Probability and Statistics in Particle Physics

Javier Magnin

CBPF – Brazilian Center for Research in Physics

Rio de Janeiro - Brazil

Outline

- Course: 4-5 ~two hours lectures

- 1st lecture:
  - Preliminary concepts
  - Why to estimate errors/Types of errors
  - Distributions: general concepts
  - Combining errors/results
  - Probability
  - Statistics

Outline

- 2nd lecture:
  - Special distributions
  - Sampling distributions
  - Moments of a distribution
  - Error matrix
- 3rd lecture:
  - The resolution function
  - Detection efficiency
  - Statistical inference
  - Estimation of parameters

Back

Outline

- 4th lecture:
  - Parameter fitting
  - Maximum likelihood method
  - Least squares method
  - Hypothesis testing
    - χ^2 test
    - Errors of the first and second kind
  - Minimization
  - Examples: maximum likelihood fit
  - Final remarks

Back

1st lecture

Back

Preliminary concepts

Scientific method:

- ✓ procedure consisting in the collection of data through observation and experimentation and the formulation and testing of hypothesis:

  i. observation → a problem arises
  ii. hypothesis formed
  iii. a experiment is made to test the hypothesis
  iv. analysis of results → conclusions

Experiment:

- ✓ step in the scientific method that arbitrates between competing models or hypotheses.
- ✓ used to test existing theories or new hypotheses in order to support or disprove them.
- ✓ used also to test previous experimental results.

Experiment:

- ✓ step in the scientific method that arbitrates between competing models or hypotheses.
- ✓ used to test existing theories or new hypotheses in order to support or disprove them.
- ✓ used also to test previous experimental results.

An experiment **can never prove** a hypothesis

An experiment **can disprove** a hypothesis

Experiment:

- ✓ it must be repeatable.
- ✓ results should be as accurate as possible
- ✓ all factors intervening in the result must be known:

  - ✓ dependent and independent variables.
  - ✓ correlations among dependent variables.
  - ✓ capabilities of the apparatus used to perform the measurement.
  - ✓ etc.

Experimental results:

- ✓ Two different types:
  - i. Determination of the numerical value of some physical quantity } Parameter determination
  - ii. Testing whether a particular theory is consistent with data } Hypothesis testing
- ✓ In real life there is a degree of overlapping between both types above

Why to estimate errors ?

An experimental measurement has no sense if it is not accompanied by its error.

Note that:

- ✓ An error is not a mistake, it is the difference between the measured and the true value of a given quantity.
- ✓ Errors cannot be avoided, but the experimentalist can make them as small as possible.

Why to estimate errors ?

Example: The fine structure constant α

- ✓ Definition: $\alpha = \dfrac{e^2}{(4\pi\varepsilon_0)\hbar c}$
- ✓ Experimental value: $\alpha = \dfrac{1}{137.035999084(51)}$

(measured through the relationship between the magnetic moment of the electron and α predicted by QED, by far the most precise measurement of α possible today - (PRL 100 (2008) 120801)).

Question: is the fine structure constant α truly constant over time ?

(a varying α may help to solve cosmological and astrophysical problems. It is also of theoretical interest motivated by String theory and Beyond the Standard Model physics)

Question: is the fine structure constant α truly constant over time ?

(a varying α may help to solve cosmological and astrophysical problems. It is also of theoretical interest motivated by String theory and Beyond the Standard Model physics)

✓ Define $$\frac{\Delta\alpha}{\alpha} = \frac{\alpha_{prev} - \alpha_{now}}{\alpha_{now}}$$

✓ In 1999, J.K. Webb *et al.*, using the Keck telescopes and a data set of 128 quasars at red-shifts $0.5 < z < 3$ found that (PRL **82** (1999) 884, **87** (2001) 091301)

$$\frac{\Delta\alpha}{\alpha} = (-0.57 \pm 0.10) \times 10^{-5}$$

consistent with a slight increase of a over the last 10 - 20 billion years

Question: is the fine structure constant α truly constant over time ?

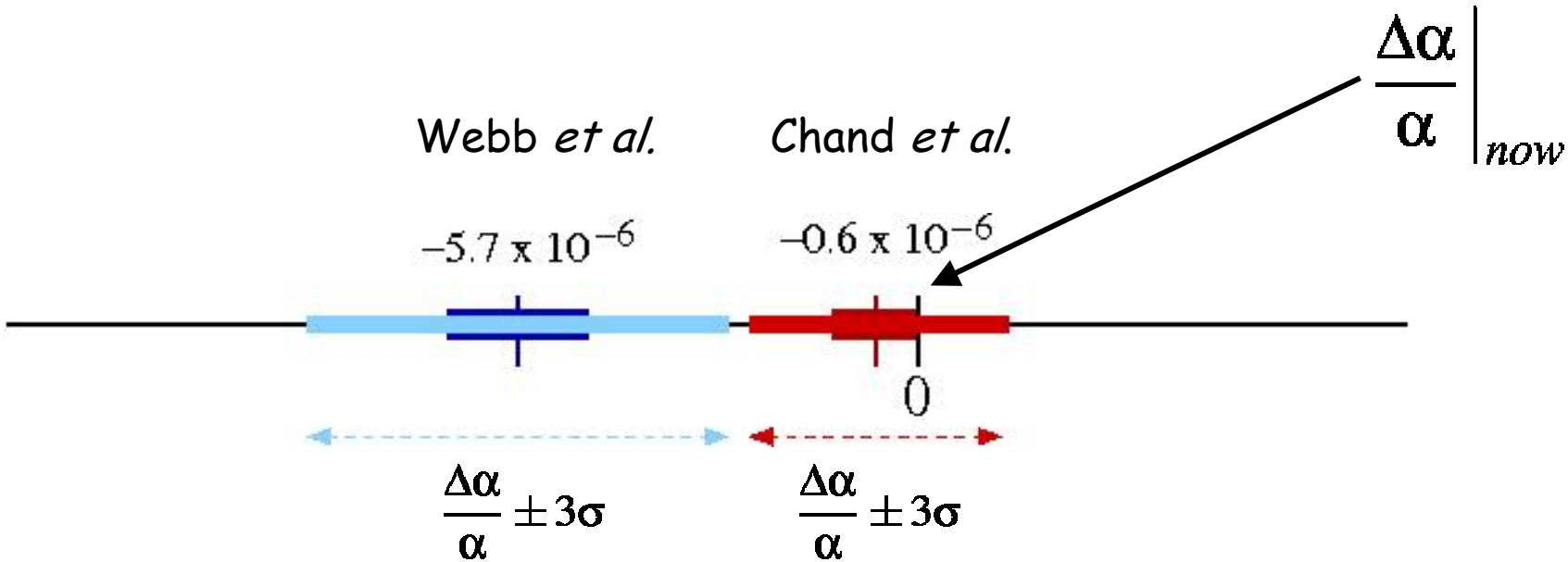
(a varying α may help to solve cosmological and astrophysical problems. It is also of theoretical interest motivated by String theory and Beyond the Standard Model physics)

✓ Define $$\frac{\Delta\alpha}{\alpha} = \frac{\alpha_{prev} - \alpha_{now}}{\alpha_{now}}$$

✓ In 2004, H. Chand *et al.*, using the Very Large telescope found (Astron. Astrophys. **417** (2004) 853)

$$\frac{\Delta\alpha}{\alpha} = (-0.6 \pm 0.6) \times 10^{-6}$$

consistent with no measurable variation



✓ Even at a 1σ level ($\sigma = 0.6 \times 10^{-6}$), the measurement of Chand *et al.* is compatible with no change in α.

✓ The measurement by Webb *et al.* is compatible with a change of α with time at 3σ level ($\sigma = 0.1 \times 10^{-5}$).

✓ With an error $\sigma = 2 \times 10^{-6}$ the measurement by Webb *et al.* ➔ is compatible with no change of α with time at 3σ !

Different type of errors

There are two different kind of errors:

1. Statistical (or random) errors:
   difference between the measured and the true value of a given quantity caused by random, and inherent unpredictable fluctuations in the measurement apparatus or in the system being studied.
2. Systematic errors:
   difference between the measured and the true value of a given quantity caused by non-random fluctuations. Once identified the source of a systematic error, it can usually be eliminated.

Example: measurement of the decay constant λ of a radioactive source

Assume that you have a radioactive source of a given mass (let's say, a given amount of Am (Americium)), then

✓ Count how many decays are observed in a time interval Δt

✓ Weight the sample (to obtain the number n of decaying nuclei present)

✓ Calculate the decay rate, λ, from the formula

$$-\frac{dn}{dt} = \lambda n$$

n = number of decaying nuclei

Errors in the measurement

- **Statistical errors**:
  - ✓ inherent statistical error in counting random events
  - ✓ uncertainty in the mass of the sample
  - ✓ timing of the period for which the decay are observed

- **Systematic errors**:
  - ✓ efficiency of the counter used to detect the decays
  - ✓ background (i.e. particles coming from other sources)
  - ✓ purity of the radioactive sample
  - ✓ calibration errors

Statistical errors:

- ✓ produce different individual results in a series of measurements of the same physical quantity.
- ✓ can be reduced performing repeated measurements and combining the results.
- ✓ average to zero.

Systematic errors:

- ✓ if the source is known, they can be corrected or at least, reduced (eg. hidden variables, like effects of the temperature in the measurement apparatus).
- ✓ there are no simple rules to eliminate them. The "repeated measurements" approach does not work
- ✓ in average they are different to zero

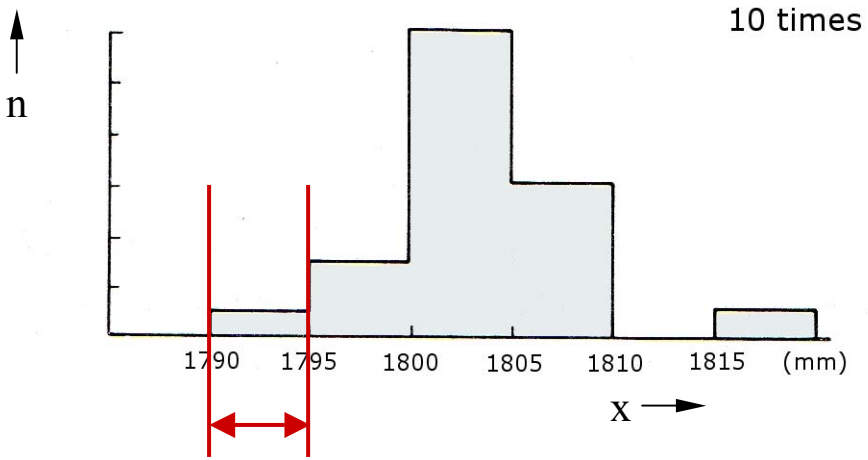
Distributions

In general, the result of repeating the same measurement many times does not lead to the same result

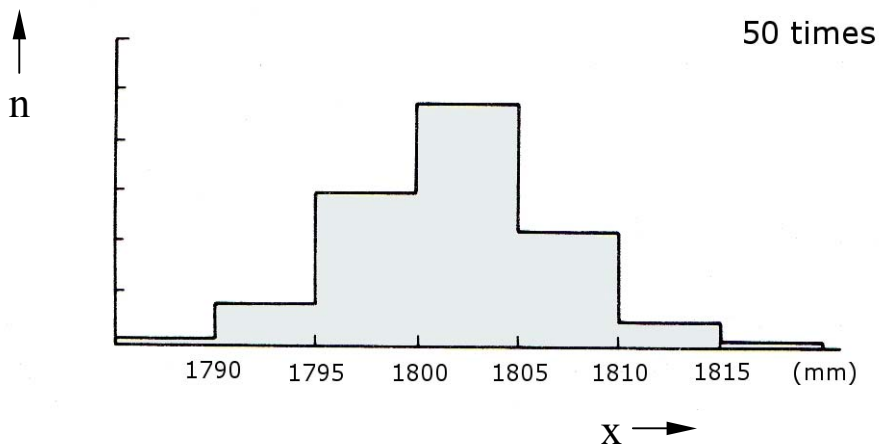
Experiment:

- Measure the length of one side of your table 10 times and display the results in a histogram.
- What happens if you repeat the measurement 50 times ?

Measuring the table...



Bin size

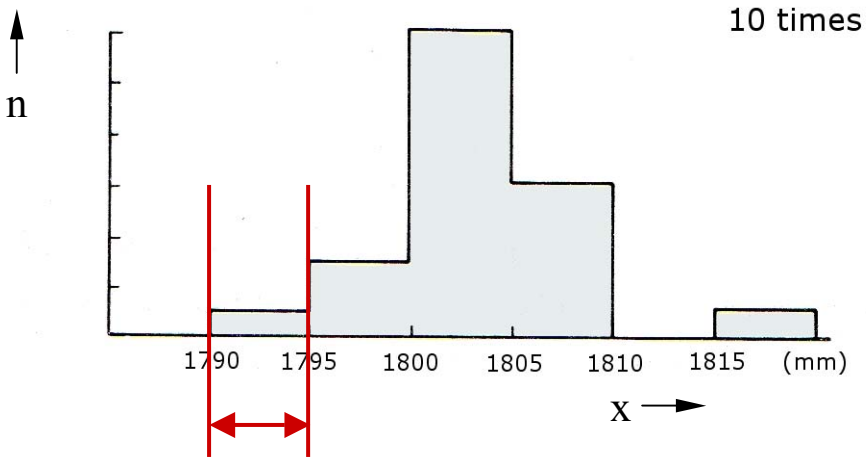


x = length of the side of the table

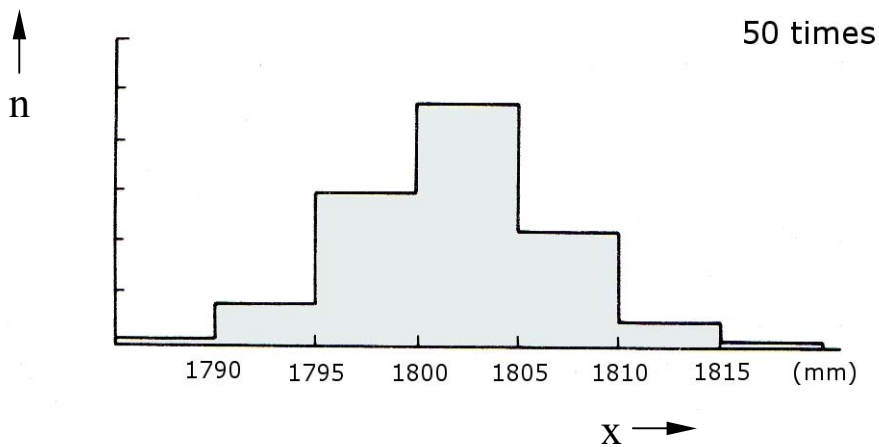
n = number of observations per unit of Δh

$n_i \Delta h$ = number of observations in the i-th bin

Measuring the table...



Bin size = Δh



Distribution n(x): describes how often a value of the variable x occurs in a sample

Sample: the total number of measurements of the side of the table, N

x = length of the side of the table

n = number of observations per unit of Δh

$n_i \Delta h$ = number of observations in the i-th bin

- ✓ The random variable, x, can be discrete or continuous
  - ✓ A discrete random variable is a variable whose probability distribution is discrete
  - ✓ A continuous random variable is a variable whose probability distribution is continuous
  - ✓ If x is discrete, n(x) gives the probability of observing the value x of the random variable
  - ✓ If x is continuous, n(x)dx is the probability of observing the random variable in the range between x and x + dx
- ✓ A distribution is defined in the full range of the random variable. It may be finite or infinite. In a more technical terminology, the support of a distribution is the smallest interval or set whose complement has probability zero.

Probability distributions of real-valued random variables

✓ The cumulative distribution function (CDF) completely describes the probability distribution (pdf) of a real-valued random variable

✓ **Definition**: for every real number x, the CDF of a real-valued random variable X is

$$x \longrightarrow F_X(x) = P(X \leq x)$$

where the right hand side represents the probability of X being les than or equal to x. The probability of X being in the interval (a,b) is therefore $F_X(b) - F_X(a)$ if $a < b$.

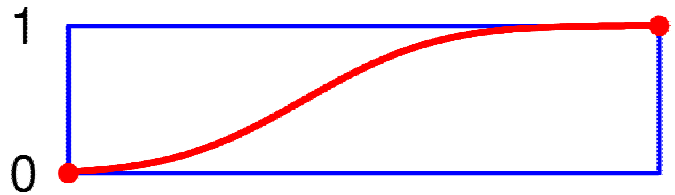
✓ A discrete pdf has a CDF which increases by jumps

✓ A continuous pdf has a continuous CDF

Cumulative distribution function: properties

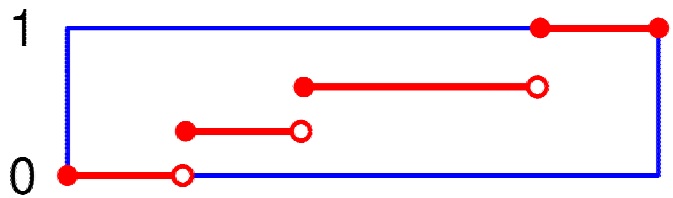
- ✓ If f(x) is a continuous pdf, then its CDF is

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



- ✓ If p(x) is a discrete pdf, then its CDF is

$$F(x) = \sum_{x_i \le x} p(x_i)$$



- ✓ $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1$
- ✓ $F(b) - F(a) = P(a \le x \le b) = \int_{a}^{b} F(t)dt$

Probability distribution function: general properties

Assume you have a set of N separate measurements of a given physical quantity. Then these N separate values are distributed according to a given pdf, which can be discrete ($p(x_i)$) or continuous ($f(x)$). In order to characterize the pdf we define:

Discrete case:

Mean $$\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i$$

(Stat. Infer.)

In practice, however, it is not possible to perform an infinite number of measurements, then we better use

Mean $$\bar{x} = \frac{1}{N}\sum_i x_i$$

Alternatively, we can define the mean in terms of the pdf as

$$\bar{x} = \sum_{i=1}^{n} x_i \; p(x_i)$$

where we have excluded the repeated values of x_i (then the sum runs over i until i=n < N)

(Stat. Infer.)

Other useful information is how the measured values deviate from the mean, i.e. $d_i = x_i - \mu$. Note however that

$$\lim_{N\to\infty} \bar{d} = \lim_{N\to\infty} \left[\frac{1}{N} \sum (x_i - \mu) \right] = (\mu - \mu) = 0$$

A parameter which better satisfy our needs is then the standard deviation squared or variance, σ^2, defined as

$$\sigma^2 = \lim_{N\to\infty} \left[\frac{1}{N} \sum (x_i - \mu)^2 \right]$$

Again, σ^2 depends on m, which is in general unknown, then is better to use

$$s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

or its equivalent

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 p(x_i)$$

where the sum runs over i excluding equal values of x_i

Again, σ^2 depends on m, which is in general unknown, then is better to use

$$s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

degrees of freedom

or its equivalent

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 p(x_i)$$

where the sum runs over i excluding equal values of x_i

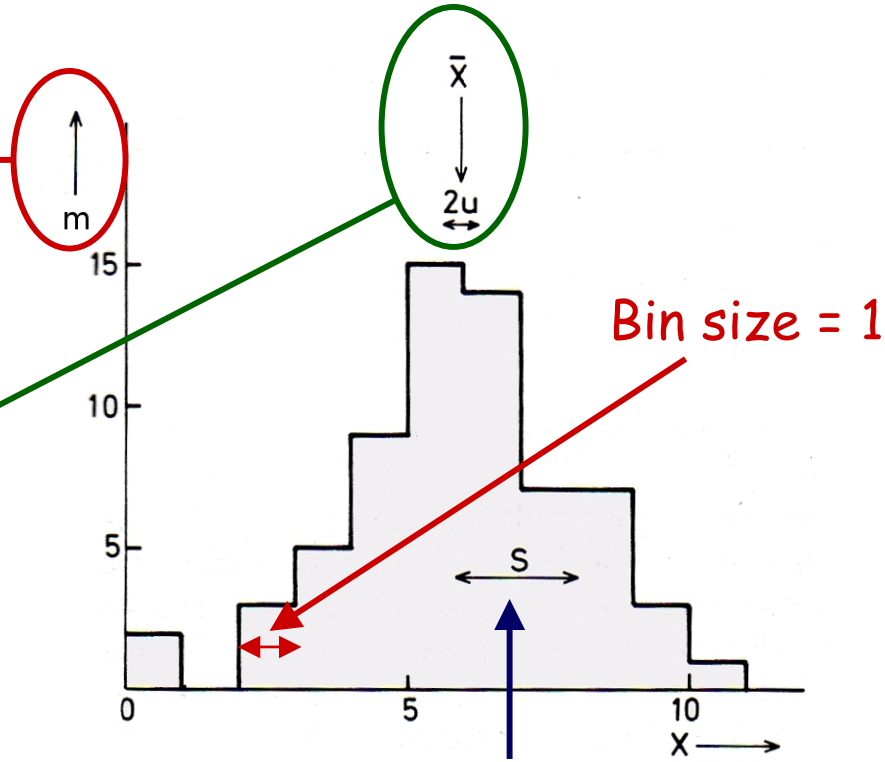
The generalization for continuous distributions is

$$\bar{x} = \frac{1}{N} \int f(x) x \, dx$$

$$s^2 = \frac{1}{N-1} \int f(x)(x_i - \bar{x})^2 dx$$

Convention: μ and σ^2 are reserved for the true values of the mean and variance respectively

Entries in each x bin (66 in total)



Bin size = 1

The mean $\bar{x}$ is known to an accuracy of $s/\sqrt{N}$

$$s^2(\bar{x}) = s^2\left(\frac{1}{N}\sum_i x_i\right)$$

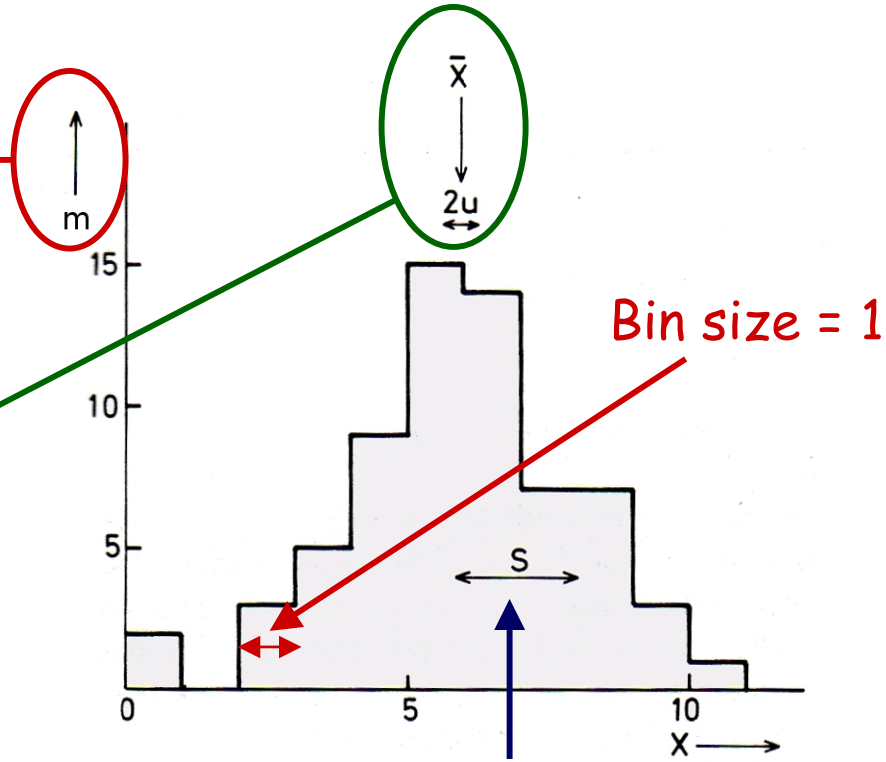
$$= \frac{1}{N}\sum_i s^2(x_i)$$

$$= \frac{s^2}{N}$$

s^2 is a measurement of how the distribution spreads out and not a measurement of the error in $\bar{x}$!

Stat. Infer.

