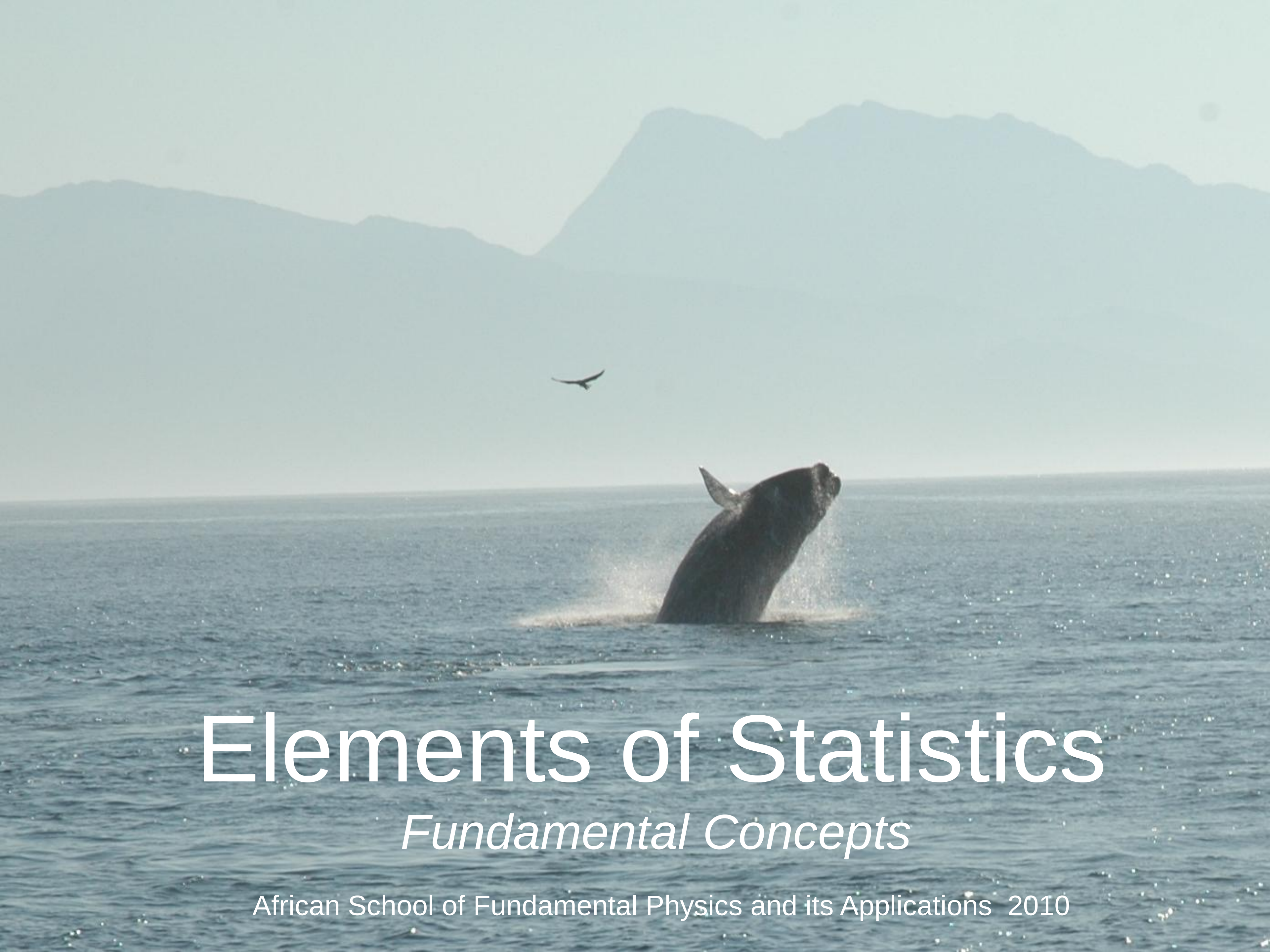


Elements of Statistics

Fundamental Concepts

African School of Fundamental Physics and its Applications 2010

Disclaimer :
What this lecture is not going to be about…

- It will not be a lecture on the fundamental theory of statistics

- Multivariate techniques

- Bayesian confidence intervals

- Goodness of fit theory

- In depth discussion of systematics and their treatment

- Bayesian vs. Frequentist diatribe

Goal(s) of the Lecture

- Review the basic knowledge in statistics as needed in HEP

All you need to know are the basic four operations

- Give concrete Root-based macro examples for hands-on experience

Just starting from a simple uniform random number generator

- Finally apply these fundamentals to a concrete hot topic in today's experimental physics…

The search for the Higgs boson

Why are Statistics so Important in Particle Physics ?

Because we need to give quantitative statements about processes that have some inherent randomness…

… May this randomness be of measurement nature or quantum …

How did it all start ?

Liber de ludo aleae

QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

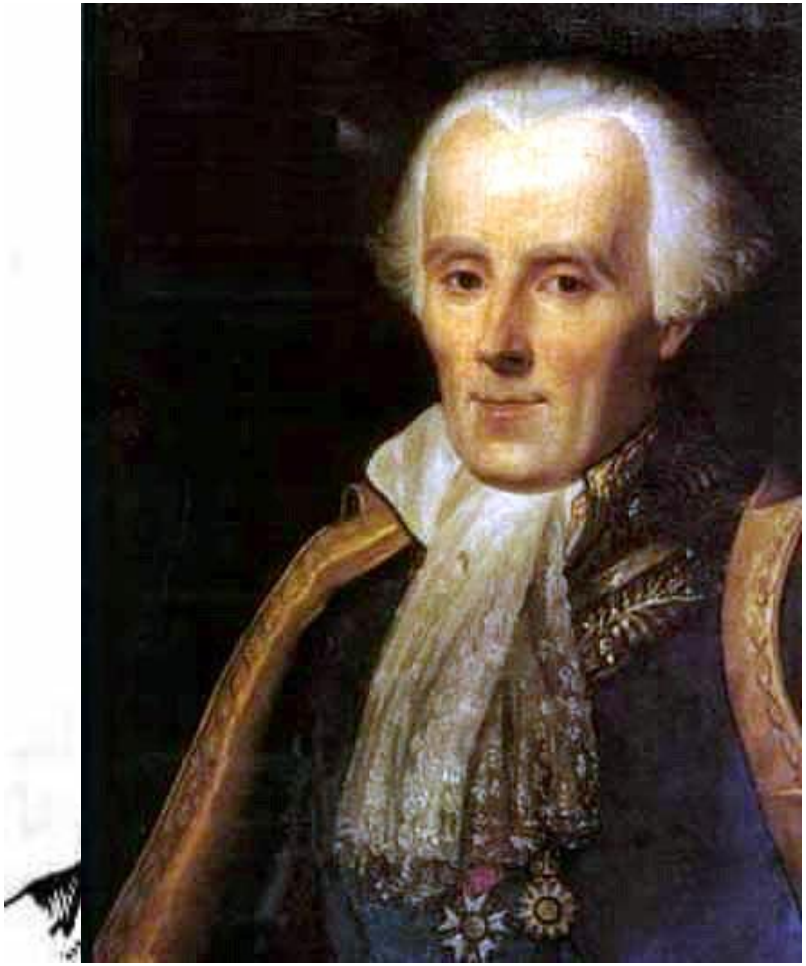
To study games of chance !

G. Cardano (1501-1576)

And many others to follow (Pascal, Fermat, etc..)

“La theorie des probabilités n’est, au fond, que le bon sens reduit en calcul”

P. S. Laplace (1749-1824)



What is a Statistical Error ?

Imagine I have a billion white ○ and blue ● golf balls

I decide to through one million of them into a well and decide an admixture of 15 out of one hundred blue ones…

I then know PRECISELY the probability that if you pick one at RANDOM, it will be blue…

$$p = 15\%$$

QuickTime™ et un décompresseur sont requis pour visionner cette image.

