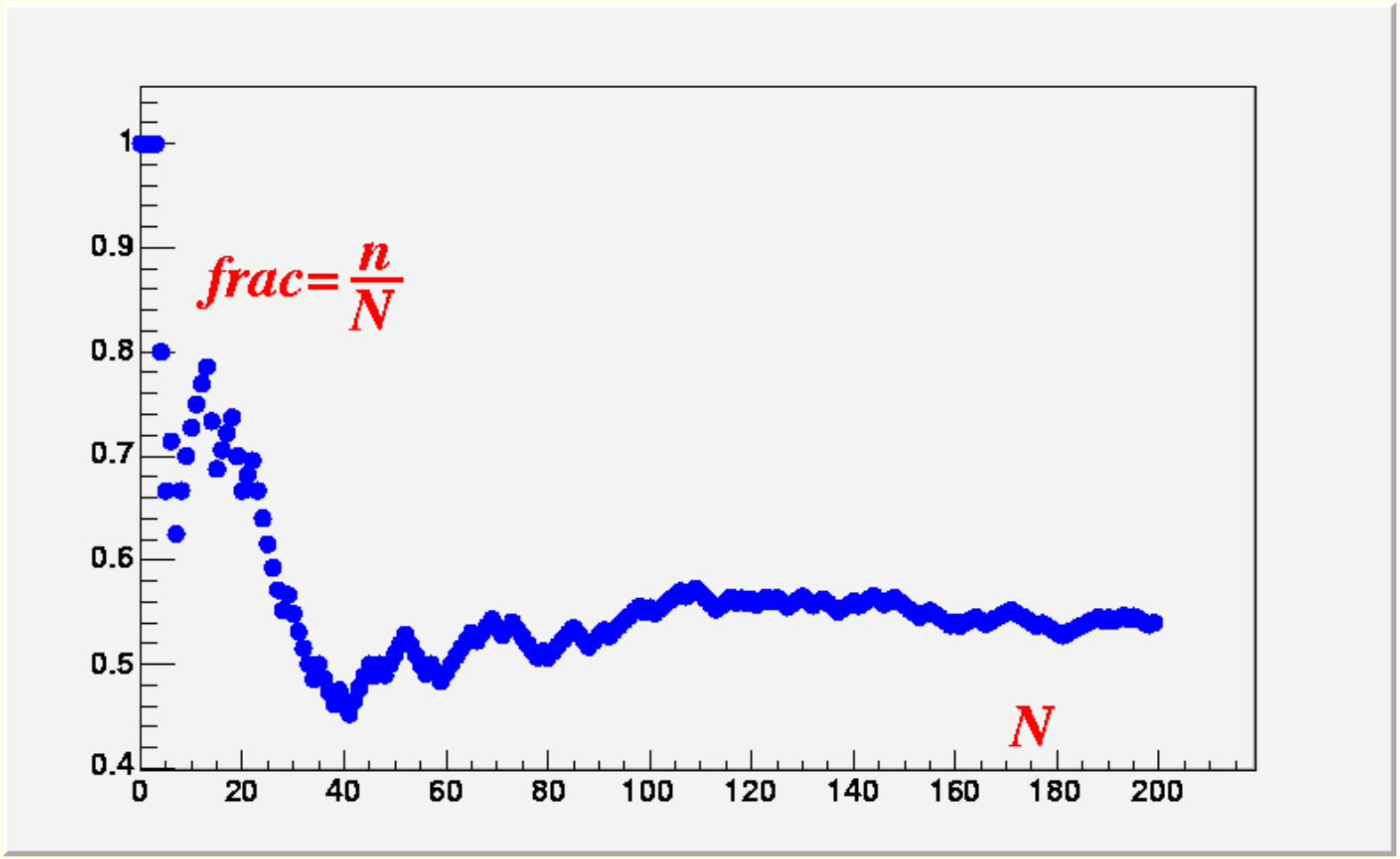
In spite of the irregular behavior of the individual events, their frequency, in a long sequence of random experiments, performed in uniform conditions, is rather stable and seems to converge to a constant value as the number of experiments tends to infinity.

This limit is called *probability* of the event.

The mathematical abstraction leads to the concept of probability as a non negative measure of an event P_E, in some space (see later).

An experiment done in class

I asked students to throw a die several times and measure the frequency of success; (success: the die lands less then 4.)



Measuring Probabilities

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

$$\boxed{P(E) = \lim_{N_{tot} \to \infty} \frac{N_E}{N_{tot}}}$$

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.
$P(E)$ is a property of the system.

Random Event

In principles classical physics is deterministic...

When we flip a coin...

if we would know exactly its starting position, velocity, angular momentum, coin tensor of inertia... — Initial Conditions

If we would know also air temperature and viscosity, wind velocity, elasticity of the table where the coin is going to land — Boundary Conditions

Then we would have a perfectly defined deterministic problem, we could compute for each flip which would be the result and we leave no space to the game of chance.

Random Event

But usually we do not control all the initial conditions and we know very little of the coin and of the initial and boundary condition which determines the motion.

A very little change of the initial conditions has a dominating influence on the result which becomes unpredictable in each coin flip.

QM events are at all unpredictable, even in principle, and are the best example of a Random Event.

Random Event

The result of individual events cannot be predicted with certainty.
However the situation changes if we study a sequence of events.
The average of a long run sequence of events performed in uniform conditions, shows a regular behavior.
Historically this behavior of the frequency was first observed in the field of the games of chance (coin, dice, cards, etc.). The frequency of a given result seemed to converge to the same precise value when the game was repeated very many times.

The Population

This regularity can be interpreted by considering the *samples* as part of a very large *parent population*.
Consider a town consisting of N individuals.

The Population

We are interested in some property B of the town's population
i.e. people higher than 160 cm and we want to know
$P(h > 160cm)$
In principle we could measure all N people and know the number N_B higher that 160cm.
But this would be expensive and time consuming.
We observe n $(\ll N)$ individuals (a sample of individual from the town's population) and count a number n_B of people higher than 160 cm.

The Population

The frequency $\nu = n_b/n$ is a measure of $P(h > 160cm)$, it will not be exact and it will change if we repeat the sampling.

If the sample size increases, the frequency will tend to the true value $\nu^* = n_B/N$, since we leave less and less space to the fluctuation. If we sample the whole population ($n = N$) then $n_b = n_B$ and we have measured the true fraction ν^*.

Population and Sample

The town's population:

- The population is characterized by (unknown) parameters $\theta_1, \theta_2, \cdots$ that we want to estimate from measurements.
- We sample *n* elements (we make n measurements)
- from those we infer the values of $\theta_1, \theta_2, \cdots$.
- Sampling has to avoid *bias*.

The mathematical abstraction of the real population is a finite (or infinite) population whose members have a probability $P(X, \vec{\theta})$.
The parameters $\vec{\theta}$ are the subject of our interest.

Sampling

The sampling is difficult! each element of the population must have the correct probability of being sampled.

Sometimes a bias is unavoidable. In these cases we have to know it else the sample is useless and cannot be used to estimate $\vec{\theta}$.

The methods that we will study are optimal in average (frequency statistics).
Unfortunately we have only ONE experiment.

Sampling and parameter estimation

In the long run, we're all dead.
J.M. Keynes
It is better to be roughly right than precisely wrong

J.M. Keynes

The frequency methods are designed to provide an unbiased and consistent estimator.

They are based on the assumption of a fair sampling. If the sample is not fair we must know the bias introduced by the sampling to proceed to estimation.

In the long run and with a proper sampling, the mean value of these estimators would *converge* to the true value of the signal.

Frequency Theory and Sampling

The properties of the population are the object of the scientific measurement.

In the example of the town the population was a real entity. The population is then abstracted to a mathematical object populated by all the possible events, each in proportion of the its probability. A measurement consists of drawing a sample from this population.

The interpretation of probability as the limit of the frequency gives a degree of reality to those samples that have not been measured but can eventually be drawn from the population by virtual experiments.

In practice

The aim of the frequency statistics is to infer from the data to the value of the unknown parameters.
This is performed by constructing functions of the data $a(\vec{x})$ with properties:

- $E(a) = \theta$ (unbiased). Repeated measurements should cluster around the true value of the parameter.
- $Var(a)$ decreases (consistency); the precision increases if we increase the number of measurements,
- (Sufficiency).
- (Efficiency).

The function a (called estimator) is good in average for all the samples. It is NOT optimal for the particular sample that we have made.

Short Summary (Frequency Theory)

The basic assumptions in the Frequency Theory can be summarized by the following lines:

- $P(A) = \lim_{N\to\infty} \frac{N(A)}{N}$
- The limit has to be possible at least in principle. There is no need to evaluate it, unless we want to measure $P(A)$.
- The probability can be measured directly.
- $P(A)$ is restricted to repeatable experiments.
- $P(A)$ is an intrinsic property of the system. This is why we can repeat the sampling.
- This probability is the one used in QM:

$$P(x \in S) = \int_S |\psi(x)|^2 \, dV$$

Subjective Theory I I

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 (1936)).
According to this theory, to any proposition on which there is no certainty we associate a numerical value, the 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.
This probability $P(E)$ expresses one's opinion on the proposition (E) and depends on the information available to whoever evaluates E.

Subjective Theory II

Consider propositions like:

- The mass of the Higgs boson is less than 200 GeV.
- The absolute value of the charge of electrons and protons are equal.
- I will swim tomorrow

These proposition are uncertain and we can associate a probability to each of them.
Each of us can have his/her own opinion on the specific topic.
Hence each of us will assign a different probability to each statement.
Since no objective rule can be constructed to express one's opinion, this theory of Probability is called *subjective*.

Subjective Theory III

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.

- There is NO random event.
- There is NO repeated samples.
- The probability is NOT an intrinsic property of the system but rather it is assigned to the event/proposition by the observer.
- The probability cannot be measured, it is assigned by the experimenter.

Short Summary (Bayes Theory)

The basic assumptions in the Subjective (or Bayes, since it relies a lot on the use of Bayes theorem) Theory can be summarized by the following lines:

- $P(A)$ is NOT an intrinsic property of A, it depends on the state of Information to whoever evaluates $P(A)$.
- It is always conditional to some *background* information.
- It is used very often in everyday life (*...he is probably right... ...m_H is probably less than 200 GeV... etc. etc.*)
- Often it is subjective and cannot be falsified.

We will see in short in more details how all this works.

Outline

Part 3

- Abstract probability.
- Bayes theorem.
- How Bayes statistics works.
- Example (and Exercises!)

Mathematical Probability

Fortunately both definition of probability (from the Frequency and Subjective schools) are fitting in the abstract definition of probability given by Kolmogorov.
Assume that in a given experiment the possible outcomes are:

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

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

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

Each e_i is called *an elementary event*. The collection of all e_i is called *sample space (S)*.

Mathematical Probability

We shall use the language of set theory:
an *event* E is a subset of S, i.e. the ensemble of elementary events that share the same property.

$$E \equiv \bigcup_{\{k\}} e_k$$

Example

In a dice game $E = \textit{die lands even}$ than $E = \{2, 4, 6\}$.

Class of E: it is made by subsets of S. There is no need to define the class of events of as the ensemble of all subsets of S. This generates often mathematical monstrosities.
$\rightarrow$ Use the smallest event space compatible with our problem.

Exercising with events

Example

In the previous exercise, let us use an "Indicator" to avoid non numeric symbols: $I = 1$ *if* H, 0 *if* T. If we flip the coin n times the outcome is a ordered sequence of "0" and "1".
The sample space is $S_n = S_1 \times S_1 \cdots S_1$ and is composed of all possible ntuples of "1" and "0"; its dimension is 2^n.
If we write the outcomes as 0 and 1, in order, at the right of the decimal point, then each event is represented by a rational number (in binary) in $[0, 1]$:

$$s_i = 0.00110... \qquad \text{there are } 2^n \text{ of such numbers}$$

If we let $n \to \infty$. The number so constructed is a real number.
Continuous variables belongs to infinite dimensional sample space.

Exercises!

In many instances it is possible to enumerate all possible events (N) and also those in which E occurs (m). If all the cases have the same probability, than the probability of the event E is $P(E) = m/N$.

Exercise

What is the probability that throwing n times a die we get at least once 6?

The algebra of sets

Let us start from the set of elementary events: $S \equiv \{sample\ space\}$. We will call *event class E* the set of all subsets of S having the following properties (closure under union):

- 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 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*.

The metrix on the sets

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):

- $\forall\ A \subseteq E \qquad 0 \leq P(A) \leq 1$
- The probability of occurrence of any event is one: $P(S) = 1$.
- $\forall (A_i, A_j) \subseteq E :\ A_i \bigcap A_j = 0\ (i \neq j) \qquad P\left(\bigcup_i A_i\right) = \sum_i P\left(A_i\right)$ (countable additivity).

Exercises!

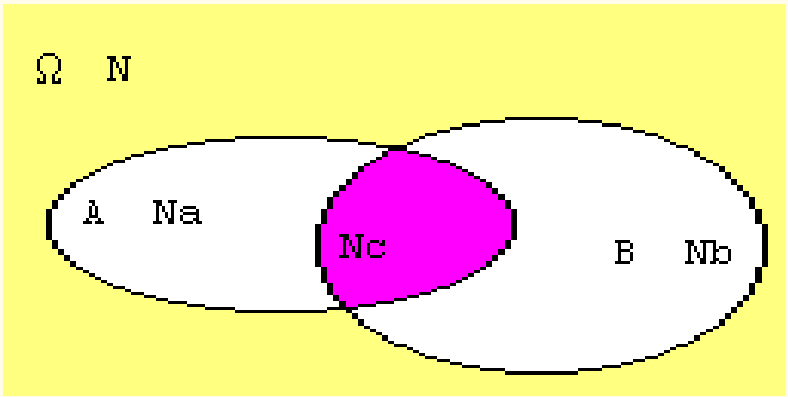
Exercise

With the use of the axioms and the properties of set theory prove: :

- $\forall A \subseteq S: \qquad P(\overline{A}) = 1 - P(A)$
- *If* $B \subseteq A$ *then* $\qquad P(B) \leq P(A)$
- $\forall A, B \subseteq S: \qquad P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability

Suppose $A,\ B \subseteq S$ with probabilities $P(A)$ and $P(B)$. Suppose that we know that A is true. In this case the probability of B is *relative* to the sample space A. This new probability will be called *conditional probability*. The following Venn diagram illustrates the meaning of conditional probability.



$$P(A) = \frac{N_A}{N}$$

$$P(A \cap B) = \frac{N_C}{N}$$

$$P(B|A) = \frac{N_C}{N_A}$$

$$\boxed{P(B|A) = \frac{P(A \cap B)}{P(A)}}$$

Independence

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, for independent events:

$$\begin{aligned} P(E_2|E_1) &= P(E_2) \qquad \text{hence} \\ P(E_1 \cap E_2) &= P(E_1) \cdot P(E_2) \end{aligned}$$

Conditional Probability

Example

A card is drawn from a pack of 52.

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

Since $P(\textit{ace of spade}) = 1/52 = P(A) \cdot P(B)$, the events are independent.

Exercise

We throw two dice. What is the probability that the sum is 9, if the first die gave 5?

Exercise

Consider two events A and B. We know that:

$$\begin{aligned} P(A) &= 0.1 \\ P(B) &= 1.0 \end{aligned}$$

are they independent?

Exercise

Consider two events A and B which are mutually exclusive. Are they independent?

Exercise

If $A \subseteq B$ what is $P(A \mid B)$?

Is it so simple and clear?

Exercise

A friend of mine has two children. One is male. What is the probability that both are male?
(The probability that a child is male is roughly equal to the probability to be a female)

Is it so simple and clear?
The prosecutor fallacy: the Sally Clark case

Sally Clark, a British woman, was accused in 1998 of having killed her first child at 11 weeks of age, then conceived another child and allegedly killed him at 8 weeks of age.
The defense claimed that these were two cases of sudden infant death syndrome.
The prosecution had expert witness Sir Roy Meadow testify that the probability of two children in the same family dying from sudden infant death syndrome is about 1 in 73 million.
Based only on this probability statement, Mrs Clark was convicted in 1999.
The Royal Statistical Society, in a press release, pointed out the mistake.
A higher court later quashed Sally Clark's conviction but on other grounds, on 29 January 2003.

Is it so simple and clear?
The prosecutor fallacy: Sally Clark case (cnt.)

The argument can be analyzed using conditional probability. let us call:

- E= the observed evidence;
- I= the accused is innocent;

Hence:

- $P(E|I)$ is the probability of the *evidence* if the accused is innocent; this is the quoted 1/73 millions.
- $P(I|E)$ is the probability that the accused is innocent given the evidence.

The prosecutor wrongly confuses P(I|E) with P(R|I), two rather different concepts. We will need Bayes theorem to compute one from the other. (We will resume later the discussion on Sally Brown case.)

Is it so simple and evident? The Monty Hall Problem I

Monty Hall problem involves a classical game show situation and is named after Monty Hall, the long-time host of the TV game show Let's Make a Deal.

There are three doors labeled 1, 2, and 3. A car is behind one of the doors, while goats are behind the other two: The rules are as follows:

- The player selects a door.
- The host selects a different door and opens it.
- The host gives the player the option of switching from her original choice to the remaining closed door.
- The door finally selected by the player is opened and she either wins or loses.

Is it so simple and evident? The Monty Hall Problem II

You can either change the door or keep the door you selected first. The aim is to optimize your chances of winning the car.
At the beginning of the game nothing is known of the "good" door and your chances were 1:3.
After the host has made his choice you should know something more...
(The problem was first proposed as three box paradox by J. Bertrand (1889), and then as the three card paradox by W. Weaver (1950).)

Try to work out the solution by yourselves first, we will discuss the solution of the problem off-line during our coffee-break!

Bayes' Theorem

We have seen how to compute the conditional probability $P(A \mid B)$, the probability of A knowing that B has occurred.

How can we express $P(B \mid A)$ using the knowledge of $P(A \mid B)$? i.e. how can we interchange the A with B in the conditional probability?

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 Bayes work is subtle and points to fundamental concepts at the basis of the probability theory.

Bayes' Theorem

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

$$\forall (B_i, B_j) \subseteq S \quad B_i \bigcap B_j = 0 \quad i \neq j \qquad \text{disjoint sets}$$
$$\bigcup_i B_i = S \qquad \text{complete sets}$$

then:

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

If A is also a set belonging to S then:

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

hence:

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

Bayes' Theorem

Given the definition of B_i we have (marginalization):

$$A = A \bigcap \left(\bigcup_j B_j \right) = \bigcup_j (A \bigcap B_j) \qquad \text{union of disjoint sets}$$

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

Finally:

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

This is Bayes theorem.

A classical example I

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 and take a coin. It is a gold coin.
This event has increased our information on which is U_1.
Let us use Bayes Theorem!
The event $U_i \equiv$ *the chosen urn* $U_M = U_i$

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

A classical example II

- A (event) pick up a gold coin.
- $P(A|U_i)$ is called Likelihood. By construction:

$$\begin{cases} P(A \mid U_1) &= 1 \\ P(A \mid U_2) &= 1/2 \\ P(A \mid U_3) &= 0 \end{cases}$$

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

A classical example III

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), hence:

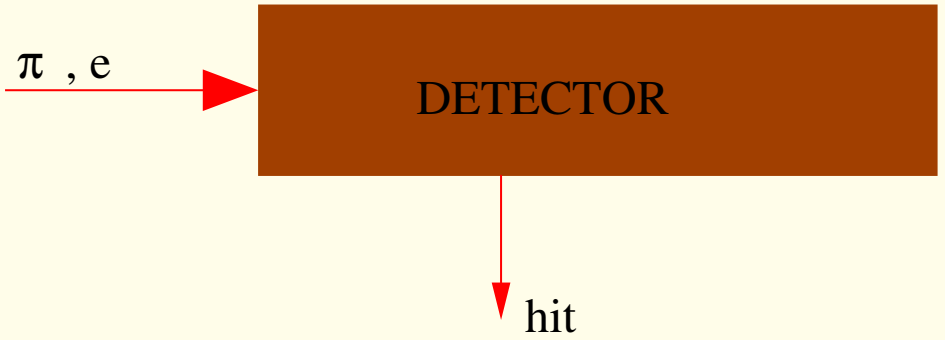
$$
\begin{aligned}
P(U_1|A) &= \frac{1 \cdot 1/3}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3} \\
P(U_2|A) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{3} \\
P(U_3|A) &= 0
\end{aligned}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

The urn that we have randomly selected and in which we have found a gold coin has the larger probability of being U_1.

The observation has changed the prior probabilities and we have gained information on the system.

A HEP application of Bayes theorem

A beam of particles contains pions and electrons. A detector is designed to respond to π and e with a hit with:



$$P(hit \mid e) = 0.9$$
$$P(hit \mid \pi) = 0.05$$

A particle enters the detector and produce a hit. What is the probability it is an electron?:

$$P(e \mid hit) = \frac{P(hit \mid e) \cdot P(e)}{P(hit \mid e) \cdot P(e) + P(hit \mid \pi) \cdot P(\pi)}$$

To compute $P(e \mid hit)$ we have to know $P(\pi)$ and $P(e)$, the priors (i.e. the beam composition).