Entries in each x bin (66 in total)



Bin size = 1

The mean $\bar{x}$ is known to an accuracy of $s/\sqrt{N}$

$$s^2(\bar{x}) = s^2\left(\frac{1}{N}\sum_i x_i\right)$$

$$= \frac{1}{N}\sum_i s^2(x_i)$$

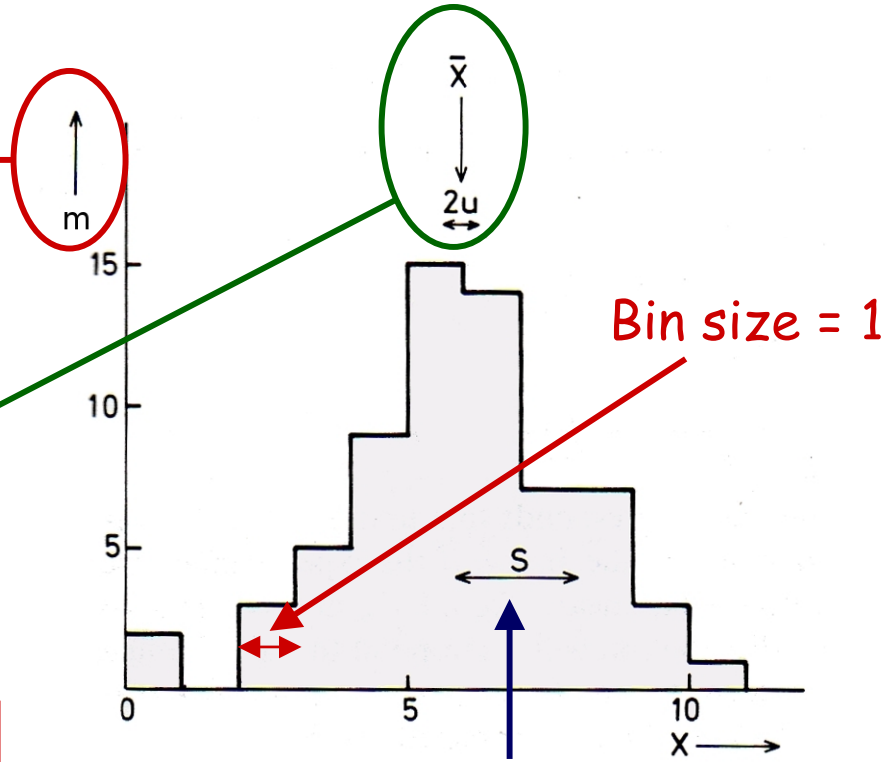
$$= \frac{s^2}{N}$$

s^2 is a measurement of how the distribution spreads out and not a measurement of the error in $\bar{x}$!

increasing N increases the accuracy in $\bar{x}$

Entries in each x bin (66 in total)

The mean $\bar{x}$ is known to an accuracy of $s/\sqrt{N}$



s^2 and $\bar{x}$ are properties of the distribution and they do not change with N. The accuracy (variance) of the mean increases with N

s^2 is a measurement of how the distribution spreads out and not a measurement of the error in $\bar{x}$!

$$\lim_{N\to\infty} \bar{x} = \mu \qquad \lim_{N\to\infty} s^2 = \sigma^2$$

Combining errors / results

- ✓ Very often we are confronted with a situation where the result of an experiment is given in terms of two or more variables.
  - What we want to know is what is the error of the final result in terms of the errors of the measured variables.
  - This is the well known problem of "propagation of errors".
- ✓ A second (related) problem is how to combine the results of two or more experiments who have made the same measurement.

Propagation of errors

General method

Suppose that you have to determine a physical quantity which is not directly accessible in the experiment, but in turns, it depends on several variables which you can really measure. The measured variables are determined with a given experimental error, then the question is: how the experimental errors in the measured variables propagates onto the physical quantity we want to determine ?

To start with, let us consider the following

Example:
Suppose you want to determine the volume V of a box with length L, width W and height H. Suppose you measure the three quantities L,W and H obtaining $L_0+\Delta L$, $W_0+\Delta W$ and $H_0+\Delta H$. The volume of the box is then

$$V_0 = L_0 W_0 H_0$$

The error in the determination of the Volume V can be estimated from ($\Delta x = x_0 - x_{true}$)

$$\Delta V \simeq \Delta L \left(\frac{\partial V}{\partial L}\right)_{W_0 H_0} + \Delta W \left(\frac{\partial V}{\partial W}\right)_{L_0 H_0} + \Delta H \left(\frac{\partial V}{\partial H}\right)_{L_0 W_0}$$

To start with, let us consider the following

Example:
Suppose you want to determine the volume V of a box with length L, width W and height H. Suppose you measure the three quantities L,W and H obtaining $L_0+\Delta L$, $W_0+\Delta W$ and $H_0+\Delta H$. The volume of the box is then

In order to obtain ΔV we have made the assumption that the measurements of L,W and H are uncorrelated

The e… …be
estim…

$$\Delta V \simeq \Delta L \left(\frac{\partial V}{\partial L}\right)_{W_0 H_0} + \Delta W \left(\frac{\partial V}{\partial W}\right)_{L_0 H_0} + \Delta H \left(\frac{\partial V}{\partial H}\right)_{L_0 W_0}$$

In the most general case, we have a quantity x = f(u;v;…) depending on a number of other variables {u;v;..}, which are measured with known variance σ_u; σ_v; ...

We can assume

$$\bar{x} = f(\bar{u};\bar{v};\ldots)$$

from where we can calculate

$$\sigma_x^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} (x_j - \bar{x})^2$$

using the approximation

$$x_i - \bar{x} \simeq (u_i - \bar{u}) \frac{\partial f}{\partial u} + (v_i - \bar{v}) \frac{\partial f}{\partial v} + \ldots$$

then

$$\sigma_x^2 \simeq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left[(u_i - \bar{u}) \frac{\partial F}{\partial u} + (v_i - \bar{v}) \frac{\partial F}{\partial v} + \ldots \right]^2$$

$$\simeq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left[(u_i - \bar{u})^2 \left(\frac{\partial F}{\partial u} \right)^2 + (v_i - \bar{v})^2 \left(\frac{\partial F}{\partial v} \right)^2 \right.$$

$$\left. + 2(u_i - \bar{u})(v_i - \bar{v}) \left(\frac{\partial f}{\partial u} \right) \left(\frac{\partial f}{\partial v} \right) + \ldots \right]$$

(Error matrix)

now, the first two terms can be rewritten as

$$\sigma_u^2 = \lim_{N\to\infty} \frac{1}{N} \sum (u_i - \bar{u})^2 \qquad \sigma_v^2 = \lim_{N\to\infty} \frac{1}{N} \sum (v_i - \bar{v})^2$$

while for the third we introduce the *covariance*

$$\sigma_{uv}^2 = \lim_{N\to\infty} \frac{1}{N} \sum [(u_i - \bar{u})(v_i - \bar{v})]$$

The final result is

$$\sigma_x^2 \simeq \sigma_u^2 \left(\frac{\partial f}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial f}{\partial v}\right)^2 + 2\sigma_{uv}^2 \left(\frac{\partial f}{\partial u}\right)\left(\frac{\partial f}{\partial v}\right) + \ldots$$

Comments:

- ✓ the first two terms are average of squared deviations
- ✓ they will presumably dominate
- ✓ the third term is the average of the cross terms involving products of the deviations of u and v
- ✓ if u and v are uncorrelated variables then the covariance terms will vanish in average
- ✓ generalization for more than two variables is trivial

Special cases:

Exercise: prove them !

$$x = au \pm bv \qquad \sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 \pm 2ab\sigma_{uv}^2$$

$$x = avu \qquad \sigma_x^2 = a^2v^2\sigma_u^2 + a^2u^2\sigma_v^2 + 2a^2uv\sigma_{uv}^2$$

$$x = \pm\frac{au}{v} \qquad \sigma_x^2 = \left[\frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} - 2\frac{\sigma_{uv}^2}{uv}\right]x^2$$

$$x = au^{\pm b} \qquad \sigma_x = b\frac{\sigma_u}{u}x$$

$$x = a\exp[\pm bu] \qquad \sigma_x = b\sigma_u x$$

$$x = a\ln(\pm bu) \qquad \sigma_x = a\frac{\sigma_u}{u}$$

averaging

Assume that you perform N independent measurements q_i of a quantity q, each one of accuracy σ

The average $\bar{q}$ of the N measurements q_i is then

$$\bar{q} = \frac{1}{N}\sum q_i$$

Using that if x = u+v then $\sigma_x^2 = \sigma_u^2 + \sigma_v^2$ for $x=N\bar{q}$, we obtain

$$N^2\sigma_q^2 = \sum \sigma^2 = N\sigma^2$$

and the error on $\bar{q}$ results

$$\sigma_q = \frac{\sigma}{\sqrt{N}}$$

Combining results of different experiments

Assume that several experiments measured the same physical quantity a and obtained the set of values $\{a_i\}$, with errors $\{\sigma_i\}$. Then the best estimates of a and σ are given by

$$a = \frac{\sum a_i/\sigma_i^2}{\sum 1/\sigma_i^2} \qquad \frac{1}{\sigma^2} = \sum \frac{1}{\sigma_i^2}$$

No proof. However, if $\sigma_i = \sigma \; \forall$ i, i=1...n, then the results above reduce to the averaging case of the previous slide !

Probability

Suppose you repeat an experiment several times. Even if you are able to keep constant the essential conditions, the repetition will produce different results



The result of an individual measurement will be unpredictable even when the possible results of a series of measurements have a well-defined distribution

Definition:

The probability p of obtaining a specific result when performing one measurement or trial is defined as

$$p = \frac{\text{Number of times on which that result occurs}}{\text{Total number of measurements or trials}}$$

Rules of probability

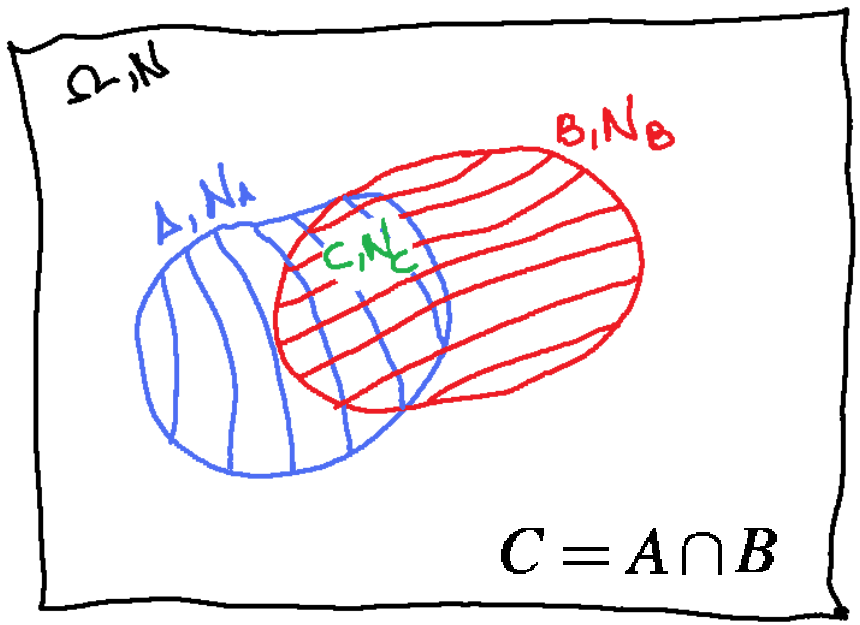
1. If $P(A)$ is the probability of a given event A, then $0 \leq P(A) \leq 1$.
2. The probability $P(A+B)$ that at least A or B occurs is such that $P(A+B) \leq P(A) + P(B)$. The equality is valid only if A and B are exclusive events.
3. The probability $P(AB)$ of obtaining both A and B is $P(AB) = P(A|B)P(B) = P(B|A)P(A)$, where $P(A|B)$ is the probability of obtaining A given that B has occurred ($P(A|B)$ is know as the conditional probability of A given B).
4. The rule 3. defines the conditional probability as $P(A|B) = P(AB)/P(B)$.

Comments (about the rules !)

1. $P(A+B) = P(A) + P(B) - P(AB)$
   to avoid double counting !

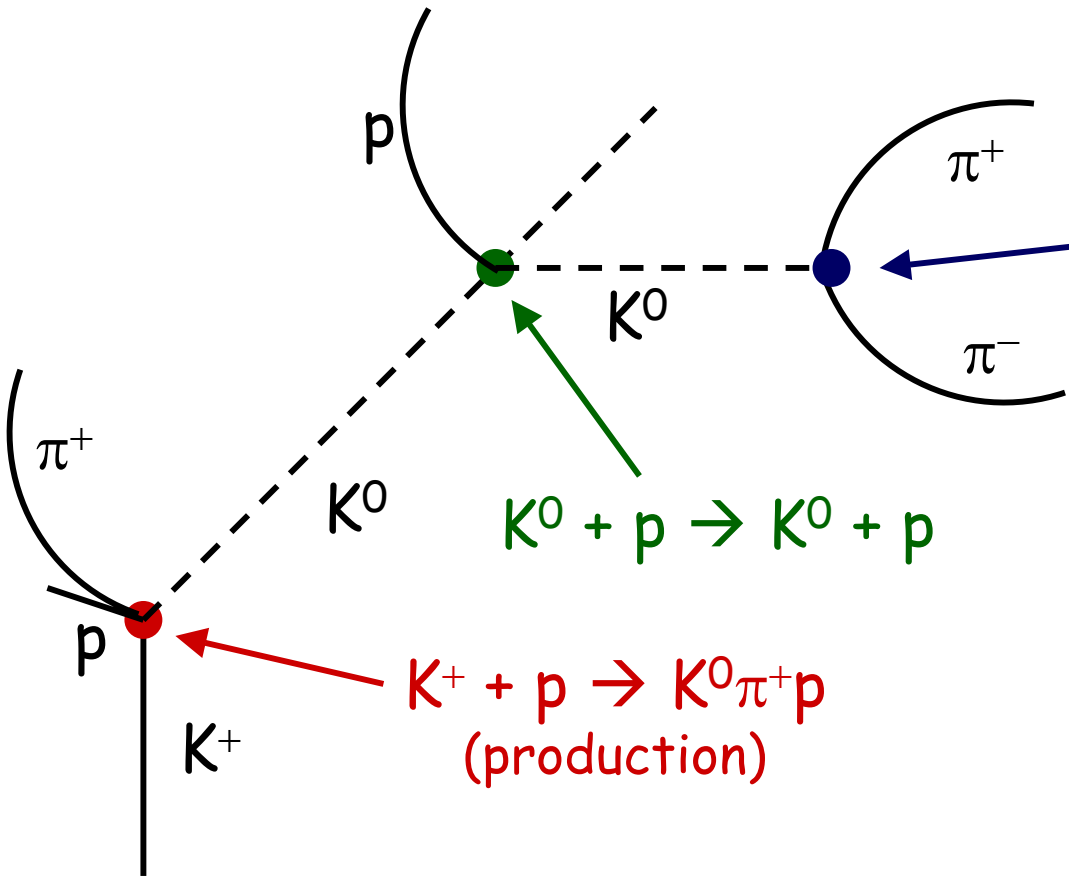
2. $P(A|B) = P(AB)/P(B)$
   $= (N_C/N)/(N_B/N)$
   $= N_C/N_B$

3. $P(A|B) \neq P(B|A)$
4. If $P(A|B) = P(A)$
   then A and B are
   independent



Example: use of conditional probability

Measurement of the mass difference $\Delta m = m(K_L) - m(K_S)$ from the K^0p and $\bar{K}^0p$ cross sections



$K^0 \rightarrow \pi^+ + \pi^-$ (decay)

$K^0 + p \rightarrow K^0 + p$

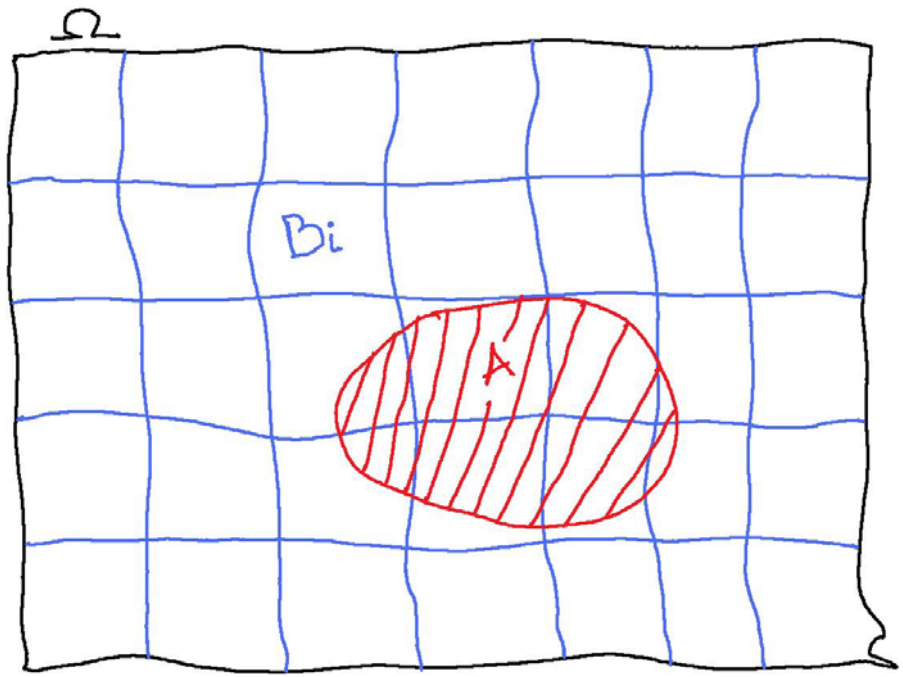
$K^+ + p \rightarrow K^0\pi^+p$ (production)

- K^0 are detected in the decay (event B)
- We want to measure $K^0p \rightarrow K^0p$ (event A)
- We are interested in $P(AB) = P(B|A)P(A)$

Bayes theorem

Let the sample space Ω be spanned by n mutually exclusive and exhaustive sets B_i, $\Sigma_i P(B_i) = 1$. If A is also a set belonging to Ω, then

$$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\Sigma_j P(A|B_j)P(B_j)}$$



Example (of the Bayes theorem)

Consider three drawers B_1, B_2, B_3, each one with two coins. B_1 has two gold coins, B_2 has one gold and one silver and B_3 has two silver coins.

- ✓ Now select a random drawer and pick a coin from it.
- ✓ Supposing that the coin is gold, what is the probability of having a second gold coin in the same drawer (or, what is the probability of having selected the drawer B_1, given that I selected a gold coin)?



B_1 B_2 B_3

If A is the event of first picking a gold coin, then

$$P(A|B_1) = 1; \quad P(A|B_2) = 1/2; \quad P(A|B_3) = 0$$

Since the drawers is selected at random,

$$P(B_1) = P(B_2) = P(B_3) = 1/3$$

Hence, from Bayes theorem follows that

$$P(B_1|A) = \frac{P(A|B_1)\,P(B_1)}{\sum_j P(A|B_j)P(B_j)} = \frac{1 \times 1/3}{1x1/3 + 1/2x1/3 + 0x1/3} = 2/3$$

If A is the event of first picking a gold coin, then

$$P(A|B_1) = 1; \quad P(A|B_2) = 1/2; \quad P(A|B_3) = 0$$

Since the drawers is selected at random

Thus, although the probability of selecting the drawer B_1 is 1/3, the observation that the first coin drawn was gold doubles the probability that the drawer B_1 had been selected

$$P(B_1|A) = \frac{P(A|B_1)\,P(B_1)}{\sum_j P(A|B_j)P(B_j)} = \frac{1 \times 1/3}{1x1/3 + 1/2x1/3 + 0x1/3} = 2/3$$

Statistics

Probability: we start with a well defined problem, and calculate from this the possible outcome of a specific experiment

Statistics: we use data from a experiment to deduce what are the rules or laws relevant to the experiment

Statistics → two different problems (with some overlap)

Parameter determination: use data to measure the parameter of a given model or theory

Hypothesis testing: use data to test the hypothesis of a model

From theory to data	From data to theory
•Toss a coin. If P(heads) = 0.5, how many heads you get in 50 tosses? •Given a sample of K* mesons of known polarization, what is the forward/backward asymmetry?	•If you observe 27 heads in 50 tosses, what is the value of P(heads)? (*) •Of 1000 K* mesons, 600 were observed that decay forward. What is the K* polarization? (**)

(*) Parameter determination: deduce a quantity and its error
(**) Hypothesis testing: check a theory

2nd lecture

Back

Special distributions

Binomial distribution

N independent trials, each of which has only two possible outcomes, success p or failure q=(1-p)

$$P(r,N,p) = \binom{N}{r} p^r q^{N-r}$$

$$= \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

Probability of failure in the remaining N-r trials

Probability of obtaining successes on r attempts

Ordering of the successes and failures

r = non negative integer
0 < p < 1, p Real

Mean:

$$\mu = \sum_{r=0}^{N} \left[r \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} \right] = Np$$

Variance:

$$\sigma^2 = \sum_{r=0}^{N} \left[(r-\mu)^2 \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} \right] = Np(1-p) = Npq$$

Exercise: evaluate the sums !

Let us re-examine the formula for P(r,N,p):

$$P(r,N,p) = \binom{N}{r} p^r q^{N-r}$$

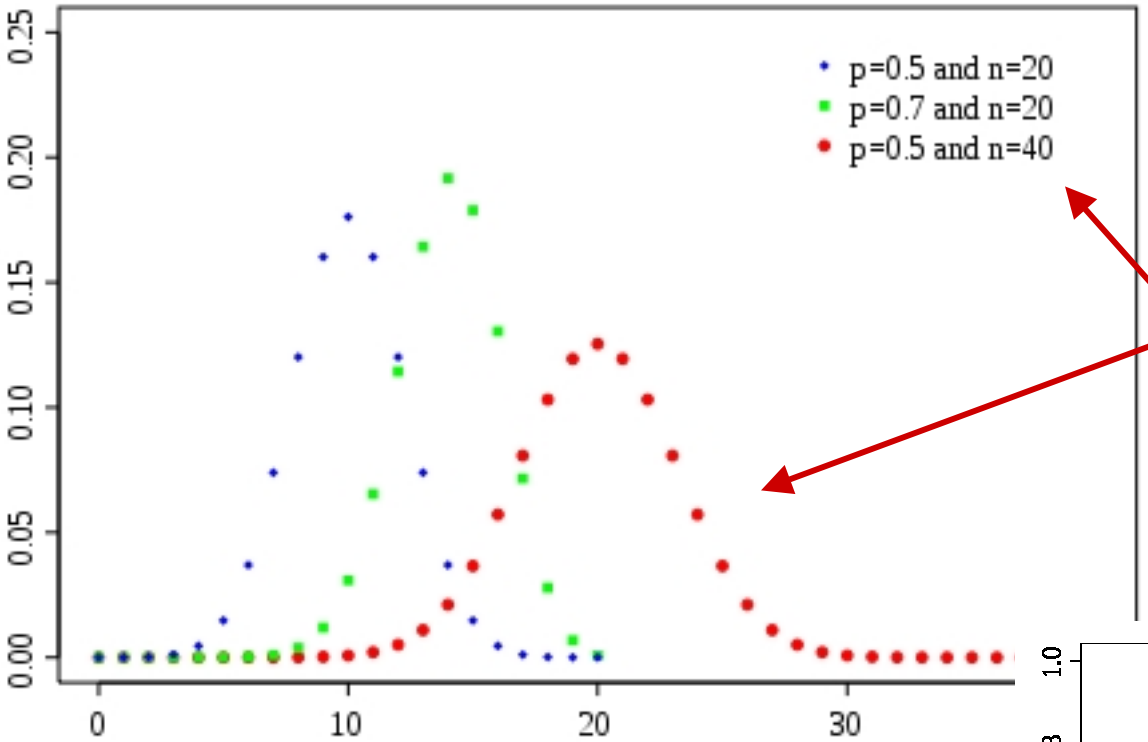
and suppose that p=q. Then the distribution is symmetric since there is nothing which can distinguish p from q

Property:

- if r and s are binomial with P(r,N,p) and P(s,M,p), then t=r+s is binomial with P(t,N+M,p)

Be careful !