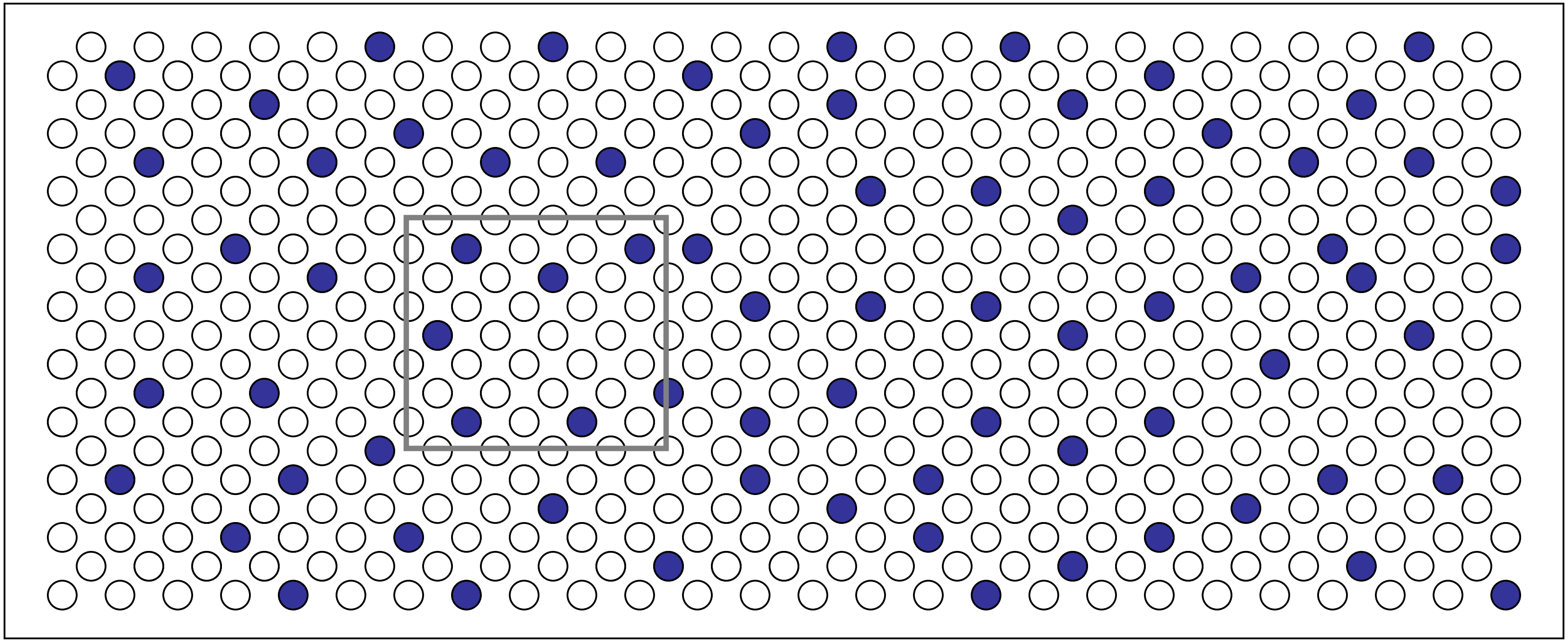
You of course don't know this number and you want to measure it…

Please Measure

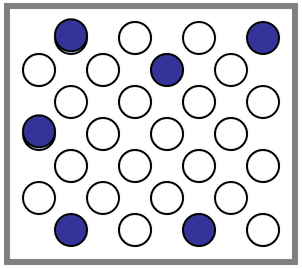
All you have is a bucket…

Which contains exactly 300 balls

This is approximately how the well looks like inside…



You throw the bucket and pull out the following outcome



$$n = 300$$

$$k = 36$$

Aha! You have a measurement!

The probability is…

$$P = 12\%$$

… But how precise is it ?

Remember you are not supposed to know the true value!

The difference between a measurement and the true value is the Statistical Error

Precise definition of statistical error

In this case it would be 3% absolute (20% relative), but since you don't know the true value you don't know at all what your statistical error really is !

Of course had you thrown your bucket on a different spot, you would have probably had a different measurement and the statistical error would be different…

What you want to know is your measurement error, or what the average statistical variation of your measurement is…

This can be done provided that you know the law of probability governing the possible outcomes of your experiment … (and the true value of p, but assume that 12% is a close enough)

You want to know what the probability for an outcome of k golf balls to be blue is.

For one specific outcome the probability is:

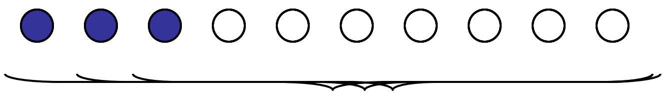
$$P = p^k \times (1-p)^{n-k}$$

What are all possible combination of outcomes of k blue balls out of n?

What are all possible combination of outcomes of k blue balls out of n?

For the first blue ball there are n choices, once this choice is made the second ball has n-1 choices,… the k^{th} ball has (n-k) choices.

In a simple case… n=10 and k=3 this can be seen as:

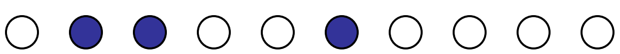


The first ball has n choices
The second ball has n-1 choices
The third ball has n-2 choices

So the number of combinations is : $n \times (n-1) \times (n-2)$

In the general case : $n \times (n-1) \times (n-2) \times (n-3) \ldots \times (n-k+1) = \dfrac{n!}{(n-k)!}$

Because we do not care about the order in which we have picked the balls



… avoid the double counting!

1	2	3
1	3	2
2	1	3
3	1	2
2	3	1
3	2	1

Each configuration is counted 6 times

This number corresponds in fact to the number of combinations of k blue balls out of k balls and therefore :

$$k \times (k-1) \times (k-2) \times (k-3) \ldots \times 1 \; = k!$$

Aka the number of re-arrangements of the k blue balls.

In order to account for each combination only once you just need to divide by the number of re-arrangements of the k blue balls.

So the number of combinations of k elements among n is given by :

$$C_n^k = \frac{n!}{k!(n-k)!}$$

The probability to pick k blue balls among n, given a probability P that the a ball is blue is thus :

$$P = C_n^k \times p^k \times (1-p)^{n-k}$$

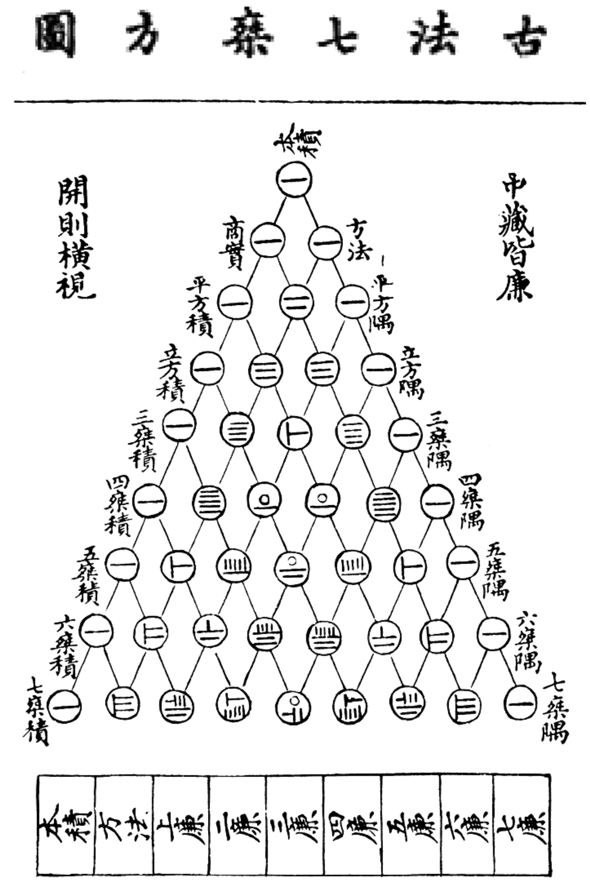
This is an absolutely fundamental formula in probability and statistics!
It is the so called Binomial Probability!

The Binomial Probability

Binomial coefficients were known since more than a thousand years…

… they were also the foundation of modern probability theory!

B. Pascal (1623-1662)



The Pascal Triangle (~1000 AD)

So what is the precision of your measurement ?

A good measure of the precision (not the accuracy) is the *Root Mean Square Deviation* (square root of the variance) of possible outcomes of the measurement.

You will compute it yourself. To do so you need two steps…
(see next slide for the full derivation)

Step 1 : Compute the mean value of the binomial probability

$$\mu = nP$$

Step 2 : Compute the variance of the binomial probability

$$Variance = nP(1-P)$$

So now you know the variance of your distribution for a given probability P…

In your case : $P = 12\%$ Assuming P is close enough to the true value, the precision is :

$$RMSD = \sqrt{nP(1-P)} = 5.6$$

The relative precision ~15% is rather poor and the accuracy questionable! (Remember, your statistical error is 45 - 36 = 9, although you are not supposed to know it !)

Step 1 : Compute mean value

Mean of the Binomial Probability

Let us denote by $\mathcal{P}_k^n(p)$ the binomial probability of an outcome k among n with single event probability p, and μ its mean value, then

$$\mu = \frac{\sum_{k=0}^{n} k\, \mathcal{P}_k^n(p)}{\underbrace{\sum_{k=0}^{n} \mathcal{P}_k^n(p)}_{=1}} = \sum_{k=0}^{n} k\, C_n^k p^k (1-p)^{n-k}$$

Note that it looks like a derivative - !

Nice trick : Start from the derivative ...

$$\frac{\partial}{\partial p}\left[\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k}\right] = 0 = \frac{\partial [1]}{\partial p}$$

so

$$\sum_{k=0}^{n} k\, C_n^k p^{k-1} (1-p)^{n-k} - \sum_{k=0}^{n} (n-k)\, C_n^k p^k (1-p)^{n-k-1} = 0$$

$$0 = \frac{\mu}{p} - n \underbrace{\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k-1}}_{1/(1-p)} + \underbrace{\sum_{k=0}^{n} k\, C_n^k p^k (1-p)^{n-k-1}}_{\mu/(1-p)}$$

thus

$$\frac{\mu}{p} + \frac{\mu}{1-p} = \frac{n}{1-p} \Rightarrow (1-p)\mu + p\mu = np$$

so $\boxed{\mu = np}$ Aha!