Interpretation

The characteristics of the detector are not enough to identify the incoming particle. We have to measure the beam composition with an ancillary experiment.
(If only 5% of the beam would be electrons, π and e would produce hits with the same rate, overcoming the different sensitivity of the detector to π and e.)
Finally we shall have $P(e \mid hit)$. How do we interpret this quantity?
Assume $P(e \mid hit) = 0.75$. Since the incoming particle is either an electron or a pion we might rephrase this number saying that, in average, in a sample of 100 hits, 75 will be electrons and 25 pions.
(This interpretation has a rather frequentistic flavor...)

Another application of Bayes theorem

Example

0.1% of the population is affected by AIDS. A test identifies sick persons with a probability of 0.98. The test 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?

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

A failure of the computerized diagnostic?

Interpretation

The prior $P(AIDS)$ is the key to understand this apparent paradox.
We 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 ($P(T \mid \overline{A})$).
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.
The use of Bayes theorem depends critically on the prior.

Exercise

However...

Exercise

You repeat a second time the test and again the test is positive. Compute the probability of being sick $P(A \mid TT)$.

Note: $P(A \mid TT..)$ improves the knowledge on the status of the individual. The same results would be obtained also in the previous example (π e discrimination) by adding another detector and using the posterior probability of the first as prior to the second.

The problem of priors

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 knowledge?
- Or, even more difficult, to convey our ignorance?

It seems that there is no general satisfactory answer to this second question.

Another example (1)

Example

A box contains 3 balls either black or white. We make three extraction with replacement. The three extracted balls are 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.

First problem: H is not a random variable, rather it is a fact: the composition of the box:
$P(H)$ is either 1 or 0 depending on which balls were put in the box, but we do not know which one is true!

Another example (2)

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.
Second problem: How is constructed the collection of boxes?

Another example (3)

Quantum balls

$$\{W\ W\ W\}$$
$$\{W\ W\ B\}$$
$$\{W\ B\ B\}$$
$$\{B\ B\ B\}$$

Classical balls

$$\{W\ W\ W\}$$
$$\{W\ W\ B\}\ \{W\ B\ W\}\ \{B\ W\ W\}$$
$$\{W\ B\ B\}\ \{B\ B\ W\}\ \{B\ W\ B\}$$
$$\{B\ B\ B\}$$

The problem is substantial: you have to know which is the *collection*.

Exercise

Work out the posterior probabilities using both collections of boxes.

Bayes and Sally Clark case

Back to Sally Clark case. The Bayes formula is needed to compute $P(I \mid E)$. We will see it is not so easy... We have called:
E = the observed evidence, I = Sally is innocent

$$P(I \mid E) = \frac{P(E \mid I)P(I)}{P(E \mid I)P(I) + P(E \mid \bar{I})P(\bar{I})}$$

$P(I)$ can be estimated from the frequency of those crimes in the population. In those years England and Walls reported 30 children killed by mother on 640000 births.
$P(I) = 1 - 30/640000 = 1 - 5 \cdot 10^{-5}$. Also $P(E \mid \bar{I}) \approx 1$. Thus:

$$P(I \mid E) \approx \frac{10^{-8} \cdot 1}{10^{-8} \cdot 1 + 1 \cdot 10^{-5}} \approx 10^{-3}$$

Small but not so small.

Bayes Theorem and the Subjective Probability

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 $(0, 1)$.

Bayes Theorem and the Subjective Probability

Both interpretations of probability already existed at the beginning of the theory of probability, in the middle of seventeenth century:

- the empirical based on the frequency of random events, widely used in the games of chance,
- the epistemological based on our (incomplete) knowledge of facts or propositions; the name of probability has even its etymology in the Latin word *probus*: praised by a wise person and shows that the cognitive meaning of the word is very rooted in the logic of everyday experience.

These so different concepts should deserve different names!
Many complications arise by the use of the same word for different concepts!

What (probably) Bayes wanted...

Rev. Bayes (1763) knew the answer to the problem:
In a Bernoulli trial of N with probability p what is the probability of n successes?
If p is known the answer is 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?
Bayes theorem suggests the answer:

$$P(p \mid n, N) \propto P(n \mid p, N) \cdot P(p) \qquad \text{Inference on } p$$

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)?

What (probably) Bayes wanted...

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.
$P(p)$ is a way of quantifying our personal believe on the value of p, before the measurement. This belief is modified by the experiment leading to $P(p \mid n, N)$.
Different persons can have different opinions and hence different $P(p)$; hence different $P(p \mid n, N)$.
The probability is not a property of the system, but depends on the information available to the experimenters.
In this scheme no frequency interpretation of $P(p)$ is possible, and no MEASURE of p by the experience.

The problem that stopped Bayes from publishing his theorem was how to convey our complete ignorance of p in the mathematical expression of $P(p)$.
Bayes' naive solution to consider all the possibilities as equally probable leads to inconsistencies.

- A parameter θ can take any of the values: $\theta_1, \theta_2, \cdots, \theta_k$ with $k \geq 3$.
- Nothing is new on θ hence: $P(\theta) = 1/k$.
- Be 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, hence $P(\phi) = 1/2$.
- 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.

- $P(E)$ is not an intrinsic property of E but depends on the state of information available to whoever evaluates $P(E)$ (D'Agostini, 1999)
- Hence it is always conditional to some background information I we should always write $P(E \mid I)$.
- It is very often used: *It is probable that the mass of Higgs boson is larger than 100 GeV*, *probably tomorrow it will rain* etc.
- Since it is subjective, cannot be verified of falsified.

Bayes Inference in Binomial Experiment

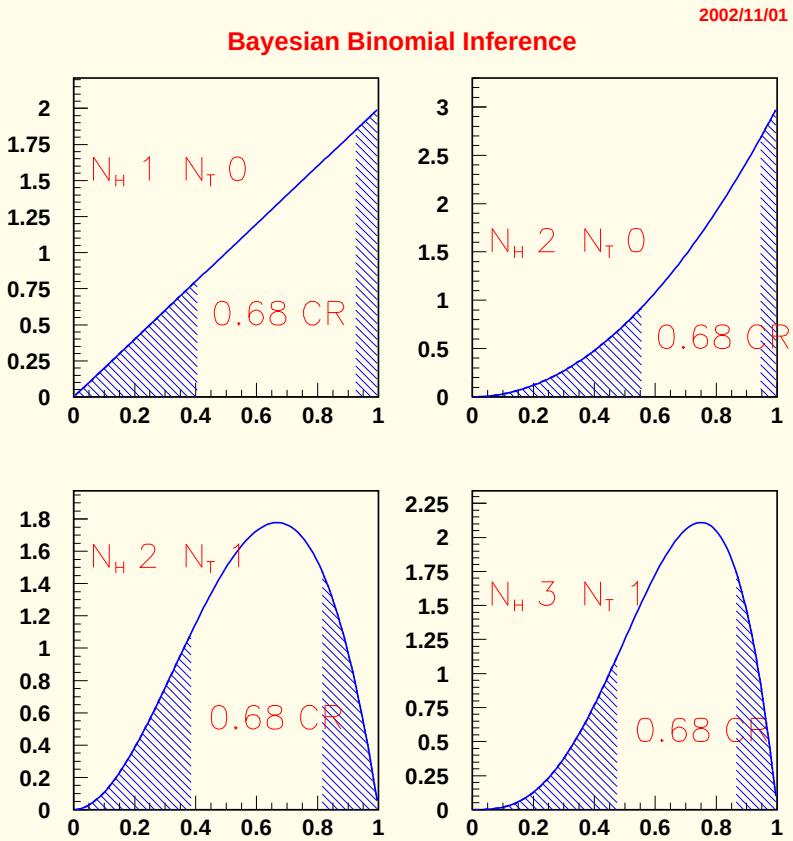
Bayes Theorem:

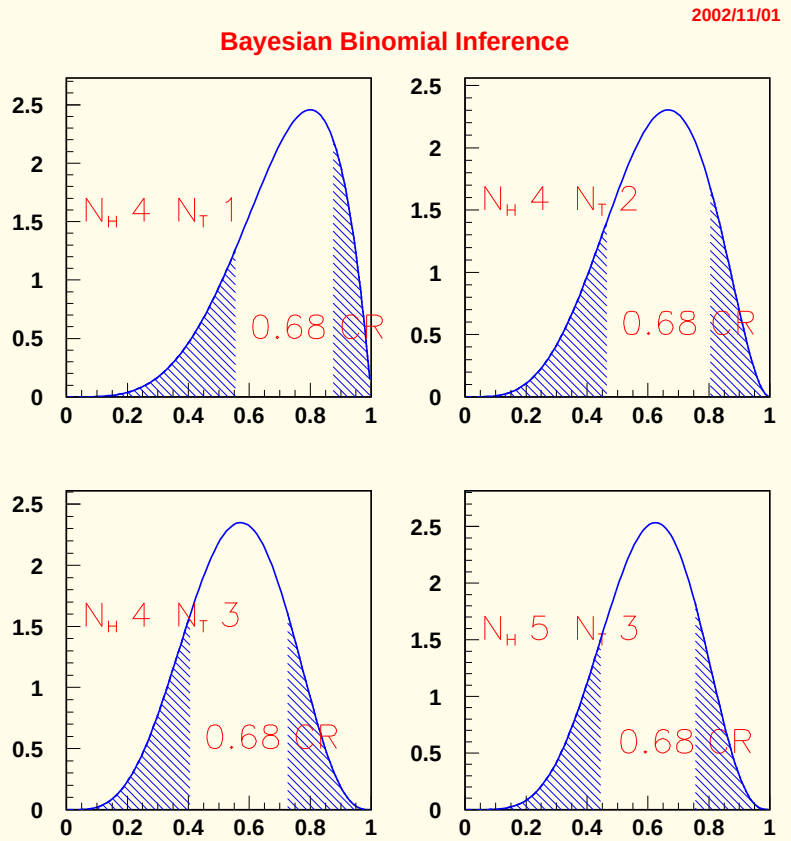
$$P(\theta \mid D, I) \propto P(D \mid \theta) \cdot P(\theta \mid I)$$

D is the *data*, I is the background information, $P(D \mid \theta)$ is the binomial formula.
Assume a flat prior $P(\theta \mid I)$ in a Binomial experiment: flip a coin.
θ is the probability of H
Bayes formula will update the information at each experiment:

$$\begin{aligned}
P(\theta \mid noData, I) &= P_0(\theta \mid I) = 1 \qquad \text{uniform prior} \\
P(\theta \mid H, I) &\propto \theta \\
P(\theta \mid H, H, I) &\propto \theta^2 \\
P(\theta \mid H, H, T, I) &\propto \theta^2(1-\theta) \\
&\dots \\
P(\theta \mid nN, mT, I) &\propto \theta^n(1-\theta)^m
\end{aligned}$$

The posterior density after each measurement





The plots are the posterior density after each measurement.
Also shown the 68% Credible Intervals.

The Likelihood is all what we need

Actually it is not necessary to use Bayes formula at each experiment, we can use the Bernoulli formula as Likelihood directly:

$$P(\theta \mid D, I) \propto K \cdot \theta^H \cdot (1-\theta)^T$$

The normalization K depends on the number of Heads and Tails (data) but not on the parameter hence it is irrelevant for Bayes Statistics.

- The maximum of the density (mode):

$$\theta_{mode} = \frac{H}{H+T}$$

- The mean is:

$$\overline{\theta} = \int_0^1 \theta \cdot P(\theta \mid D, I)\, d\theta = \frac{H+1}{H+T+2} = \frac{H+1}{N_{tot}+2}$$

Few comments

- We started from an uniform prior and, after a few experiments we lose track of the prior and are driven mainly by data. (D'Agostini: Overcoming prior anxiety (1999)).
- However in case of frontier physics, when only few events will be available the dependence from priors (or theoretical prejudices) can be very annoying.
- The mean is NOT the same object as in classical statistics; it is not the average over many experiments (over sample space), a concept unknown to Bayes school. It is the average over the posterior obtained by the experiment just performed.
- The Bayes mode coincides with the mean over sample space only because we have used a flat prior.

More comments

In Bayes Statistics all the information on the system is contained in the Likelihood.
The Likelihood is a function which describe mathematically the system with the measured data. Any inference must proceed through the Likelihood and the available prior which should consider any background information.
We must draw the same inference from two experiments having the same likelihood
Hence Bayes Statistics rejects the idea of the virtual repetition of experiments (no random events!). In fact these are not measurements and do not bring any information on the system and must not be considered

More comments

The analysis will produce a Posterior Probability density for the parameter θ. The posterior conveys all the information on the system.
Summary information (i.e. the point estimate) is produced using decision theory by minimizing a loss function. Loss functions are subjective; frequently used is a quadratic loss function which produces an estimate of the parameter equal to the mean. The mean is the average of the density of the experiment just performed. (Average over the parameter space.)
The analysis is, in a sense, optimized to the particular results obtained in the experiment. This contrary to the Frequency school whose methods work well for any sample (Average on sample space)

Least Informative Priors

The priors, mainly in case no prior information is available, is a delicate point in Bayes analysis.
If prior information exist, Bayes analysis will automatically provide the update of the previous information with the new data.
But if no information exists we should resist to the temptation of injecting in the analysis our prejudices or sympathies.
We should (at least in Physics) use priors which do not influence, or influence little, the posterior. We should let data speak by themselves!
This is mainly true in case of frontier experiments where data are few and we cannot rely too much on the independence from the prior of the large statistics.
There comes the problem of the so called least informative priors.

Least Informative Priors

The flat prior is not the only choice, in case of absence of any information before we do the experiment. The flat prior is very intuitive but there are other considerations.

- 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 $\theta \rightarrow k \times \theta$. (I want to be ready with my prior to any size of object to measure, before I know the subject of my investigation.)

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

hence:

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

This is the so called Jeffreys' prior.

Least Informative Prior

- If, on the contrary, a parameter is unbounded and we are interested in the location of the parameter, then, for the same argument, the *LIP* should be invariant under the transformation $\theta \rightarrow \theta + a$.

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

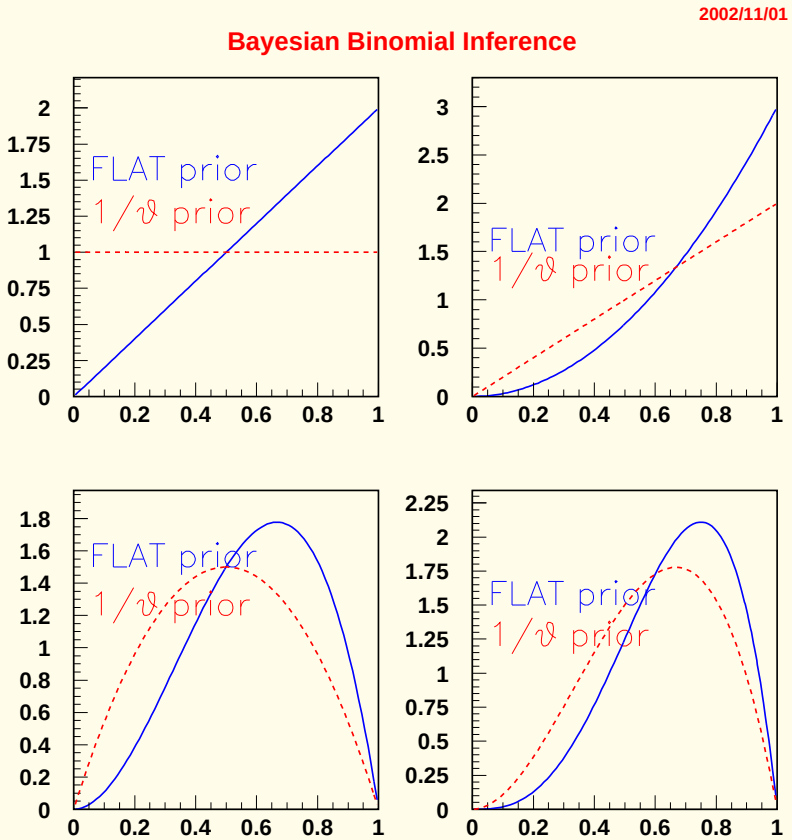
hence:

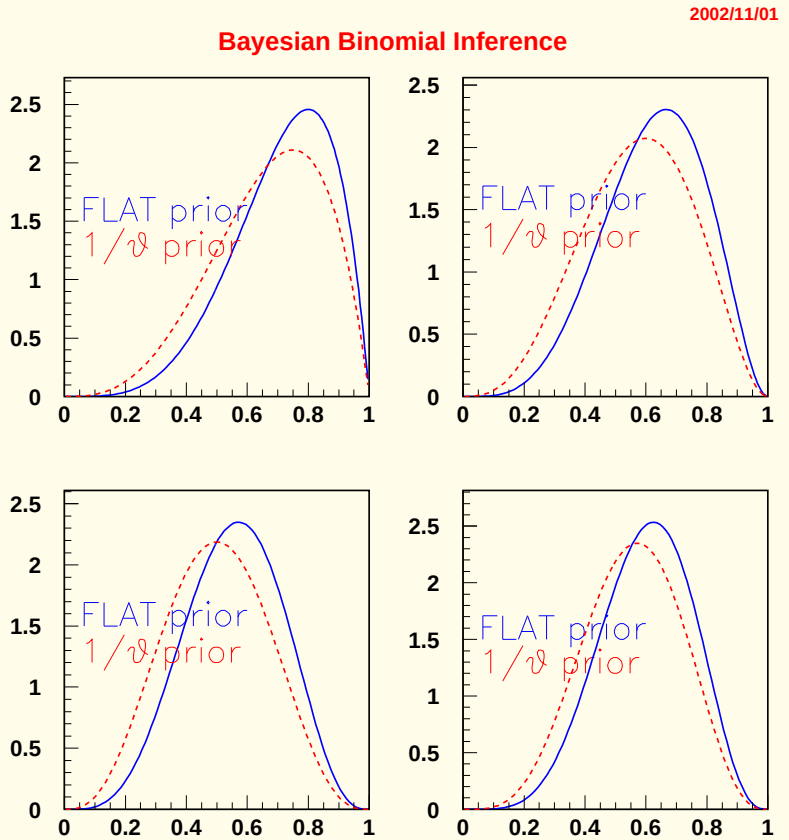
$$P(\theta \mid I) = constant$$

and we recover Bayes prescription.

Binomial model with constant and Jeffreys prior

Posterior density depends on the prior, but with increasing data it loses memory of the prior.





Bayes: search theory
The case of SSN Scorpion

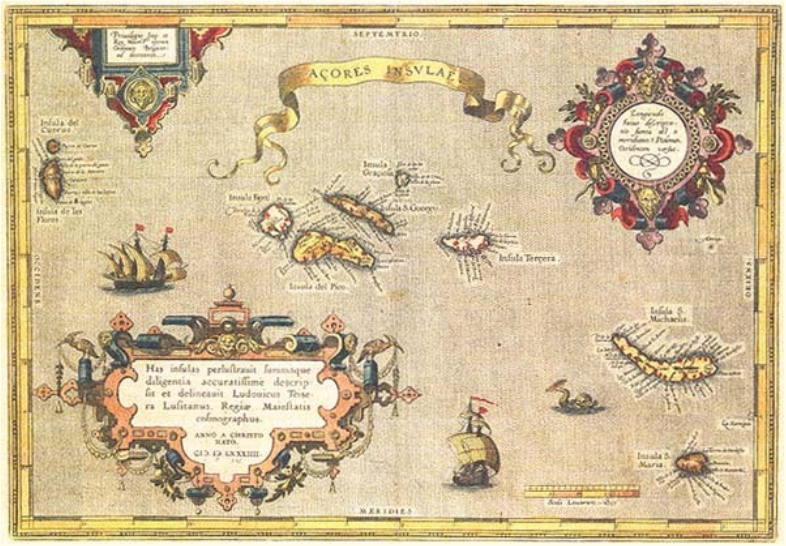
May 1968, SSN-589 failed to arrive to her home port in Virginia. First search failed to locate the wreck. A second search was organized adopting a Bayes search method.

The expected region of the loss was divided into grid squares (1 mile). In each square two probability were assigned:
$p = P(wreck \in square)$,
$q = P(find\ the\ wreck \mid wreck)$

Needed priors were obtained by submarine specialists, experts in deep water recovery, submarine commanders.

Bayes: search theory
The case of SSN Scorpion

The sea around the Azores was searched starting from the square with larger probability (total probability is P=p q)



Each time a square was unsuccessfully searched the probability of the square

$$p \rightarrow \frac{p(1-q)}{(1-q)p+(1-p)}$$

and all the squares probabilities were reassessed upward.

The use of this approach was a major computational challenge for the time, but it was successful: the Scorpion was found after five months.

Bayes: search theory
The case of SSN Scorpion

For your curiosity, Bayes theorem works like that, in this case:

$$
\begin{aligned}
A &= \textit{event : the wreck is in square } i \\
B &= \textit{event : the wreck is found} \\
P(A \mid \bar{B}) &= \frac{P(\bar{B} \mid A)P(A)}{P(\bar{B} \mid A)P(A) + P(\bar{B} \mid \bar{A})P(\bar{A})} \\
&= \frac{(1-q)p}{(1-q)p + 1(1-p)}
\end{aligned}
$$

The priors have been obtained from the experience of people working in the field and by considering several plausible scenarios for the wreck worked out with with Monte-Carlo calculations.

Part 3

- Two sampling theorems.
- Parameter estimation (frequency).
- Compare Bayes and frequency methods.
- Example (and Exercises!)