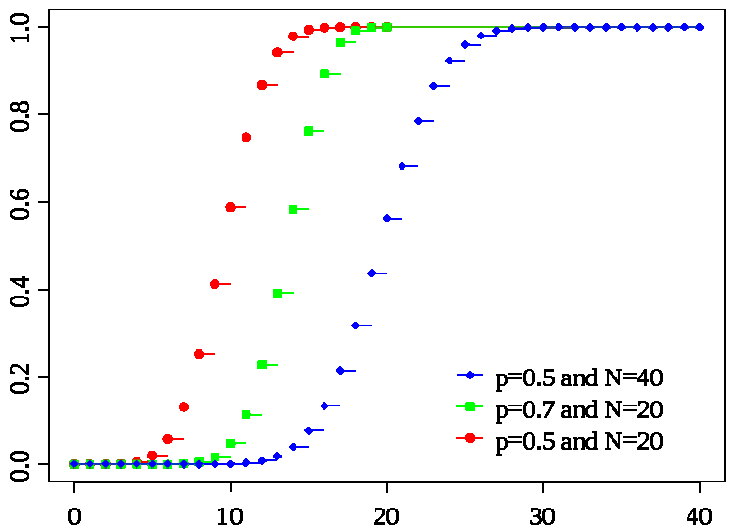
Binomial distribution for different p and N



Symmetric if p=q

Cumulative Binomial distribution

$$F(r,N,p) = \sum_{r=0}^{r} \binom{N}{r} p^r q^{N-r}$$



Where the binomial distribution apply ?

- ✓ Suppose we toss 10 coins in the air. The probability that exactly r ($\leq$ 10) coins land heads up is P(r,10,0.5), assuming that the probability of obtaining a head in a single toss is p=0.5.

- ✓ Throw a dice N times. What is the probability of obtaining a 6 on exactly r occasions ?

- ✓ The angles that the decay products from a given source make with some fixed axis are measured. If the expected distribution is known, what is the probability of observing r decays in the forward hemisphere ($\theta < \pi/2$) from a total sample of N decays ?

Poisson distribution

Consider you are interested in the radioactive decay of a given sample.

- ✓ The number of nuclei in one milligram of radioactive material is of the order of 10^{19}.
- ✓ To determine the decay rate, the number of disintegrating nuclei per unit time has to be measured.
- ✓ It is many order of magnitude smaller than the number of nuclei.
- ✓ The binomial distribution $P_B(x,n,p)$ correctly describes the probability of observing x events per unit time out of n possible events, provided each one has probability p of occurring.

However:

- ✓ The large number of possible events, n, makes exact evaluation impossible.
- ✓ In an experiment like the one of measuring the decay rate of a radioactive material neither the number of possible events, n, nor the probability, p, for each one is known.

In turn,

- ✓ We known the average number of observed events per unit time μ or its estimate, $\bar{x}$.
- ✓ $\mu \ll n$ and $p \ll 1$.
- ✓ The appropriated distribution is the binomial one in the limit $p \to 0$, $n \to \infty$ with $\mu = np = cte$.

Let us consider the binomial distribution written as

$$P_B(x,n,p) = \frac{1}{x!}\frac{n!}{(n-x)!}p^x(1-p)^{-x}(1-p)^n$$

- in the limit n → ∝

$$\frac{n!}{(n-x)!} = \underbrace{n(n-1)\ldots(n-x+1)}_{x \text{ terms}} \to n^x$$

then

$$\frac{n!}{(n-x)!}p^x \to (np)^x = \mu^x$$

Let us consider the binomial distribution written as

$$P_B(x,n,p) = \frac{1}{x!}\frac{n!}{(n-x)!}p^x(1-p)^{-x}(1-p)^n$$

- in the limit p → 0

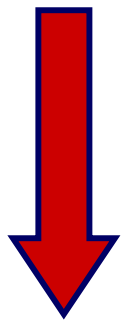
$$(1-p)^{-x} \sim (1+xp) \sim 1$$

and

$$\lim_{p\to 0}(1-p)^n = \lim_{p\to 0}\left[(1-p)^{1/p}\right]^{\mu} = \left(\frac{1}{e}\right)^{\mu}$$

Binomial distribution

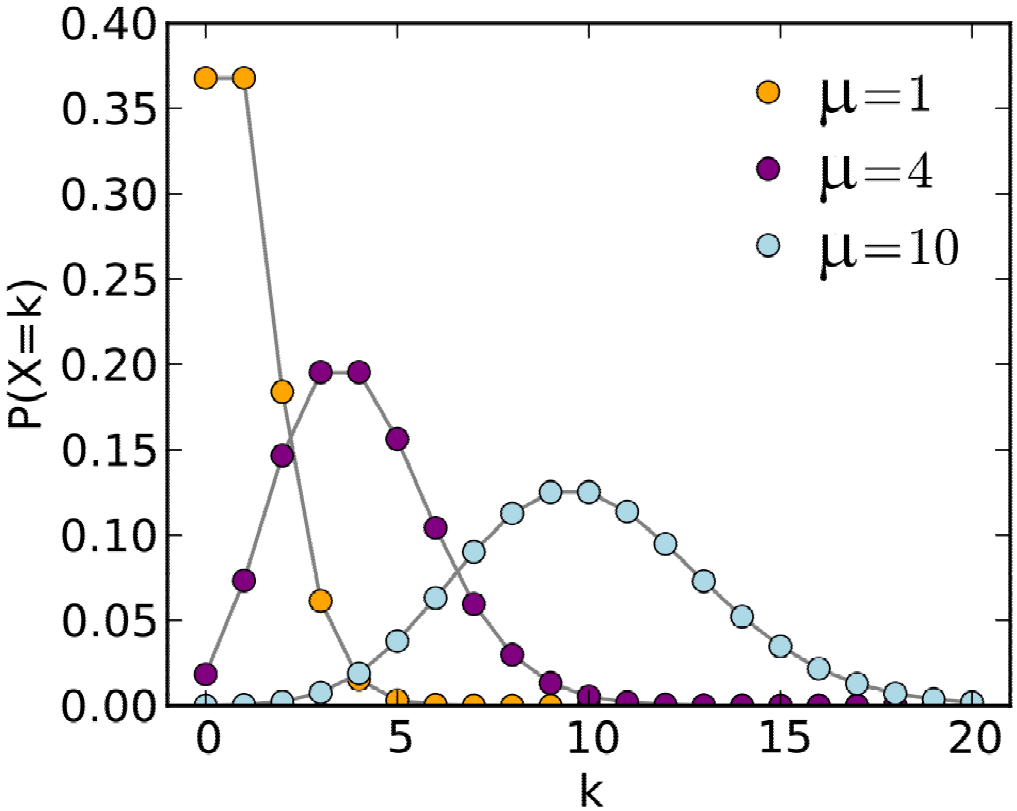
$$P_B(x,n,p) = \frac{1}{x!}\frac{n!}{(n-x)!}p^x(1-p)^{-x}(1-p)^n$$



$$\lim_{p\to 0, n\to\infty} P_B(x,n,p) = P_P(x,\mu) \equiv \frac{\mu^x}{x!}\exp\{-\mu\}$$

$\mu = np = cte.$

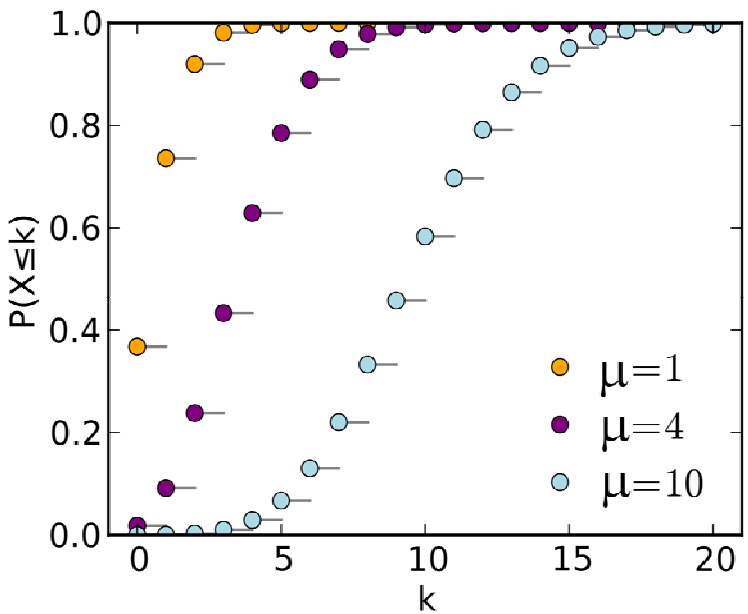
Poisson distribution



Poisson distribution for several values of μ

Cumulative Poisson distribution

$$F_P(x,\mu) = \exp\{-\mu\} \sum_{i\ 0}^{x} \frac{\mu^i}{i!}$$



Properties of the Poisson distribution

- ✓ discrete variable x
- ✓ $\mu > 0$ real
- ✓ support $x \in \{0,1,2,3,\ldots\}$
- ✓ mean $\bar{x} = \mu$
- ✓ variance $\sigma = \mu$
- ✓ sum of independent Poisson variables is Poisson

Exercise: evaluate μ and σ !

Example

Assume that you have a radioactive emitting source in a medium where there exist radioactive background emissions at a rate μ_b. The radioactive source emits at a ratio μ_x. Then

$$P_P(r_x, \mu_x t) = \frac{(\mu_x t)^{r_x}}{r_x!} \exp\{-\mu_x t\}$$

Radioactive source

$$P_P(r_b, \mu_b t) = \frac{(\mu_b t)^{r_b}}{r_b!} \exp\{-\mu_b t\}$$

Background

Notice:

- μ_x = average number of events observed per unit time
- $\mu_x t$ = average number of events observed in the time t
- similar for μ_b

Question: what is the probability of emission for the source plus background ?

Assume that $r = r_x + r_b$ = number of particles emitted by source + background

$$P_P(r_x, \mu_x t; r_b, \mu_b t) = \sum_{r_b=0}^{r} P_P(r - r_b, \mu_x t) P_P(r_b, \mu_b t)$$

$$= \sum_{r_b=0}^{r} \frac{(\mu_x t)^{r-r_b}}{(r-r_b)!} \exp\{-\mu_x t\} \frac{(\mu_b t)^{r_b}}{r_b!} \exp\{-\mu_b t\}$$

$$= \frac{1}{r!} \sum_{r_b=0}^{r} \frac{r!}{(r-r_b)!r_b!} (\mu_x t)^{r-r_b} (\mu_b t)^{r_b} \exp\{-(\mu_x t + \mu_b t)\}$$

Binomial formula

$$\sum_{r_b=0}^{r} \frac{r!}{(r-r_b)!r_b!} (\mu_x t)^{r-r_b} (\mu_b t)^{r_b} = (\mu_x t + \mu_b t)^r$$

Assume that $r = r_x + r_b$ = number of particles emitted by source + background

$$P_P(r_x,\mu_x t;r_b,\mu_b t) = \sum_{r_b=0}^{r} P_P(r-r_b,\mu_x t)P_P(r_b,\mu_b t)$$

$$= \sum_{r_b=0}^{r} \frac{(\mu_x t)^{r-r_b}}{(r-r_b)!}\exp\{-\mu_x t\}\frac{(\mu_b t)^{r_b}}{r_b!}\exp\{-\mu_b t\}$$

$$= \frac{1}{r!}\sum_{r_b=0}^{r} \frac{r!}{(r-r_b)!r_b!}(\mu_x t)^{r-r_b}(\mu_b t)^{r_b}\exp\{-(\mu_x t+\mu_b t)\}$$

$$P_P[r,(\mu_x+\mu_b)t] = \frac{1}{r!}(\mu_x t+\mu_b t)^r \exp\{-(\mu_x t+\mu_b t)\}$$

Where the Poisson distribution apply ?

✓ experiments in which the possible values of the data are strictly bounded on one side but not the other

✓ decay of a radioactive material over a period of time which is significant compared to the lifetime of the sample

✓ the number of interactions between a thin target and an intense incident pulse of particles

✓ the number of particles detected by a counter in a time t (the particle flux Φ and detector efficiency have to be independent of time, also if the dead-time is τ, $\Phi\tau << 1$)

Uniform distribution

Is the simplest distribution you can think in a single continuous variable

Definition:

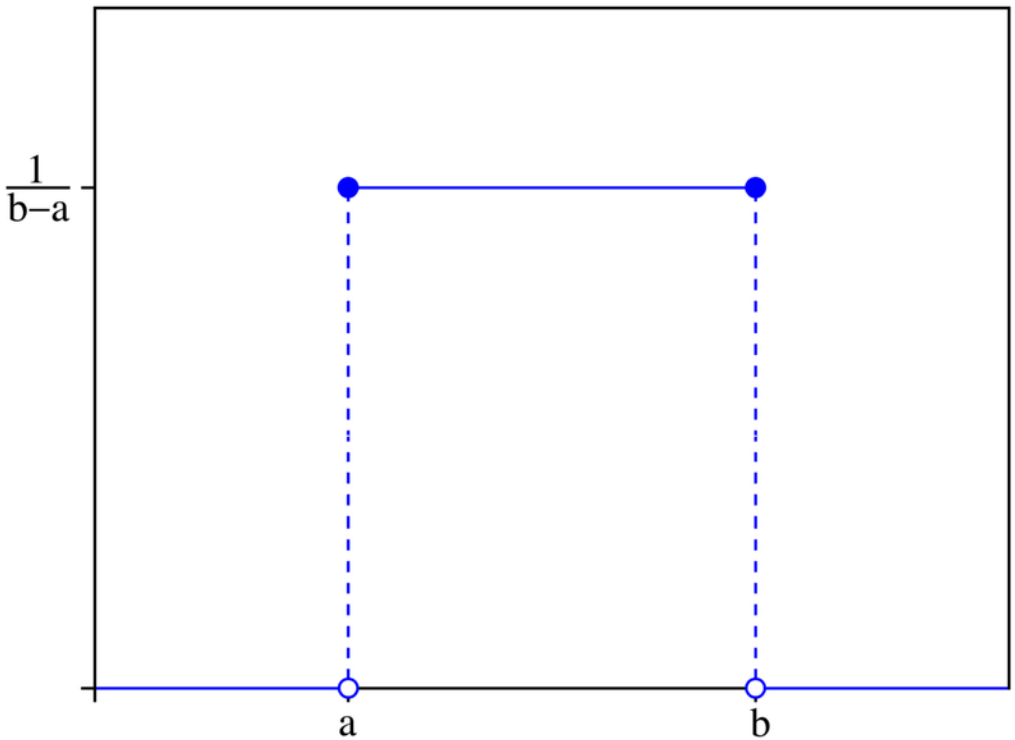
$$P(x) = \frac{1}{b-a} \qquad a,b \ \ Real;\ a \le x \le b$$

Mean:

$$\mu = \int_a^b xP(x)dx = \frac{1}{2}(a+b)$$

Variance:

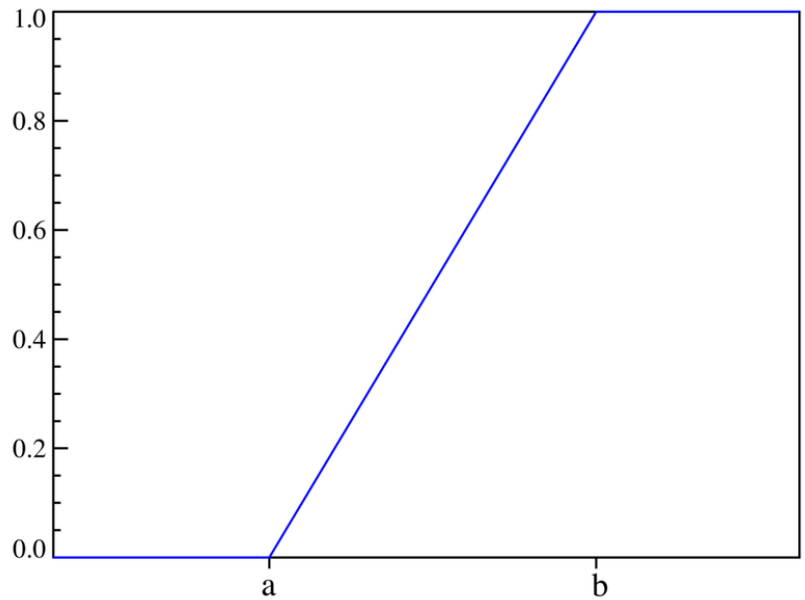
$$\sigma^2 = \int_a^b (x-\mu)^2 P(x)dx = \frac{1}{12}(b-a)^2$$



Uniform distribution

Cumulative uniform distribution

$$F(x) = \frac{x-a}{b-a}$$



Gaussian distribution

General form:

Continuous variable

Mean $\mu = \int_{-\infty}^{+\infty} xP(x)dx$

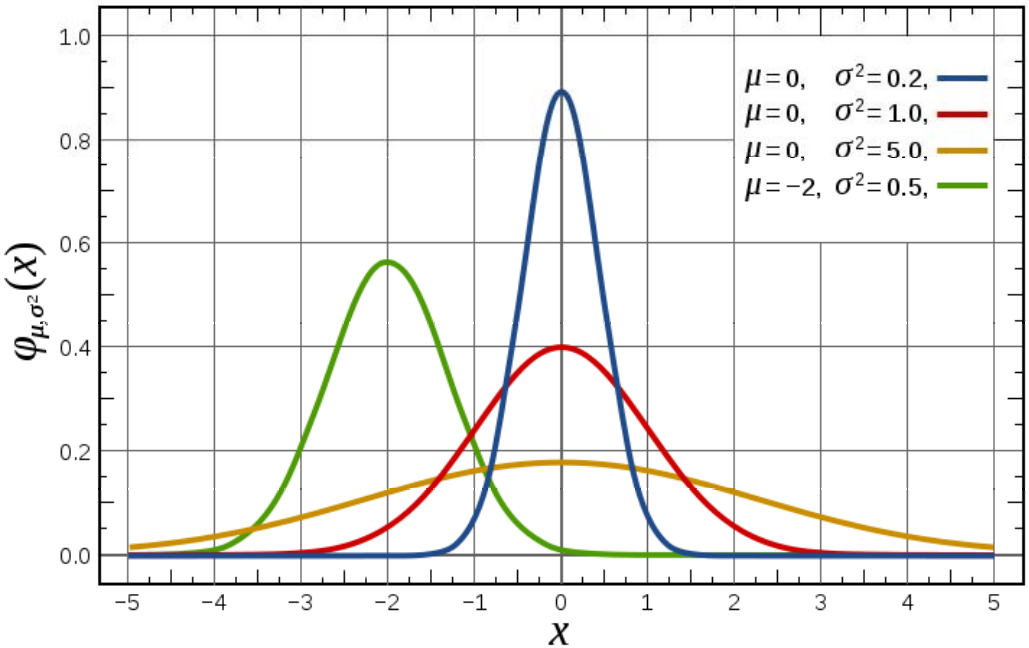
$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x-\mu)^2/2\sigma^2\}$$

Variance $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x)dx$

Normalization

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

Limiting case of the binomial distribution when n → ∝, np >>1

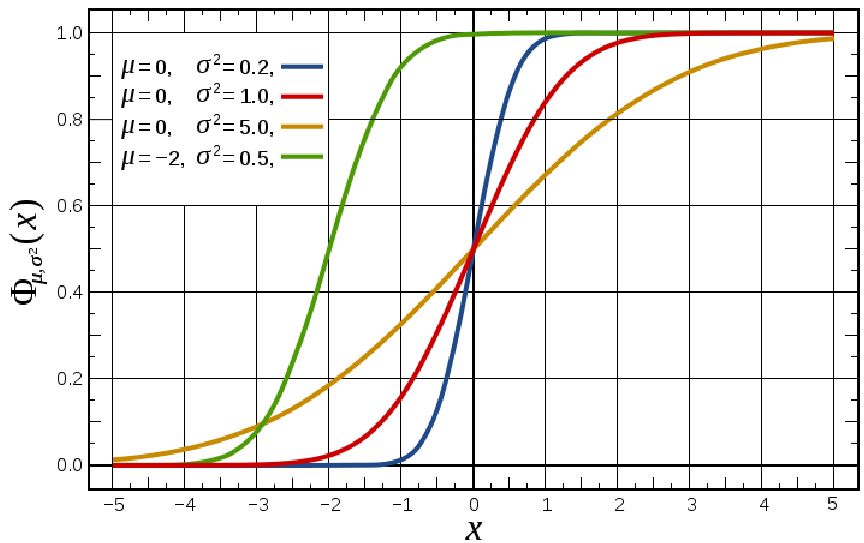


Gaussian distribution

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x-\mu)^2/2\sigma^2\}$$

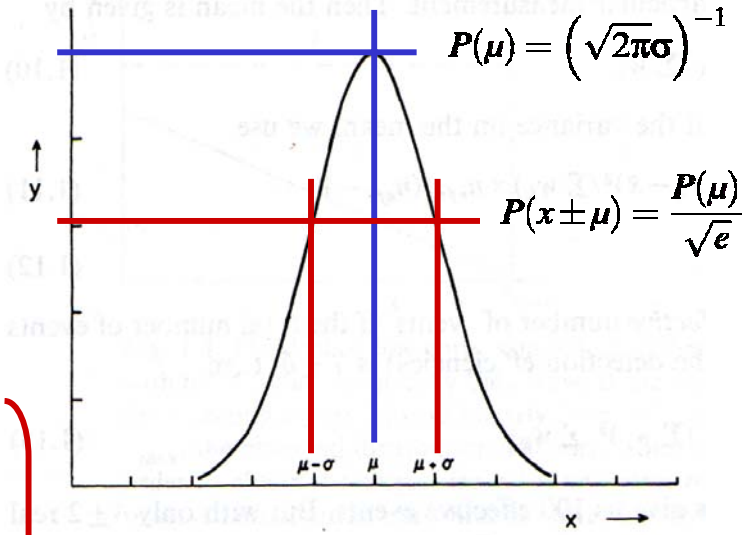
Cumulative Gaussian distribution

$$F(x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$



Properties:

- ✓ P(x) is symmetric around μ
- ✓ The inflection points of P(x) are at $x = \mu \pm \sigma$
- ✓ P(x) is infinitely differentiable
- ✓ if x is Gaussian with mean μ and variance σ^2, then ax + b is Gaussian with mean $a\mu + b$ and variance $a^2\sigma^2$
- ✓ if x, y are Gaussian with mean μ_1, μ_2 and variance σ_1^2, σ_2^2, then z = ax + by is Gaussian with mean $a\mu_1+b\mu_2$ and variance $\sigma^2 = a^2\sigma_1^2 + b^2\sigma_2^2$



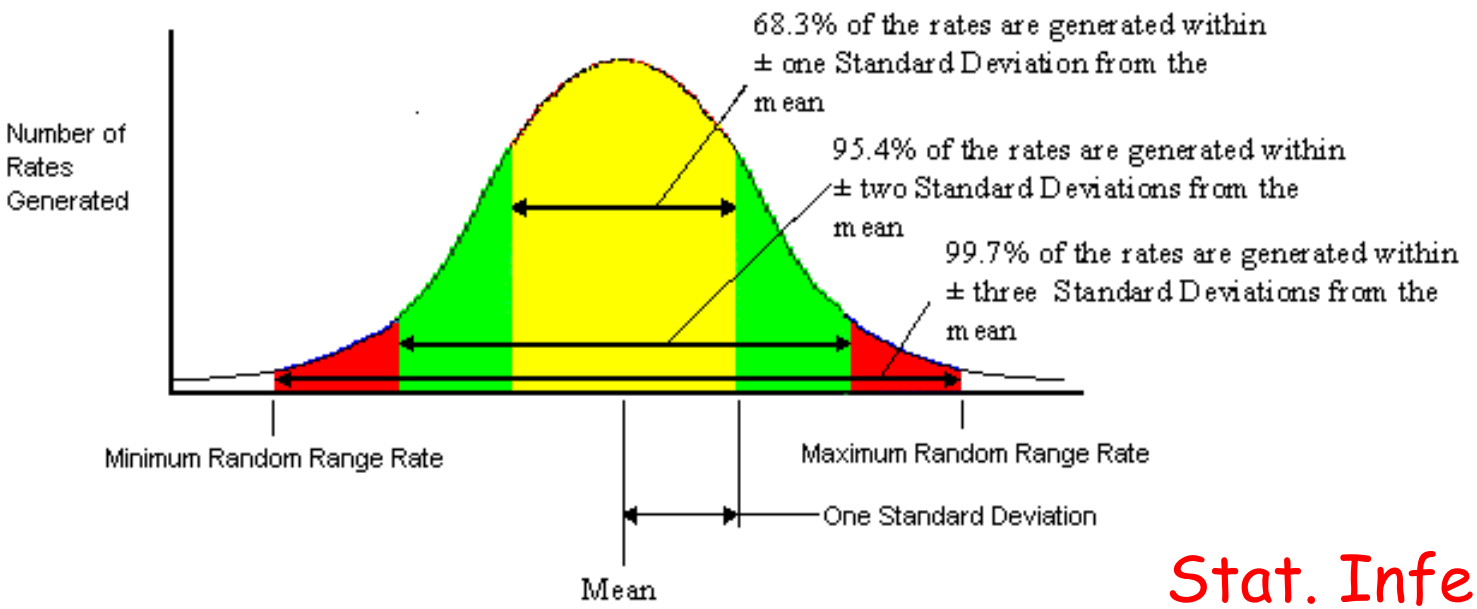
Exercise: prove it !