Step 2 : Compute variance

Variance of the Binomial Probability

Let us start from the previous formula for the mean value of the binomial probability.

$$\text{Variance} = \sum_{k=0}^{n} k^2 C_n^k p^k (1-p)^{n-k} - \underbrace{\left(\sum k\, C_n^k p^k (1-p)^{n-k}\right)^2}_{\mu^2}$$

given that $\sum_{k=0}^{n} k\, C_n^k p^k (1-p)^{n-k} = \mu = np$

then $\frac{\partial}{\partial p}\left[\sum_{k=0}^{n} k\, C_n^k p^k (1-p)^{n-k}\right] = n$ thus :

$$\sum_{k=0}^{n} k^2 C_n^k p^{k-1} (1-p)^{n-k} - \sum_{k=0}^{n} k(n-k) C_n^k p^k (1-p)^{n-k-1} = n$$

$$\frac{1}{p}\sum_{k=0}^{n} k^2 C_n^k p^k (1-p)^{n-k} - \frac{n\mu}{1-p} + \frac{1}{1-p}\sum k^2 C_n^k p^k (1-p)^{n-k} = n$$

so

$$(1-p+p)\sum_{k=0}^{n} k^2 C_n^k p^k (1-p)^{n-k} = np(1-p) + \underbrace{np\mu}_{\mu^2}$$

when $\mu = np$ so :

$$\underbrace{\sum k^2 C_n^k p^k (1-p)^{n-k} - \mu^2}_{\text{Variance !}} = np(1-p)$$

Thus $\boxed{\text{Variance} = np(1-p)}$

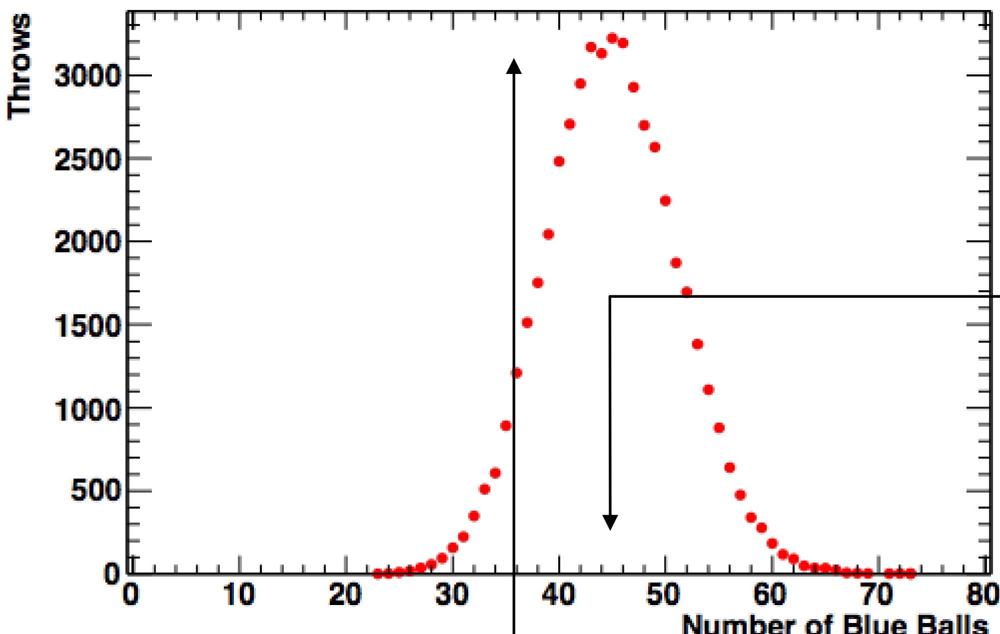
But wait…

Now you are curious to see what happens if you repeat your measurement!

You have noticed that the average binomial probability is the expected value!

Intuitively you will therefore try to repeat and average your measurements…

You will do it 50,000 times and meticulously plot the number of counts. This is what you get :



$N_{throws} = 50000$

The average number of blue balls in 50,000 throws :

$$\langle Number_{Blue} \rangle = 44.98$$

$$\langle P \rangle = 14.99\%$$

Your initial measurement (36) !

See Binomial.C

Now you decide that your measurement is the average, what is its precision ?

What is the variance of the average ?

Let's start from one straightforward property of the Variance for two random variables X and Y :

$$Var(aX+bY) = \left\langle (aX+bY-\langle aX+bY\rangle)^2 \right\rangle = \left\langle \left[a(X-\langle X\rangle)+b(Y-\langle Y\rangle)\right]^2 \right\rangle$$

$$= a^2Var(X)+b^2Var(Y)+2abCov(X,Y)$$

Where the covariance is : $Cov(X,Y) = \left\langle (X-\langle X\rangle)(Y-\langle Y\rangle) \right\rangle$

This formula generalizes to… $Var\left(\sum_{i=0}^{n} a_i X_i\right) = \sum_{i=0}^{n} a_i^2 Var(X_i) + \sum_{0\le i<j\le n} a_i a_j Cov(X_i, X_j)$

Therefore assuming that each of the bucket throws measurement N_{Blue}^k is independent from the previous one, the mean value being a simple sum of the measurements divided by the number of throws :

$$\langle Number_{Blue} \rangle = \sum_{k=1}^{N_{Throws}} N_{Blue}^k$$

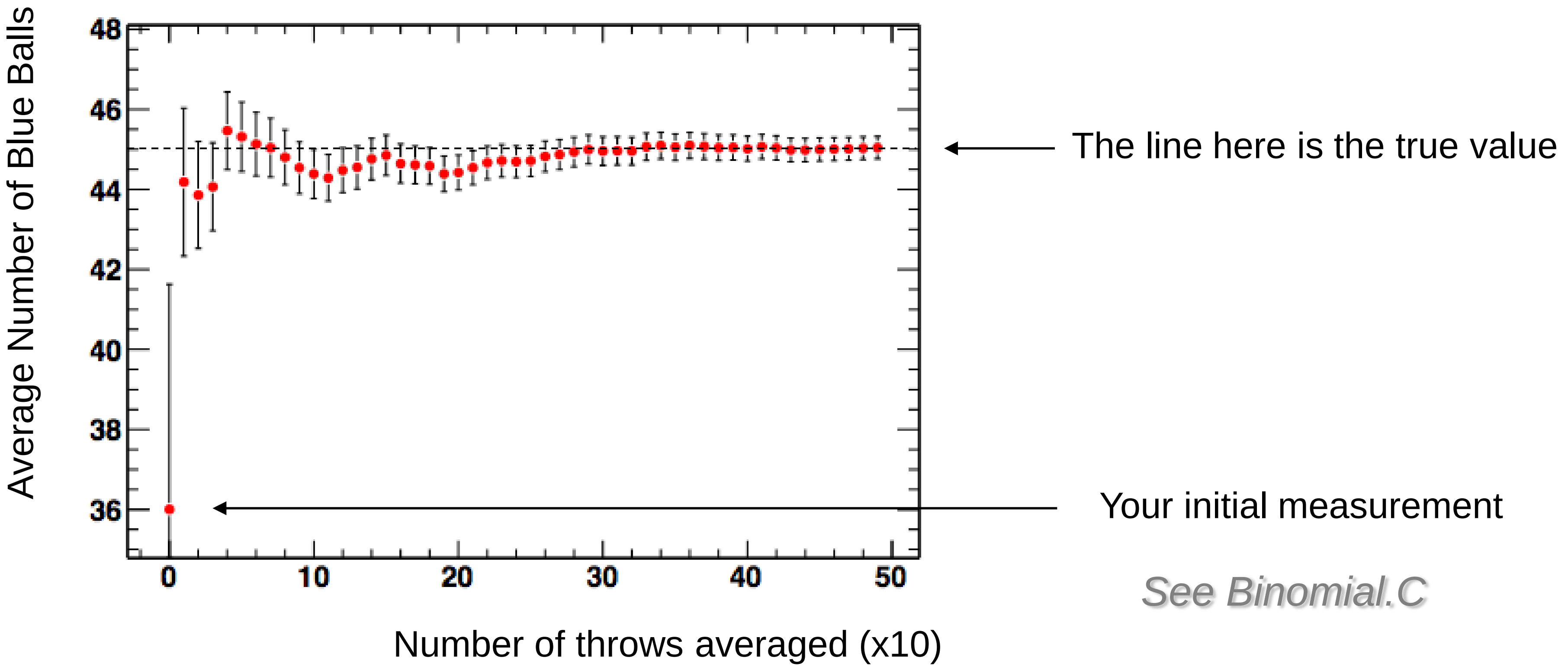
The variance then is :

$$Variance = \frac{nP(1-P)}{N_{Throws}}$$

The precision being given by the *Root Mean Square Deviation* :

$$RMSD = \sqrt{\frac{nP(1-P)}{N_{Throws}}} = \frac{RMSD_{Individual}}{\sqrt{N_{Throws}}} = 0.01\%$$

Very interesting behavior : Although you do not know the true value p, you see that the average is converging towards it with increasing precision!



See Binomial.C

This is an illustration of the *LAW of LARGE NUMBERS* ! Extremely important, intuitive but not trivial to demonstrate…

What is the meaning of our first measurement $N_{blue} = 36$?

Now that we know (after 50,000 throws) to a high precision that the probability of a blue ball is very close to 15%.

The frequency of an outcome as low as 12% is ~10% (not so unlikely!)

What difference would it make if you had known true value ?