Back to Probability and Sampling (Frequency)

Let us now discuss two important theorems in the theory of Probability which are very relevant in the Frequency School of Statistics:

- The Law of Large Numbers,
- The central Limit Theorem.

We shall not give any proof, only state the theorems and give examples.
These two theorems are very important in Frequency inference.

The Law of Large Numbers (LLN)

Let us consider n repeated measurements $(x_1, x_2 \cdots x_n)$ of a random variable

$$X \sim f(x \mid \mu) \quad \mu \text{ is the finite mean and define: } \quad \overline{x} = \frac{\sum x_i}{n}$$

$$\lim_{n \to \infty} P(|\overline{x} - \mu| \leq \epsilon) = 1 \qquad (\forall \epsilon > 0)$$

The sample mean converges, in probability, to the mean of the population.

$$\overline{x} \xrightarrow{p} \mu$$

The theorem is true even if $f(X|\mu)$ has no variance.
This is *the weak law of large numbers*.

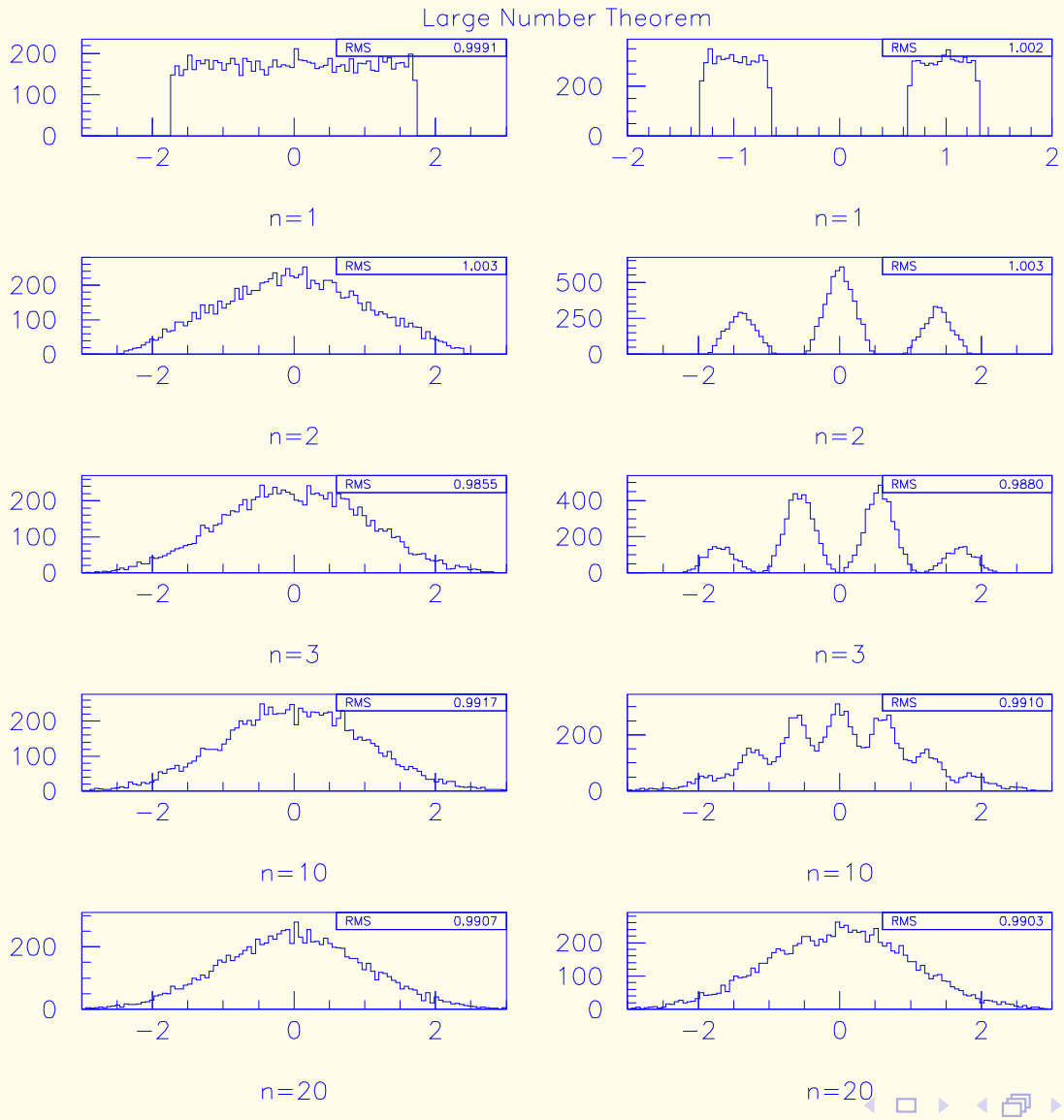
The Central Limit Theorem

Consider a variable $X \sim f(x \mid \mu \cdots)$, with mean and variance:
$E(X) = \mu \qquad Var(X) = \sigma^2$, both finite
The the distribution of the sample mean ($\overline{x} \sim g(\overline{x} \mid \mu \cdots)$) approaches the normal distribution with mean μ and variance $\frac{\sigma^2}{N}$ as $N \to \infty$.

$$\lim_{N \to \infty} g\left(\overline{x}; \mu, \sigma, N\right) = N\left(\mu, \frac{\sigma^2}{N}\right)$$

g is the finite sample size density of $\overline{x}$ which depends on the density of X.
Asymptotically $g(\overline{x}|\mu)$ becomes normal.

Example of the Central Limit Theorem



Example

My wrist watch (an old mechanical one!) in average does not systematically advances or delays, but randomly makes an error of 1/2' per day. What is the probability that in one year the error is less than $\pm 5'$?

X (daily error) $\sim U(-1/2, +1/2)$

E(X) = 0
Var(X) = 1/12

$r = \sum_{i=1}^{365} x_i$ is the error in one year:

$$P(|\sum x_i| \leq 5') = P(-5 \leq \sum x_i \leq 5) = P(-\frac{5}{365} \leq \frac{\sum x_i}{365} \leq \frac{5}{365})$$

Thus, by CLT:

$$P(-\frac{5}{365} \leq \overline{X} \leq \frac{5}{365}) \approx \int_{-5/365}^{5/365} N(0, \frac{1}{12 \cdot 365}) dx \approx 0.63$$

Estimation

We now discuss a few examples of *non standard estimation*. I hope that in this way you can grasp better the statistical principles of parameter estimation.
These principles are often hidden by the known methods (LSQ, Max Likelihood)

The German tanks problem

This happened really during WW2. It was a job of British statistician to estimate the number of German tanks from the debris of the tanks destroyed in battle.

(The same methods were used by Germans to estimate the production of arms by Soviets, by Japanese for the US weapon production...)

Also intelligence had its estimates, based on other, less evident methods.

The German tanks problems

During World War II, information about German war potential was essential in order to schedule the time of invasions and to carry out the bombing program.

In order to obtain reliable estimates of German war production, experts started to analyze markings and serial numbers obtained from captured German equipment.
Each piece of equipment was labeled with markings, which included

- (a) the name and location of the maker;
- (b) the date of manufacture;
- (c) a serial number;
- (d) miscellaneous markings such as trademarks, mold numbers, casting numbers, etc.

The German tanks problem

The serial numbers on tanks are numbers running from 1 to some unknown largest number N.
What we know is a subset of n numbers of this set obtained from destroyed or captured tanks $\{x_1, x_2 \cdots x_n\}$

This dataset is modelled as a realization of random variables $X_1, X_2, \cdots X_n$ representing n draws *without replacement* from the numbers 1, 2,$\cdots$ N with equal probability.

Sample Space	$1, 2, 3, 4, \cdots$ N
Data set	$X_1, X_2, \cdots X_n$

The objective is to estimate the total number N on the basis of the observed serial numbers.

The estimators

We propose two unbiased estimators. The first one is based on the sample mean

$$\bar{X}_n = \frac{X_1 + X_2 \cdots + X_n}{n}$$

and the second one is based on the sample maximum

$$M_n = max\{X_1, X_2, ..., X_n\}$$

Estimation based on sample mean

To construct an estimator based on sample mean, we start by computing the expected value of the sample mean

$$E\left(\bar{X}_n\right) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n}$$

This can be easily computed since $P(X_i = k) = 1/N$ for all k.

$$E(X_i) = \sum_k k \frac{1}{N} = \frac{1 + 2 \cdots + N}{N} = \frac{\frac{N(N+1)}{2}}{N} = \frac{N+1}{2}$$

Estimation based on sample mean

Thus:

$$E\left(\bar{X}_n\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{N+1}{2}$$

This directly implies that

$$\boxed{T_1 = 2\bar{X}_n - 1}$$

is an unbiased estimate of N

Exercise

Verify that $E(T_1) = N$

Estimation based on sample maximum M_n

The calculation of this statistics is a bit more elaborated.

The number of ways to draw n numbers out of N is $\left(\begin{array}{c} N \\ n \end{array} \right)$ and each combination has the same probability $\left(1 / \left(\begin{array}{c} N \\ n \end{array} \right) \right)$

In order to have $M_n = k$, one label must be equal to k and all the rest of the sample ($n-1$ numbers) out of the numbers $1, 2, \cdots k-1$.

There are $\left(\begin{array}{c} k-1 \\ n-1 \end{array} \right)$ ways to do that.

(Note: If $M_n = k$, then k has to have values $n, n-i \ldots N$.)

Estimation based on sample maximum

All these combinations have the same probability, thus:

$$P(M_n = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}} = n\frac{(k-1)!(N-n)!}{(k-n)!N!}$$

And:

$$E(M_n) = \sum_{k=n}^{N} kP(M_n = k) = n\frac{(N-n)!}{N!}\sum_{k=n}^{N}\frac{k!}{(k-n)!}$$

Exercise

Verify that: $\sum_{k=n}^{N}\frac{k!}{(k-n)!} = \frac{N+1)!}{(n+1)(N-n)!}$

Estimation based on sample maximum

Thus:

$$E(M_n) = n\frac{N+1}{n+1}$$

This directly implies that

$$\boxed{T_2 = \frac{n+1}{n}M_n - 1}$$

is an unbiased estimate of N

Exercise

Verify that $E(T_2) = N$

Which estimator should we use?

T_1 and T_2 are two unbiased estimators for the same parameter N. The estimator T_2 is called *more efficient* than T_1 if

$$\forall N \quad Var(T_1) > Var(T_2)$$

It is a bit too long and technical to compute the variances of the two estimators (the variables X_i are NOT independent). We quote the results:

$$Var(T_1) = \frac{(N+1)(N-n)}{3n} \qquad Var(T_2) = \frac{(N+1)(N-n)}{(n+2)n}$$

Thus:

$$\frac{Var(T_1)}{Var(T_2)} = \frac{n+2}{3} \geq 1 \quad \forall N,\ n$$

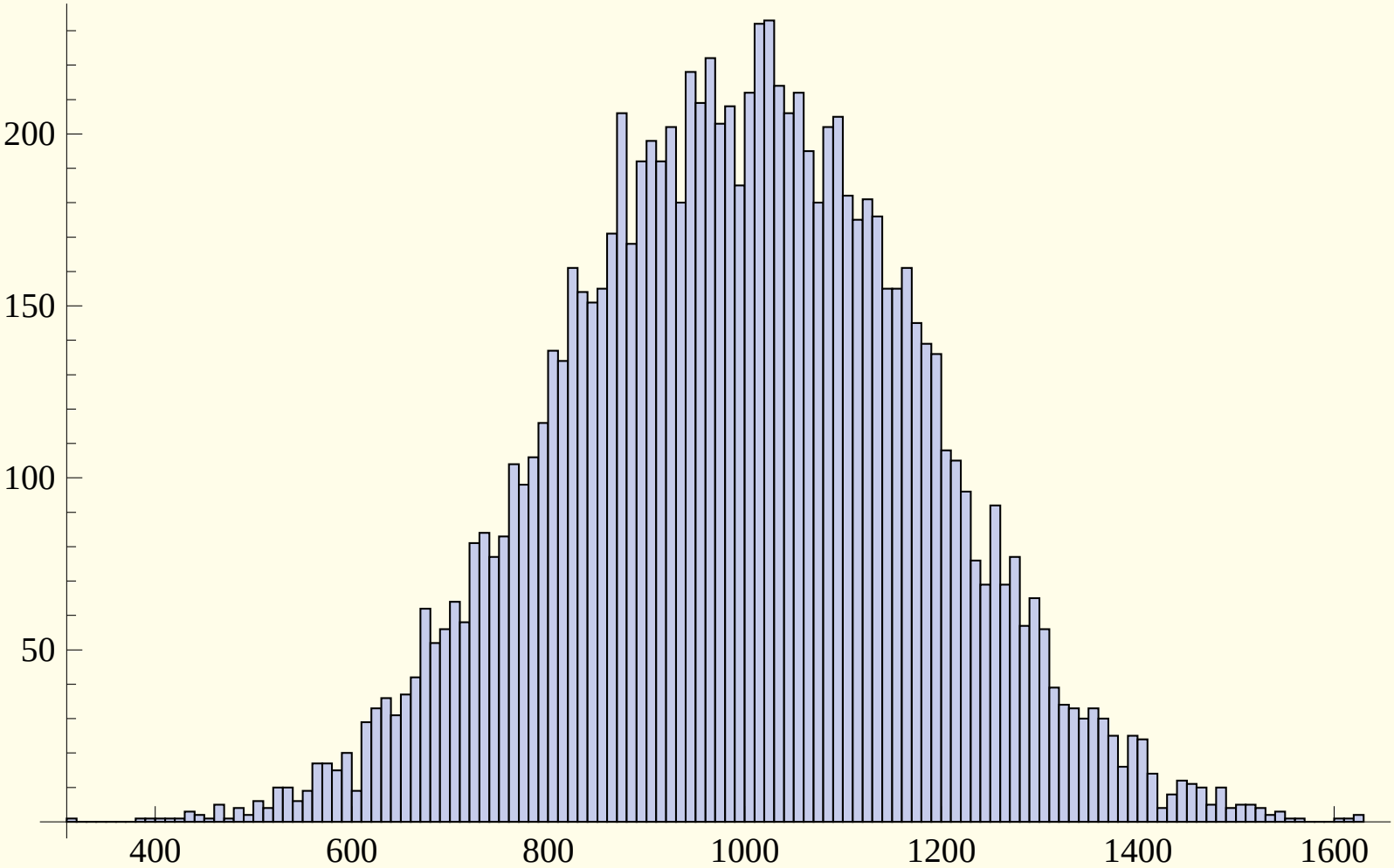
The distributions of T_1

Instead of demonstrating the formulas, we will make a small Monte-Carlo to compute *experimentally* the distribution of the two estimators.

Use this method of numerical calculation whenever a problem gets intricate and you do not have a clear feeling of the result!

The distributions of T_1

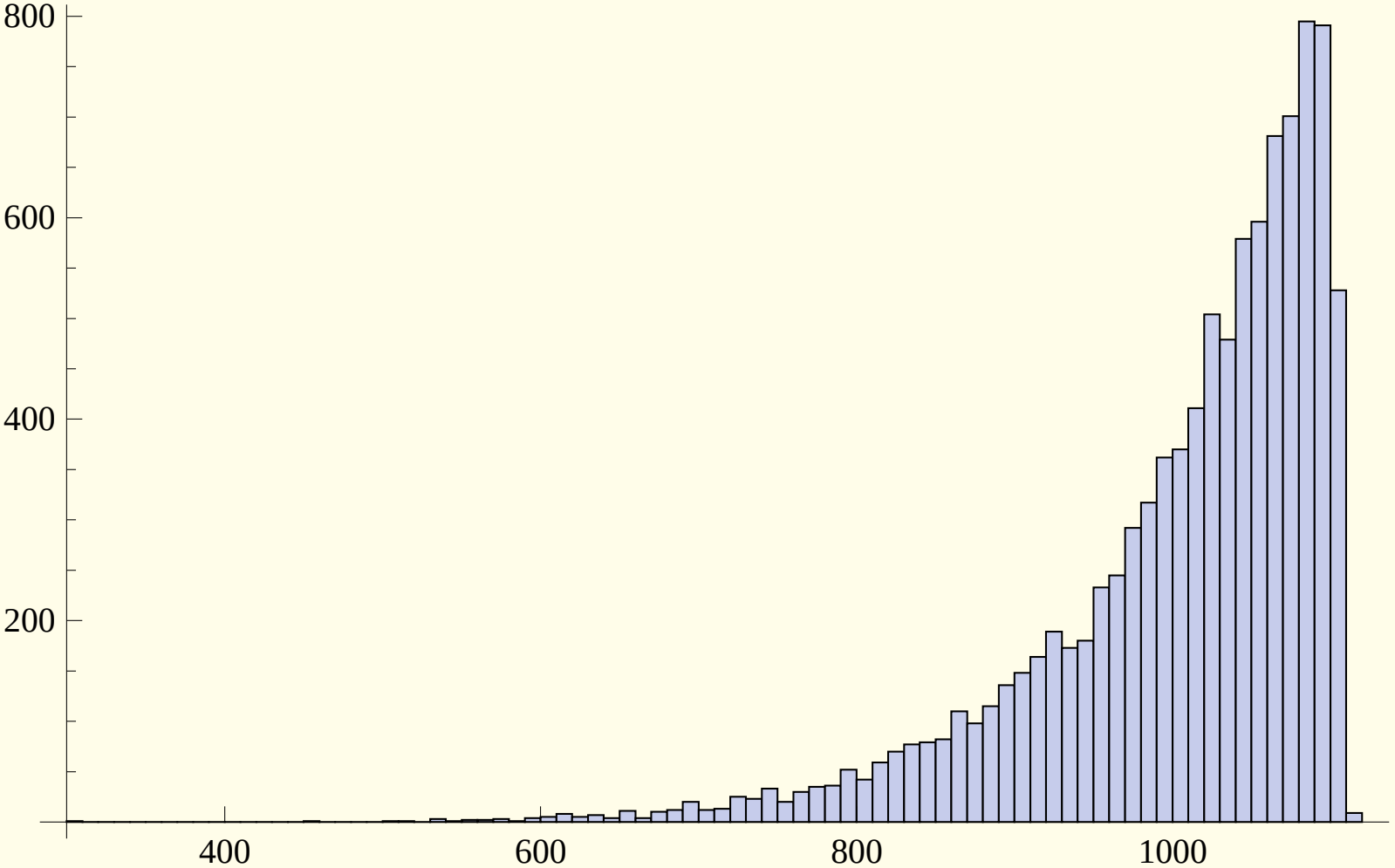
In this run $N = 1000$, $n = 10$



Mean = 995.5 Variance = 32752.

The distributions of T_2

In this run $N = 1000$, $n = 10$



Mean = 1006.3 Variance = 8342.

There is a lower limit to an estimator Variance?

T_2 should be used since its variance is smaller. Do we have a lower bound on the estimator's variance?

Yes! If you have a random sample from a continuous distribution $f(x|\theta)$, θ being our parameter and T an unbiased estimator of θ, then, under certain conditions, the variance of T has to be larger than:

$$Var(T) \geq \frac{1}{nE\left(-\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right)}$$

n is the sample size of the variable X.

Unbiased estimators that attain the minimum variance may exist.

Estimating the number of German tank

For the records, this table compares the estimated German production rate with the real one. The statistical analysis was rather accurate. Much less accurate the numbers provided by the intelligence.

Date of estimate	Statistical	Intelligence	German records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

The true production rate became available after the German surrender from the Speer Ministry.

Exercise!

Exercise

We have a random sample $X_1, X_2, \cdots X_n$ of iid variables distributed exponentially:

$$f(x|\lambda) = \lambda e^{\lambda x}$$

Verify that the following estimators:

$$T_1 = \bar{X}_n = \frac{\sum_i X_i}{n} \qquad T_2 = nM_n$$

where M_n is the maximum of the sample, are unbiased for the parameter $1/\lambda$. Which estimator would you use for estimating $1/\lambda$?

Exercise!

Exercise

In the problem of determining the number of German tanks we have found an unbiased estimator based on the maximum of the sample.
Could we have used the minimum of the sample to construct another estimator for N? Is it reasonable?
Work out the details.

Exercise!

Exercise

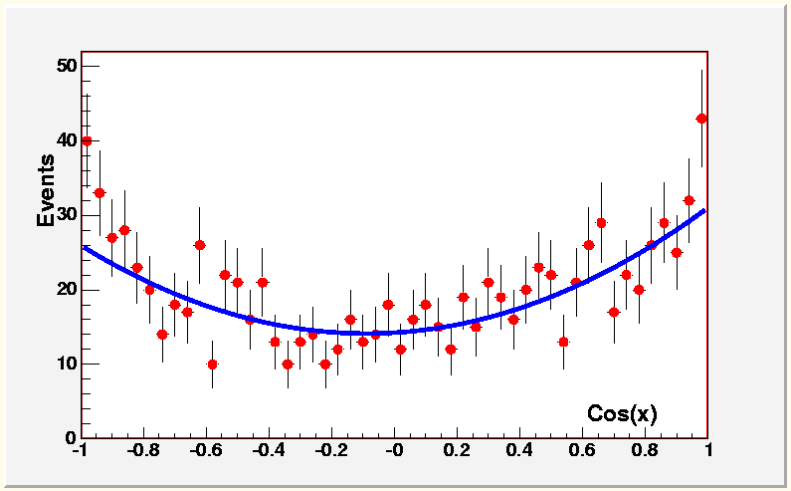
Suppose that you have found two unbiased estimators for a parameter θ. The two estimators U and V have the same Variance. We cannot decide which of the two to use on the ground of efficiency.