Standard form

$$x' = (x - \mu)/\sigma$$

$$P(x') = \frac{1}{\sqrt{2\pi}} \exp\{-(x')^2/2\}$$

Note also that as x is a continuous random variable, the probability that the value of a random observation will fall in the interval dx is dP(x) = P(x) dx



Stat. Infer.

Where the Gaussian distribution apply ?

- ✓ The result of repeating an experiment many times produces a spread of answers whose distribution is approximately Gaussian.
- ✓ If the individual errors contributing to the final answer are small, the approximation to a Gaussian is especially good.

The Central Limit Theorem

✓ It gives conditions under which the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed.

✓ Provides a partial explanation for the prevalence of the Gaussian distribution in the real world.

✓ Justifies the approximation of large-sample statistics to the Gaussian distribution in controlled experiments.

✓ The first version of the theorem is due to Abraham de Moivre. It was published in 1733.

Theorem (of the central limit)

Let $\{x_1, x_2, \ldots, x_n\}$ be a sequence of n independent and identically distributed random variables, each with mean μ and variance $\sigma^2 > 0$. Then increasing the size n of the sequence, the sample average of these random variables approaches the Gaussian distribution with mean m and variance σ^2/n, irrespective of the individual distributions of the x_i (The proof is remarkable simple using characteristic functions)

Example: Gaussian random number generator

Suppose you have an ordinary random number generator which delivers numbers uniformly distributed over the interval [0,1] (with mean = 0.5 and variance $\sigma^2 = 1/12$).

Then the variable

$$z = \frac{\sum_{i=1}^{n} x_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$$

is Gaussian with $\mu = 0$ and $\sigma^2 = 1$

Exercise: construct a computer program to generate random numbers distributed according to a Gaussian with $\mu = 4$ and $\sigma^2 = 0.5$

Law of large numbers

Let $\{x_1, x_2, \ldots, x_n\}$ a set of random variables identically distributed and with mean μ. Then the arithmetic mean, given by

$$\bar{x}_n = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

then,

$$\bar{x}_n \to \mu \quad for \quad n \to \infty$$

Two forms:

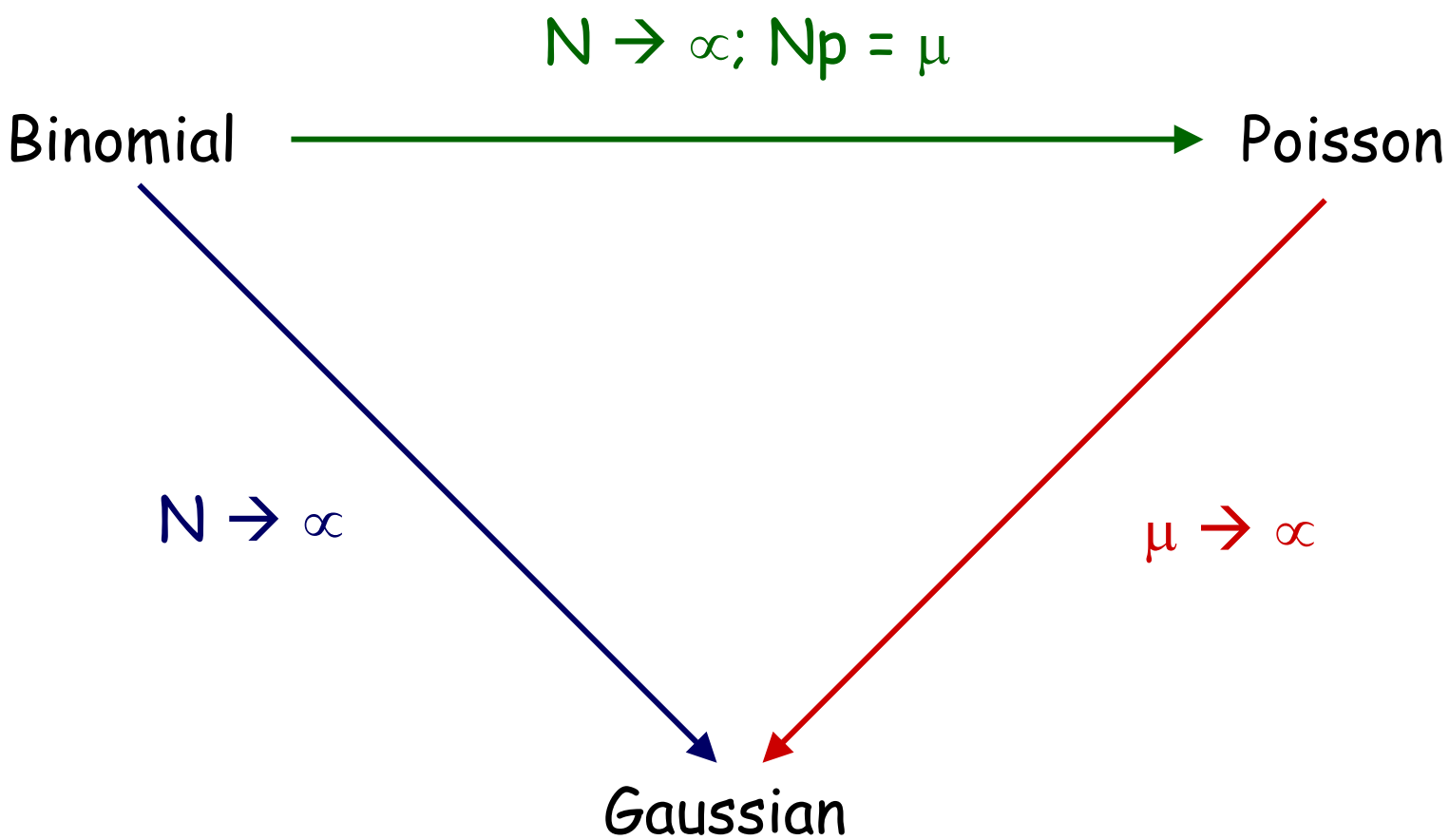
- ✓ Weak form: convergence in probability

$$\lim_{n\to\infty} Prob\left(|\bar{x}_n - \mu| > \varepsilon\right) = 0$$

- ✓ Strong form: almost sure convergence

$$\lim_{n\to\infty} Prob\left(|\bar{x}_n - \mu| > \varepsilon\right) < \delta$$

Limiting cases

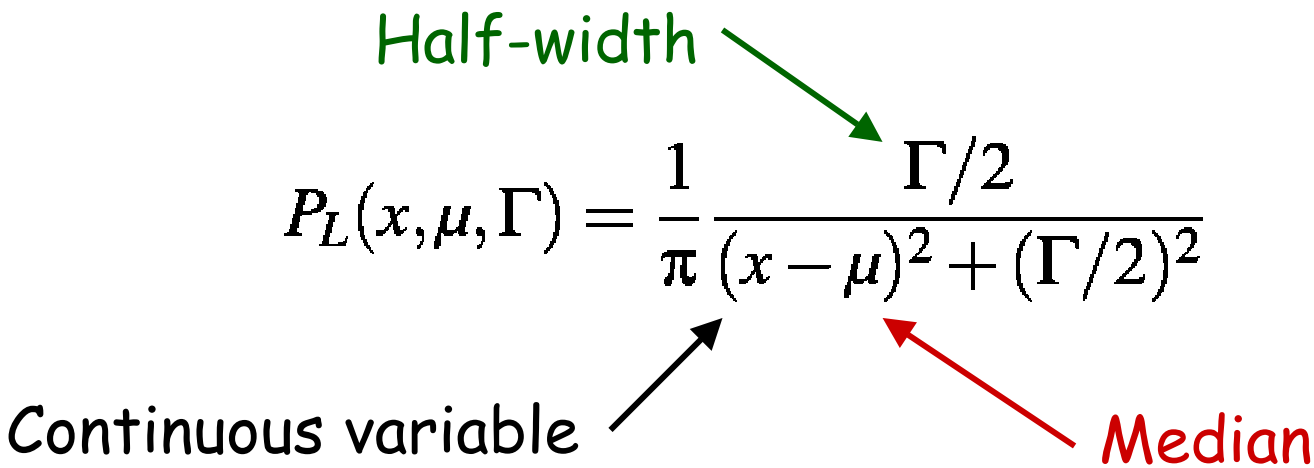


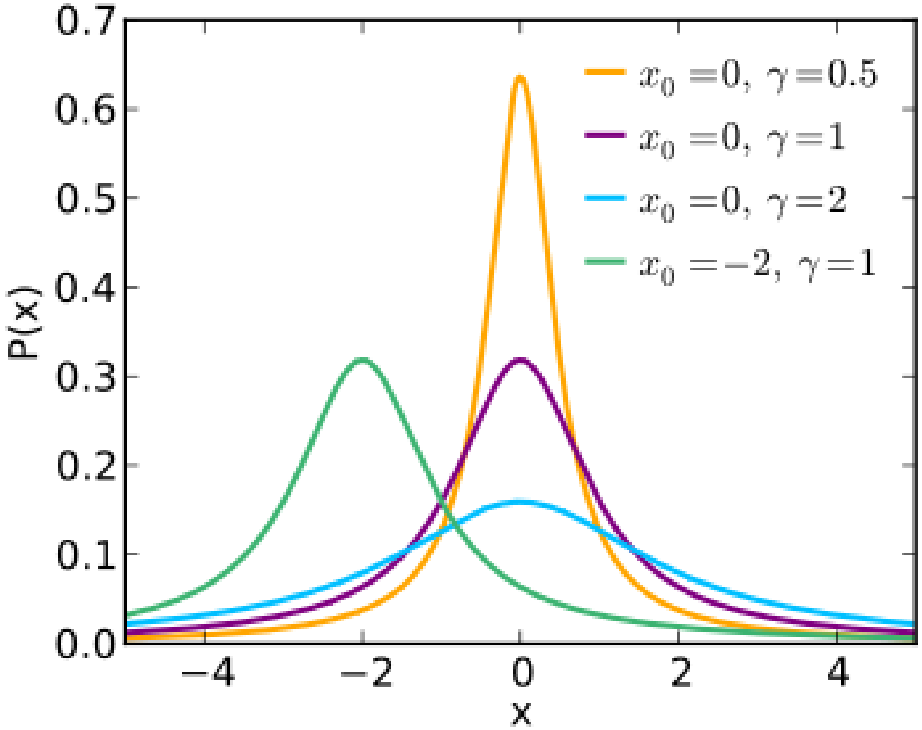
Cauchy distribution

✓Known also with the name of Lorentzian distribution or Breit-Wigner distribution (among physicists)

✓ Important in physics because it is the solution to the differential equation describing forced resonance.

✓ Appropriated distribution to describe resonance behavior.



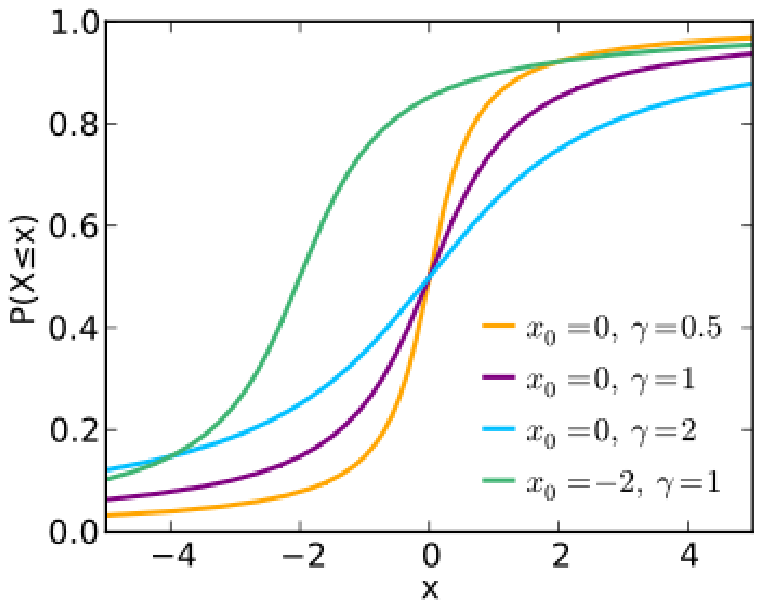


Cauchy distribution

$$P_L(x,\mu,\Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2+(\Gamma/2)^2}$$

Cumulative Cauchy distribution

$$F(x,\mu,\Gamma) = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\Gamma/2}\right)+\frac{1}{2}$$



Properties:

- ✓ symmetric around the median μ
- ✓ has no mean, variance or higher moments defined.

Consider for instance:

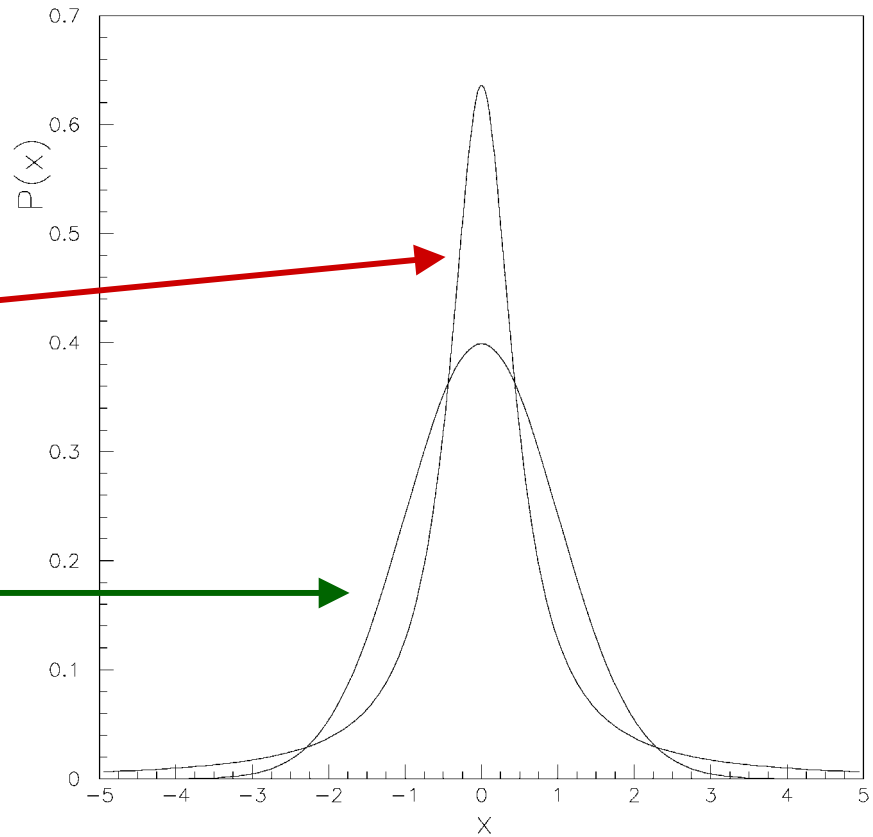
$$\int_{-\infty}^{+\infty} xP_L(x,\mu,\Gamma)\,dx = \int_{0}^{+\infty} xP_L(x,\mu,\Gamma)\,dx - \int_{-\infty}^{0} |x|P_L(x,\mu,\Gamma)\,dx$$

If at most one of the two terms in the right hand side of equation above is infinite, then you can define it to be the same than the other. But in the case of the Cauchy distribution, both integrals on the right are infinite. The same behavior is seen for each moment of the distribution

Properties:

Cauchy distribution

Gauss distribution



$$\sigma = \Gamma = 1;\ \mu_G = \mu_L = 0$$

Sampling distributions

χ^2 - distribution:

✓ χ^2 – distribution with k degrees of freedom is the distribution of a sum of squares of k independent gaussian normal random variables.

✓ It is widely used in connection with the problems of goodness of fit, hypothesis testing and the construction of confidence intervals, and many other statistical tests.

Definition:

If $X_1, \ldots, X_k$ are independent standard normal random variables (i.e. gaussian variables with $\mu = 0$ and $\sigma^2 = 1$), then the sum of their squares

$$\chi^2 = \sum_{i=1}^{k} X_i^2$$

is distributed according to

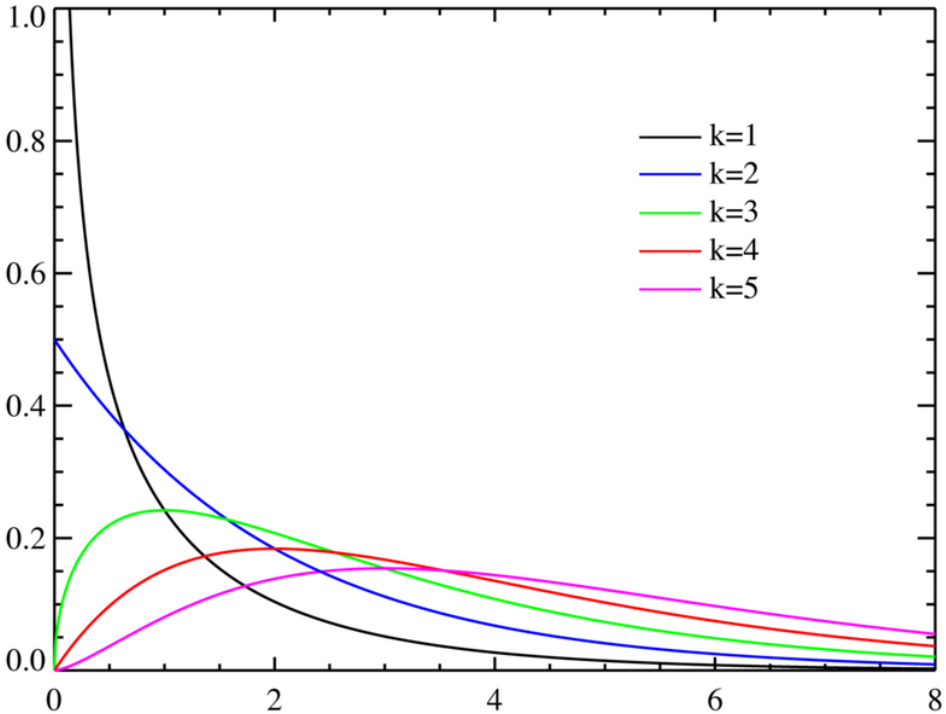
degrees of freedom

$$P_{\chi^2}(\chi^2, k) = \frac{1}{2^{\frac{1}{2}k}\Gamma(k/2)} (\chi^2)^{\frac{1}{2}k-1} \exp\left\{-\frac{1}{2}\chi^2\right\}$$

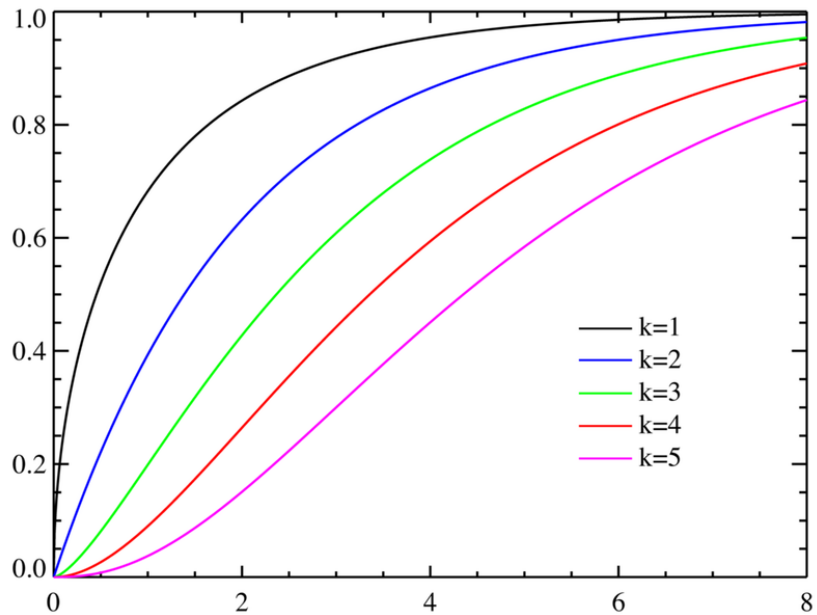
$\Gamma(z) = (n-1)!$ n integer

$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t}\, dz$ z complex

Gamma function



χ^2-distribution for k degrees of freedom



Cumulative χ^2-distribution

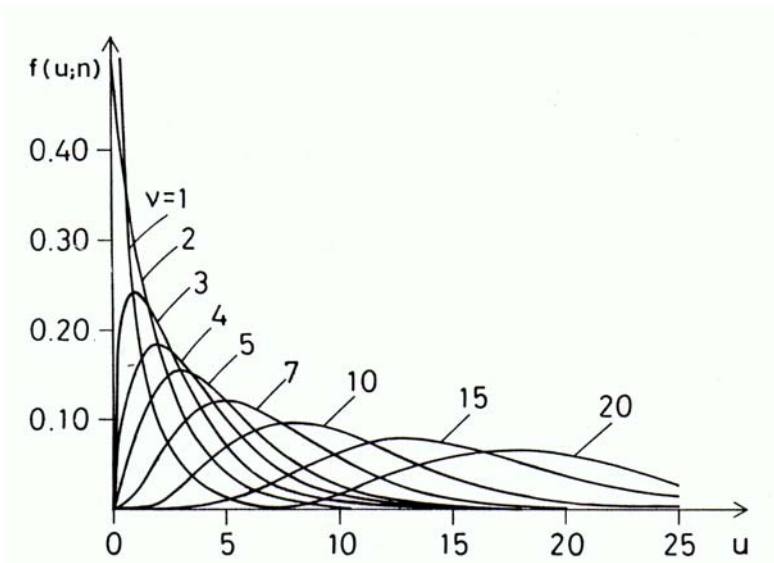
$$F_{\chi^2}(x) = \frac{1}{\Gamma(k/2)}\gamma(k/2, x/2)$$

$$\gamma(s,x) = \int_0^x t^{s-1} e^{-t}\, dt$$

Properties:

- ✓ The mean of the χ^2 distribution is $\mu = k$
- ✓ The variance of the χ^2 distribution is $\sigma^2 = 2k$
- ✓ the sum of independent χ^2 variables is also a χ^2 variable
- ✓ The χ^2 distribution approaches the Gaussian distribution as $n \rightarrow \infty$

for k = 20, the χ^2 distribution is quite close to a Gaussian distribution



Student´s t distribution

✓ Used to estimate the mean of a gaussian distributed population when the sample size is small.
✓ Used in the Student´s test for the statistical significance of the difference between two sample means and for confidence intervals for the difference between two population means.
✓ Useful also when the population standard deviation is unknown and has to be estimated from data.