Frequency at which the measurement is within the precision as estimated from the truth :

$$\left|P_{meas} - p\right| \leq \sqrt{np(1-p)}$$

$\Rightarrow$ 70% (of the cases the measurement is within the true statistical RMSD)

Frequency at which the true value is within the precision as estimated from the measurement :

$$\left|P_{meas} - p\right| \leq \sqrt{nP_{Meas}(1-P_{Meas})}$$

$\Rightarrow$ 67% (of the cases the true value is within the measured error)

See Coverage.C

The true value coverage is similar in the two cases, keep these values in mind…

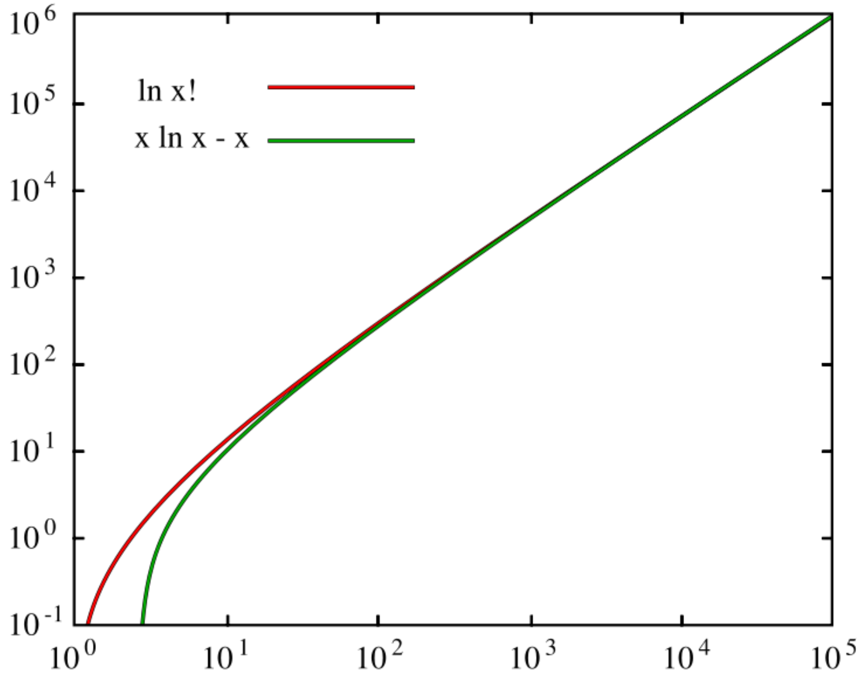
Here all results are derived from a simulation in terms of frequencies…
Computing Binomial probabilities with large numbers of N can be quite difficult !

The Gaussian or Normal Probability

Is there a way to simplify the computation ? Not so trivial to compute 300! directly…

A very nice approximation of the Binomial Probability can be achieved using Stirling's Formula !

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$



Formula is valid for large values of n…

$$C_n^k p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k-\langle k\rangle)^2}{2\sigma^2}}$$

$$\sigma = \sqrt{np(1-p)}$$

(See derivation in the next slide)

Binomial convergence towards Normal

Derivation of the Normal Probability

... Again start from the binomial probability, but this time use the Stirling formula: $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$

$$\ln(k!) \sim k\ln k - k + \frac{1}{2}\ln(2\pi k)$$

Very useful!!

then $\ln C_n^k = \ln n! - \ln k! - \ln(n-k)!$

thus

$$\ln\left[C_n^k p^k (1-p)^{n-k}\right] = n\ln n - n + \frac{1}{2}\ln(2\pi n) - \left[k\ln k - k + \frac{1}{2}\ln(2\pi k)\right] - \left[(n-k)\ln(n-k) - n + k + \frac{1}{2}\ln(2\pi(n-k))\right] + k\ln p + (n-k)\ln(1-p)$$

Again using a derivation trick

$$\frac{\partial}{\partial k}\left[\ln\left(C_n^k p^k (1-p)^{n-k}\right)\right] = \underbrace{-\ln k + \ln(n-k) + \ln p - \ln(1-p)}_{\ln\left(\frac{p(n-k)}{k(1-p)}\right)} \underbrace{-\frac{1}{2k} + \frac{1}{2(n-k)}}_{O\left(\frac{1}{k}\right)}$$

We see that the maximum value will occur at

$p(n-k) = k(1-p)$ or $\boxed{k = np}$ Which is also the average value.

Then looking at the second derivation:

$$\frac{\partial^2}{\partial k^2}\left[\ln\left(C_n^k p^k (1-p)^{n-k}\right)\right] = -\left(\frac{1}{n-k} + \frac{1}{k}\right) + O\left(\frac{1}{k^2}\right)$$

then take at maximum value $k = np$:

$$\frac{\partial^2}{\partial k^2}\left[\ln\left(C_n^k p^k (1-p)^{n-k}\right)\right] \sim \frac{-1}{np(1-p)} = -\frac{1}{\sigma^2}$$

where $\sigma = \sqrt{np(1-p)}$

Therefore with a second order Taylor series expansion:

$$\ln\left(C_n^k p^k (1-p)^{n-k}\right) \sim \underbrace{\ln A}_{\text{constant}} + \frac{(k-np)^2}{2}\cdot\left(\frac{-1}{\sigma^2}\right)$$

therefore $C_n^k p^k (1-p)^{n-k} \sim A\, e^{-\frac{(k-np)^2}{2\sigma^2}}$

introducing the notation $\tilde{k} = np$ corresponding to the maximum of the probability - then

$$C_n^k p^k (1-p)^{n-k} \sim A\, e^{-\frac{(k-\tilde{k})^2}{2\sigma^2}}$$

then using the normalization of the probability

$$\int_{-\infty}^{\infty} A\, e^{-\frac{(k-\tilde{k})^2}{2\sigma^2}}\, dk = 1$$ and the Gauss integral

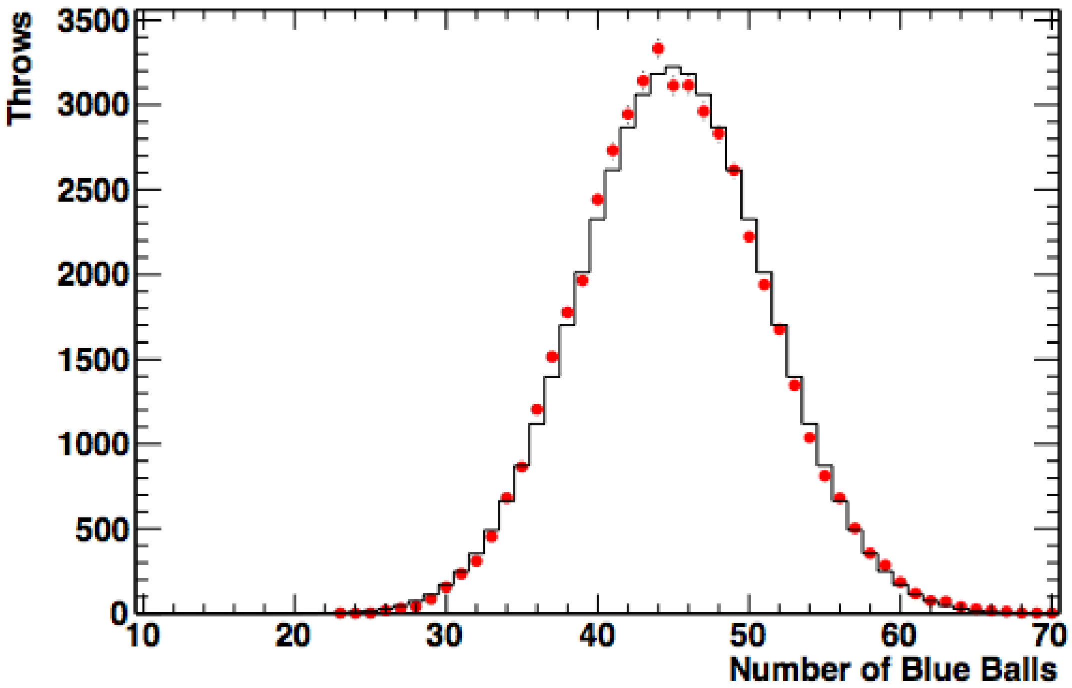
$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\sigma^2}} = \sqrt{2\pi\sigma^2}$$ so:

$$C_n^k p^k (1-p)^{n-k} = \frac{e^{-\frac{(k-\tilde{k})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Validity of the Normal Convergence (Approximation)

Does the approximation apply to our bucket experiment (n=300 and p=15%) ?

See NormalConvergence.C



C. F. Gauss (1777-1855)

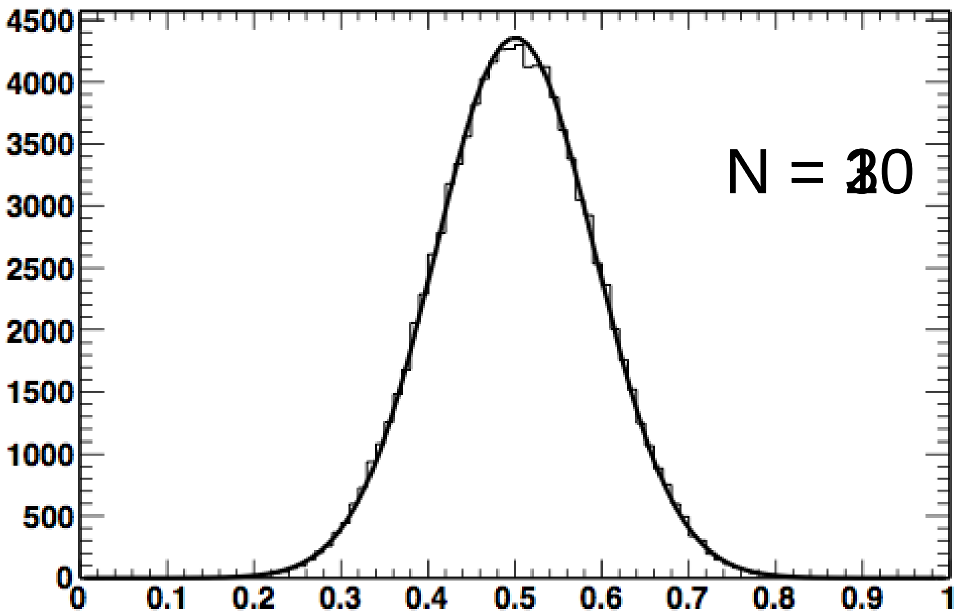
Not bad (although not perfect) !