However we could do better and use a third estimator $W = (U + V)/2$.
Verify that even W is unbiased and show that the variance of W is smaller than that of U and V, even if the two estimators are not independent.

Estimation if Frequency theory
An example

The reaction: $e^+e^- \rightarrow \mu^+\mu^-$ has the cross section:

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4\ s}(1 + \cos^2\theta + \alpha_W \cos\theta)$$



We make n measurements of θ to estimate α_W. Before you rush to make a LS fit, lets think to different methods, may be easier.

Example III

- Let us call $x = \cos(\theta)$
- transform the x-sect to a *p.d.f.* (must be normalized to one):

$$f(x;\alpha) = \frac{3}{8}(1 + \alpha_w x + x^2) \qquad \text{Assume } x \in (-1, 1) \text{ with full eff.}$$

- Find the mean:

$$E(x) = \int_{-1}^{+1} x f(x) dx = \frac{\alpha_w}{4}$$

- Hence a good estimator for α_w is:

$$\boxed{\widehat{\alpha_w} = 4\overline{x}}$$

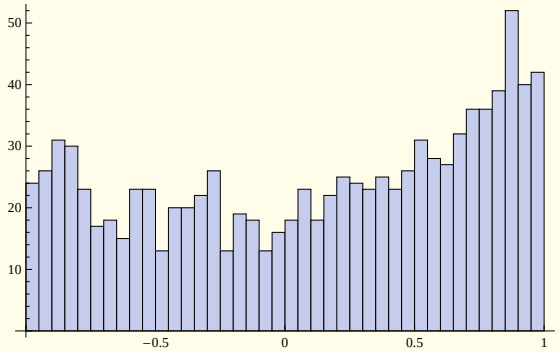
Example IV

This function of the data (statistics) is good to estimate α_W since:

- $$E(\widehat{\alpha}_W) = \alpha_w \qquad \text{NO bias}$$
- $$Var(\widehat{\alpha}_w) = \frac{16\,Var(X)}{n} \qquad \text{Consistent}$$
- *p.d.f.* of $\widehat{\alpha}_W$ is asymptotically Normal (CLT).
- These are direct consequences of the very general theorems of the Large Numbers and of the Central Limit.
- The method works even if the number of events is so low that the Least Square cannot be used.

Estimators in frequency theory

If your favorite method is still the Least Square, let us see how you would have done. We have $n = 1000$ events classified in 20 bins.



Let us call:

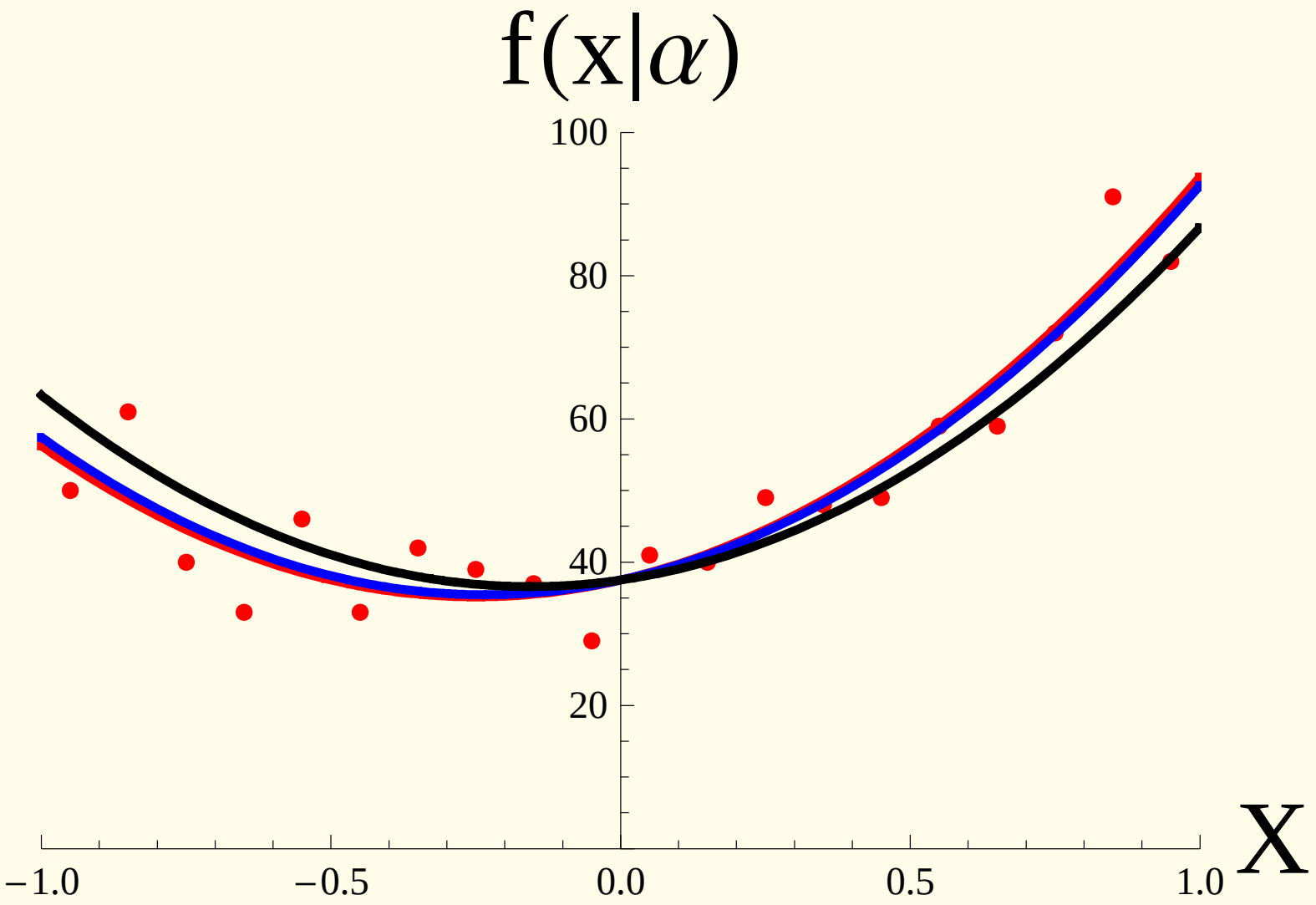
- r_i the event number in each bin.
- x_i the center of each bin
- w_i the weight.

The method consists in minimizing:

$$X^2 = \sum_i w_i \left(r_i - K\frac{3}{8}(1 + x_i^2 + \alpha x_i \right)^2$$

$$\frac{\partial X^2}{\partial \alpha} = 0 \quad \rightarrow \quad \boxed{\widehat{\alpha} = \frac{\sum_i w_i x_i + \sum_i w_i x_i^3 - \frac{8}{3K}\sum_i w_i r_i x_i}{\sum_i w_i x_i^2}}$$

Comparison of the two estimators



Estimators in frequency theory

The previous examples indicates the way to build an estimator:

- Find a *convenient* function $a(x)$ (in previous example $a(x) = x$)
- From the Law of Large Number we know that:

$$\overline{a} = \frac{\sum_i a(x_i)}{n} \rightarrow E(a) = h(\theta)$$

- Assume that the inverse exists:

$$\theta = h^{-1}\left(E(a)\right)$$

Estimators in frequency theory

An estimator for θ is

$$\boxed{\hat{\theta} = h^{-1}(\overline{a})}$$

This function is, by construction:

- $E(\hat{\theta}) = \theta$ (Asymptotically)
- $Var(\hat{\theta}) \propto \frac{1}{n}$ (Consistent)
- Asymptotically Normal.

Max-L in short

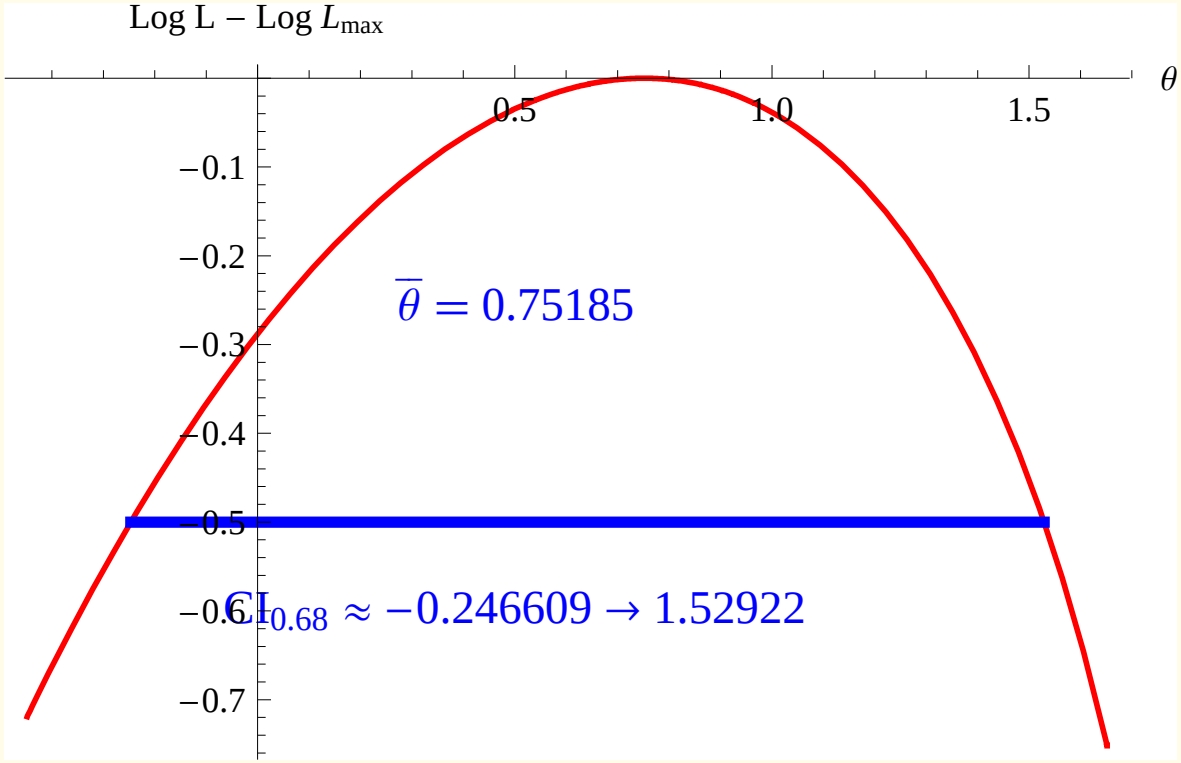
In the example discussed above, assume you have only 10 detected events, too few for LS method. We build the Likelihood (independent events):
The Max-L method (Fisher): among all possible values of a_W one should choose the one which maximizes the joint probability of the observations.

$$L = \prod_i f(x_i, a_W) \rightarrow LogL = \sum_i Log\ f(x_i, a_W)$$

L is NOT a probability: $X_i \rightarrow x_i$ not RV but data. L is a function of a_W.

Max-L analysis of $e^+e^- \to \mu^+\mu^-$ experiment

Ten events for $f \propto 1 + a_W x + x^2 \quad L = \prod f(x_i)$



If the *pdf* would be N, the line at $\boxed{Log\ L = Log\ L_{max} - 0.5}$ would cut a 0.683 *CI* for a_W (why?). In this case this is true only approximately. (later on exact CI.)

Comments on the inference in Frequency statistics

- These frequency methods are good in average.
- They all produce an estimator (a function of the data) whose value is an estimation of the parameter. Properties are: unbiasedness (asymptotically), consistency, asymptotic Normality.
  (Sufficient and efficient estimators, if exist, are also found.)
- The inference is performed only through the data, not on the parameter directly.
- No prior information on the parameter is needed. In fact we do not know how to use any previous information.

Comparison of Bayes and Frequency: Inference on a Parameter

Bayes and frequency methods produce similar results; let us analyze a very simple experiment:
A box contains N_R and N_W red and white balls...

Given	Frequency	Bayes
$N_R \geq 0$ $N_W \geq 0$	$P(R) = \frac{N_R}{N_W+N_R}$ *unknown*	$P(R) = \frac{1}{2}$ *Insufficient reason*
1 Ball R	$P(R) > 0$	$P(R) = 0.66$
2 Balls R+W	$P(R) = 0.5$	$P(R) = 0.5$
100 Balls, 40 R	$P(R) = 0.4 \pm 0.05$	$P(R) = 0.402$

We have assumed a flat prior for Bayes and used the mean value of the probability to quote the inference.

$$P(R) = \frac{N_R + 1}{N_{tot} + 2}$$

The Hypothesis test

After the inference on the parameter ("the fit" to determine α_W) we still miss two important steps:

- test of the hypothesis to "validate" the model: how well the data and the theory agree;
- determine the precision of the estimate. In the case of large statistics the CLT makes it easy: the distribution of estimators is Normal; The standard deviation define intervals with well defined probability content ($\pm 1\sigma \rightarrow 0.683$ *probability content*).
  In the case of low statistics experiments we will have to compute intervals of expected probability content with more refined techniques (Confidence Intervals)

Part 4

The methods of hypothesis test

- Two schemes: Significance test and Decision test.
- Significance test or Goodness of fit (Fisher).
- Decision test (Neyman and Pearson).
- Example (and Exercises!)

Hypothesis test (introduction)

One of the first step in the analysis of an experiment is to ask ourselves if the measurements are in agreement with the model that has motivated our work.

The methods of inference used to support or reject claims based on sample data are known as tests of hypothesis.

We will consider only those hypothesis that are able to make predictions on the way a random variable is distributed (statistical hypothesis).

The theories exposed here are bases on the frequency statistics and due largely to R.A. Fisher and J. Neyman.

Hypothesis test (general)

A test of hypothesis starts with the definition of the null hypothesis H_0, the model that will be subject to experimental test. It must be a statistical hypothesis since it is used to make statistical predictions.

The null hypothesis is assumed true and used to compute the distribution of the random variable on which we base the test.

Hence we can only disprove H_0. The hypothesis test methods work always against H_0.

(Even if the probability of the measurement, under H_0, is large,we cannot logically conclude that we have proved that H_0 is true.)

Hypothesis test (general) I

If the outcome of the measurement has a probability, under H_0, too small then

- we have encountered a rare event;
- the model is wrong.

At this point two main schools exist:

- Significance test
- Decision making test

We will start discussing the first, more used by *scientists working on the field* and then the more mathematically structured Acceptance test.

The Significance test

The first to use this test was K. Pearson. Fisher was the most remarkable founder of this way of testing.

Significance test: The lady and the tea I

A lady declares that, by testing a cup of tea made with milk she can discriminate whether the milk or the tea was first added to the cup.(This example is due to R.A.Fisher)

Significance test: The lady and the tea II

We will prepare four cups of tea prepared in one way and four in the other way and present them to the lady in random order. The lady is asked to divide the eight cups in two sets of four, accordingly with the way they are prepared.

The first step is to define the null hypothesis (H_o): the lady has no skill in choosing the correct preparation.

With this hypothesis we can make the analysis of the results of the experiment.

There are 70 ways of choosing 4 objects out of 8: $N = \binom{8}{4}$.

If the lady had no skill in choosing the correct preparation, the preparations would be indistinguishable to her and she would be correct (correctly assign all eight cups) once in seventy.

Significance test: The lady and the tea II

This means that if the experiment would be repeated very many times in uniform conditions, she would be correct with a frequency of 1/70; or the probability to be correct would be 1/70.

In the experiment the lady could be right with all the four cups, an event rather rare if the null hypothesis would be true. We say that the result is *significant*. This means that we have encountered a rare event, if H_0 is true; or the claim of the lady has some foundation...

Significance test: The lady and the tea I

In case the lady would have been correct three times and one wrong, what would be our conclusion?

We compute the possible number of times this event occurs in this way: the 3 correct cups are drawn from the four available, and the number of ways this occurs is $\begin{pmatrix} 4 \\ 3 \end{pmatrix} = 4$, while the wrong cup is drawn also from four and the number of ways is $\begin{pmatrix} 4 \\ 1 \end{pmatrix} = 4$. Thus the total number of ways is 16.
Similarly in the other cases.

Significance test: The lady and the tea I

If the lady would have had 3 successes and one failure we could not claim a statistical significant result. In fact the frequency in chance would be 16 in 70 (or 22%) for the obtained result.

Moreover in computing its significance we must take into account not only its frequency, but also the frequency of any better (against $H0$) result.

Significance test: The lady and the tea

The reason for considering the cases better than the one observed is clear with this example

Suppose that the case *3 correct 1 wrong* would have one chance in 70 and the case *4 correct and none wrong* would have a 16 chances in 70 and assume also that the lady would have guessed correctly only three cups.

The rare case *3 correct and 1 wrong* cannot not be judged significant even if its probability is low.

In repeated experiments the event *4 correct and none wrong* would have occurred more frequently by mere chance. Hence the chance of having obtained the *3 correct 1 wrong* cannot be judged significant without considering also the more extreme event *4 correct none wrong*.

The P-value

The probability of the data, under the hypothesis H_0 are thus defined as the probability of what has been observed and all possible more extreme outcomes, computed assuming H_0 is true:

$$\boxed{P - \text{value} = P\left(N \geq n \,|\, H_0\right)}$$

This quantity was introduced first by K. Pearson (1900) and it was then elaborated by R.A.Fisher (1922).

The P-value

The $P - value$ is a random variable.

If H_0 is true, then it is easy to prove that $P - value$ is distributed uniformly:

$$P - value \sim U(0, 1)$$

The concept of P-value is clearly very frequentistic, since it makes use of hypothetical, not observed results.

In Bayes statistics nothing similar exists: you cannot make significance test.

The utility of these methods should not hide its philosophical implications: we are making use of unmeasured quantities.

Significance test: The lady and the tea III

To avoid the noise produced by chance, it is convenient to disregard all experiments that have a large probability, for they do not reject the claim of the lady (H_0).

Fisher suggests, as a practical rule, that we should ignore results with a probability greater than five per cent (one in twenty). These results can be produced by chance effects and are irrelevant to assess H_0

By convention we agree that effects which can occur by chance once in seventy trials are *significant*. The value of 5% is convenient, but there nothing sacred in it.
It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. (Fisher)

Example

If, in the example of $e^+e^- \rightarrow \mu^+\mu^-$, we would have tested a theory without the parity violating term versus the data we would have obtained:

$$X^2_m = 81 \qquad Dof = 49 \qquad \text{hence P-value} = 3.4 \cdot 10^{-3}$$

The result is "very significant". Only 3 in a thousand experiments (performed in uniform conditions) would have performed worse. What should we do?
We have to go back to the experiment, check the apparatus, etc. Eventually we would be obliged to admit that the theory cannot account our results.
If these results would be obtained before the P-violation discovery (1954) we hardly could announce, with this significance only, a P-violation discovery.
The relevant significance level of the test depends on the issue: the announce of an important discovery needs an overwhelming evidence... and often is not enough!

Significance test

R.A. Fisher on the significance test:

- The significance of a result is used in day-to-day work in experimental research, in natural sciences. It is a tool used to distinguish real important effects from those produced by the chance.
- The significance test is a way to gain knowledge. Fisher: *inductive inference,* : from the empirical evidence to the general laws. The P-value assumes an epistemic value in Fisher's statistics.
- Significance test is applicable to single experiment.
- The $P - value$ is just the beginning to judge the hypothesis. No automatic rejection must be based on the $P - value$ only. The researcher must exercise his judgment and make decision on the specific issue.

Examples of Significance tests

The method should be already clear:

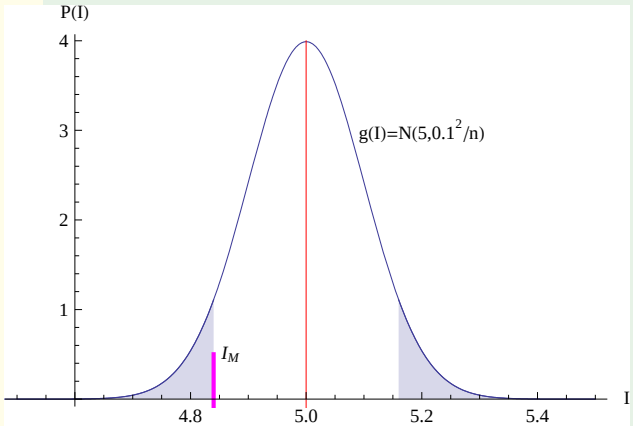
- a null hypothesis H_0 is identified. This must be capable of making quantitative predictions.
- a statistics t (a function of the observations only) is built; this is a Random Variable.
- the density of t is computed, $g(t \mid H_0)$.
- we compute the value of t from the observation: t_0
- the probability that t is even more extreme than what observed (the P-value) is, $\alpha = \int_{t_0}^{\infty} g(t)dt$

If α is larger than a a threshold (5% is the value suggested by Fisher) H_0 is accepted, without further analysis. If it is smaller, then H_0 is questioned and further analysis is needed (significant result)

The z-test

Example

An electric line should draw a current i_0= 5 A (H_0)
Measurements are done with an instrument having a Normal noise *noise* $\sim N(0, (0.1\ A)^2)$.