Definition:

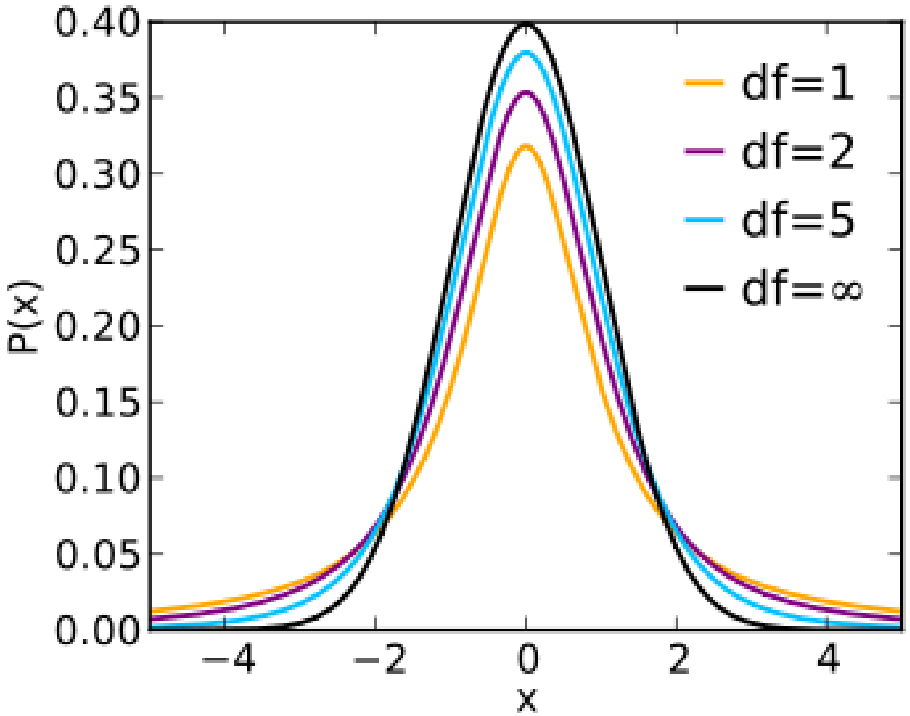
Let us define the variable

$$t = \frac{x}{\sqrt{u/\nu}}$$

where x is normally distributed with mean zero and variance 1, and u is χ^2 with ν degrees of freedom. Then, if x and u are independent, t has a p.d.f. given by

$$P_t(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{1}{2}\nu\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

If x is normal with mean μ, P_t is the non-central t-Student's distribution

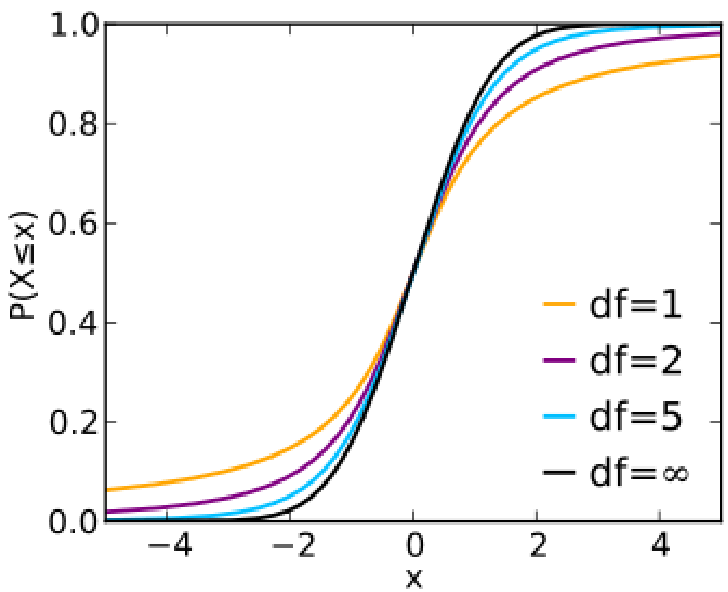


Student's t-distribution

if $\nu \to \infty$,
$P_t(t,\nu) \to$ Gaussian
if $\nu = 1$,
$P_t(t,n) \to$ Cauchy

Cumulative t-Student's distribution

$$F_t(t,\nu) = \frac{1}{2} + t\Gamma\left(\frac{\nu+1}{2}\right)$$



Moments of a distribution

The n^{th} moment of a distribution is defined as

$$\mu_n = \int_{-\infty}^{+\infty} (x-\mu)^n P(x) dx$$

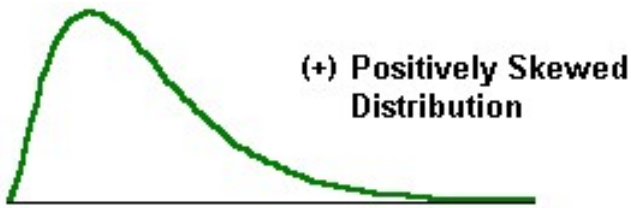
From them there are a few which are particularly interesting because they can be used to quantify properties of the distributions. For instance

$$\mu = \int_{-\infty}^{+\infty} xP(x)dx$$ mean = first central moment

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x)dx$$ variance = second moment

Skewness

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{1}{\sigma^3}\int_{-\infty}^{+\infty}(x-\mu)^3 P(x)dx$$



$\gamma_1 > 0$

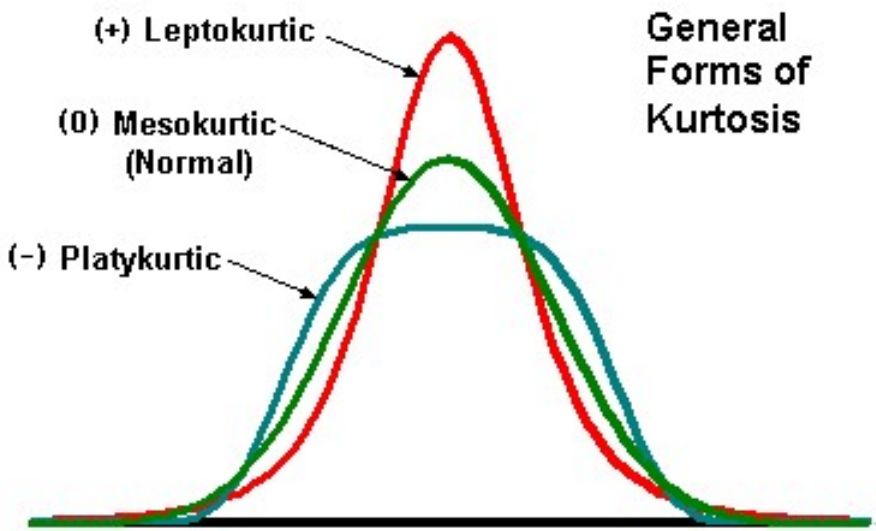


$\gamma_1 < 0$

Kurtosis

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{1}{\sigma^4}\int_{-\infty}^{+\infty}(x-\mu)^4 P(x)dx - 3$$

where the (–3) at the end is to make the kurtosis of the Gaussian distribution equal to zero



Error matrix

1- Assume that x and y are two uncorrelated gaussian variables, then

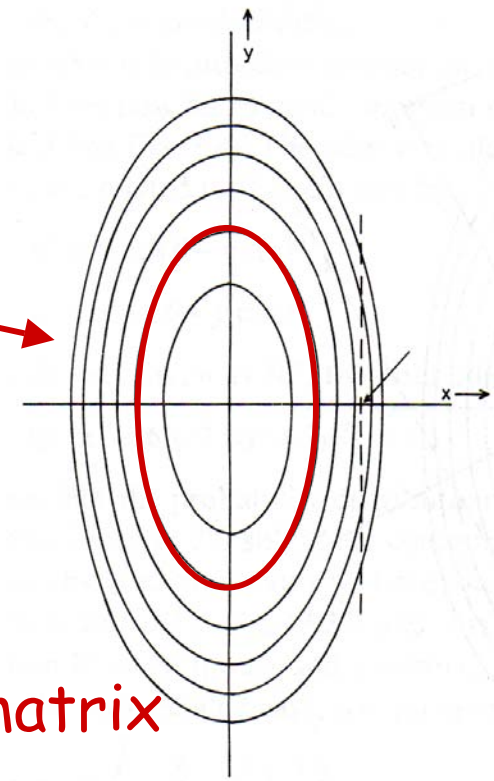
$$\left.\begin{aligned} P(x) &= \frac{1}{\sqrt{2\pi}\sigma_x}\exp\{-\frac{1}{2}\frac{x^2}{\sigma_x^2}\} \\ P(y) &= \frac{1}{\sqrt{2\pi}\sigma_y}\exp\{-\frac{1}{2}\frac{y^2}{\sigma_y^2}\} \end{aligned}\right\}$$

Given that x and y are independent variables, it follows that

$$P(x.y) = P(x)P(y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right\}$$

with

$$\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}=1$$



or, in matrix form

$$\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1$$

Inverse error matrix

Error matrix

$$\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

- The diagonal term are the variance of x and y respectively
- Off-diagonal terms are the covariance. Zeroes indicate no correlation among x and y
- The error matrix is a symmetric matrix
- The general definition, even for non-gaussian distributions, is

$$\mathcal{M}_{ij} = \int_{-\infty}^{+\infty} (x_i - \bar{x}_i)(x_j - \bar{x}_j) P(x) dx$$

2- Correlated variables

1- Start with the uncorrelated variables case and perform a clockwise rotation of an angle θ, then

$$\left.\begin{aligned} x' &= x\cos\theta - y\sin\theta \\ y' &= x\sin\theta + y\cos\theta \end{aligned}\right\}$$

2- once you rename the variables to x and y, you obtain the general form of a Gaussian in two variables (Exercise: do it !)

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\frac{1}{\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y}\right)\right\}$$

Correlations

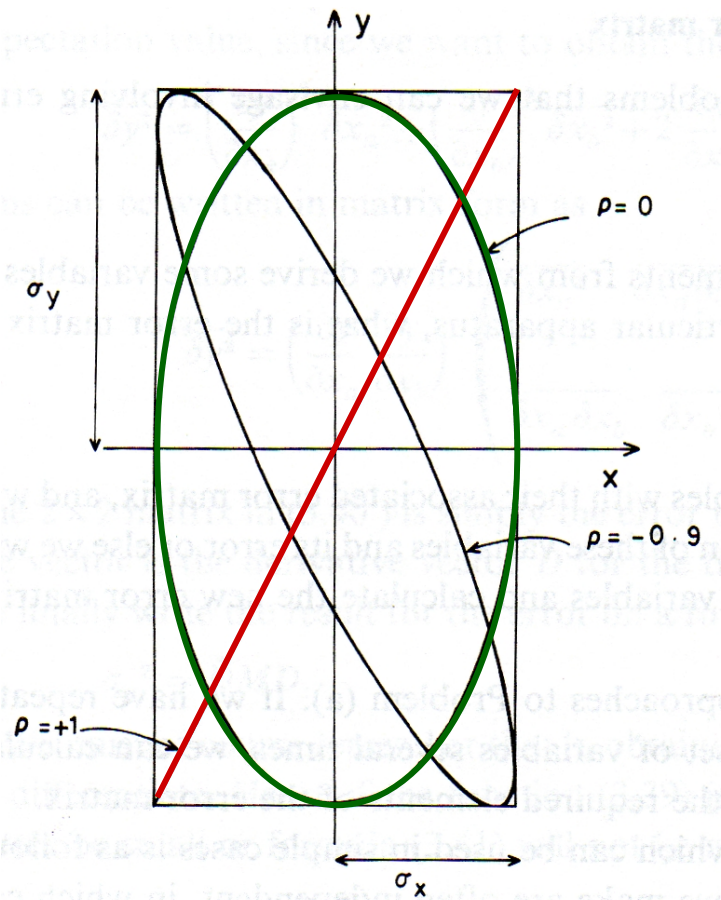
Error matrix

ρ "measures" the correlation among the variables x and y

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} \qquad |\rho| \leq 1$$

✓= 0 → no correlation (independent variables)

✓ = ± 1 full correlation (ellipse → straight line)



Error Matrix

$$\begin{pmatrix} \sigma_x^2 & cov(x,y) \\ cov(x,y) & \sigma_y^2 \end{pmatrix}$$

Example of the use of the error matrix:

Consider the case in which we want to determine the quantity $y = y(x_1,x_2)$ and its error, given in terms of x_1 and x_2 for which we know their error matrix. We have to calculate the value of $\langle (y - \overline{y})^2 \rangle = \delta y^2$. Then using (errors) we obtain

$$\delta y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \delta x_1^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \delta x_2^2 + 2\frac{\partial y}{\partial x_1}\frac{\partial y}{\partial x_2}\delta x_1 \delta x_2$$

or, in matrix form

$$\delta y^2 = \left(\frac{\partial y}{\partial x_1} \; \frac{\partial y}{\partial x_2}\right) \begin{pmatrix} \delta x_1^2 & \delta x_1 \delta x_2 \\ \delta x_1 \delta x_2 & \delta x_2^2 \end{pmatrix} \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{pmatrix}$$

Error matrix

3rd lecture

Back

The resolution function

The data which we have to deal are subject to observational errors

- a variable whose true value is x can be observed at x'
- events can be observed where ideally there should be none
- etc.

Then, before comparing with data, the ideal theoretical distribution has to be adjusted to correspond to some modified or expected distribution for the measurable quantity.

Definition:

Suppose that the true p.d.f. for the variable x is f(x;Θ), where Θ is a given set of parameters. A function r(x′,x) which describes the distribution of the measured quantity x′ for a given true value of x is called resolution function and the p.d.f. for x′ is given by

$$f'(x') = \int f(x;\Theta) r(x',x) dx$$

f′(x′) is called the resolution transform.

➡ If the resolution function is strongly varying about the peak of the true p.d.f. in such a way that $r(x',x)$ can be approximated by a $\delta(x'-x)$ then

$$f'(x') = f(x; \Theta)$$

so that, for a experiment with very good precision, the shape of the p.d.f. is unaffected.

➡ On the other extreme, if the precision is so poor that the $f(x; \Theta)$ can be approximated by a $\delta(x-x_0)$, with x_0 the true value of x, then

$$f'(x') = r(x',x_0)$$

meaning that the observable is the resolution function itself.

In most cases it is quite reasonable to assume that repeated measurements of the variable x would lead to observations x′ which are gaussian distributed about x. Then the resolution function is

$$r(x',x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(x'-x)^2}{\sigma^2} \right\}$$

where σ measures the experimental uncertainty and is often called resolution width.

The resolution function depends exclusively on the apparatus used for measuring. Its determination may be complicated and has to be done experimentally.

Example:

Suppose you are measuring a random variable x whose distribution is exponential,

$$f(x;\lambda) = \lambda \exp\{-\lambda x\}\ ,\ \ 0 \leq x < \infty$$

Assume also that the resolution function is a Gaussian with resolution width σ,

$$r(x',x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(x'-x)^2}{\sigma^2} \right\}$$

The effect of the resolution function on the true p.d.f. is given by

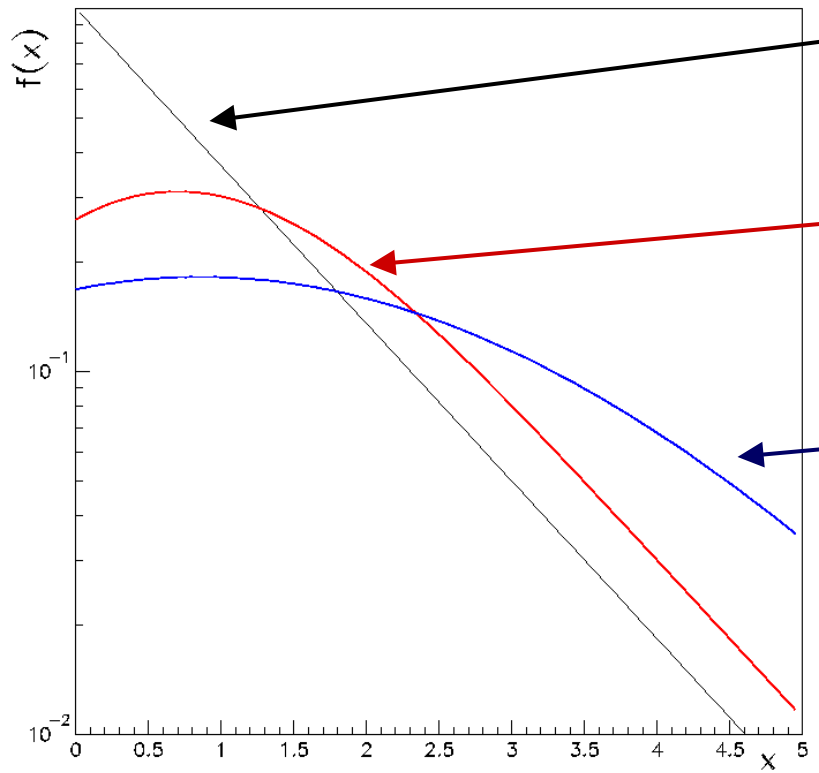
$$f(x') = \int_0^\infty \lambda \exp\{-\lambda x\} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(x'-x)^2}{\sigma^2} \right\}$$

then, after integration,

$$f(x') \sim \exp\left\{ \frac{1}{2}\lambda^2\sigma^2 \right\} \cdot G\left(\frac{x'}{\sigma} - \lambda\sigma \right) \cdot \lambda \exp\{ -\lambda x' \}$$

Cumulative Gaussian distribution

Effect of the resolution function



$\lambda \exp[-\lambda x]$

$\sigma = 1$

$\sigma = 2$

Detection efficiency

In many situations the measurement apparatus does not have the same sensitivity for all type of measurements.

Consider as an example a detector of the kind used in high energy physics experiments. They are usually huge detectors composed by several subsystems and the detector efficiency depends on:

- location of the interaction point
- momentum of the particles
- emission angles of the particles
- etc.

✓ The inability for perfect detection can be serious for many experiments.

- It always implies a loss of information.
- and consequently less reliable information.

✓ However, a perfect detector can not be achieve

- cost limitations
- time consuming
- technological limitations
- etc.

✓ As a result, all possible sources of information loss and systematic effects must be checked.

Symptoms of inefficiency problems are (among others)

- ✓ Distributions mathematically defined over non-observable values of the physical variable.
- ✓ Problems associated to the finite geometry of the detector.
- ✓ Non-uniformity in the detection of events over all the detector volume.
- ✓ etc.

Detector inefficiency can be handled according to two essentially different approaches

- ✓ modification of ideal p.d.f.
- ✓ weighting of observed events

Modification of ideal p.d.f.

- Suppose we observe a variable x to estimate the parameters of the ideal p.d.f., $f(x;\Theta)$.

- Suppose also that what we observe is some distorted p.d.f., $f'(x;\Theta)$, which is related to $f(x;\Theta)$ through the detector efficiency, $D(x,y)$.

Modification of ideal p.d.f.

- Suppose we observe a variable x to estimate the parameters of the ideal p.d.f., $f(x;\Theta)$.

- Suppose also that what we observe is some distorted p.d.f., $f'(x;\Theta)$, which is related to $f(x;\Theta)$ through the detector efficiency, $D(x,y)$.

The detector efficiency depends on the variable x, but also on additional variables, y, which may be dependent on x

Modification of ideal p.d.f.

- Suppose we observe a variable x to estimate the parameters of the ideal p.d.f., $f(x;\Theta)$.

- Suppose also that what we observe is some distorted p.d.f., $f'(x;\Theta)$, which is related to $f(x;\Theta)$ through the detector efficiency, D(x,y).

- Then, in general

$$f'(x,\Theta) = \frac{\int f(x;\Theta) D(x,y) P(y|x) dy}{\int\int f(x;\Theta) D(x,y) P(y|x) dy dx}$$

Modification of ideal p.d.f.

- Suppose we observe a variable x to estimate the parameters of the ideal p.d.f., $f(x;\Theta)$.

- Suppose also that what we observe is some distorted p.d.f., $f'(x;\Theta)$, which is related [illegible] detector efficiency, D(x,y).

- Then, in general

conditional probability of y, given x

$$f'(x,\Theta) = \frac{\int f(x;\Theta) D(x,y) P(y|x) dy}{\int\int f(x;\Theta) D(x,y) P(y|x) dy dx}$$

to ensure correct normalization

- The detection efficiency, D(x,y), or acceptance, is a property of the detecting system and can, in principle be assumed known or measurable.
- The conditional probability P(y|x) depends on physical hypothesis that may or may not be known. If not known, must be inferred from data.

In essence, the idea behind this method is to apply corrections to the theoretical model, leaving data as they are observed

Example:

suppose you want to determine the decay constant of a given unstable particle. The decaying particles decay according to the distribution

$$f(x;\lambda) = \exp\left\{ -\lambda t \right\}$$

Assume also that the detector efficiency is unity in the interval $t_{min} < t < t_{max}$, then the observed p.d.f. will be

$$f'(x;\lambda) = \frac{\exp\left\{ -\lambda t \right\}}{\int_{t_{min}}^{t_{max}} \exp\left\{ -\lambda t \right\} dt} = \frac{\exp\left\{ -\lambda t \right\}}{\exp\left\{ -\lambda t_{max} \right\} - \exp\left\{ -\lambda t_{min} \right\}}$$

Weighting of observed events

- only approximately correct, but more frequently used in practice.
- weights individual observed events by the reciprocal of the detection efficiency.
- the p.d.f. of the weighted events is compared to the true p.d.f.
- for a event observed at a particular value x_i of the variable, the corrected number of events is w_i, with

$$w_i = \frac{1}{D(x_i, y_i)}$$

where D is the inverse of the detection probability (acceptance)

Example:

The E791 experiment at FNAL measured the asymmetry in the production of $\Lambda_c^{\pm}$ both as a function of x_F (=$2p_l/\sqrt{s}$) and p_T^2, in the decay channel $\Lambda_c \rightarrow pK\pi$. The asymmetry is defined as

$$A = \frac{N_{\Lambda_c^+} - N_{\Lambda_c^-}}{N_{\Lambda_c^+} + N_{\Lambda_c^-}}$$

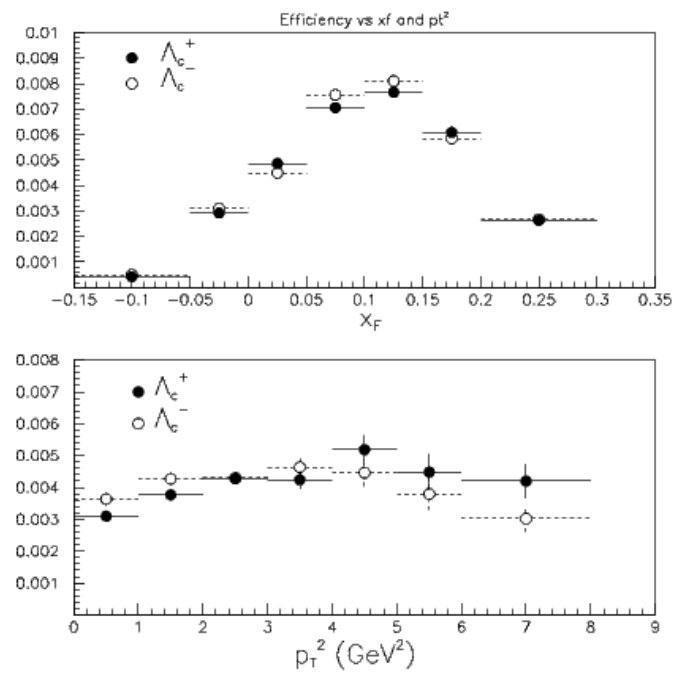
then, in principle, if the detection efficiency is the same for particle than for anti-particle, it is independent of the acceptance.

However, it was found that the acceptance of the E791 detector was different for Λ_c^+ than for Λ_c^-, implying that acceptance corrections were necessary.