In practice you can use the normal law when approximately n>30 and np>5

What is so “Normal” About the Gaussian?
The Central Limit Theorem… … at Work !

When averaging various independent random variables *(and identically distributed)* the distribution of the average converges towards a Gaussian distribution

See CLT.C



$$\text{RMS} = \frac{[0,1]}{\sqrt{12}} \times \frac{1}{\sqrt{10}}$$

At N=10 an excellent agreement with a gaussian distribution is observed

The CLT is one of the main reasons for the great success of the Gaussian law…

On the one hand the CLT is very powerful to describe all those phenomena that result from the superposition of various other phenomena… but on the other hand it is just a limit…

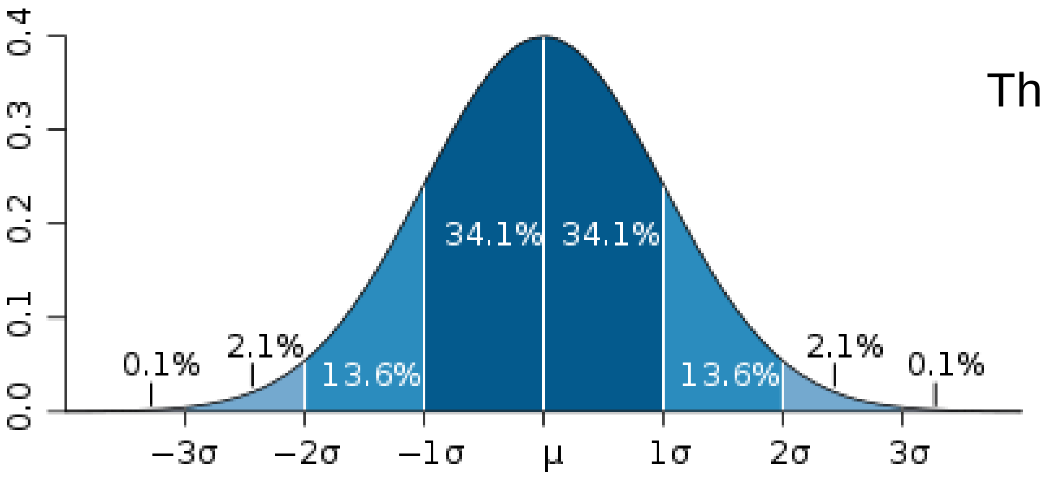
The Notion of Standard Error

Starting from the gaussian PDF :

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

Let's give a first definition of a central confidence interval as the deviation from the central value…

$$P(a\sigma) = \int_{\mu-a\sigma}^{\mu+a\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Then for :
- a = 1 : P(aσ) = 68.3%
- a = 2 : P(aσ) = 95.4%
- a = 3 : P(aσ) = 99.7 %

See NormalCoverage.C

If you knew the true value of the "error" (σ) then you could say that the in the gaussian limit that the true value has 68.3% probability to be within the 1s, but in many practical examples (such as the well) the true value of the error is not known…

How does the Bucket Experiment Relate to Particle Physics?

The bucket experiment is the measurement of an abundance (blue balls)…

This is precisely what we call in particle physics cross sections…

… except that the bucket contains all collisions collected in an experiment so…

- We try to fill it as much as possible (N is very large and not constant!)
- The processes we are looking for are very rare (p is very small)

The very large N makes it difficult to compute the binomial probability…

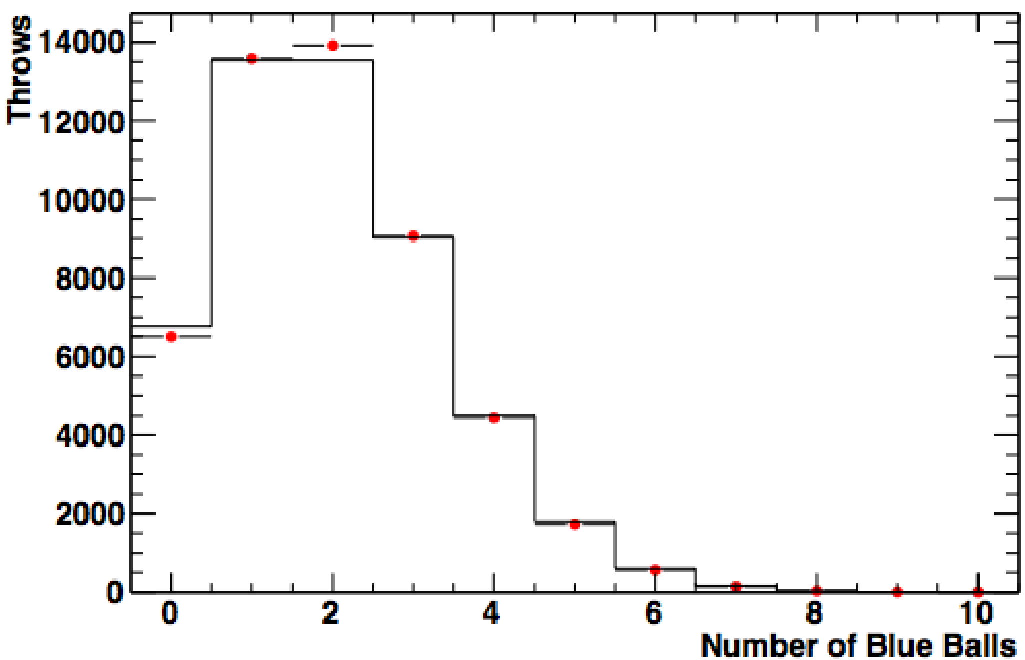
The Poisson Probability

In the large *n* and small *p* limit and assuming that *np* = μ is finite you can show (see next slide) that …☐☐

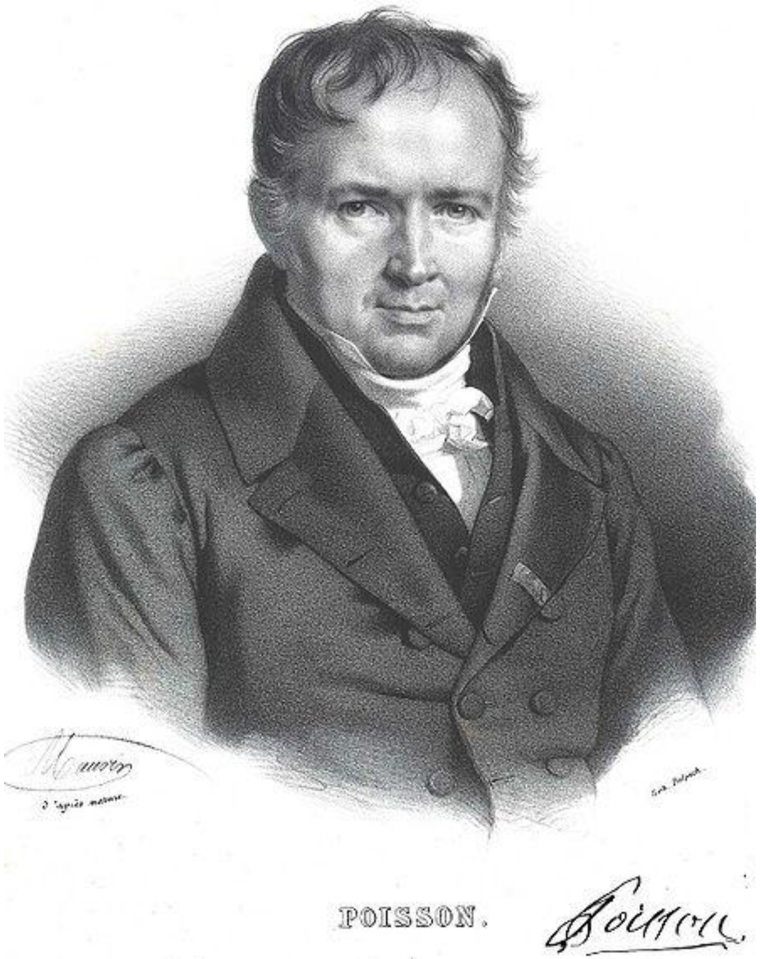
$$C_n^k p^k (1-p)^{n-k} \approx \frac{(np)^k}{k!} e^{-(np)} = \frac{\mu^k}{k!} e^{-\mu}$$

Much simpler formulation! In practice you can use the normal law when approximately n>30 and np<5

See PoissonConvergence.C



N=100 and p=25%



S. D. Poisson (1781-1840)

Interesting to note that Poisson developed his theory trying not to solve a game of chance problem but a question of Social Science !