If (H_0) is true: $I \sim N\left(5, 0.1^2\right)$
$i_i = 4.58, 4.91, 4.93, 4.95\ A$
The statistics is $\bar{I} \sim N(5, 0.1^2/2)$
in our case $\bar{i} = 4.84$
The P-Value is the shadowed area.

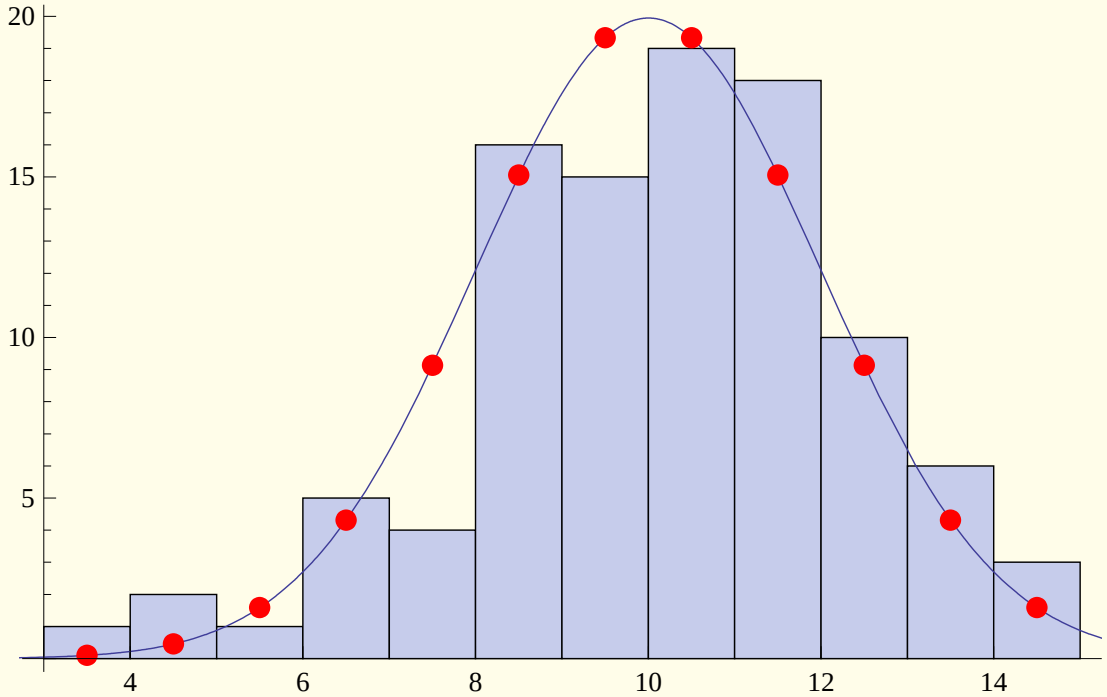
Better: construct the variable:

$$Z = \frac{\bar{I} - \mu_0}{\sigma_{\bar{I}}} \sim N(0,1) \rightarrow P\text{-}value = 2\int_{|-3.16|}^{\infty} N(0,1)dx = 0.16\%$$

The result is *significant*, H_0 is questioned. The experimenter has now the initiative.

Significance test: the Pearson test

This method is due to K. Pearson.
Let us consider a random variable X, measured in n independent experiments yielding: $x_1, x_2 \cdots x_n$. The histogram of the x_i is shown in:



$$\begin{cases} r_i & \text{events in bin } i \\ m_i & = p_i n \text{ expectation } H_0 \\ n & \text{total events} \\ k & \text{number of bins} \end{cases}$$

(The bis size can also be different bin to bin, it works also for two and higher dimension histograms)

The Pearson test

H_0 specifies the distribution of the X:

$$X \sim f(x \mid \theta_0) \quad \rightarrow \quad p_i = \int_{X \in bin\ i} f(X|\theta_0)dx$$

The statistics is:

$$u = \sum_{i=1}^{k} \frac{(r_i - n \cdot p_i)^2}{n \cdot p_i}$$

To compute the distribution of u we have to know the distribution of r_i.

$$r_i \sim B(p_i; n) \xrightarrow[n \to \infty]{} N(p_i n, p_i(1 - p_i)n) \approx N(p_i n, p_i n)$$

The Pearson test

Assume that this approximation holds (n large), then the variable:

$$z_i = \frac{r_i - n \cdot p_i}{\sqrt{n \cdot p_i}} \sim N(0,1)$$

Hence

$$u = \sum_{i=1,k} \frac{(r_i - m_i)^2}{m_i} = \sum_{i=1,k} z_i^2 \sim \chi^2_{k-1} \tag{1}$$

The distribution is independent of H_0

The Pearson test

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. For instance we could compute r_k as

$$r_k = n - \sum_{i=1}^{k-1} r_i.$$

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.

The tests that need no specifications of the parent distribution are called *distribution free* or *non-parametric*

These formulas are correct but the argument that we have followed is not mathematically sound.

When Pearson is not so simple...

Simple hypothesis and occurs seldom in practice. In most of the cases some parameters have to be estimated from the sample. If the value of the parameter is determined by the ensemble $\{r_i\}$ then is possible to prove (Fisher-Cramer) that the effective number of *DoF* is decreased by the number of parameters determined from the data.

It often happens 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 variable u is not distributed as a χ^2_{k-1}. u density has to be computed numerically often with Monte Carlo methods. The distribution depends not only on the number of classes but also on the model; the test is not any longer *distribution free*.

Pearson test example: Mendel peas

Modern genetics begins with the work of Gregor Mendel, an Austrian monk whose breeding experiments with garden peas led him to formulate the basic laws of heredity.

Mendel published his findings in 1866, but his discoveries were ignored till 1900 when a number of researchers independently rediscovered Mendel's work and grasped its significance.

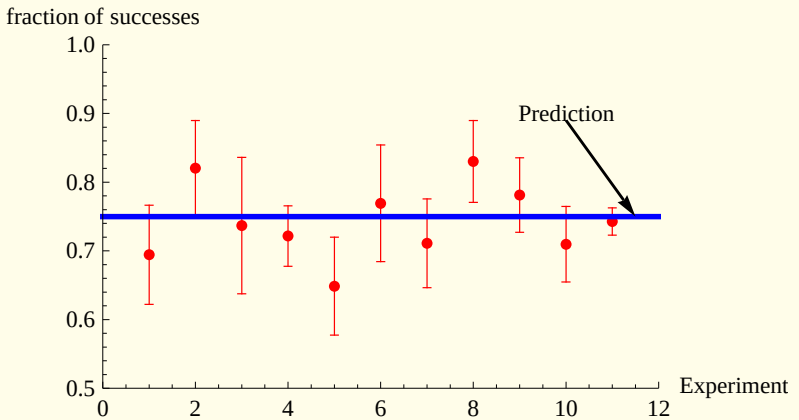
In a famous experiment, Mendel cross-pollinated smooth yellow pea plants with wrinkly green peas. Mendel then counted the number of times he got green and yellow peas. He expected $p = 0.75$.

So $H_0 : p = 0.75$ and we will try to disprove it.

Pearson test example: Mendel peas

The data are summarized in the following table. For each experiment we report: the number of observations, the number of yellow peas (successes), u (assuming H_0) and the $P - values$.

Experiment	N.O tests	yellow	u	$P - value$
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



No result is significant, thus no reason to reject H_0.

Rutherford's α

In 1912 Lord Rutherford (Phil. Mag. 1912) was investigating whether the decay process of a radioactive atom is independent from the status of all the other atoms.

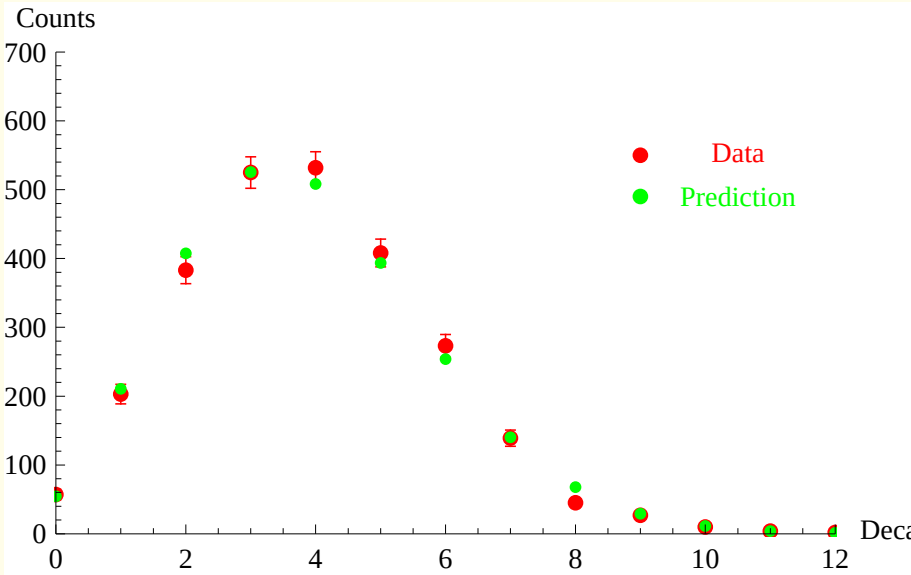
If such is the case, the number of events in a fixed time interval follows the Poisson statistics. However the mean is not known and has to be determined from the data. The hypothesis is not *simple*.

The experiment consisted in measuring the number of alpha particles emitted from a radioactive source in time intervals 7.5 *sec* long. The number of measured intervals was $n = 2608$ (see the next Table).

The hypothesis to test is (H_0): the population is Poisson distributed with (unknown) average μ.

Rutherford's α, results

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	273	253	1.44
7	139	140.325	0.0125
8	45	67.882	7.7132
9	27	29.189	0.1642
10 11 12 Total	10 4 2 2608	17.075 2608	0.0677 12.8849



From columns 1 and 2: $\overline{x} = \dfrac{\sum_i i \cdot N_i}{n} = 3.870$ The P-value is:

$P(u = 12.885) = \displaystyle\int_{12.885}^{\infty} \chi^2_9(u)\, du = 0.17$ not significant: no reason to discard the Poisson distribution.

More on the P-value

There are common misunderstandings in the interpretation of the $P - value$.

- It is not correct that large values of $P - value$ is evidence for H_0. It is true that large values of $P - value$ means little or no evidence against H_0.
- It is not correct that small values of $P - value$ is evidence of important effects. It is true that small values of the $P - value$ is evidence against H_0.
- $P - value$ is not the probability that H_0 is true. In frequency statistics there is no such a thing as the probability of an hypothesis.
- $P - value$ is not the probability of the alternative hypothesis is true. In significance test there is NO alternative hypothesis.

Significance tests

The significance test tool box is rich!:
Z-test
T-test
χ^2 (Pearson)
Kolmogorov-Smirnov
Cramer-Smirnov-von Mieses
Run test
Sign test
Wilcoxon rank test
....
Some are specific test some are of much wider applicability. Most are *parametric*. You do not need to invent your own test, just take off shelf the most appropriate. None is always superior to the others. If two tests are independent they can be combined (Pearson-Run).

The Acceptance test, Discussion

Unfortunately physicists mainly have created a “mixed” method where the frame of Neyman-Pearson (see next) is implemented: H_0, H_A and size are defined and the statistics computed with its $P - values$, which is then interpreted both as type I error rate and as evidence against H_0. A sharp decision is then made based on the $P - value$: if the $P - value$ is below an assigned threshold (a sort of size of the test) H_0 is rejected. Very often the tests follow formally the acceptance-rejection rules but then the $P - value$ is computed. In terms of accept-reject hypothesis tests, the $P - value$ has no relevance and should not be quoted.

The Decision making test

This test of hypothesis was elaborated by Neyman and Pearson, originally as an attempt to improve on the Significance test. Fisher strongly disagree with the principles governing the test claiming that those methods were unsuited to scientific research.
The decision test is what is now called “hypothesis test” in statistical mathematics and has totally replaced the Significance test.

The Decision making test, Introduction

In addition to the null hypothesis (H_0) we have an alternate hypothesis (H_A) (*"the rational human mind did not discard a hypothesis unless it could conceive at least a plausible alternative hypothesis"* (Pearson). (Often we have only one model to test...)

A statistics $t(X)$ is defined and its distribution $g(t \mid H_0)$ is computed under the assumption that H_0 is true. A Critical region w_α, (α test's size) is defined with the property

$$\boxed{P(t \in w_\alpha \mid H_0) = \alpha}$$

If $t \in w_\alpha$ we accept H_A else we accept H_0.

Brief considerations on the two approaches

Fisher's view of *inductive inference* based on significance test and focused on the rejection of the null hypothesis as a way of gaining knowledge, is completely dismissed by Neyman and Pearson.

They introduce instead the concept of *rules for making decision between two hypothesis*: the *inductive behavior*: *"The term inductive behavior means simply the habit of humans and other animals (Pavlov's dogs etc) to adjust their behavior to noticed frequencies of events, so to avoid undesirable consequences."* (Neyman)

Neyman-Pearson theory is *non evidential but behavioral*, as discussed by Fisher: *"No particular thought is given to each case as it arises, no tester's capacity for learning exercised"* (Fisher).

How the decision making test is formulated

The test may fail in assessing H_0 or H_A. By construction there is a probability α of wrongly rejecting H_0. We distinguish two type 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_A is true (probability β).

Test are prepared by fixing α (the losses) and searching a test which minimizes β.

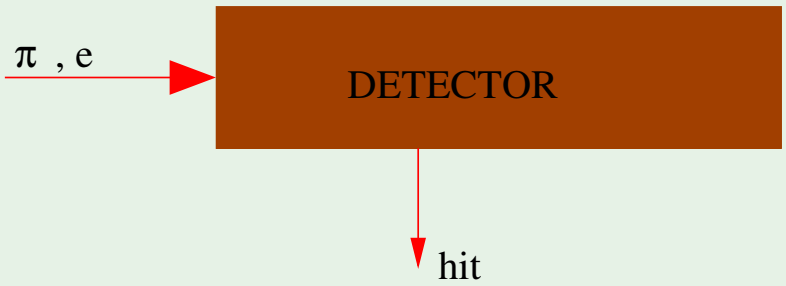
	H_0 is true	H_A is true
H_0 is accepted	correct	Type II error (β)
H_A is accepted	Type I error (α)	correct

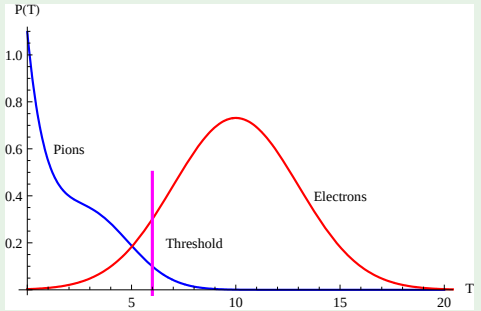
It is clear that we would like a test which could allow both α and β equal to zero or as small as possible.

Errors I,II in Physics

Example

In a beam of e, π we have a detector to tag electrons. The beam contains also pions.





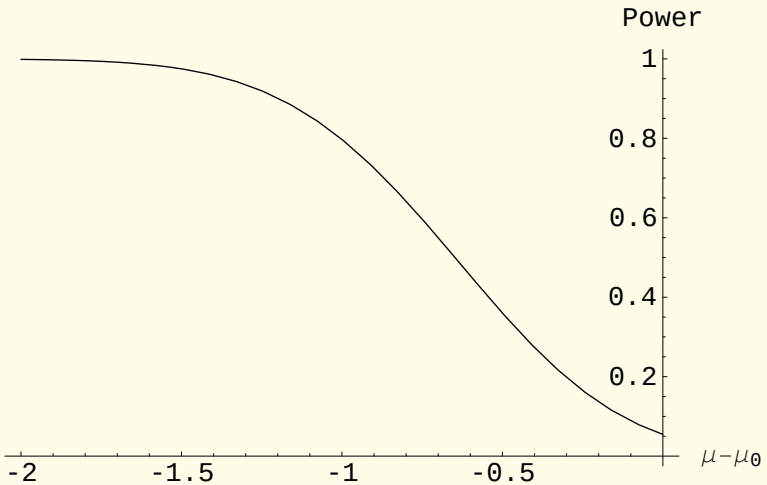
H_0: the hit is an electron,
H_A: the hit is a pion
Type I Error: $P(T \in w_\alpha \mid H_0) = \alpha$ is the loss of electrons,
Type II error: $P(T \in \bar{w}_\alpha \mid H_A) = \beta$ is the contamination of pions in the electron sample.

How the acceptance-rejection test is formulated

The quantity:

$$P(X \in w \mid H_1) = 1 - \beta \tag{2}$$

is called the *power* of the test and represents the probability of correctly classify H_1.
The power of the test depends in general on H_1. As an example consider the test on the value of a parameter θ: H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$. figure)



If $\theta_1 = \theta_0$, the limit where $H_0 = H_1$ we have: $1 - \beta = \alpha$.
The power will usually increase with the distance between θ_1 and θ_0.

The LR test for simple hypothesis is UMP: the most powerful on all the parameter space.

The acceptance test in the case of simple hypothesis

The method is the following:
Decide the size of the test α. This will set the *critical region* w_α.

$$\int_{w_\alpha} f_0(x)\, dx = \alpha$$

$$1-\beta \;=\; power \;=\; \int_{w_\alpha} f_1(x)\, dx = \int_{w_\alpha} \frac{f_1}{f_0} f_0\, dx = E_{w_\alpha}\left(\frac{f_1}{f_0} \mid H_0\right)$$

The power is maximized if in all w_α we maximize $\frac{f_1}{f_0}$. Therefore we have to choose w_α such that:

$$t = \frac{f_1}{f_0} \geq t_\alpha \qquad \text{in all } w_\alpha$$

Hence H_0 is rejected if $t \geq t_\alpha$

t_α defines w_α and must fulfill: $P(t \in w_\alpha \mid H_0) = \alpha$

The Decision test: Simple hypothesis I

The $\Delta S = 1/2$ predicts that Ξ^0 proper lifetime is twice the known Ξ^- lifetime. We have made measurements and we want to check the theoretical predictions.
Measured: $\vec{t} = \{t_1, t_2 \cdots t_n\}$. Let us call τ_0 and τ_1 the predicted lifetimes. The hypothesis are simple, the two lifetimes are known.

$$H_0 : \tau = \tau_0 \qquad H_A : \tau = \tau_1$$

The Decision test: Simple hypothesis I

The likelihood ratio is:

$$L = \frac{L(\vec{t} \mid H1)}{L(\vec{t} \mid H0)} = \frac{\prod_i \frac{1}{\tau_1} e^{-t_i/\tau_1}}{\prod_i \frac{1}{\tau_0} e^{-t_i/\tau_0}} = \left(\frac{\tau_0}{\tau_1}\right)^n exp\left(-n\bar{t}(1/\tau_1 - 1/\tau_0)\right)$$

$L > c_\alpha$ is the condition to reject H_0 in favor of H_1

(c_α depends on the size of the test.) Instead of solving directly we take the logarithm:

$$n(\log(\tau_0) - log(\tau_1)) - n\bar{t}(1/\tau_1 - 1/\tau_0) > \log(c_\alpha)$$

The acceptance test: example I

This can be rewritten as:

$$\begin{aligned} \bar{t} &> T_n^\alpha \qquad \text{if } \tau_1 > \tau_0 \\ \bar{t} &< T_n^\alpha \qquad \text{if } \tau_1 < \tau_0 \end{aligned}$$

There is no need to solve for T_n^α. T_n^α is found by solving:

$$P(\bar{t} > T_n^\alpha \mid H_0) = \alpha$$

If $\tau_1 > \tau_0$, else the inequality must be reversed.
To solve the previous equation we have to know the density of $\bar{t}$.

The acceptance test: example

We will consider only the simple case of $n = 1$; in this case $\bar{t} = t$ and the density is the original exponential density. Let us consider only the case $\tau_1 > \tau_0$, the other case is analogous. T_1^α is computed by solving:

$$\alpha = \int_{T_1^\alpha}^{\infty} f(t \mid H_0)dt = e^{-T_1^\alpha/\tau_0} \qquad \text{hence} \qquad T_1^\alpha = -\tau_0 \log \alpha$$

If $t > -\tau_0 \log \alpha$ we reject H_0

The Power of the test is:

$$1 - \beta = \int_{T_1^\alpha}^{\infty} f(t \mid H_1)dt = e^{-T_1^\alpha/\tau_1} = e^{-\tau_0 \log \alpha/\tau_1} = (\alpha)^{\tau_0/\tau_1}$$

The acceptance test: example I

Exercise

Perform the test in the limit n very large, using the CLT.

Exercise

Compute the density of $\bar{t}$ (use MGF). Then make the exact test.

The exponential density that we have discussed in the previous examples has sufficient estimators. The likelihood function is function of any sufficient estimator. This is why we could greatly simplify the calculation of the critical region.

Symmetry of the two hypotheses

There is an asymmetry between H_0 and H_1. We can understand it from the Mendel example. Mendel had two hypothesis to test (in the decision framework):

- H_A: p = 0.75 This is what Mendel wanted to prove.
- H_0: p = 0.5. This is the alternative hypothesis, no preference in the offsprings of peas.