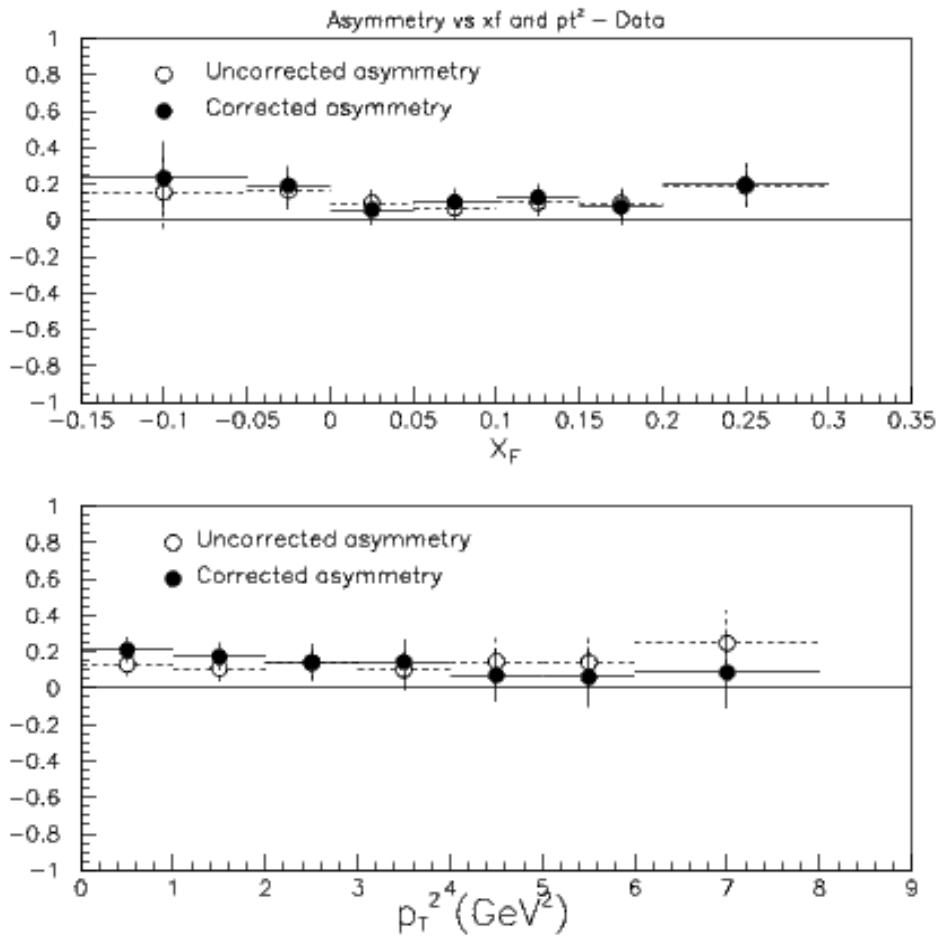
The acceptance was determined using MC data and a detailed simulation of the detector. It is given by

$$acc = \frac{N_{rec}(\Lambda_c^+)}{N_{gen}(\Lambda_c^+)}$$

both as a function of x_F and p_T^2. Here N_{gen} represent the MC generated events, while N_{rec} is for the MC generated events after passing by the detector simulation.



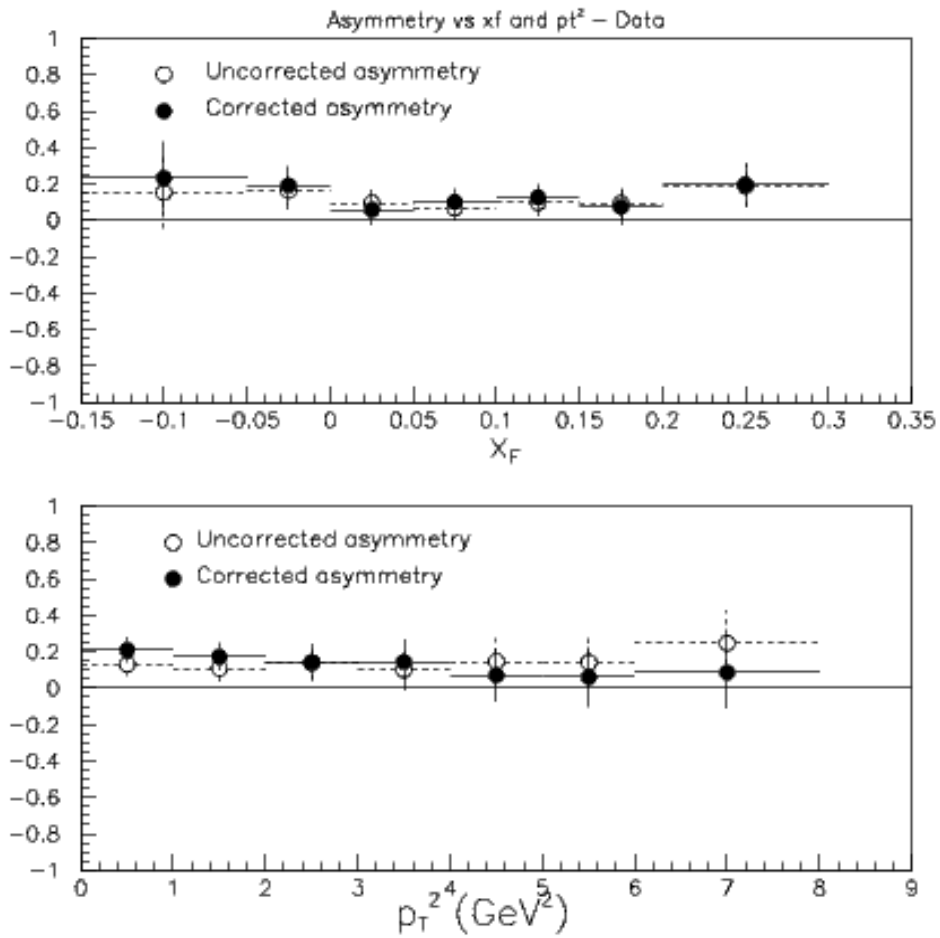
Once the acceptance was determined, it was used to correct the observed data, bin by bin in x_F and p_T^2.



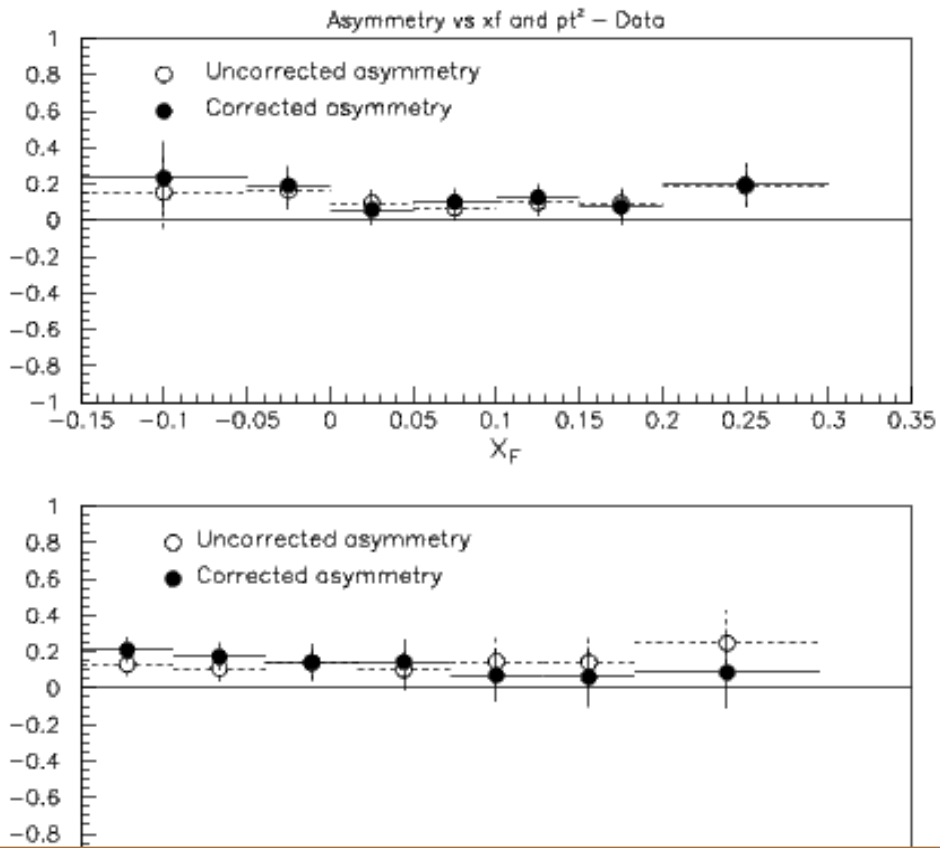
Detection efficiency

Once the acceptance was determined, it was used to correct the observed data, bin by bin in x_F and p_T^2.

The philosophy of the method is to correct the data, rather than the theoretical model.



Once the acceptance was determined, it was used to correct the observed data, bin by bin in x_F and p_T^2.



The philosophy of the method is to correct the data, rather than the theoretical model (Weighting the data).

In essence, the idea behind this method is to apply corrections to the theoretical model, leaving data as they are observed (Modification of ideal p.d.f.)

Statistical inference

In the following we will assume normal samples

Confidence intervals for the mean

Suppose you are measuring a give physical quantity μ.

- In many cases it is reasonable to assume that the result of the measurement is a random variable x which is gaussian distributed about μ.
- In these conditions, μ is an unknown for which $\overline{x}$ is an estimator, as long as $\overline{x} \rightarrow \mu$ when $n \rightarrow \infty$. (see dist.)
- After n measurements you are left with a random sample $\{x_1;\ldots;x_n\}$ from a normal population $N(\mu,\sigma^2)$.

Case 1: σ is known

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is distributed as $N(\mu,\sigma^2/n)$

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

is distributed as $N(0,1)$

Gauss dist.

$$P\left(-2 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le +2\right) = \int_{-2}^{+2} g(y)dy = 0.954$$

states that the random variable $y = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has probability

0.954 of having some value between -2 and +2.

Case 1: σ is known

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is distributed as $N(\mu,\sigma^2/n)$

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

is d

Cumulative gaussian distribution

$$\int_{-\infty}^{+2} g(y,0,1)dy - \int_{-\infty}^{-2} g(y,0,1)dy$$

$$P\left(-2 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le +2\right) = \int_{-2}^{+2} g(y)dy = 0.954$$

states that the random variable $y = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has probability

0.954 of having some value between -2 and +2.

The preceding statement can be rewritten as

$$P\left(\bar{x}-2\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{x}+2\frac{\sigma}{\sqrt{n}}\right)=0.954$$

meaning that μ has probability 0.954 of being in the interval $\bar{x} \pm \sigma/\sqrt{n}$. Note however that as $\bar{x}$ is a random variable, the interval above is a random interval, then

Prior to the repeated, independent measurements there is a probability 0.954 that the random interval $[\bar{x} - \sigma/\sqrt{n}; \bar{x} + \sigma/\sqrt{n}]$ will include the unknown, but fixed value μ. That is the symmetric 95.4% confidence interval for μ.

Case 2: σ is unknown

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$ is N(0,1), where $$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\frac{(N-1)s^2}{\sigma^2}$$ is χ^2_{N-1}, where $$s^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

estimator for σ^2

$$t=\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(N-1)s^2}{\sigma^2}/(N-1)}}=\frac{\bar{x}-\mu}{s\sqrt{n}}$$

is t-Student with (N-1) degrees of freedom

Analogous to case 1, we can write down the probability statement

$$P\left(a \le \frac{\bar{x}-\mu}{s\sqrt{n}} \le b\right) = \int_a^b P_t(t, N-1) = \gamma$$

which means that the variable t has a probability γ of being in the interval [a,b].

Given that P_t is symmetric about t=0 it is customary to choose symmetric intervals, then

$$P\left(\bar{x} - b\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + b\frac{s}{\sqrt{n}}\right) = \gamma$$

which is larger for small values of (N-1) but approaches the size of the corresponding N(0,1) interval as N grows

Confidence intervals for the variance

As before, assume you have a random sample $\{x_1, x_2, \ldots, x_n\}$, which is Gaussian distributed with mean μ and variance σ^2. We want now to make inferences about the parameter σ^2.

Case 1: μ known

May correspond to the experimental situation in which repeated measurements of the known quantity μ are performed using a measuring device of unknown precision

$$\sum_{i=1}^{N} \left[\frac{x_i - \mu}{\sigma} \right]^2 \longrightarrow$$

- related to the variance
- distributed as χ^2_n

For a χ^2_n distribution, it is possible to find numbers a and b such that

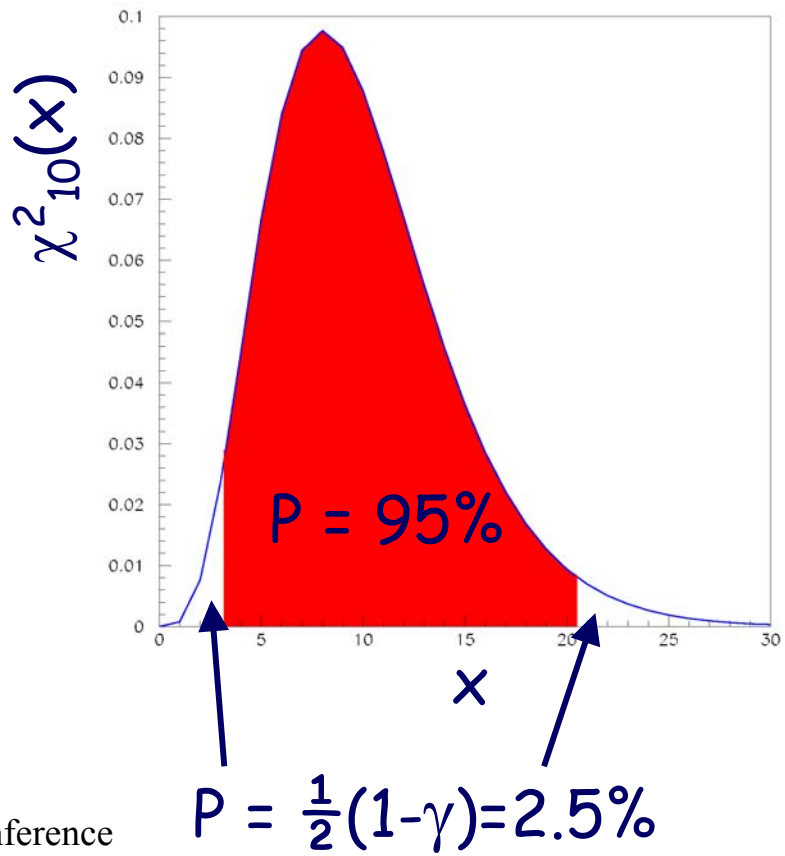
$$P\left(a \leq \sum_{i=1}^{N} \left[\frac{x_i - \mu}{\sigma}\right]^2 \leq b\right) = \int_a^b P_{\chi^2}(x)dx = \gamma$$



$$P\left(\frac{\sum (x_i - \mu)^2}{b} \leq \sigma^2 \leq \frac{\sum (x_i - \mu)^2}{a}\right) = \gamma$$

Example: suppose that from 10 independent measurements you obtain $\{x_i\} = \{1002, 1000, 997, 1001, 999, 998, 999, 1000, 1003\}$, which is a random sample from a Gaussian distributed variable with mean $\mu=1000$ ($\sum(x_i-\mu)^2 = 30$)

If we demand a 95% confidence interval for σ^2, then for 10 degrees of freedom we determine the symmetric interval a = 3.247, b=20.483 giving the 95% confidence interval (1.46, 9.23) for σ^2



Case 2: μ unknown

- ✓ If we don't know the mean μ, the sample $\{x_1, x_2, \ldots, x_n\}$ still provides the arithmetic mean $\bar{x}$, then we can repeat the previous procedure but for s^2.

- ✓ The main difference is that, as we are using s^2, and not σ^2, in the present case we have to use a χ^2 distribution for n-1 degrees of freedom.

- ✓ The consequence is that the confidence interval for σ^2 will be wider than in the case in which the mean is known.

Confidence regions for the mean and variance

✓ Assume again you have a sample $\{x_1, x_2, \ldots, x_n\}$ from a Gaussian population with mean μ and variance σ^2.

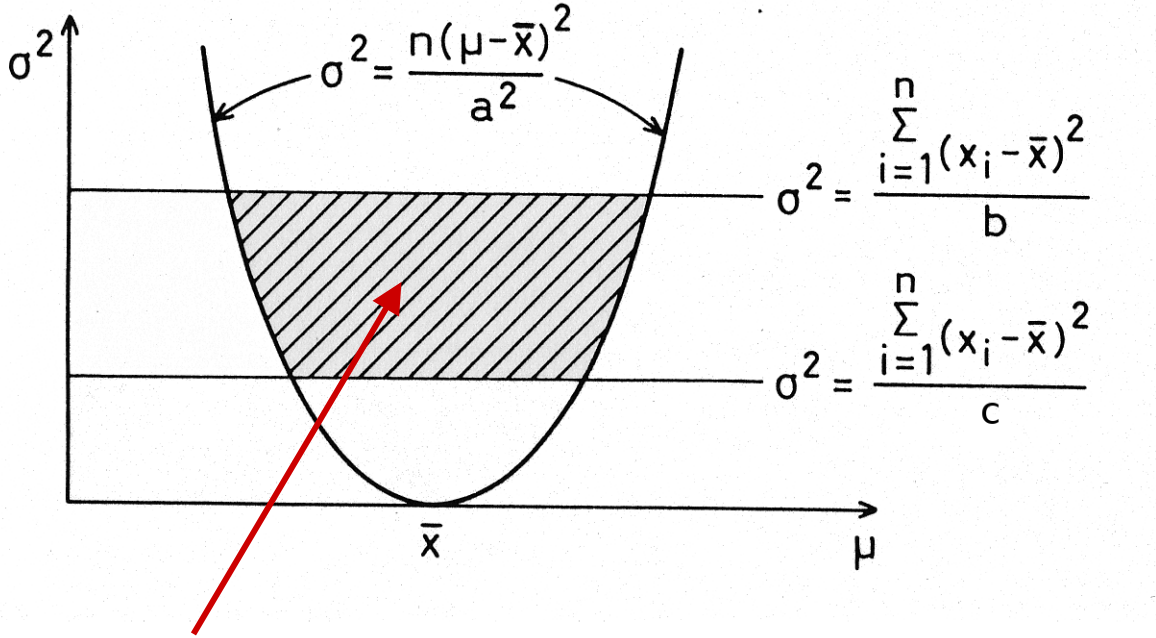
✓ The variables $\bar{x}$ and s^2 are independent (prove !), then, for instance, the 95% confidence interval for the variables

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \qquad \sum\left[\frac{x_i-\bar{x}}{\sigma}\right]^2$$

is given by the joint probability

$$P\left(-a \le \frac{\bar{x}-\mu}{\sigma\sqrt{n}} \le a,\ b \le \sum\left[\frac{x_i-\bar{x}}{\sigma}\right]^2 \le c\right) = 0.95$$

The values a, b and c have to be determined by requiring confidence intervals of $(0.95)^{\frac{1}{2}}$ for the independent variables $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ and $\sum\left[\frac{x_i-\bar{x}}{\sigma}\right]^2$



95% confidence region

Estimation of parameters

General problem:

From a restricted number of observations from a random sample one wants to obtain information about the universe from which the sample emanated.

In general, the mathematical form of the parent distribution is well-known, but it involves a certain number of parameters which have to be determined by measurements.

Estimation of parameters

General problem:

From a restricted number of observations from a random sample one want to obtain information about the universe from which the sample emanated.

Measurements should provide the largest possible information about the unknown parameters.

Estimation of parameters

General problem:

From a restricted number of observations from a random sample one want to obtain information about the universe from which the sample emanated.

Methods to estimate parameters should provide point (numerical) estimated and confidence intervals.

Definitions:

- An estimator is a function of the observations.
- An estimate is a numerical value of a given parameter obtained from the estimator, for a particular set of observations.
- A statistic is a function of the random variables that does not depends on any unknown parameter.

Example:

$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$ are estimators for the mean and variance. They are also statistics.

Properties of estimators:

Observations are random variables. Any function of the observables is also a random variable.

1. If a statistic is used as an estimator for a parameter Θ, it should provide a distribution for the estimates $\hat{\Theta}$.
2. A good estimator should, in the long run, provide estimates which do not systematically deviate from the true parameter value.
3. An estimator is consistent if it converges in probability to the true value θ of the parameter,

$$P\left(\left|\hat{\Theta}_n - \theta\right| > \varepsilon\right) < \eta$$

4. An estimator is unbiased it it does not systematically shifts from the true value of the parameter

example:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \bar{x}' = \frac{1}{N-a}\sum_{i=1}^{N} x_i \quad \text{are both}$$

estimators for the mean μ, but the second is biased (prove it!).

4th lecture

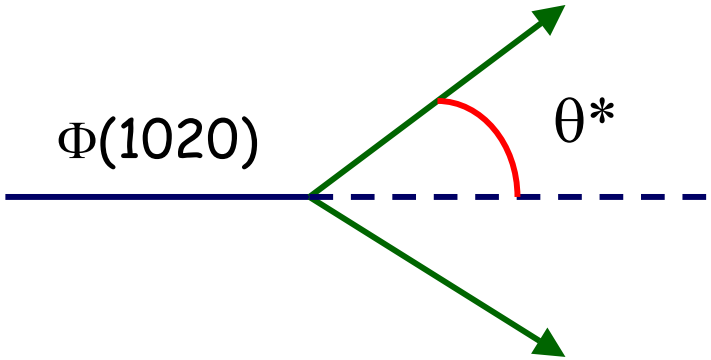
Back

Parameter fitting

Use the data to determine the value of free parameter(s)

Example:

- Suppose you want to measure the spin-alignment of the vector meson $\phi(1020)$ which has been produced in p + p interactions at some c.m. energy.
- The spin-alignment ρ is described by a 3 x 3 matrix, the spin-density matrix
- The only measurable coefficient is the ρ_{00}



- $\phi(1020)$ decays via strong interactions
- ρ_{00} can be measured by measuring the angular distribution of the decay products (which is known as a function of the parameter ρ_{00})

$$\frac{dN}{d(\cos\theta^*)} = N_0 \times [(1-\rho_{00}) + (3\rho_{00}-1)\cos^2\theta^*]$$

Now the question is: what value of ρ_{00} provides the best description of data ? And how accurately ρ_{00} can be determined ?

Comments

- Hypothesis testing precedes parameter fitting: if hypothesis are incorrect, then there is no point in determining free parameters.

- In practice, one often does parameter fitting first anyway. It may be impossible to make a test of hypothesis before fixing free parameters to their optimum values.

- In this lecture we will consider two methods: Maximum Likelihood and Least Squares

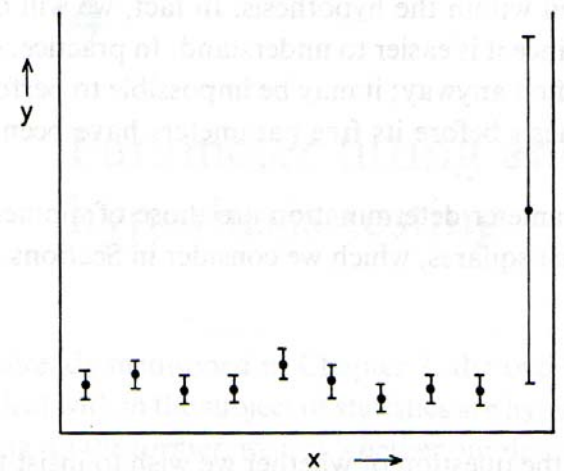
Comments II

Normalization:

- In many cases is desirable to normalize the theoretical distribution to the data. Normalization reduces the number of free parameters by one.

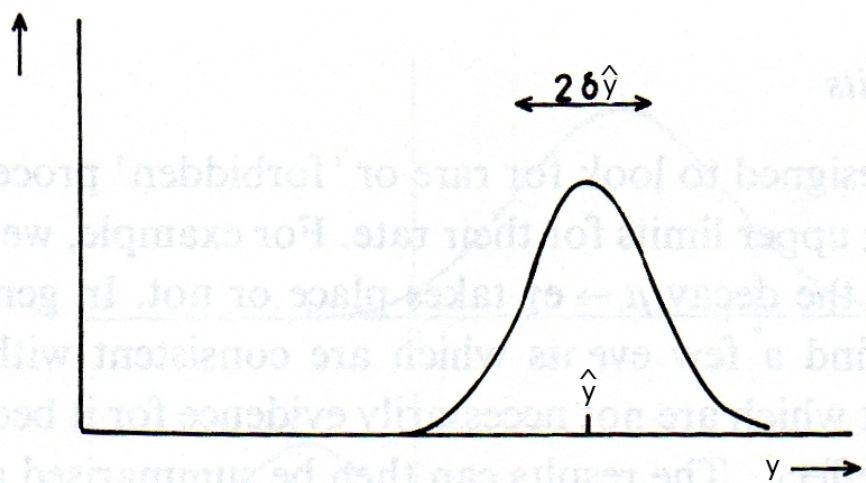
- In some cases, normalization is undesirable due to the introduction of distorting effects

Example: **fit a straight line to data**. Normalization involves the calculation of Σy_i. The large error of the last point makes it useless. Normalization will introduce distortions because all of them are equally weighted



Interpretation of estimates

Assume that a free parameter has been determined as $\hat{y} \pm \delta\hat{y}$. Assume also that our estimate $\hat{y}$ is Gaussian distributed and that the true value (unknown) is y_0.