Derivation of the Poisson Probability

... from the binomial probability : $C_n^k p^k (1-p)^{n-k}$
which can be written in the limit when

p is very small and np is finite.

$$C_n^k = \frac{n!}{k!\,(n-k)!} = \frac{1}{k!}\left[n(n-1)\dots(n-k+1)\right] \sim \frac{n^k}{k!}$$

and for p small

$$(1-p)^{n-k} = 1 - (n-k)p + \frac{(n-k)(n-k-1)}{2!}p^2 + \dots \simeq 1 - np + \frac{(pn)^2}{2!} + \dots \approx e^{-pn}$$

therefore

$$C_n^k p^k (1-p)^{n-k} \approx \frac{n^k}{k!} p^k e^{-np} = \frac{(np)^k}{k!} e^{-np}$$

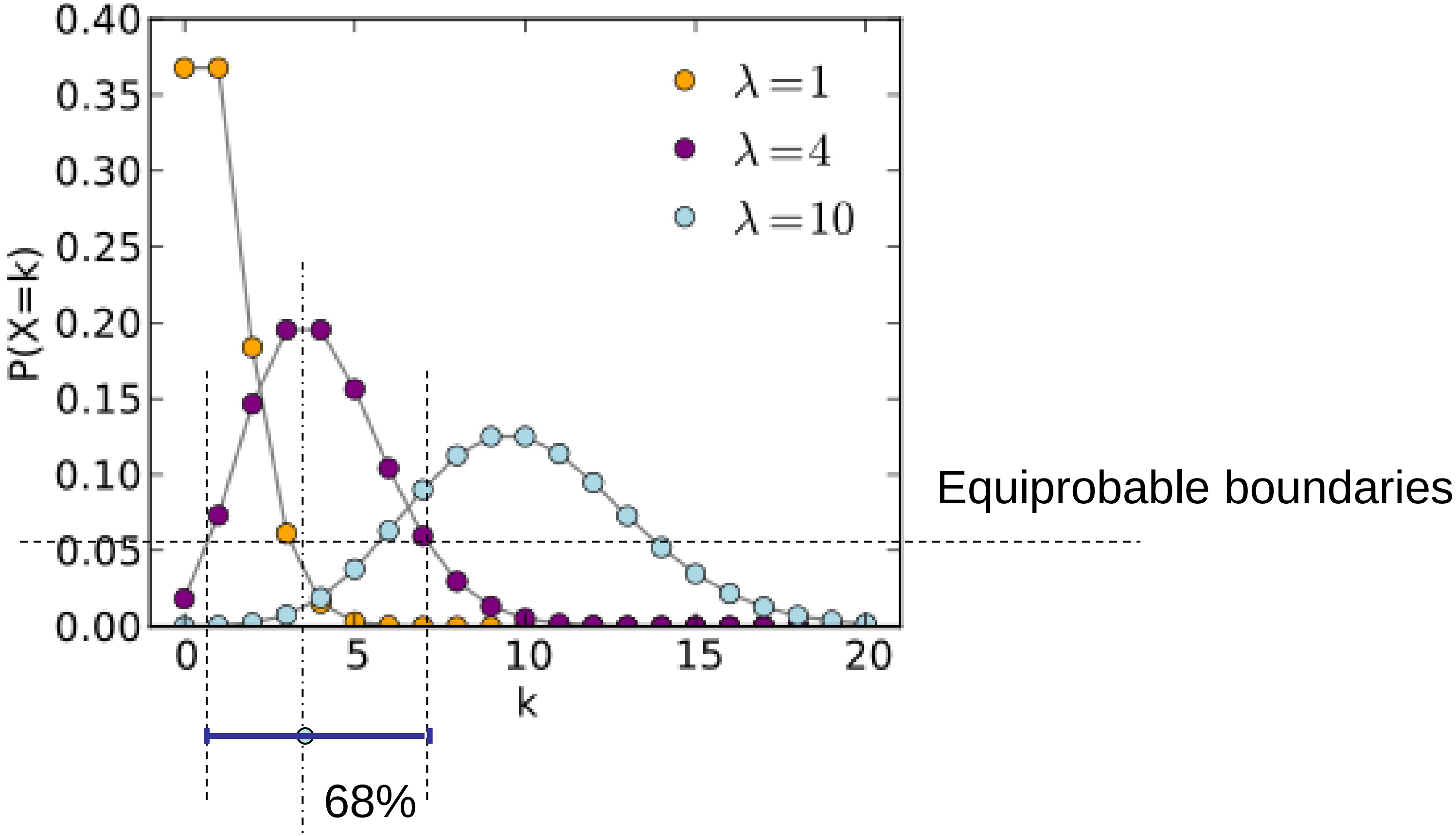
thus denoting $\mu = np$ the poisson probability of an outcome k given an expectation of μ is

$$\boxed{P(k,\mu) = \frac{\mu^k}{k} e^{-\mu}}$$

This approximation works essentially when ($n > 50$ and $p < 0.1$).

Poisson Intervals (or Errors)

Now how will you define a central confidence interval in a non symmetric case ?



The integration needs to start from the most probable value downwards…

Here is our first encounter with the necessity of an ordering !

What have we learned ?

…and a few by-products…

1.- Repeating measurements allows to converge towards the true value of an observable more and more precisely …

But never reach it with infinite precision !!!

Even more so accounting for systematics…
(what if the balls do not have an homogeneous distribution ?)

2.- Binomial variance is also useful to compute the so-called binomial error, mostly used for efficiencies :

$$\sigma_\varepsilon = \frac{\sigma_\mu}{N} = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} \qquad \mu = np$$

For an efficiency you must consider n fixed !

3.- We came across a very important formula in the previous slides

$$Var\left(\sum_{i=0}^{n} a_i X_i\right) = \sum_{i=0}^{n} a_i^2 Var(X_i) + \sum_{0\le i<j\le n} a_i a_j Cov(X_i, X_j)$$

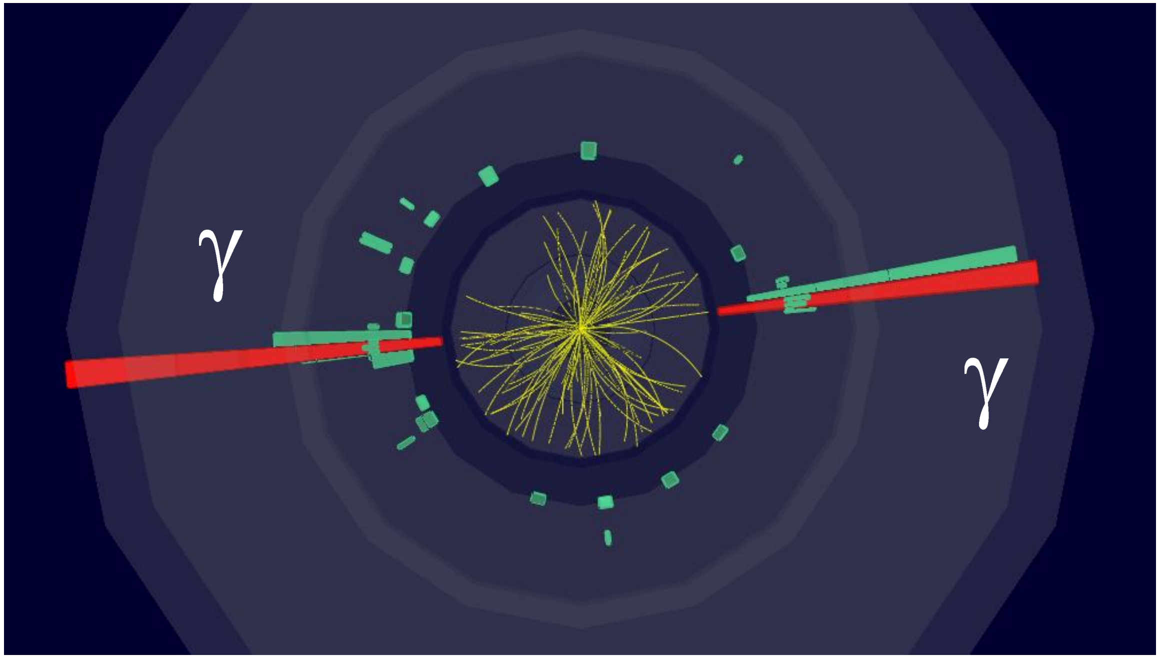
That generalizes (with a simple Taylor expansion) to…

$$\mathrm{var}(f(x_1,\ldots,x_n)) = \sum_{i=0}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \mathrm{var}(x_i) + \sum_{0\le i<j\le n} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j} \mathrm{cov}(x_i, x_j)$$

Likelihood

Unfortunately in High Energy physics experiments, events (balls) don't come in single colors (white or blue) … Their properties are not as distinct !

For instance take this simple event :



Could be many things …

Higgs ?

QuickTime™ et un décompresseur sont requis pour visionner cette image.

Background ?

QuickTime™ et un décompresseur sont requis pour visionner cette image.

Let alone that they can be indistinguishable (quantum interference)

How can we distinguish between the two ?

Very vast question, let's first start with how to measure their properties

(Which is also a very vast question!)

One clear distinctive feature is that the signal is a narrow mass resonance, while the background is a continuum !

QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

To measure properties in general *(a.k.a. parameter estimation)* among the most commonly used tools is the maximum likelihood fit…

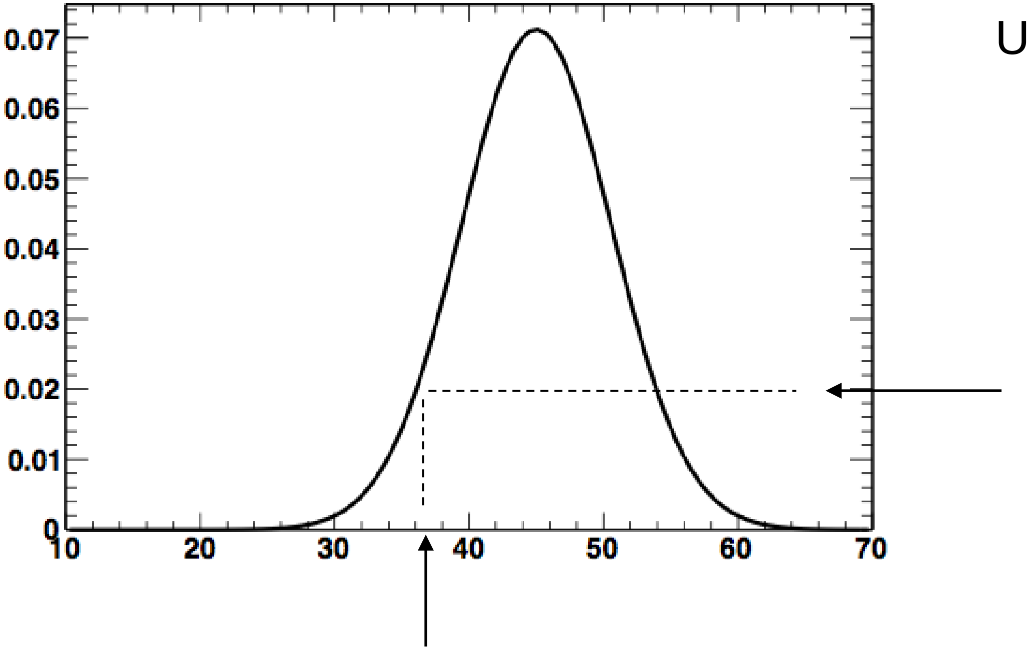
What is a Likelihood ?

A simple way of defining a Likelihood is a Probability Density Function (PDF) which depends on a certain number of parameters…

Simplistic definition is a function with integral equal to 1…

Let's return to the well experiment but under a different angle this time…

(but this applies to any parameter estimate)



Under certain hypothesis :

- Gaussian centered at 45 (p=15%)
- Width equal to error for 1 bucket (~6.2 blue balls)

Here is its probability !

or Likelihood

Not so likely !

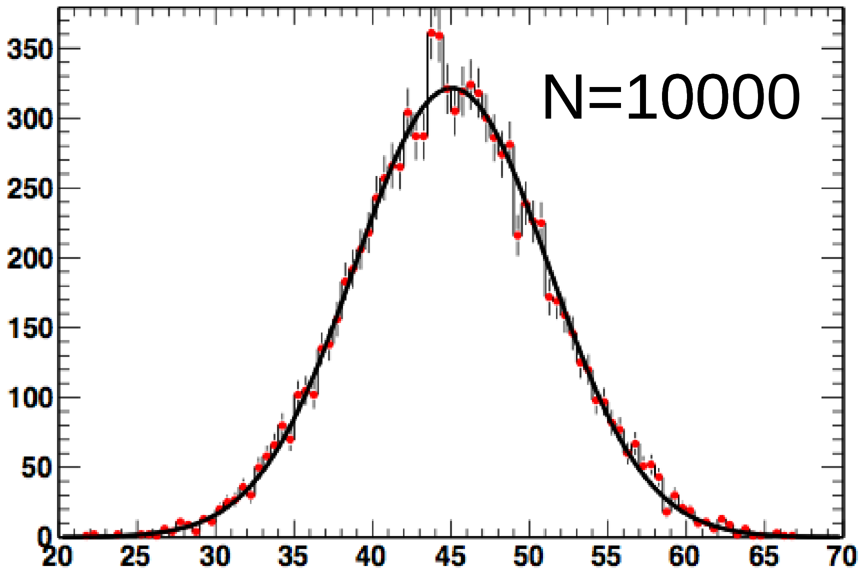
Here is your first measurement (36) !

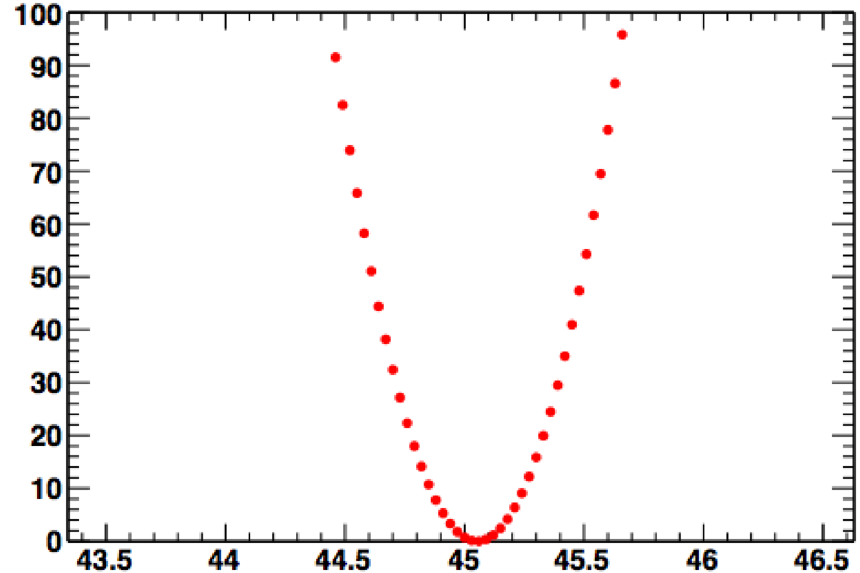
What happens when we throw more buckets ?

$$L(\mu) = \prod_{i=1}^{N} f_{\mu}(n_i)$$

Then the probability of each bucket can be multiplied!

See Fit.C





This probability will soon be very very small $(O(0.1))^{100\ldots}$ It is easier to handle its log :

$$\ln(L(\mu)) = \sum_{i=1}^{n} \ln(f_{\mu}(n_i))$$

Then tp estimate a parameter one just has to maximize this function of the parameter μ (or minimize -2lnL you will see why in a slide)…

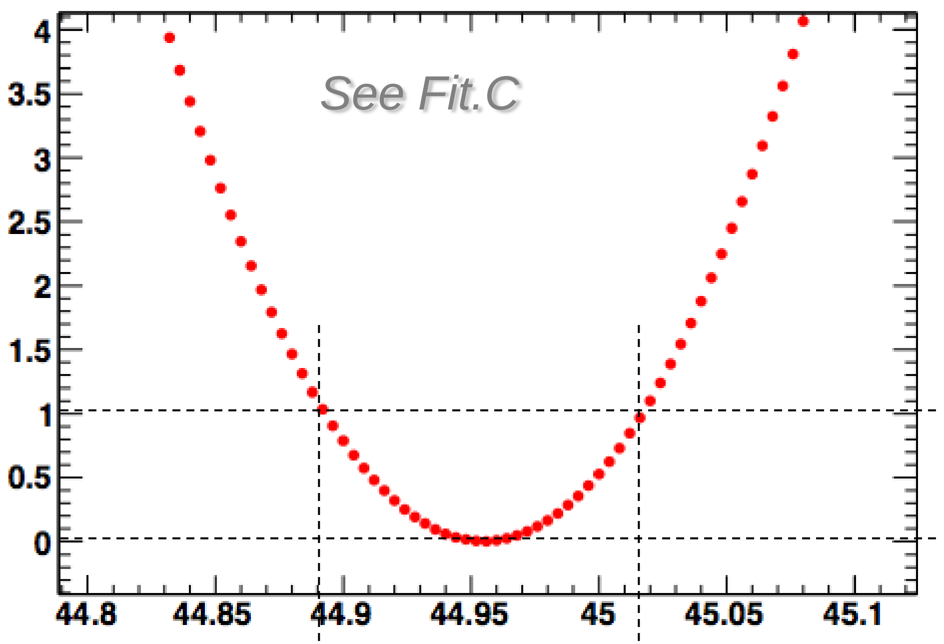
See how the accuracy translates in the sharpness of the minimum!

In our simple (but not unusual) case we can see that :

$$-2\ln(L(\mu)) = -2\sum_{i=1}^{n}\ln(f_\mu(n_i)) = -2\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(n_i-\mu)^2}{2\sigma^2}}\right) = \underbrace{\sum_{i=1}^{n}\frac{(n_i-\mu)^2}{\sigma^2}}_{} + cste$$

This is also called χ^2

There is an exact equivalence between maximizing the Likelihood or minimizing the χ^2 (Least Squares Method) in the case of a gaussian PDF



You can also see that the error on the measured value will be given by a variation of -2 ln L of one unit :

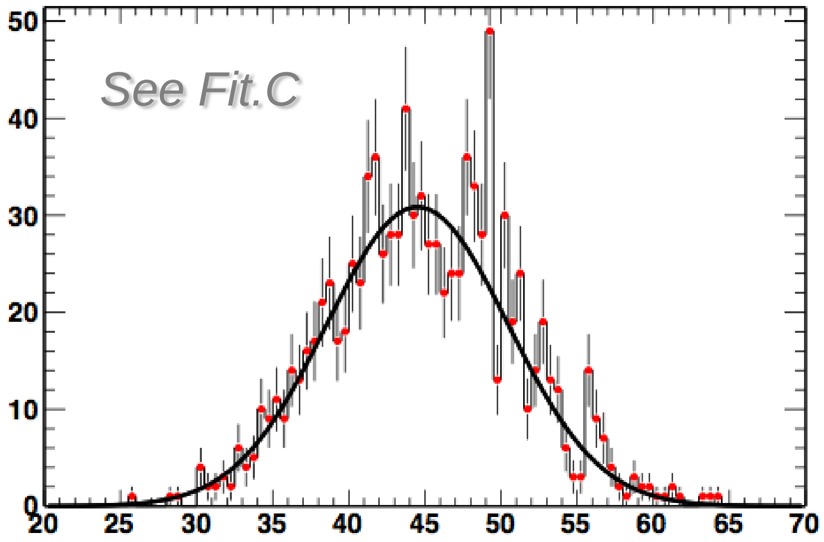
$$\Delta(-2\ln(L(\mu))) = 1$$

$$\bar{\mu} = 44.95 \pm \underbrace{0.06}_{}$$

Which is precisely $\frac{\sigma}{\sqrt{n}}$

What have we learned?