H_0 is standard science. H_A is what the scientist believes. Type I and II error probability mean:

- α wrongly accept H_A. We must be critical in accepting new theories. α is small (usually $\leq$ 0.05) since a false theory should not be supported, funded etc.
- β wrongly reject H_A (the discovery). Of course the researcher wants β to be as small as possible.

Symmetry of the two hypotheses

The damage of an error of type I could be large to the community: a wrong theory/model becomes part of the body of science.

On the contrary, if an error of type II occurs, a good theory has been rejected with frustration of the experimenter; but also a good theory would be rediscovered later with little or no impact on the community.

Even worse are the consequences if the test did not regard pure science but aimed to asses the effects of a new drug or a new fertilizer proposed by a chemical industry as can be easily guessed.

The work of Mendel is an example of how a scientist must work; on the contrary the symmetric Lysenko shameful story should teach us how bad can be to behave differently.

The Decision test, NON simple hypothesis I

The case where H_0 or H_1 or both are composite is more frequent. Typically this happens when testing a theory or model with free parameters that have to be determined from the data.

We will consider only tests on parameters that specify a distribution function.

In general no UMP test exists.

The acceptance test: example I

Assume that the distribution has the form $f(X;\vec{\theta})$. The likelihood function, for independent measurements is:

$$L = \prod_i f(X_i;\vec{\theta})$$

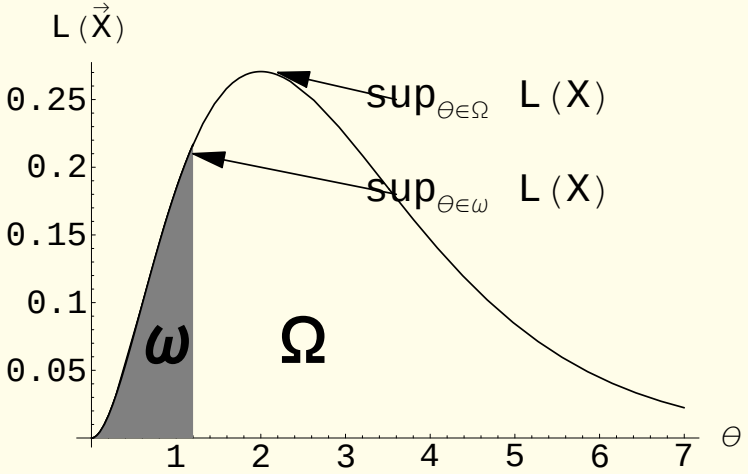
Let Ω be the space of the parameters. The hypothesis H_0 put restrictions on some of the $\vec{\theta}$ for instance:

$$\begin{cases} H_0 &= \vec{\theta} \in \omega \subset \Omega \\ H_A &= \vec{\theta} \in \Omega \end{cases}$$

The acceptance test: example I

This means that H_0 fixes the value of a part of the parameters while H_1 does not put restriction on the parameters.

Our statistics id the likelihood ratio:



$$\lambda = \frac{\sup_{\theta \in \omega} L(\vec{X} \mid \theta)}{\sup_{\theta \in \Omega} L(\vec{X} \mid \theta)}$$

$$0 \leq \lambda \leq 1$$

The test statistics that we will use consists in finding the maximum of the likelihood function in the parameter space allowed by the two hypothesis and then making the ratio.

The acceptance test: example I

The difficult part of the problem is the calculation of the critical region. For this we have to compute the density of $\lambda(H_0)$.

The usual procedure is to consider the asymptotic distribution. Wilks proved that, for $n \to \infty$ the limiting distribution of $-2\ ln(t)$ is, under H_0,:

$$-2\ ln(\lambda) \sim \chi^2_r$$

where r is the number of parameters fixed by H_0.

The acceptance test: example

An instrument was sent to the firm for upgrade. Assume the readings are $N(\mu, \sigma^2)$ and σ is our measure for the instrument precision.
Before the improvement $\sigma = \sigma_0$, after $\sigma = \sigma_1$. We want to test if the (costly) operation improved the instrument. $H_0 : \sigma_1 = \sigma_0$
$H_1 : \sigma_1 < \sigma_0$
we perform the LR test. H_1 is composite.

$$LR = \lambda = \frac{L_1(\vec{X})}{L_0(\vec{X})} = \left(\frac{\sigma_0}{\sigma_1}\right)^n \frac{exp(-\sum(x_i - \mu)^2/(2\sigma_1^2))}{exp(-\sum(x_i - \mu)^2/(2\sigma_0^2))}$$

The condition to reject H_0 is ($s^2 = \sum(x_i - \mu)^2$):

$$\log \lambda = n \log(\frac{\sigma_0}{\sigma_1}) + ns^2 \left(\frac{1}{\sigma_0} - \frac{1}{\sigma_1}\right) \geq \log c_\alpha$$

The acceptance test: example

The complicate expression is re-written as:

$s^2 \leq K_\alpha$ The inequality sign is reverse since $\sigma_1 \leq \sigma_0$

Since $\frac{s^2}{\sigma_0^2} \sim \chi^2_n$, H_0 is rejected if $\frac{s^2}{\sigma_0^2} \leq X^n_\alpha$, where X^n_α is the α-quantile of the χ^2_n distribution.

The power of the test is:

$$Power = P(s^2/\sigma_0^2 \leq K_\alpha \mid H_1) = P(\underbrace{s^2/\sigma_1^2}_{\sim\chi^2_n} \leq (\sigma_0^2/\sigma_1^2)K_\alpha \mid H_1)$$

Why Bayesian cannot do Goodness of Fit test

Bayes formula:

$$P(H_0 \mid Data) = \frac{P(Data \mid H_0) \cdot P(H_0)}{P(Data)}$$

If H_0 is a parameter the method works. In fact we can write: $P(Data) = \sum_i P(Data \mid H_i)P(H_i)$. H_i is just a possible value of the parameter.
If H_0 is an hypothesis we cannot normalize to anything meaningful, we cannot list ALL possible alternatives.
Bayesian can only compare two hypothesis, by comparing the relative probabilities:

$$\frac{P(H_0 \mid Data)}{P(H_1 \mid Data)} = \underbrace{\frac{P(Data \mid H_0)P(H_0)}{P(Data \mid H_1)P(H_1)}}_{Bayesfactor}$$

Why Bayesian cannot do Goodness of Fit test

When comparing Data to an hypothesis, both Freqentistic and Bayesian would like to compute

$$P(Data \mid H_0)$$

This is always zero, since it is a density!
Frequentistic solve the problem by identifying a *critical volume* w_c of the sample space where to integrate the density.
In the case of X^2 (Pearson) test:

$$w_c = X^2 > X^2_{obs}.$$

This volume is *on non observed data.*
Bayesian cannot do it, since all inference must be done on observed data.

Is all what we need to make a decision?

We have discussed how to perform an hypothesis test and chose H_0 or H_1 according to the value of a statistics.

	H_0 is true	H_1 is true
Chose H_0	OK	α
Chose H_1	β	OK

α and β are the probabilities to make the wrong choice. We have elaborated the procedures to optimize the correct choice, but we do not have specified the costs of a wrong choice.
This aspect may be a relevant, in practical issues.
Including the gain/loss concept into the problem will shift our methods from the usual *information theoretical* approach to the *decision theory* approach.
Let us see how it started and how it works in a simple example.

Pascal Wager (le pari de Pascal)

Pascal introduces in the great debate whether God exists or not, a new concept: the gain and loss. In his *Pensees* Pascal dismantles the notion that we can thrust reason in matters of religion. Since we are not sure if God exists, then we have to rely on probability, as in a game of chances, to compute pros and cons.

	God Exists	God does not exists
Pious life	$+\infty$	-N
Libertine life	$-\infty$	+N

In this simple version (see the Pensees, note 233, “Infinie- rien” to understand the full argument), the decision is living as if God exists is obvious, since you have nothing to lose and everything to gain.

Decision theory, by a simple example I

A detector runs with a maximum daily rate Θ_M but may drift from its optimal setting. Each morning we have to decide if we reset the apparatus or not. The decision is made from the previous day number of events t. The adjustment (d_2) takes a fraction p of the day. If we do not adjust (d_1) the efficiency is lower and we lose $\Theta_M - \theta$.
We want to make that decision which minimizes the losses.

Decision	Loss function
d_1	$\Theta_M - \theta$
d_2	$p\,\Theta_m$

We reset the system if:

$$L(d_2) \leq L(d_1) \rightarrow \boxed{\theta \leq \Theta_m(1-p)}$$

.... but θ is unknown...

Decision theory, by example II

We need to know θ. The only way is to use Bayes approach: In Bayes school θ is a "random variable" with $\pi(\theta)$ its prior. from the observations we compute the posterior for θ and we average over θ

$$\textit{posterior loss is:} \quad E_\theta \left[L(\theta, d_i) \right]$$

The Bayes rule is the one giving the smallest posterior loss. Assume that the distribution of t is a Poisson:

$$P(t|\theta) = \frac{\theta^t}{t!} e^{-\theta}$$

and also assume that the prior for θ is uniform up to the maximum production rate Θ_M:

$$\pi(\theta) = \frac{1}{\Theta_M}$$

Decision theory, by examples III

The posterior distribution is:

$$P(\theta|t)d\theta \propto \frac{1}{\Theta_M}\frac{\theta^t}{t!}e^{-\theta}d\theta$$

With the change of variables $u = 2\theta$ ($0 \leq u \leq 2\Theta_M$) we obtain:

$$P(u|t)dt \propto \left(\frac{u}{2}\right)^t e^{-u/2}du = \chi^2_{2(1+t)}$$

The *posterior losses* in case of d_1 are:

$$\begin{aligned} E_\theta(L(d_1|t) &= \int_0^{\Theta_M} (\Theta_M - \theta)P(\theta|t)\ d\theta \\ &= \Theta_M \int_0^{2\Theta_M} P(u|t)du - \frac{1}{2}\int_0^{2\Theta_M} uP(u|t)du \end{aligned}$$

Decision theory, by examples IV

$$E_\theta(L(d_1|t) = \Theta_M P(\chi^2_{2t+2} < 2\Theta_m) - (1+t)P(\chi^2_{2t+2} < 2\Theta_m)$$

This simplify if $\Theta_M \gg 2(1+t)$:

$$E_\theta(L(d_1|t) = \Theta_M - (1+t)$$

In case of d_2 calculation are simple since the loss does not depend of θ :

$$E_\theta(L(d_2|t) = p\,\Theta_M$$

Thus the Bayesian rule is: chose d_1 if $\boxed{E_\theta(L(d_1|t) \leq E_\theta(L(d_2|t)}$,

If Θ_M is large ($\gg t+1$) the rule simplify:

Reset the system if $t \leq (1-p)\Theta_M - 1$ else continue.

Outline

Part 5

- Confidence Intervals.
- Example (and Exercises!)

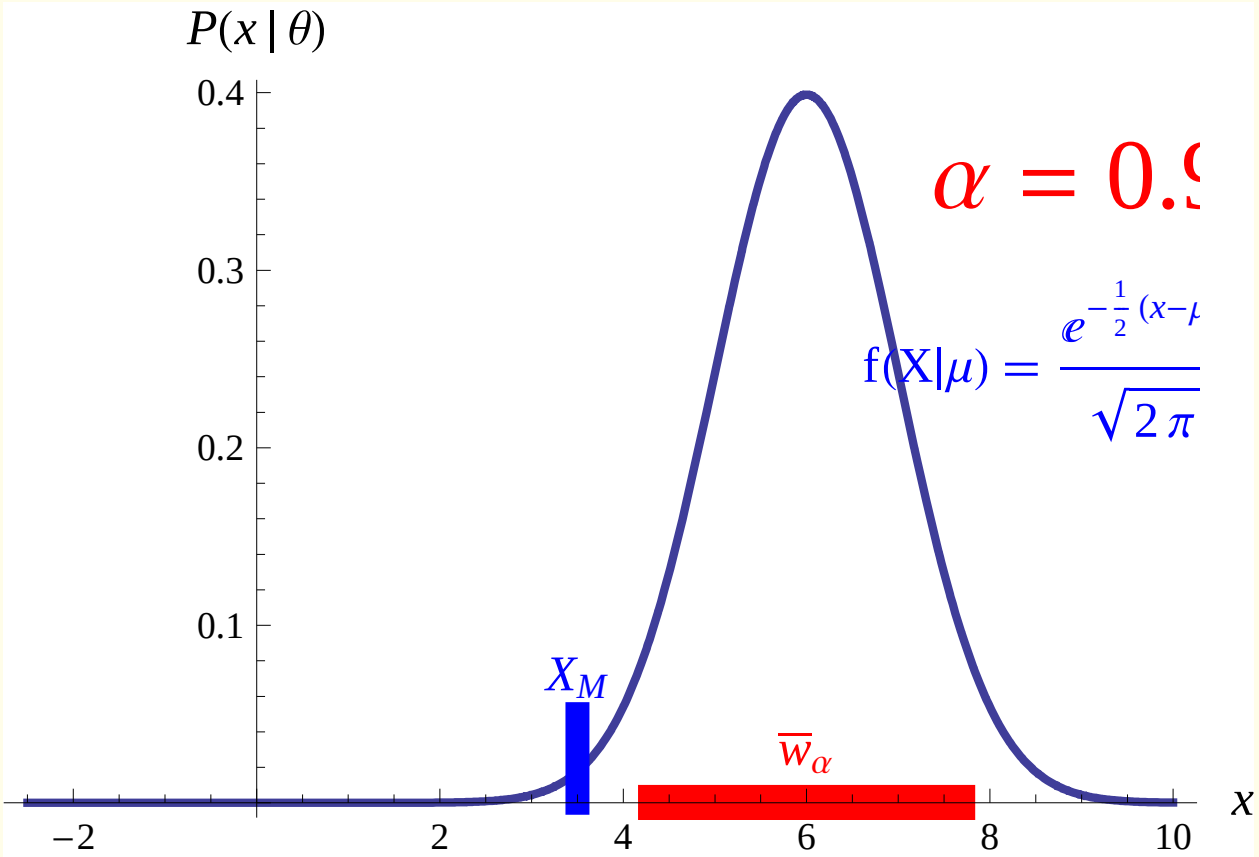
Definition of Confidence Interval

The problem of estimating a parameter by intervals was faced and solved in a general way by Neyman in 1935 and 1937 with two classical papers.

It is a very frequentistic concept: the interval estimation. Estimate the parameter by an interval which indicates the precision of the measurement.

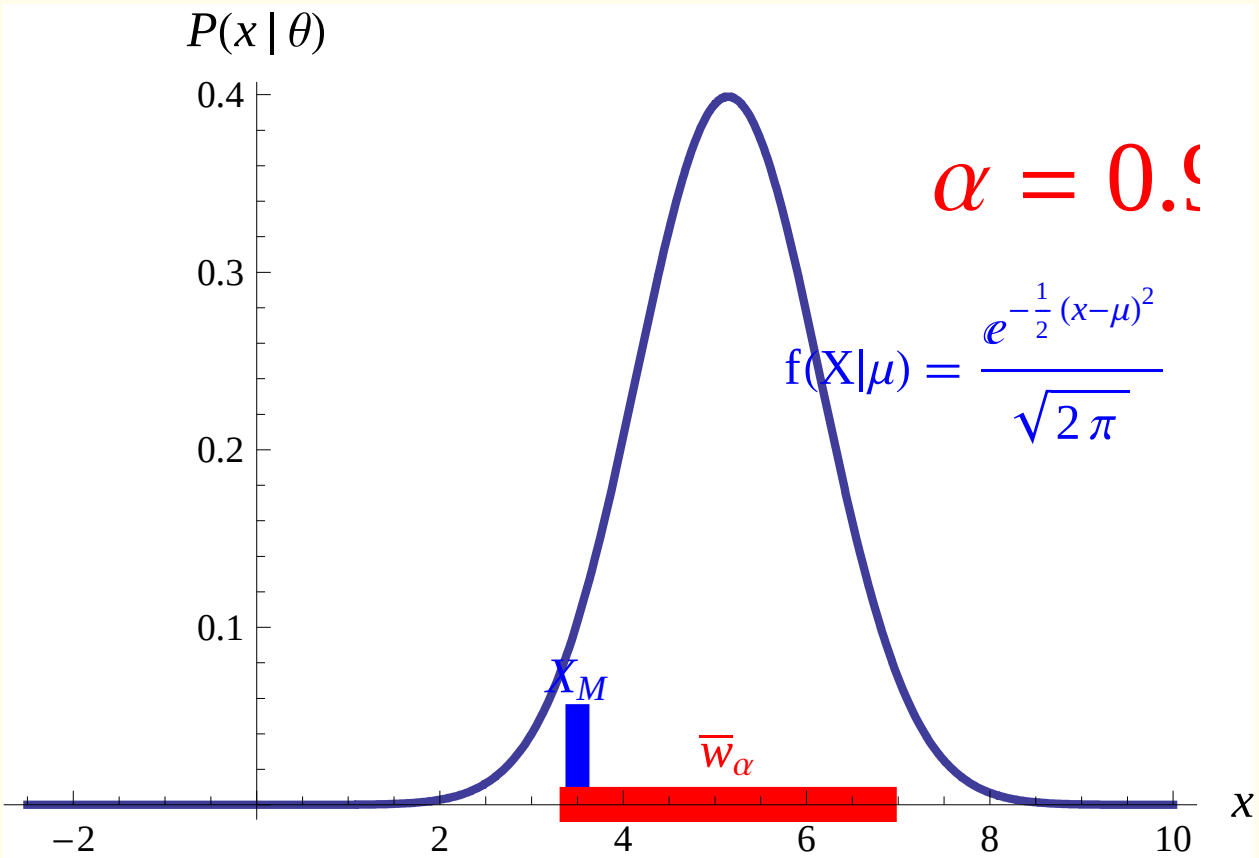
Definition of Confidence Interval

A CI is built with a sequence of hypothesis test: A measurement X_M is tested against $H_0 : \mu = \mu_0$, for all conceivable μ



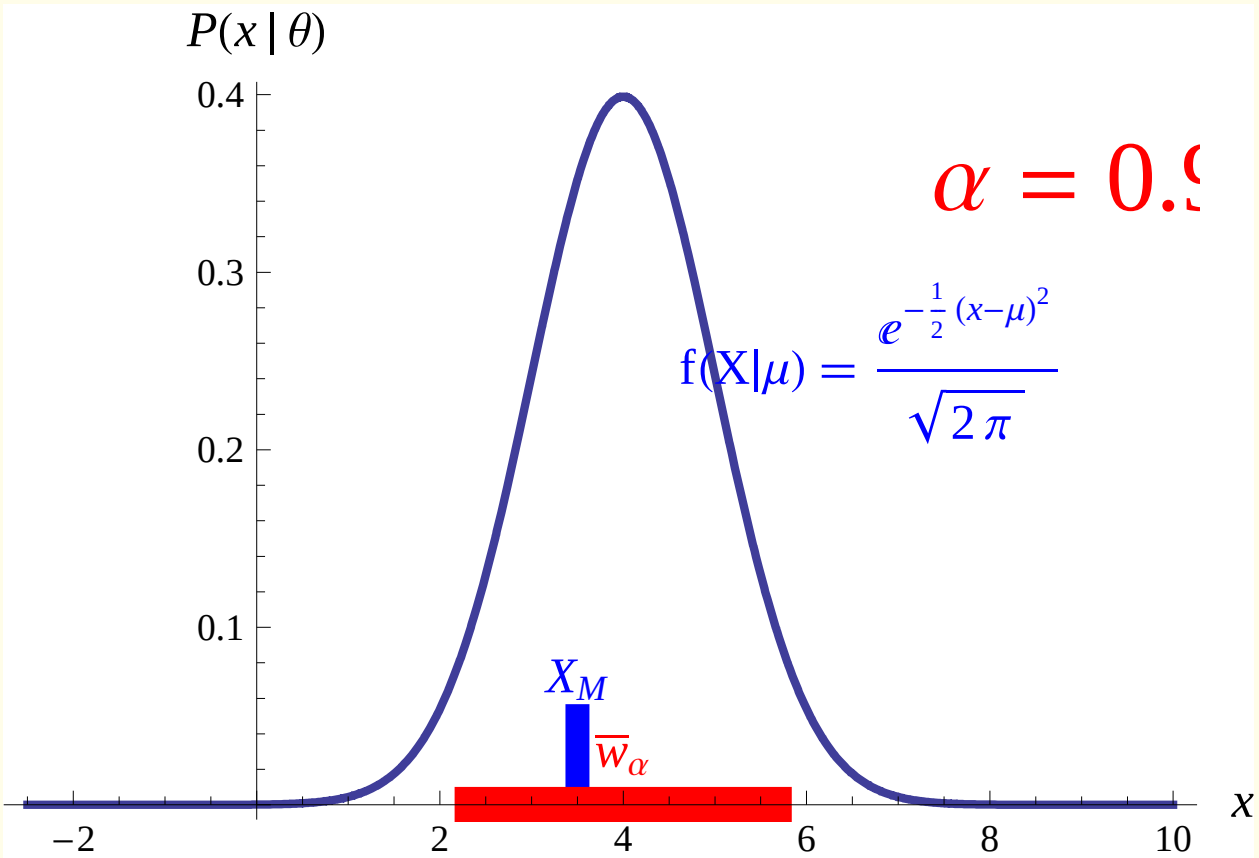
Definition of Confidence Interval

A CI is built with a sequence of hypothesis test: A measurement X_M is tested against $H_0 : \mu = \mu_0$, for all conceivable μ