- The probability that a measurement gives an answer in a specific range of y is the area under the relevant part of the gaussian
- For $\delta\hat{y} = \sigma$, the probability is ~68%
- Having an estimate $\hat{y}$, it is usual to write

$$\hat{y} - \delta\hat{y} \leq y_0 \leq \hat{y} + \delta\hat{y}$$

where $[\hat{y} - \delta\hat{y};\ \hat{y} + \delta\hat{y}]$ is the confidence range for y_0

Maximum likelihood method

Powerful method to find values of unknown parameters

Example:
Consider the following angular distribution, depending on the parameters a,b

$$\frac{dn}{d\cos\theta} = a + b\cos^2\theta$$

Normalize (if not, the method does not work !)

$$\int_{-1}^{+1} y d\cos\theta = 1 \quad \Rightarrow \quad N = \frac{1}{2}\left[1 + (b/3a)\right]$$

then

$$y = \frac{dn}{d\cos\theta} = N\left[1 + \frac{b}{a}\cos^2\theta\right]$$

behaves as a probability distribution

- For the event i we calculate

$$y_i = N\left[1 + \frac{b}{a}\cos^2\theta_i\right]$$

which is the probability density of observing the event i as a function of b/a.

- We define now the likelihood $\mathcal{L}$ as the product of the y_i's

$$\mathcal{L}\left(\frac{b}{a}\right) = \Pi_{i=1}^{n} y_i$$

- Then, for a specific value (b/a), $\mathcal{L}$ is the joint probability density for obtaining the particular set of cos θ_i we observed in the experiment.

$\mathcal{L}$ is the probability density for obtaining the particular set of observations in the ordering in which we observe them. Since the ordering is irrelevant, a factor of 1/n! should be included but, as we are interested in how the function $\mathcal{L}$ varies as a function of (b/a), that factor is irrelevant

bserving the

s the product of

the y_i

$$\mathcal{L}\left(\frac{b}{a}\right) = \Pi_{i=1}^{n} y_i$$

- Then, for a specific value (b/a), $\mathcal{L}$ is the joint probability density for obtaining the particular set of cos θ_i we observed in the experiment.

- For the event i we calculate

$$y_i = N\left[1 + \frac{b}{a}\cos^2\theta_i\right]$$

Finally maximize $\mathcal{L}$. Note the importance of the normalization: without the factor N, $\mathcal{L}$ can be as large as you want by simply increasing the value of (b/a), then $\mathcal{L}$ would not have absolute maximum !

the y_i

$$\mathcal{L}\left(\frac{b}{a}\right) = \Pi_{i=1}^{n} y_i$$

- Then, for a specific value (b/a), $\mathcal{L}$ is the joint probability density for obtaining the particular set of cos θ_i we observed in the experiment.

The logarithm of the likelihood function

Sometimes is most convenient to use the logarithm of the likelihood function

$$l = \log \mathcal{L} = \sum_i \log y_i$$

For a large number of experimental observations n, $\mathcal{L}$ tends to a Gaussian distribution at least in the vicinity of the maximum of the distribution:

$$\begin{aligned} l &= l_{max} + \frac{1}{2!} l'' \delta p^2 + \ldots \\ &= l_{max} - \frac{1}{2c} \delta p^2 \end{aligned}$$

$$\mathcal{L} \sim exp\left[-\frac{(p - p_0)}{2c} \right]$$

l'' = -1/c

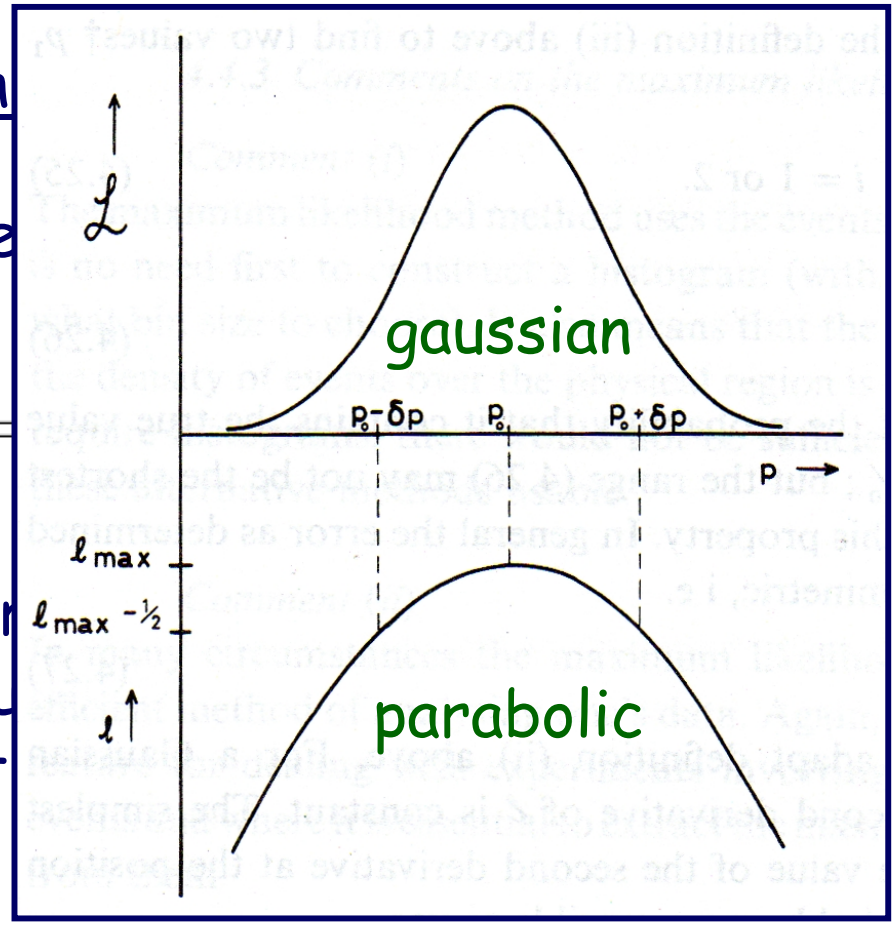
value of l'' at the maximum

The logarithm of the likelih

Sometimes is most convenie
the likelihood function

$$l = \log \mathcal{L} =$$

For a large number of exper
tends to a Gaussian distribu
of the maximum of the dist



$$l = l_{max} + \frac{1}{2!} l'' \delta p^2 + \ldots$$

$$= l_{max} - \frac{1}{2c} \delta p^2$$

$$\mathcal{L} \sim exp\left[-\frac{(p - p_0)}{2c} \right]$$

l'' = -1/c

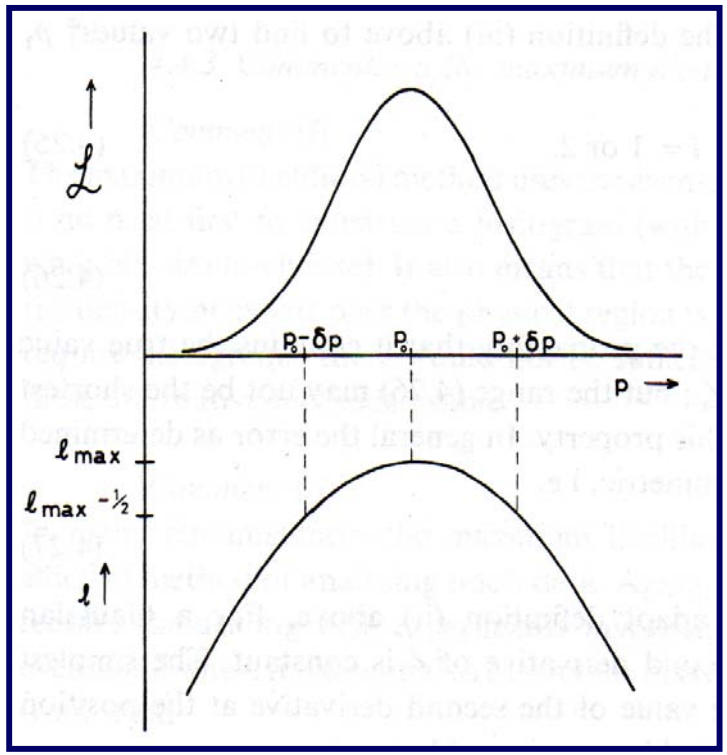
value of l'' at the maximum

When $\mathcal{L}$ is Gaussian, then the following quantities are identical and can be used as the definition of the error on p:

- the root means square deviation of $\mathcal{L}$ about its mean
- $(-\partial^2 l/\partial p^2)^{-\frac{1}{2}}$
- $l(p_0 \pm \delta p) = l(p_0) - 1/2$

δp equiv to σ

p_0 maximizes $\mathcal{L}$ and l



Clearly Gaussian variables are better than non Gaussian. Make an adequate choice of variables, e.g. in decay processes, is better to measure the decay rate $1/\tau$ than the lifetime τ !

Comments

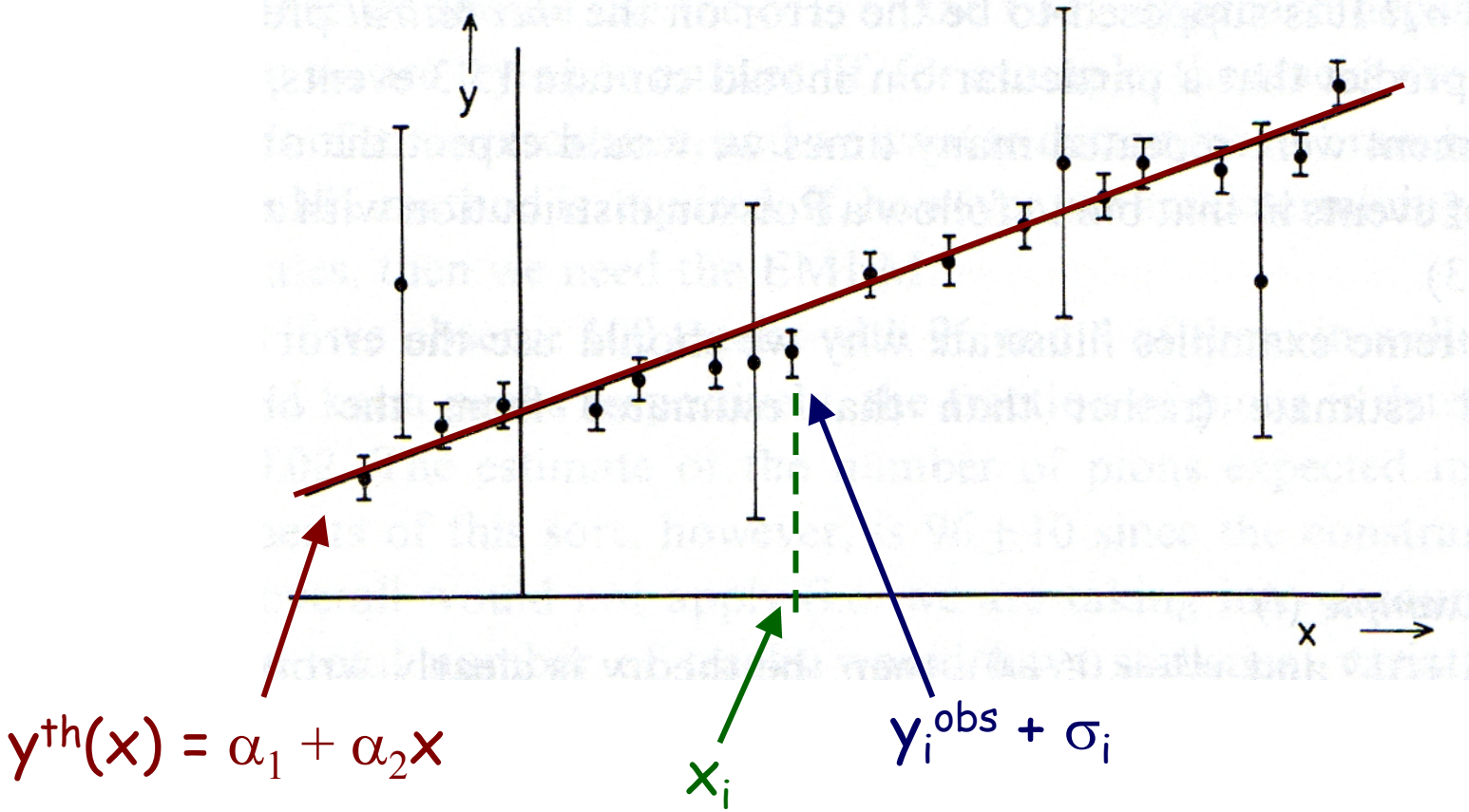
- Maximum likelihood method uses the events one at a time → no need to construct histograms → no problems associated to the binning.
- Functions of implicit variables are very easily handled.
- Data are used in the form of complete events rather than projections on various axes → powerful tool to determine unknown parameters.
- In some situations, the maximum likelihood and the least squares methods are equivalent.
- Easy to handle bounded parameters.
- One serious drawback is the large amount of computation required very often.
- Extension to several parameters is trivial.

Least squares method

- Assume you have an experimental distribution (say an histogram). The histogram represents the number of events $y_i^{obs} + \sigma_i$ as a function of a given variable x_i.
- Assume you want to describe the experimental data by a functional form $y^{th}(x;\alpha_j)$, then we construct

$$\mathcal{S} = \sum_{i=1}^{bins} \left(\frac{y_i^{obs} - y_i^{th}(x_i, \alpha_j)}{\sigma_i} \right)^2$$

- If the theory is in good agreement with data, then y_i^{obs} and y_i^{th} do nor differ too much and $\mathcal{S}$ will be small.



Bin size has to be chosen such that the number of events is large enough to ensure that the error in the number of event in then bin is approximately gaussian (remember that Poisson → Gaussian for n → ∝).

Comments

✓ Start first by choosing a suitable bin size. Hopefully results will be approximately independent of the bin size. Bins may be also of different sizes.

✓ It is desirable to avoid bins with too few events → better if the number of events is large enough to ensure gaussian errors. Also, as we use the experimental error σ_i, we have to avoid situations arising from the fact that usually few events means large errors.

✓ Easy to generalize for several variables.

✓ If $y^{th}(x,\alpha_j)$ is linear in the parameters, then the minimum of S can be found analytically.

✓ S_{min} is a measure of how well the theoretical hypothesis describes the data.

Least squares with correlated errors

We will consider now the modifications necessary in order to deal with the case in which the errors in the y_i^{obs} are correlated one another.

Let us start with the two variables uncorrelated case

$$\mathcal{S} = \left(\frac{z' - z'_0}{\sigma_{z'}} \right)^2 + \left(\frac{y' - y'_0}{\sigma_{y'}} \right)^2$$

and perform a rotation of angle θ

$$\left. \begin{aligned} z' &= z\cos\theta - y\sin\theta \\ y' &= z\sin\theta + z\cos\theta \end{aligned} \right\}$$

then

$$\mathcal{S} = \frac{1}{\sigma_z^2\sigma_y^2 - cov(z,y)^2}\left[\sigma_y^2(z-z_0)^2 + \sigma_z^2(y-y_0)^2 - cov(z,y)(z-z_0)(y-y_0)\right]$$

where the errors were transformed also to

$$\left.\begin{aligned}\sigma_{z'}^2 &= \sigma_z^2\cos^2\theta + \sigma_y^2\sin^2\theta - 2cov(z,y)\sin\theta\cos\theta \\ \sigma_{y'}^2 &= \sigma_z^2\sin^2\theta + \sigma_y^2\cos^2\theta + 2cov(z,y)\sin\theta\cos\theta\end{aligned}\right\}$$

with the condition that errors in z′ and y′ are independent

$$cov(z',y') = (\sigma_z^2 - \sigma_y^2)\sin\theta\cos\theta + cov(z,y)(cos^2\theta - \sin^2\theta) = 0$$

Now write

$$\begin{aligned}\mathcal{S} &= \frac{1}{\sigma_z^2\sigma_y^2 - cov(z,y)^2}\left[\sigma_y^2(z-z_0)^2 + \sigma_z^2(y-y_0)^2\right. \\ &\quad \left. -cov(z,y)(z-z_0)(y-y_0)\right] \\ &= H_{11}(z-z_0)^2 + H_{22}(y-y_0)^2 + H_{12}(z-z_0)(y-y_0)\end{aligned}$$

where

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_z^2 & cov(z,y) \\ cov(z,y) & \sigma_y^2 \end{pmatrix}$$

is the inverse of the error matrix

Now write (in matrix form)

$$\mathcal{S} = \sum_{i,j} \Delta_i^T \mathcal{H}_{i,j} \Delta_j \qquad \Delta_i = y_i^{obs} - y_i^{th}$$

where

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_z^2 & cov(z,y) \\ cov(z,y) & \sigma_y^2 \end{pmatrix}$$

is the inverse of the error matrix

Comparison

	Maximum likelihood	**Least squares**
How easy ?	Normalization and maximization can be difficult	Usually easy
Efficiency	Usually most efficient	Sometimes equiv. to ML
Input data	Individual events	Histograms
Estimate of goodness of fit	Very difficult	Easy

Comparison

	Maximum likelihood	**Least squares**
Constraint among parameters	Easy	Can be imposed
N-dimensional problems	Normalization and maximization can be difficult	Problems associated to the choice of the distribution
Weighted events	Can be used	Easy
Background subtraction	Can be problematic	Easy
Error estimate	$(\partial^2 \mathcal{L}/\partial p_i \partial p_j)^{-1/2}$	$\frac{1}{2}(\partial^2 S/\partial p_i \partial p_j)^{-1/2}$

Hypothesis testing

Are our hypothesis (model) consistent with the data ?

Simple case: the hypothesis consist of a particular value for one parameter.

- Measure the parameter and obtain its error
- Compare the measured value with the hypothesized
- Make a numerical statement on the probability of obtaining a result equal or worse assuming that the hypothesis is true

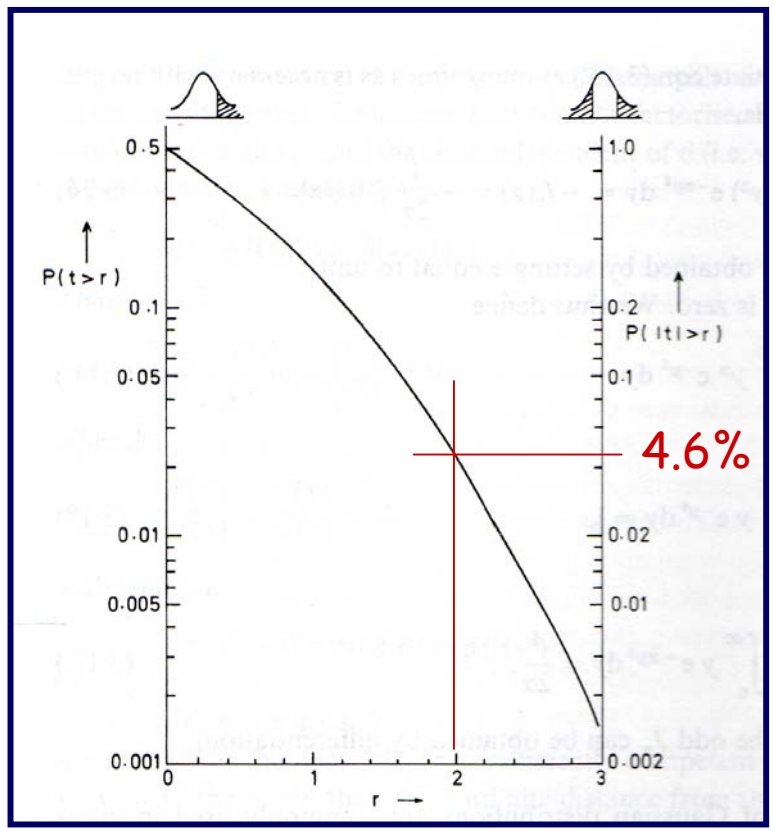
Example

Suppose that a theory predicts that the lifetime of the neutron is 910 secs. In a experiment is measured that the lifetime of the neutron is 950 ± 20 secs. Are these two numbers in agreement ?

Construct the variable

$$t = \frac{x - \mu}{\sigma}$$

In the example above t = 2,
P(|t|> 2) = 4.6%
Decide if the hypothesis is rejected or not (you make the decision)



Then

$$P(|t| > 2) = 0.46$$

meaning that in 1000 experiments of the same type, we expect only 46 of them to differ from the predicted value of the lifetime of the neutron.

- Note that we are assuming that σ is the true value of the experimental resolution.
- If we don't know σ, we have to use s and t is t-Student
- If the theoretical value has an uncertainty σ', then

$$t = \frac{y_{obs} - y_{th}}{\sqrt{\sigma^2 + \sigma'^2}}$$

Distribution testing

Consider the example of the angular distribution

$$y = \frac{dn}{d\cos\theta} = N\left[1 + \frac{b}{a}\cos^2\theta\right]$$

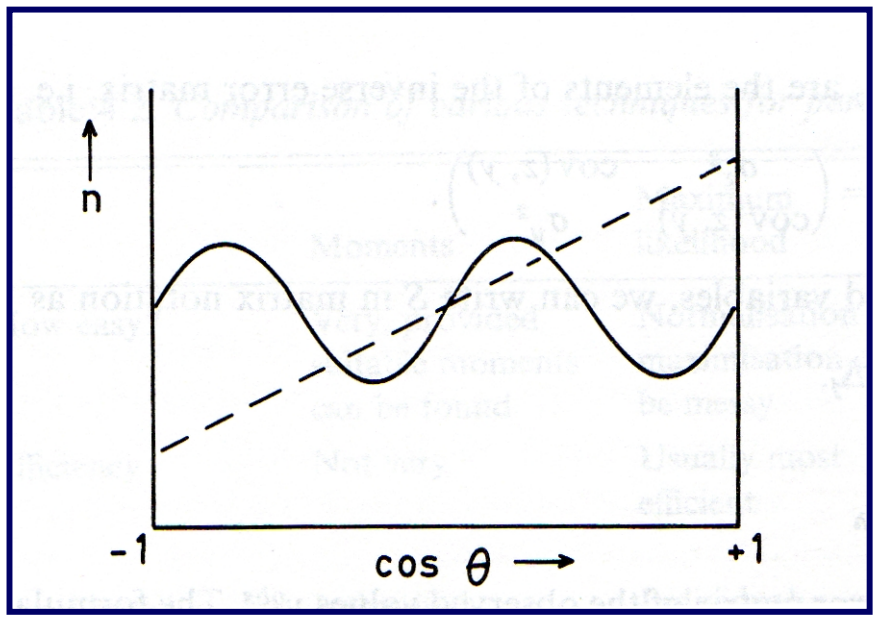
Hypothesis testing: is the angular distribution isotropic (i.e. is b/a=0 ?)

Distribution testing

Consider the example of the angular distribution

$$y = \frac{dn}{d\cos\theta} = N\left[1 + \frac{b}{a}\cos^2\theta\right]$$

Hypothesis testing: is the angular distribution isotropic (i.e. is b/a=0 ?)



Fitting the distributions in the figure to the form above, we will obtain b/a=0 but clearly distributions are not isotropic

Distribution testing

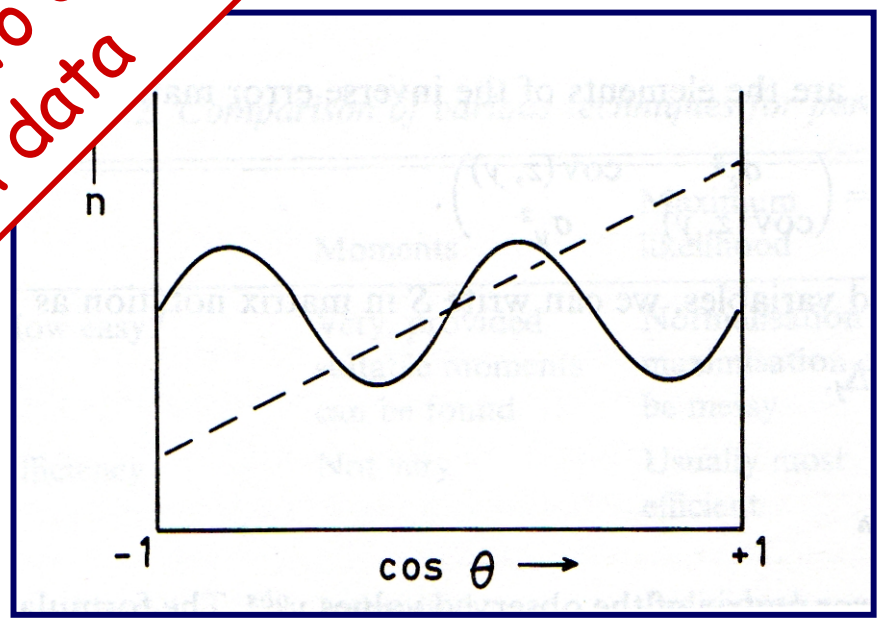
Consider the example of the angular distribution

$$y = \frac{dn}{d\cos\theta} = N\left[1 \qquad \theta\right]$$

Hypothesis testing: is the angular distribution isot
(i.e. is b/a=0 ?)

Fitting the distri h the figure to above, we wi b/a=0 but clearly dis utions are not isotropic

We need a criteria to compare distributions with data



χ^2 test

Distributions are tested by the χ^2 method

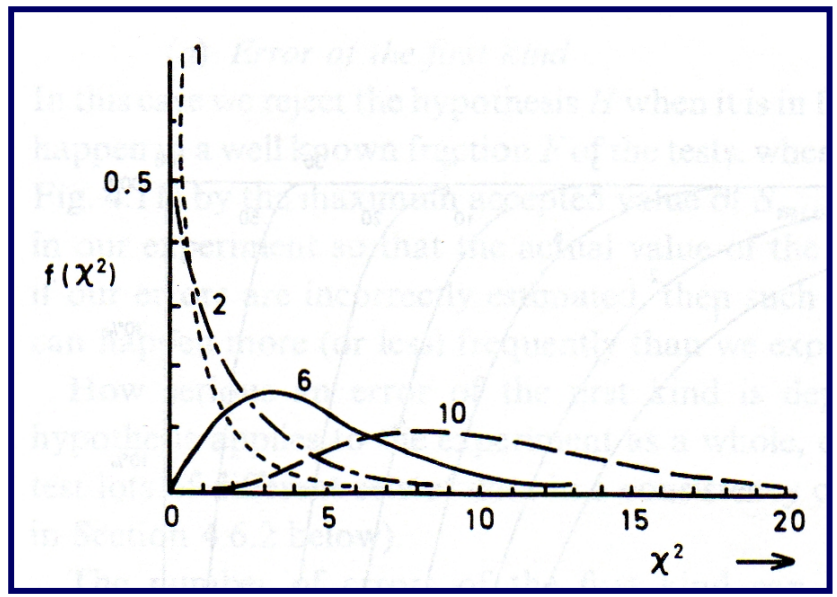
if y_i^{obs} in each bin are Gaussian distributed with mean y_i^{th} and variance σ_i^2, then

$$S = \sum_i \left(\frac{y_i^{obs} - y_i^{th}}{\sigma_i} \right)^2$$

is distributed as χ^2

Testing hypothesis by the χ^2 method

1. Construct S and minimize with respect to the free parameters.
2. Determine the number of degrees of freedom ν
$$\nu = b - p$$
where b is the number of bins (data) included in S and p is the number of free parameters.



3. Look in the tables the probability that, for ν degrees of freedom, χ^2 is greater or equal to our observed value S_{min}.

Comments

1. Remember that $\overline{\chi}^2 = \nu$ and $\sigma^2(\chi^2) = 2\nu$
2. Large values of S_{min} are unlikely ➔ hypothesis wrong
3. Small values of S_{min} are also unlikely ➔ data suspicious

Typically it means that our errors are too big, then S is artificially small

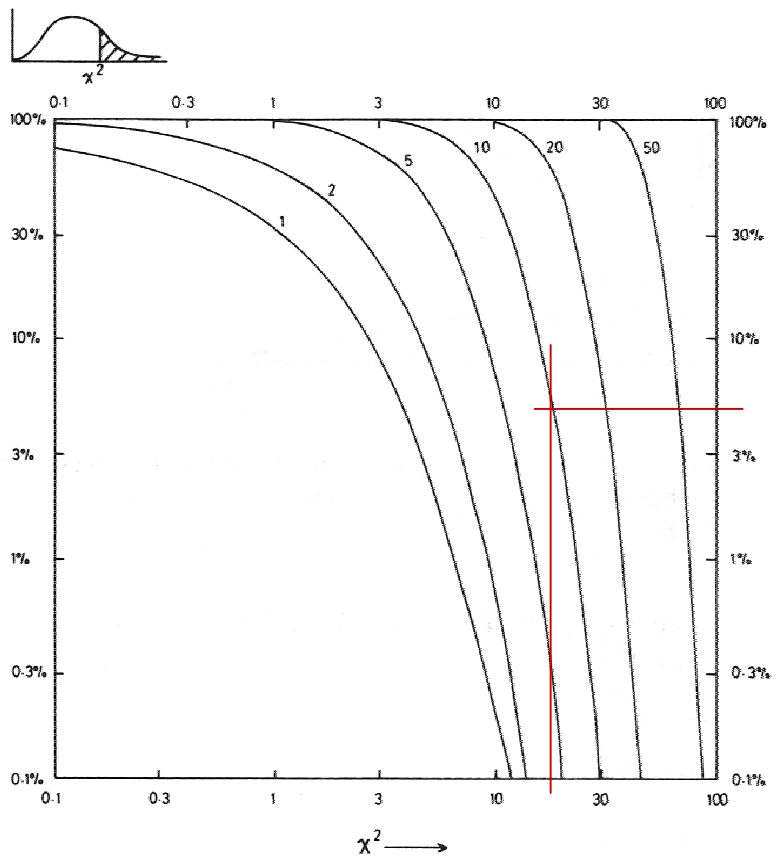
$$S = \sum_i \left(\frac{y_i^{obs} - y_i^{th}}{\sigma_i} \right)^2$$

Most useful than the χ^2 distribution itself it is the cumulative χ^2 distribution

Example

Suppose that you have 12 bins and that when fit the expression N(1+b/a cos θ)) to data you obtain S_{min} = 20.0

- $\nu = 12 - 2 = 10$
- $F_\chi(20) = P(\chi^2 > 20) \sim 4\ \%$
- Decide if the hypothesis is to be rejected or not (this is up to you…)



Errors of the first and second kind

In deciding when to reject or not a hypothesis, we can make two sort of incorrect decisions:

1. Error of the first kind:
   the hypothesis is rejected when it is in fact correct (S_{min} is big due to fluctuations or our cut in S_{min} is too low)
2. Error of the second kind:
   the hypothesis is NOT rejected when it is indeed false (S_{min} is accidentally small or the cut in S_{min} is too high)

Discussion: consequences of errors of the first or second kind are different, depending on the situation

a) Hypothesis relates to the whole experiment

- ✓ In testing hypothesis which relates to the experimental data at all, both the first and second kind errors are equally serious. In both cases the answer is wrong.
- ✓ We should be able to estimate how often we will make an error of the first or second class by testing alternative hypothesis (not easy most of the time).
- ✓ Dependent on the class of hypothesis test - sometimes 95% of confidence level is Ok, sometimes a bigger confidence level is needed.

Discussion

b) Hypothesis used as data selector

- ✓ Suppose you have a large experimental sample and you want to test the hypothesis that the data correspond to a given kind of event. An error of the first kind means that you are rejecting a small fraction of genuine events.
- ✓ In the above conditions, errors of the second kind means that you have a contaminated sample of events.
- ✓ In both cases you have to revise the cut on $S_{\min}$ you are imposing.

Minimization

- Both the least squares and maximum likelihood methods require that we minimize a function (S and $\mathcal{L}$ respectively) of several variables.

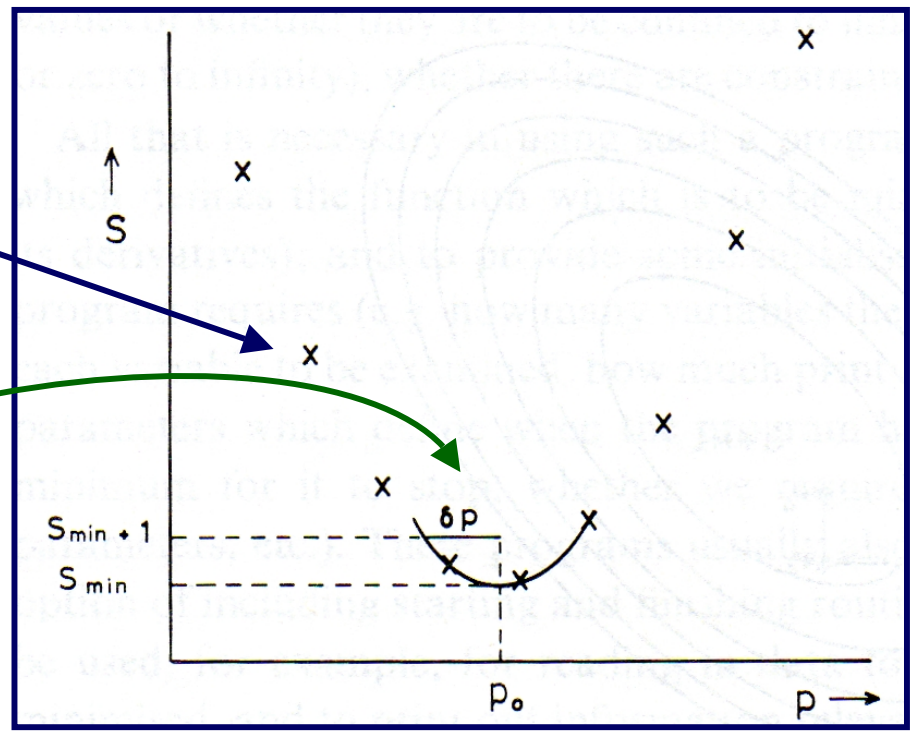
- Minimization, especially when we have to deal with functions of several variables, is a problem by itself.

- Several approaches are possible in general.

1. S is linear in the parameters → find the minimum explicitly by matrix inversion. Problem suitable for analytical solution.
2. For non linear problems in only one variable → write a computer program for calculating S at several values of the variable. Find the minimum by inspection.

Calculated values of S as a function of the parameter p

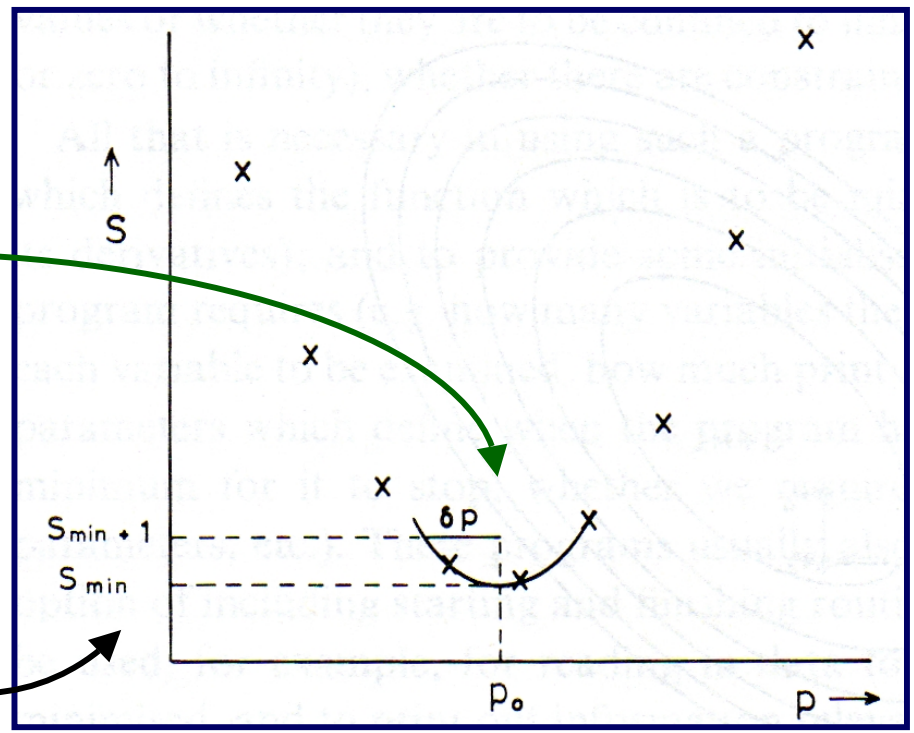
Determine a parabola around the minimum (use three points near the minimum)



1. S is linear in the parameters → find the minimum explicitly by matrix inversion. Problem suitable for analytical solution.
2. For non linear problems in only one variable → write a computer program for calculating S at several values of the variable. Find the minimum by inspection.

The minimum of the parabola gives the best estimate p_0 of the parameter

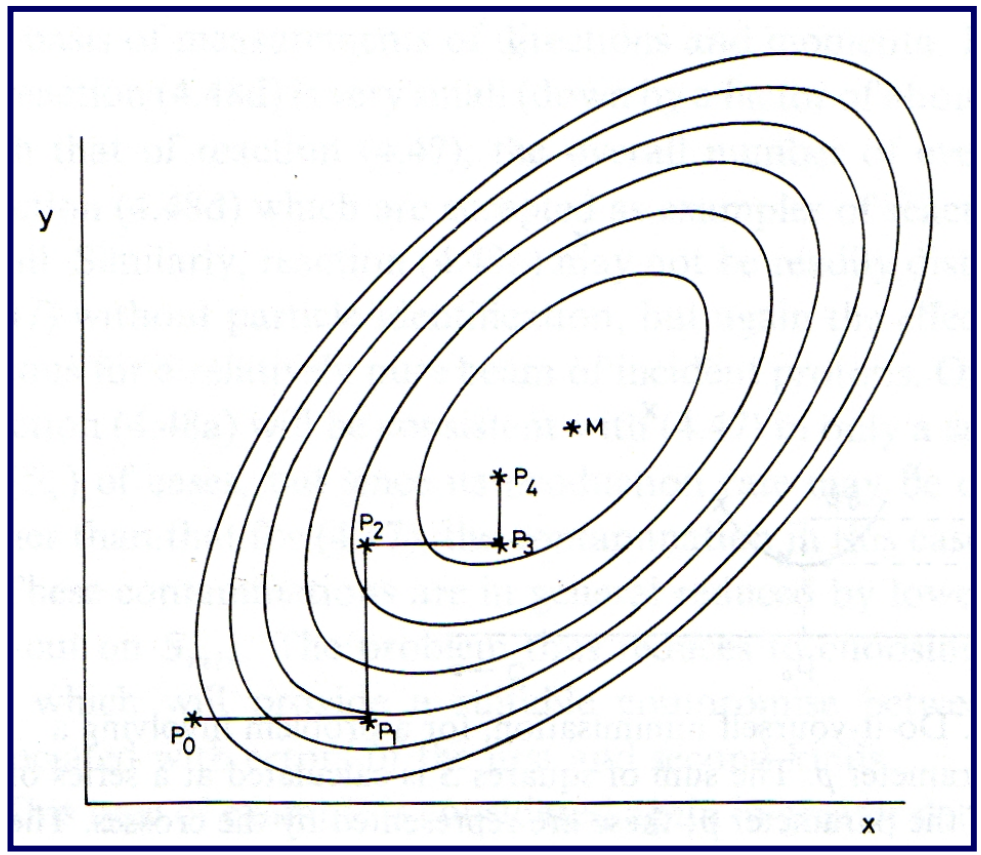
Estimate the error δp using the width of the parabola



3. The approach of 2. can be extended to two variables problems.
4. For more complicated situations, the best solution is the use of general minimization programs.

Contour plot of f(x,y)

Try to find the minimum M starting from an arbitrary position $P_0(x_0,y_0)$. The search progresses through the points $P_1, P_2, \ldots, P_n$, until the minimum is reached.
Slow convergence.



Very often we are faced to minimization problems in which there exist constraints among the parameters. If that is the case, we have at least the following two methods:

1. Use Lagrange multipliers to incorporate the constraints.

2. Change the variables to another smaller set, for instance, if your unknowns are various fractions f_i satisfying that

$$\sum f_i = 1$$

define f_k such that

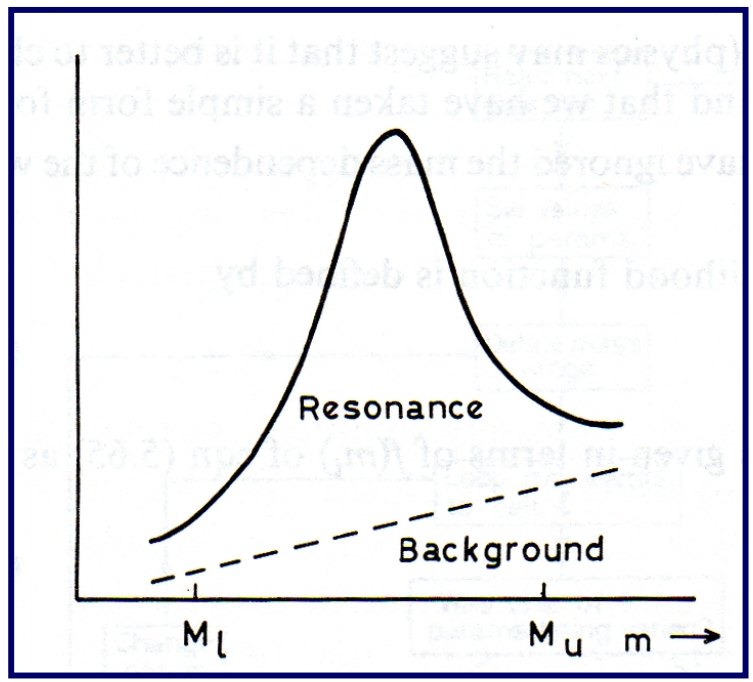
$$f_k = 1 - \sum f_i \; ; \; i \neq k$$

then the new set is an independent set.

Example: ML fit

Assume we have a set of experimentally determined mass values m_i, which we wish to fit with a Breit-Wigner resonance function together with some background

For instance, we want to measure the mass M_0 and width Γ of a resonance recently discovered.
The fitting procedure is carried out over the range M_l to M_u



For the fitting function we choose

$$f(m) = BW(m) + BGD(m)$$

where

$$BW(m) = \frac{\Gamma/2}{(m - M_0)^2 + (\Gamma/2)^2}$$

$$BGD(m) = a\left\{1 + b\left(m - \overline{M}\right) + c\left(m - \overline{M}\right)^2\right\}$$

with M_0 and Γ the mass and width of the Breit-Wigner and a, b, c the parameters characterizing the background

Then:

1. Define the logarithm of the likelihood function

$$\mathcal{L} = \sum_{events} \log y_i$$

where

$$y_i = \frac{f(m_i)}{\int_{M_l}^{M_u} f(m)dm}$$

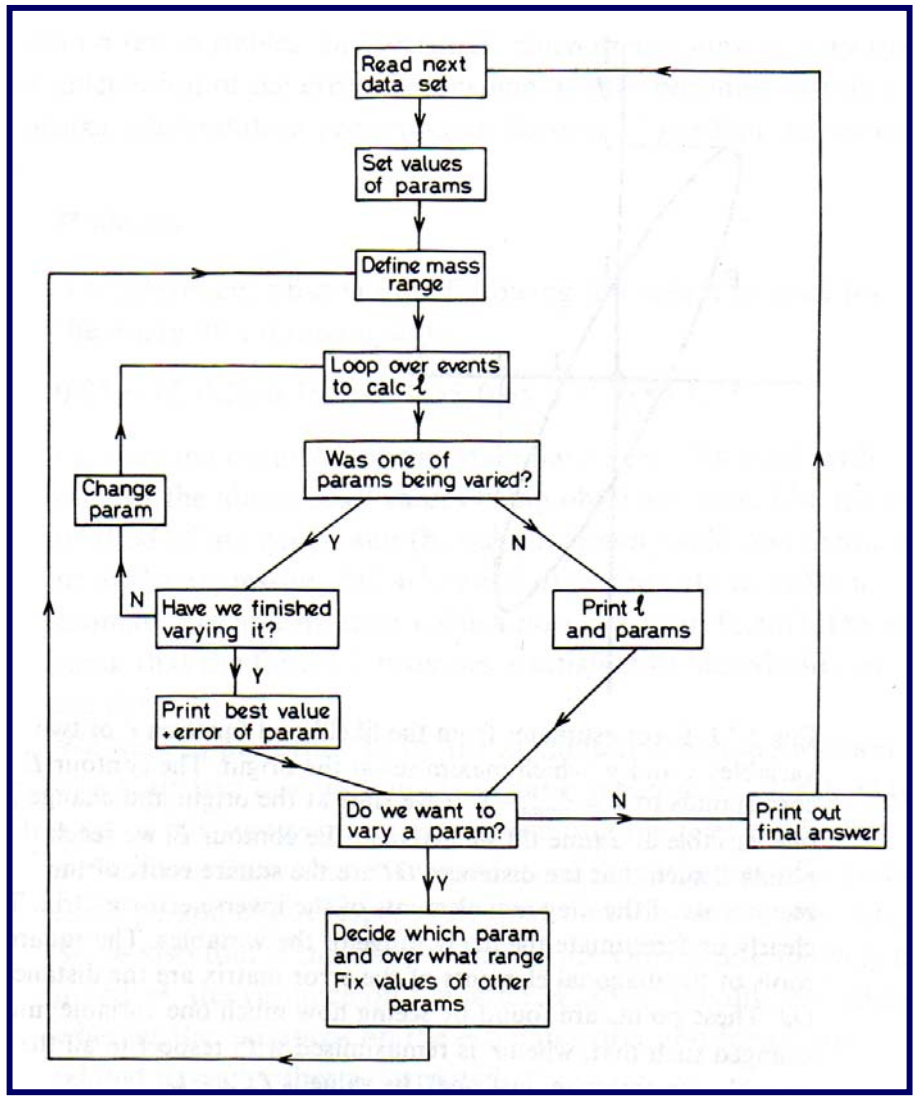
2. The integral of f can be obtained analytically

$$\int_{M_l}^{M_u} f(m)dm = \tan^{-1}\left(\frac{M_u - M_l}{\Gamma/2}\right) + \tan^{-1}\left(\frac{M_0 - M_l}{\Gamma/2}\right) + a(M_u - M_l)\left(1 + \frac{c}{12}(M_u - M_l)^2\right)$$

Find the best values of the parameters $\{M_0, \Gamma, a, b, c\}$

Step by step

1. Read the events
2. Fix 4 of 5 parameters and let one of them free
3. Vary over the range M_l to M_u and find the best value for the free parameter
4. Choose another parameter to be free and fix the first one.
5. Repeat.



- After the above procedure you eventually find the maximum of $\mathcal{L}$.
- You still have to find the error in the determination of parameters.

Determination of errors

Vary the parameters one at a time to find out by how much they must change to make

$$\mathcal{L} = \mathcal{L}_{max} - \frac{1}{2}$$

This will provide us the diagonal elements of the error matrix. What we have to find is by how much we must vary the corresponding parameter such that, when $\mathcal{L}$ is remaximized with respect to all other parameters, the condition above is satisfied.

Final remarks

Bibliography

- Statistics for nuclear and particle physicists
  Louis Lyons; Cambridge University Press.
- Probability and statistics in particle physics
  A.G. Frodesen, O. Skjeggestad and H. Tofte; Universitetsforlaget.
- Data reduction and error analysis for the physical sciences
  Philip R. Bevington; McGraw-Hill, Inc.