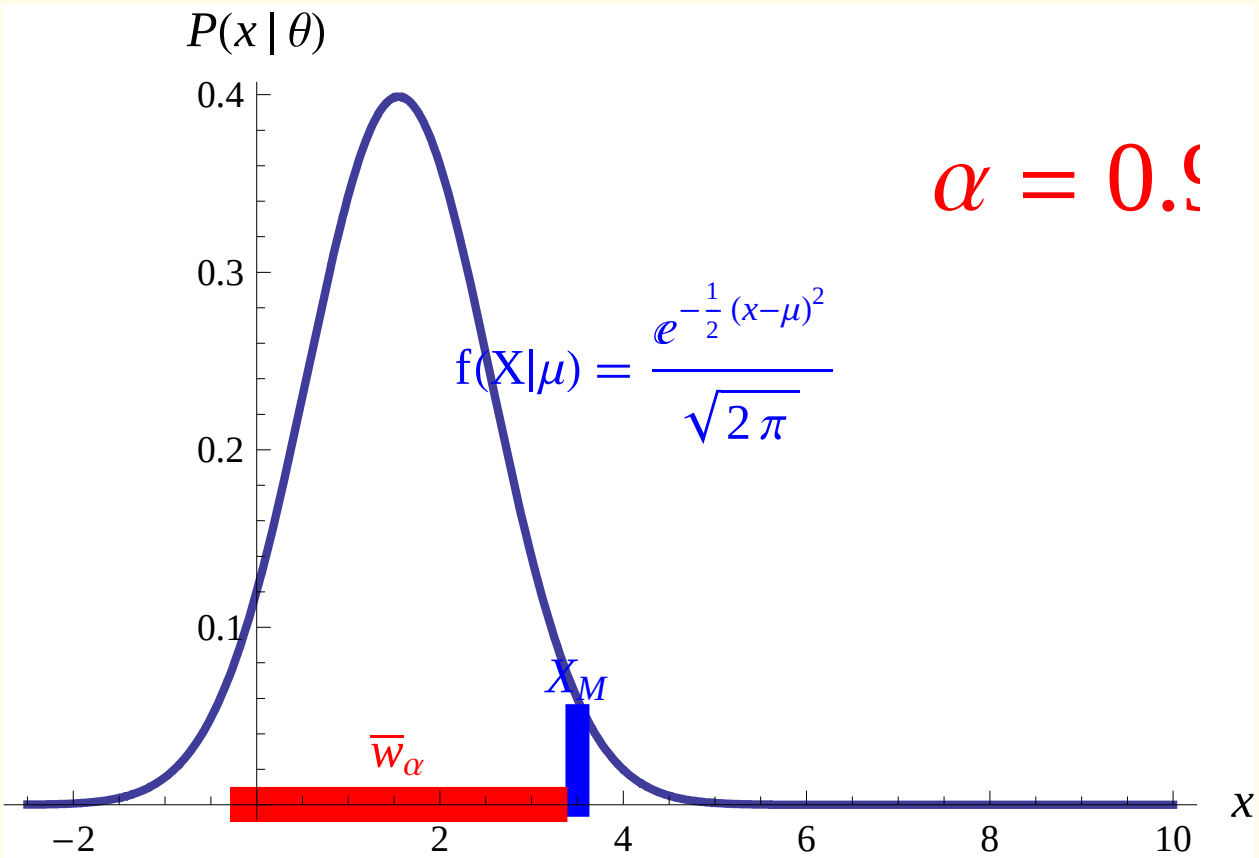
Definition of Confidence Interval

A CI is built with a sequence of hypothesis test: A measurement X_M is tested against $H_0 : \mu = \mu_0$, for all conceivable μ



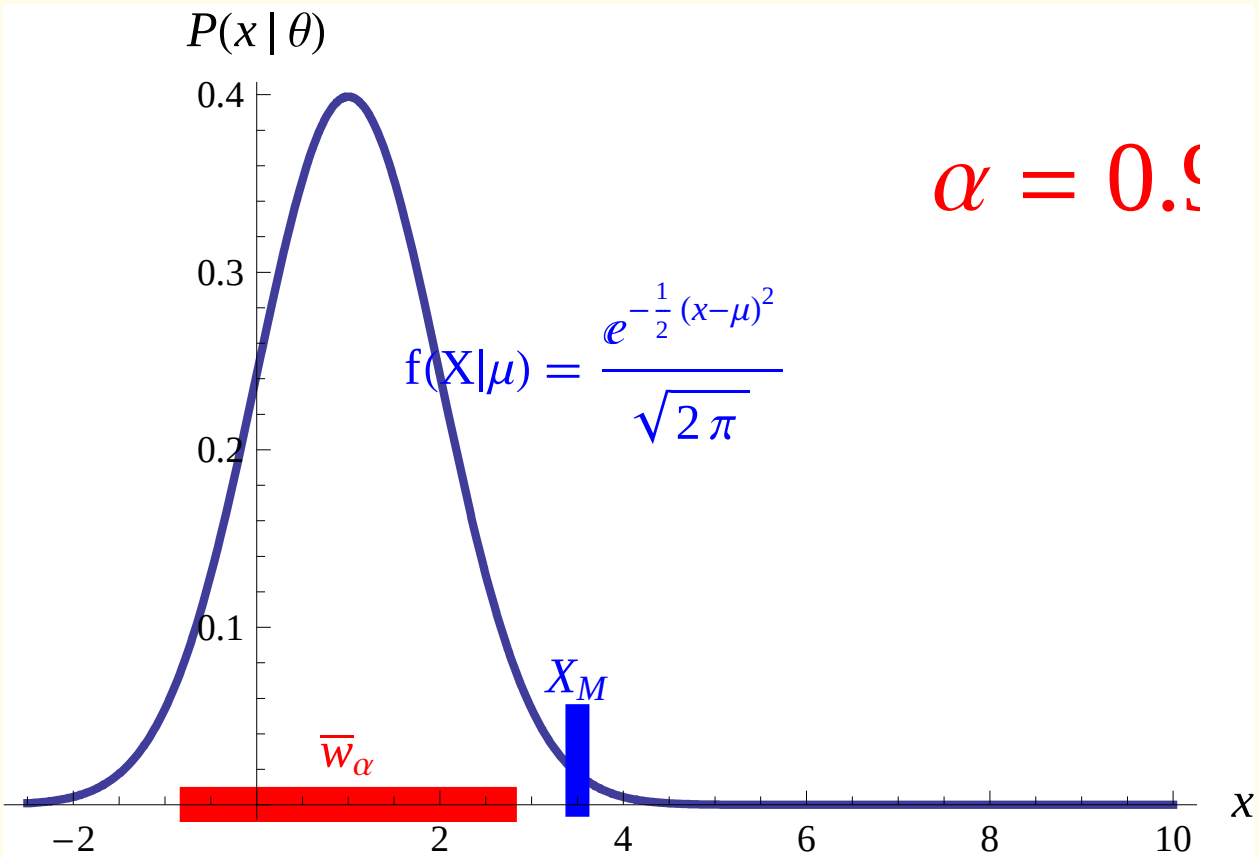
Definition of Confidence Interval

A CI is built with a sequence of hypothesis test: A measurement X_M is tested against $H_0 : \mu = \mu_0$, for all conceivable μ



Definition of Confidence Interval

A CI is built with a sequence of hypothesis test: A measurement X_M is tested against $H_0 : \mu = \mu_0$, for all conceivable μ



Definition of Confidence Interval

The CI is the ensemble of the value of the parameter θ for which H_0 is true; i.e. such that $X \in \bar{w}_\theta$:

$$I_\theta(X) : \{\theta : X \in \bar{w}_\theta\}$$

By construction, $X \in \bar{w}_\theta$ implies that $\theta \in I(X)$ and vice-versa.

$$X \in \bar{w}_\theta \longleftrightarrow \theta \in I(X)$$

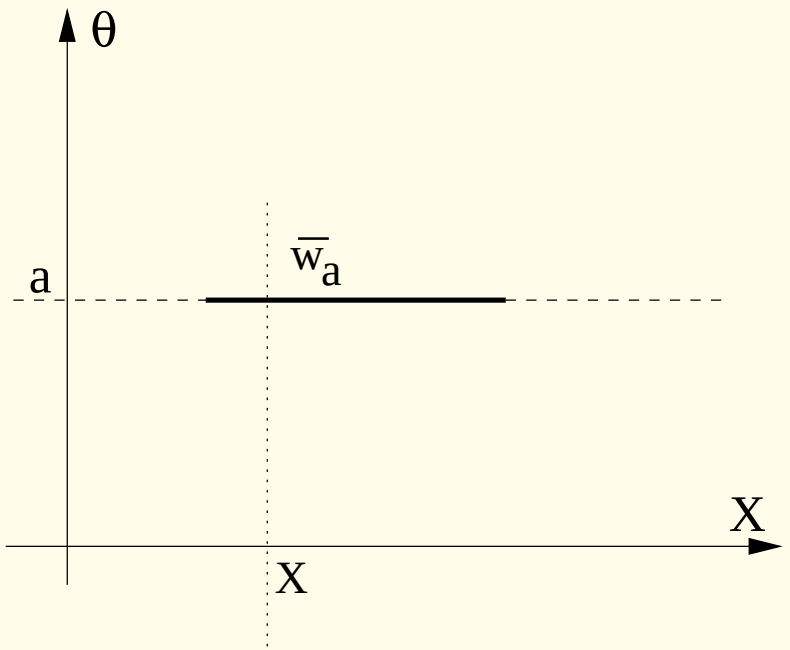
From this follows that:

$$\boxed{\alpha = P(X \in \bar{w}_\theta \mid \theta) = P(X : \theta \in I_\theta(X))}$$

The left side is equal to α by construction, hence the interval $I(X)$ is a α-CI.

General method of construction a CI

Perform a hypothesis test on the simple hypothesis: $H_0 : \theta = a$.
The solution of this problem consists in finding a *critical* region w_a of size α, that is : $P(X \in w_a \mid \theta = a) = 1 - \alpha$.)



$\bar{w}_\alpha$ is the acceptance region of size α.

$$I_\theta(X) = \{a : X \in \bar{w}_a\}$$

If the sample point X falls in w_a, the hypothesis is rejected with a significance α.
a is NOT in our $\alpha - CI$.

X is a random variable, and $I_\theta(X)$ is a random interval
The set of all possible acceptance region defines a region of the $X - \theta$ plane called *the acceptance belt or* $\alpha -$ *belt*

Meaning of a CI

Up to now the intervals $I(X)$ are random variables, function of X. Now we make a measurement:

$$\{X_1, X_2 \cdots X_n\} \longrightarrow \{x_1, x_2 \cdots x_n\} \quad \text{and} \quad I(X) \to i(x)$$

$i(x)$ is NOT a random variable, it is a segment of the real axis, a CI, the one produced by the measurement. The property

$$\alpha = P(X \in \bar{w}_a \mid \theta = a) = P(X : a \in I(X))$$

translates into: In a long run of experiments the intervals $i(x)$ cover the unknown parameter θ in a fraction α of cases.
We cannot say that our $i(x)$ is the one which covers the unknown parameter; it is produced so that in a long series of experiments the fraction of CI covering the unknown θ is α.
We cannot speak of probability: $P(\theta \in i(x))$ since there is NO random variable in this proposition.

General method of construction a CI

There are several ways to construct the acceptance region and conversely several definition of Confidence Intervals

Very used are the *central* acceptance region. This means:

$$P(T \leq w_L \mid \theta) = P(T \geq w_H \mid \theta) = \frac{1-\alpha}{2} \qquad \bar{w}_\alpha : \{w_L \leq T \leq w_H\}$$

Left and right tails have the same probability.
The upper and lower acceptance regions consider only one side of the distributions:

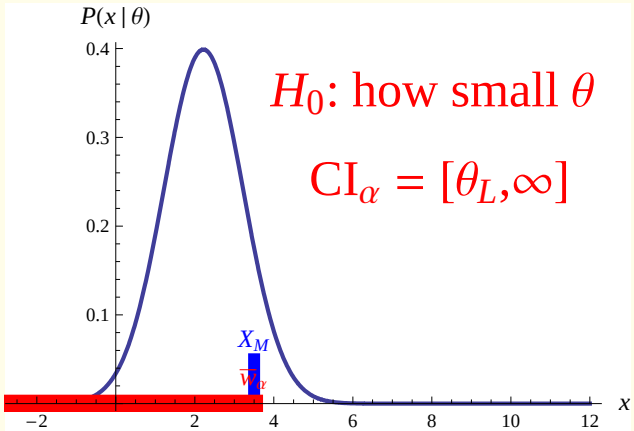
$$P(T \leq w_L \mid \theta) = \frac{1-\alpha}{2} \qquad \bar{w}_\alpha : \{T : w_L \leq T\}$$

$$P(T \geq w_H \mid \theta) = \frac{1-\alpha}{2} \qquad \bar{w}_\alpha : \{T : w_H \geq T\}$$

In all cases, the length of a CI decreases as n increases.

Example lower limit, for $N(\mu, \sigma^2)$, σ known

we have to setup a one-sided test:



$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$w_\alpha = \left\{ \bar{X} : \frac{\sqrt{n}(\bar{X} - a)}{\sigma} \leq k_\alpha \right\}$$

the corresponding α-CI for μ is:

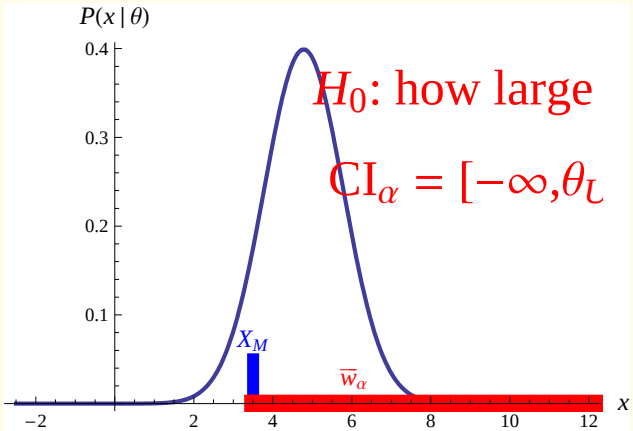
$$I(X) = \{a : X \in \bar{w}_a\} = \left\{ a : \frac{\sqrt{n}(\bar{X} - a)}{\sigma} \leq k_\alpha \right\} = \left\{ a : a \geq \bar{X} - \frac{\sigma k_\alpha}{\sqrt{n}} \right\}$$

$$\theta_L = \bar{X} - \frac{\sigma k_\alpha}{\sqrt{n}} \text{ is the lower limit}$$

We have answered the question: How small can be the parameter?

Example lower limit, for $N(\mu, \sigma^2)$, σ known

we have to setup a one-sided test:



$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$w_\alpha = \left\{ Z : \frac{\sqrt{n}(\bar{X} - a)}{\sigma} \geq k_\alpha \right\}$$

the corresponding α-CI for μ is:

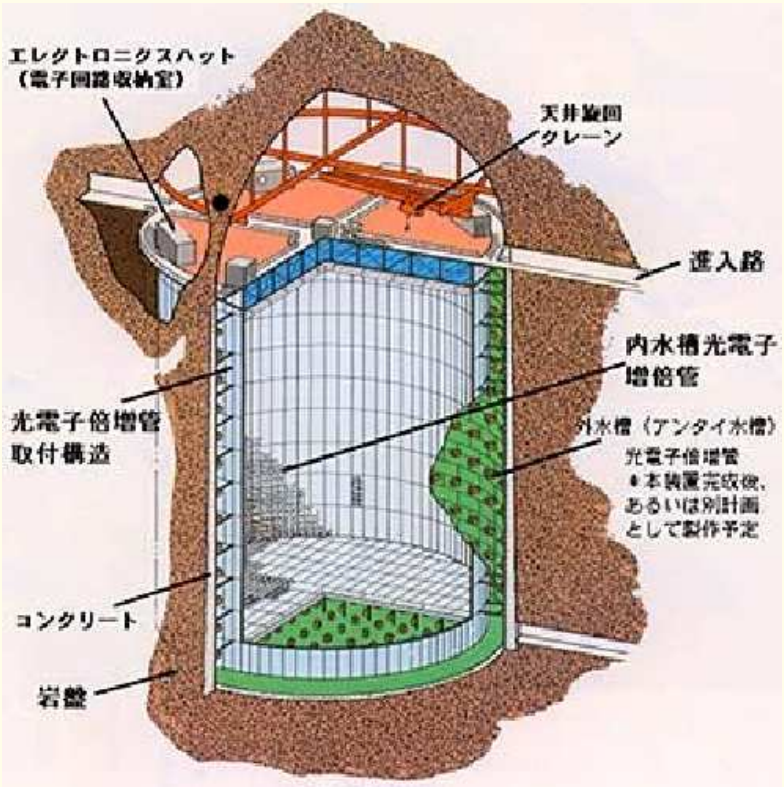
$$I(X) = \{a : X \in \bar{w}_a\} = \left\{ a : \frac{\sqrt{n}(\bar{X} - a)}{\sigma} \geq k_\alpha \right\} = \left\{ a : a \leq \bar{X} - \frac{\sigma k_\alpha}{\sqrt{n}} \right\}$$

$$\theta_U = \bar{X} - \frac{\sigma k_\alpha}{\sqrt{n}} \text{ is the upper limit}$$

We have answered the question: How large can be the parameter?

Upper Limit, an example

With the aim of measuring the lifetime of the proton an experiment is performed at an underground laboratory. In a large mass of water, the process: $p \to \pi^0 + e^+$ is searched.



In one year NO event is observed. The statistics is Poisson:

$$P(n \mid \mu) = \frac{\mu^n e^{-\mu}}{n!}$$

Assume NO background, the rate μ is the rate of the signal.

Upper Limit, an example

The upper limit of the rate of decay (.90CL) is given by:

$$1 - \alpha = 0.1 = P(N \leq 0 \mid \nu^U) = e^{-\nu^u} \rightarrow \nu^u = 2.3 \; evt/y$$

(or $\nu \leq 2.3$). A larger rate would make the probability of our event too small (too many events predicted). On the contrary a smaller rate will give a large probability to the event: $P(n = 0 \mid \mu = 0) = 1$.

The rest is "kinematics": the "flux" (number of protons participating) is

$$N_p = V(cm^3)\rho(g/cm^3)(Z/A)N_A \approx 3 \; 10^{34} \text{ for } V = 10^4 \; m^3$$

hence

$$p^u_{decay} = \frac{\nu^u}{N_p} \qquad \tau^L = 1/p^u_{decay} \approx 10^{34} \; y$$

Exercises!

Exercise

The random variable $X \sim N(0, \sigma^2)$ is sampled n times. Find 95% CI for σ^2

Exercise

The random variable $T \sim e^{-T/\tau}/\tau$ is sampled n times. Find 95% CI for τ in the following cases:

- *n=1*
- *$n \to \infty$*
- *n finite.*

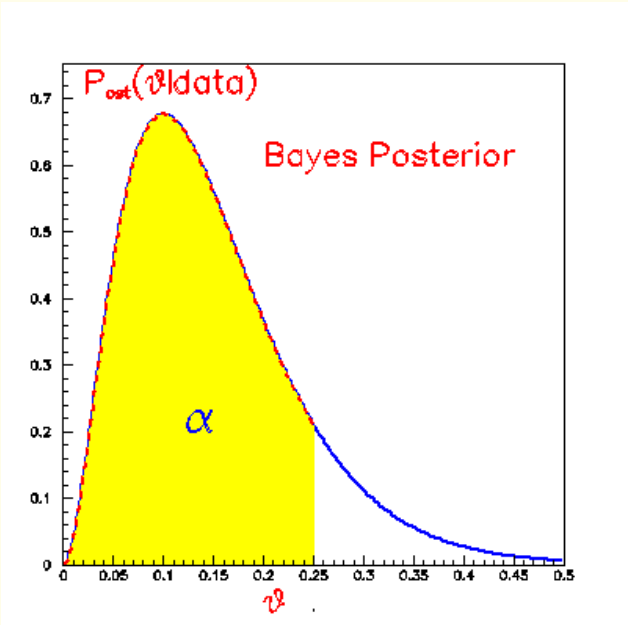
Bayes Confidence Intervals

The precise meaning of CI is different in Bayes and Frequency statistics, but for both the issue is the precision of the inference. When we say that the CI at 95% for μ is $(1, 5)$:

- In the Bayes statistics we mean: $P(\mu \in (1, 5)) = 95\%$ (computed on the posterior probability)
- In the Frequency school neither μ or $(1, 5)$ are random variables and NO probability statement is possible. But it exists a algorithms (pivotal variable, Neyman construction) which produce intervals which include the true, unknown, parameter in a fraction α of experiments performed in uniform conditions. The interval (a, b) is one of those. Hence no Probability but Confidence.

Bayes Upper Limit

Bayes Statistics: The problem is only mathematical: we start from the Posterior probability:



The upper limit θ_U (at a CL α) is that value of θ for which:

$$P(\theta \leq \theta_U \mid data, I_B) = \alpha$$

(Integration of the Posterior density ($P(\theta \mid Data)$ on θ!).

In Bayes Statistics we make direct inference on θ.

Outline

Part 6

- Some examples from real life.

More on Frequency vs Bayes I

A very strong point of Bayes statistics is all contained in Bayes theorem:

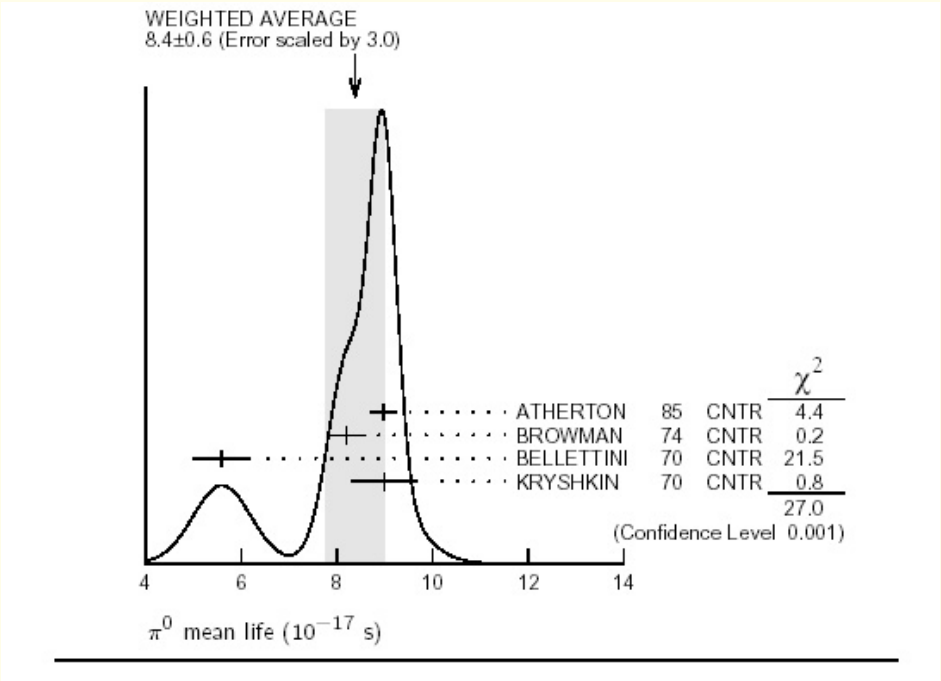
$$P(\theta \mid Data, I_B) \propto P(Data, I_B \mid \theta) \cdot P(\theta)$$

The previous formula is the mathematical description of the the inferential process: how we can step from data to parameters. It makes use of any existing prior knowledge $P(\theta)$, hence it is also a tool to update the information on the parameter θ at each measurement.

Why we should not make blind updates

In physics the update of information is made with great care and usually only after having checks of consistency among the various experiments.

As an example we report the results of measurements of π^0 meson lifetime performed with different techniques by four experiments (source PDG 2004):



One of the experiments is not compatible with the others. Care must be taken when computing the average.

More on Frequency vs Bayes III

In the analysis of a physics experiment we always assume ignorance of previous results and perform measurements and analysis ignoring any previous results. (Or at least this is the way we should do...)
Only afterwards the data of different experiments are combined, after we have judged that the results are homogeneous.
The assessment of the compatibility of the results is a feature of Frequency Statistics (Goodness if Fit).
Bayes Statistics can only make the update of the information whenever a new measurement is made. The update makes the problem of prior distribution less severe, but forbids any comparison.
If we would insist in an analysis independent of previous results, we would again fall in the ambiguity of priors.

Michel parameter story (as told by TD Lee)

In the β decay, the Michel parameter ρ determines the distribution of the electron spectrum, mainly the end point.

Before the discovery of Parity violation it was assumed to be zero. The V-A theory predicts: $\rho = 0.75$ The plots shows ρ, as a function of the year of the measurement.

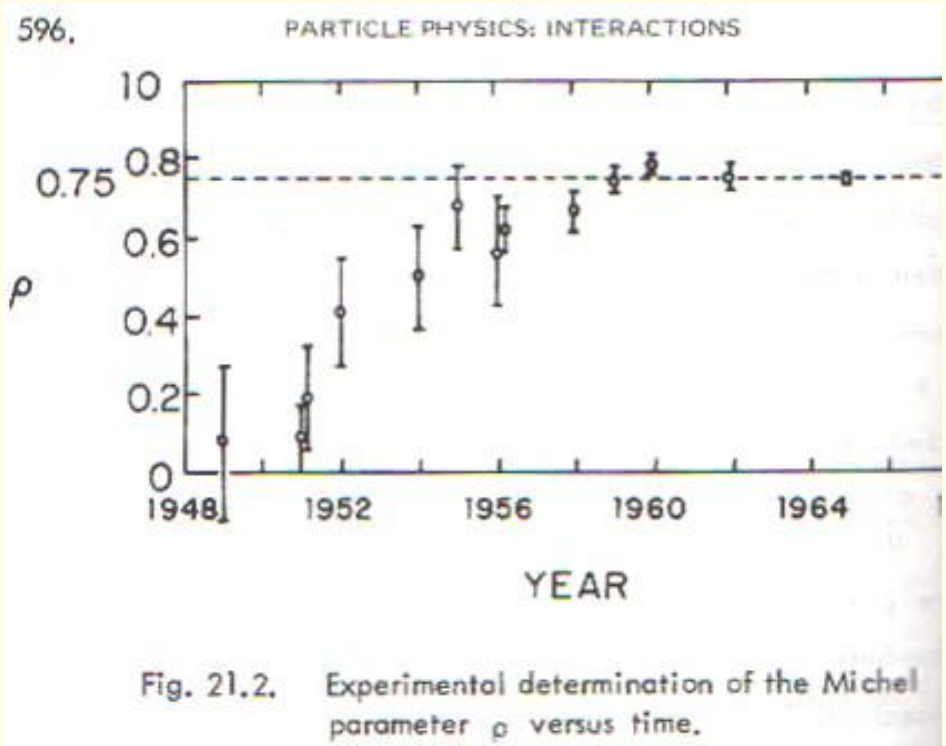


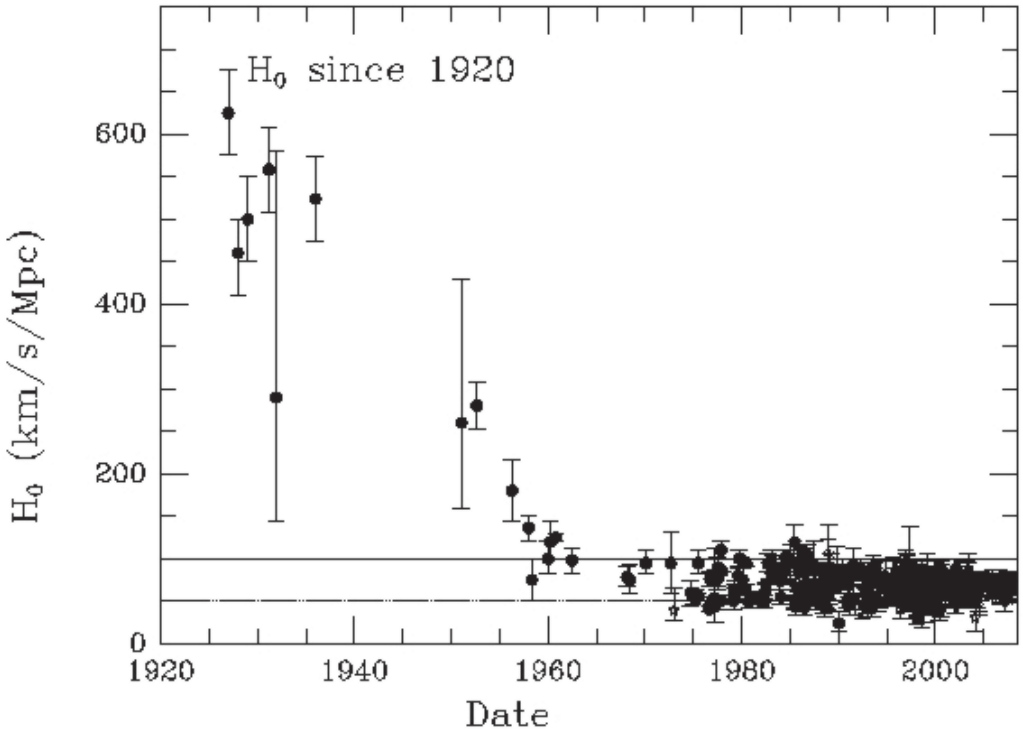
Fig. 21.2. Experimental determination of the Michel parameter ρ versus time.

It is instructive to see the slow drift upwards that only after 1957 converged to the expected value.
Also interesting that at no time the "new" values lie outside the error bar of the previous one...

Hubble constant

The recession velocity of far away objects (quasar, galaxies...) is proportional to the distance: $V_{rad} = Hd$. The first evidence is due to E. Hubble. The value of the "constant" has changed from the first evidence to nowdays. The distance span has also enormously changed since the first observations.

The most recent measurements are more precise and their value is more stable, but the relative fluctuations is largely inconsistent with quoted errors.

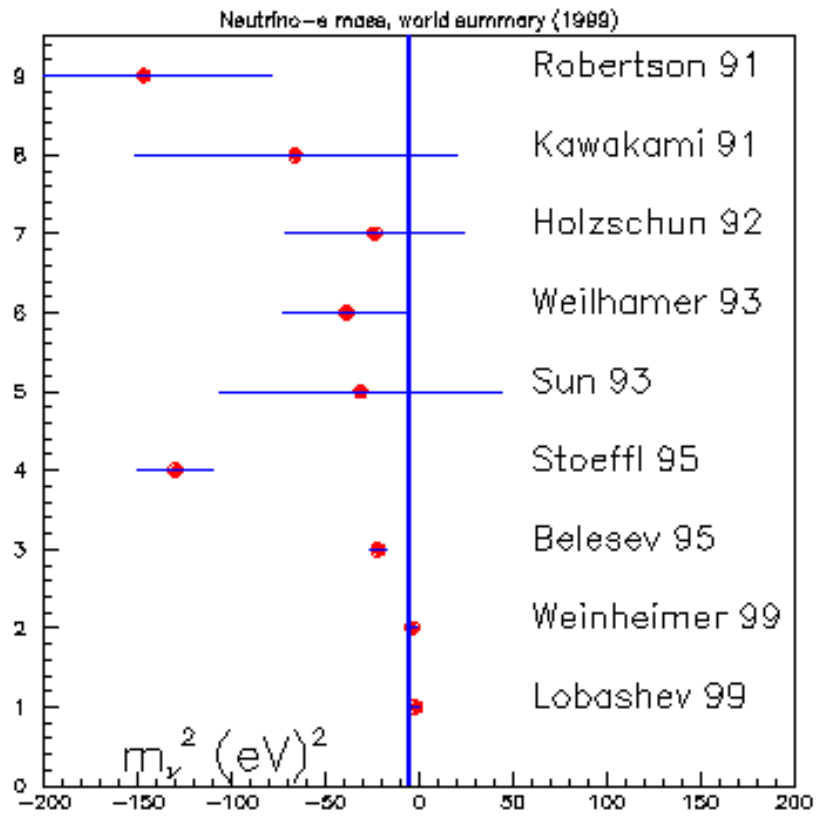


The measurement of ν mass. (Physical boundary)

Frequency Statistics cannot easily incorporate boundary conditions, which are easily handled by Bayes statistics.

Neutrino-e mass (squared) is determined by the shape of the end point of electron momentum in tritium decay.

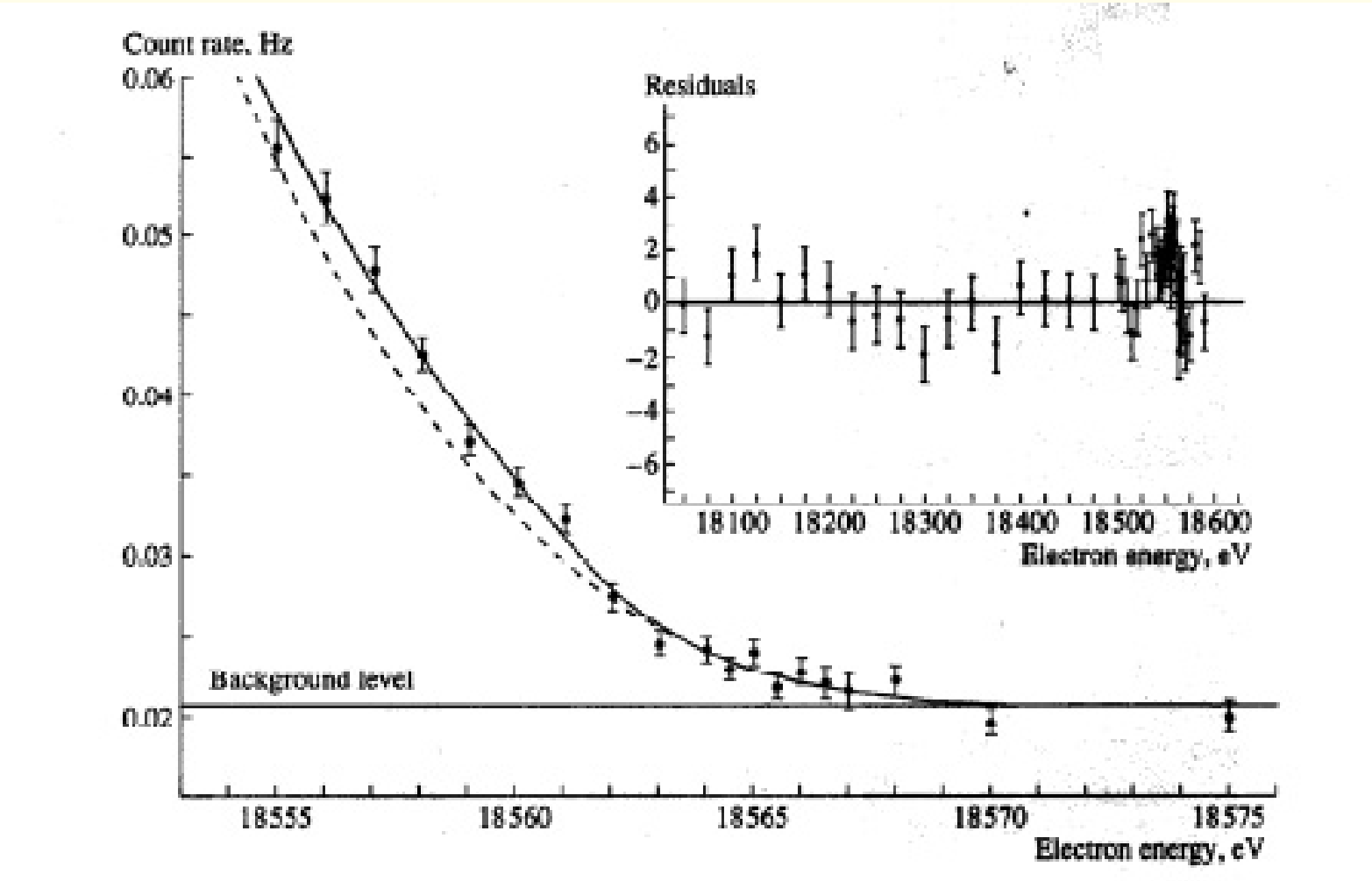
An excess of events at the head of the spectrum (unknown origin) has produced a shift of the measured m_ν^2 toward negative, unphysical values in all the experiments.



The plot shows a compilation of m_ν^2 data.

The Troitsk neutrino mass experiment.

The end point of the integral electron energy distribution, with the two anomalies.



The measurement of ν mass.

If we consider the results of Stoeffel who quotes:

$$m_\nu^2 = (-130 \pm 20 \pm 15)\ eV^2$$

in the frequency theory we should quote a Confidence Interval (at 68%) for the neutrino mass squared is (assume Normal Distribution):

$$(-105, -155)\ eV^2$$

which is clearly wrong: this interval has coverage probability zero!
Has the Frequency Statistics failed in this case?(!).

The measurement of ν mass.

Bayes statistics has an easy way out: even in absence of prior experimental information the square of neutrino mass cannot be negative, hence the prior is a step function:

$$P(m_\nu < 0) = 0 \quad P(m_\nu \geq 0) = Const.$$

All confidence intervals are always meaningful.
As an example in the case of Stoeffel results:

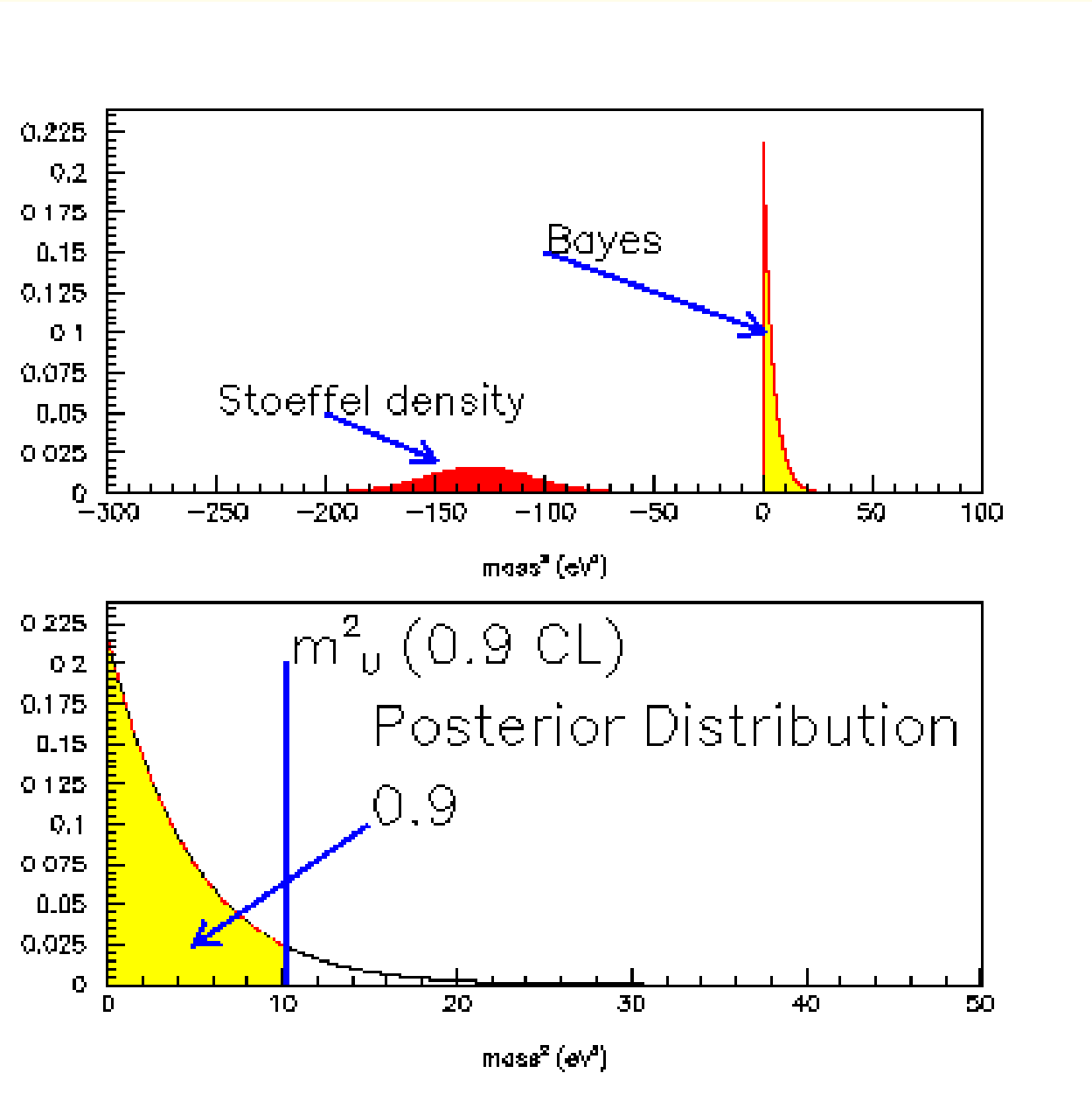
$$\theta = m_\nu^2 = (-130 \pm 20 \pm 15)\ eV^2$$

the upper limit for the neutrino mass square would be determined by:

$$P(\theta \leq \theta_U \mid Data, I_B) = 0.9 \rightarrow \theta_U \approx 10 eV^2$$

... but let's see how it is obtained...

Bayes' analysis of ν mass limit



More on Frequency vs Bayes (7)

This “politically correct” answer hides the experimental problem: there is an unknown background, experienced by all experiments, and which distorts the shape of the electron spectrum.
In these cases the most fair way of presenting the results of an experiment is to show the data themselves.

For Further Reading

There exist a lot of good books on probability and statistics with different level of difficulty.

W. Feller, *An Introduction to Probability theory and its applications*, Volls 1 and 2, John Wiley and Sons 1968

M.G. Kendall, *Advanced theory of Statistics*,Volls 1 and 2 Griffin 1958

W.T. Eadie et al,*Statistical methods in experimental Physics*,North Holland 1971

A. Rotondi, P. Pedroni e A. Pievatolo: *Probabilita', Statistica e Simulazione.* `Springer Italia`

My lessons at the University of Pisa:

T.Del Prete: *Methods of Statistical Data Analysis in High Energy Physics.*

`www.pi.infn.it/atlas/documenti/note/statistica.ps.g`