How to perform an unbinned likelihood fit :



For n=1000 the fit yields

$$\bar{\mu} = 44.91 \pm 0.19$$

Using a simple binned fit (as shown here with 100 bins) in the same data yields :

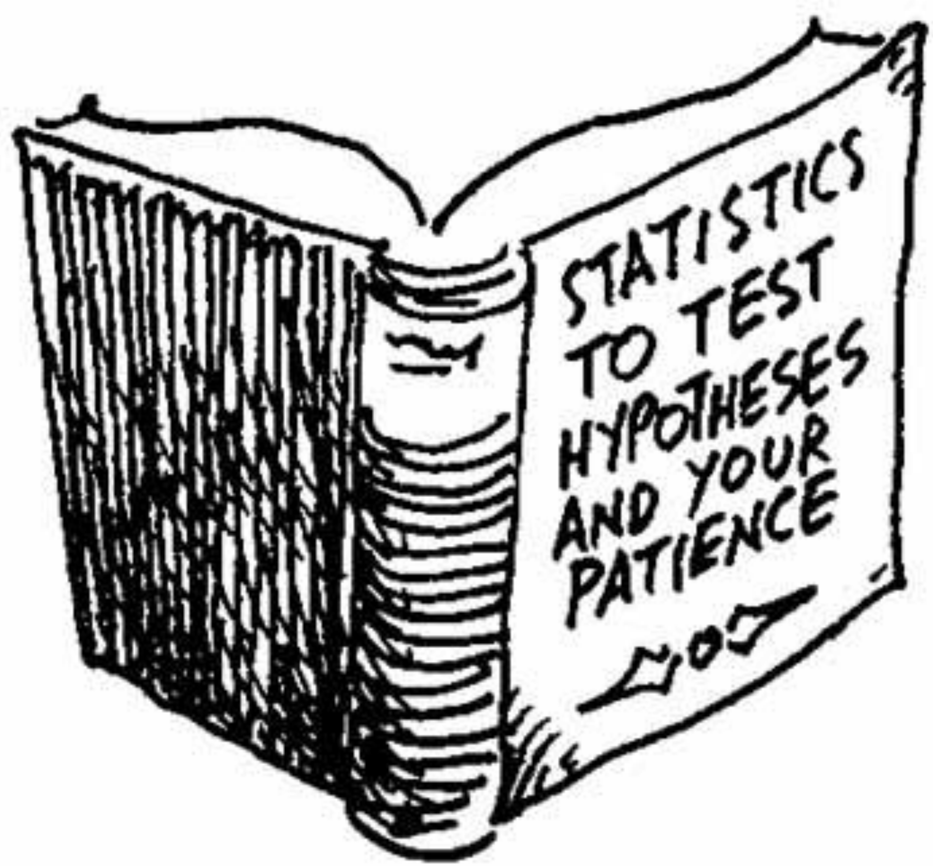
$$\bar{\mu} = 44.81 \pm 0.20$$

LSM between the PDF and the bin value

This can of course be applied to any parameter estimation, as for instance the di-photon reconstructed mass !

Hypothesis Testing

How to set limits or claim discovery ?



Hypothesis Testing in HEP Boils Down to One Question :
Is there a Signal ?

Exclusion, Observation or Discovery ?

The goal here is to assess quantitatively the compatibility of our observation with two hypotheses :

No-Signal (H_0) and presence of Signal (H_1)…

First we need to be able to have estimate whether an experiment is more Signal-like or Background-Like. In other words we need to have an ordering of our experiments.

Neyman construction (1933)

The first obvious way of ordering possible outcomes of experiments is the number of observed events…

Imagine a simple experiment which is *almost* background free.

Let's define the following estimator : $E = n_{evts}$

If you observe 0 events… then your experiment will be background like!

What if you observe one event or more ?

What would the exclusion limit on a signal be ?

What does setting an exclusion limit really means ?

When an experiment is made, excluding a given signal means that the probability for the hypothetical signal process to yield the outcome is small.

Typically exclusions are made when a signal will not yield an outcome more background like than the one observed more than ~5% of the times.

In our simple “Event Counting” experiment, using the Poisson probability, the probability that a signal experiment yields an outcome more background like than observed is given by :

$$CL_s = e^{-s} \sum_{i=0}^{n_{obs}} \frac{s^i}{i!}$$

In such case the limit on the signal yield will be given by the simple equations :

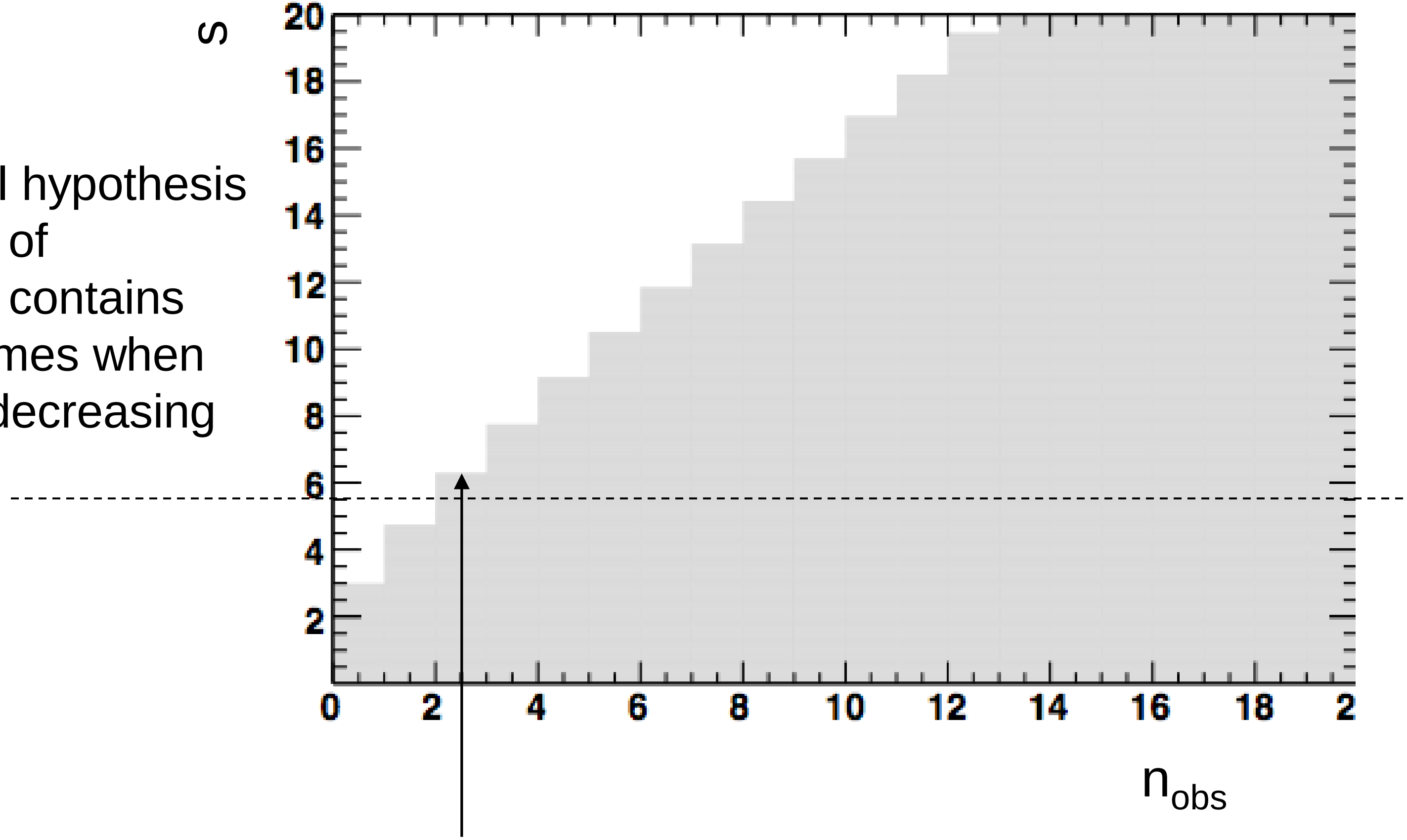
$$\begin{cases} n_{obs} = 0 \Rightarrow \overline{N}_{95} = 3 & (e^{-s} = 0.05) \\ n_{obs} = 1 \Rightarrow \overline{N}_{95} = 4.23 & (e^{-s}(1+s) = 0.05) \\ n_{obs} = 2 \Rightarrow \overline{N}_{95} = 6.30 & (e^{-s}(1+s+\frac{s^2}{2}) = 0.05) \end{cases}$$

Neyman Construction

Another way of looking at this is the Neyman construction…

See Neyman.C

For a given signal hypothesis what is the range of observations that contains 95% of the outcomes when accounted for in decreasing order (E)

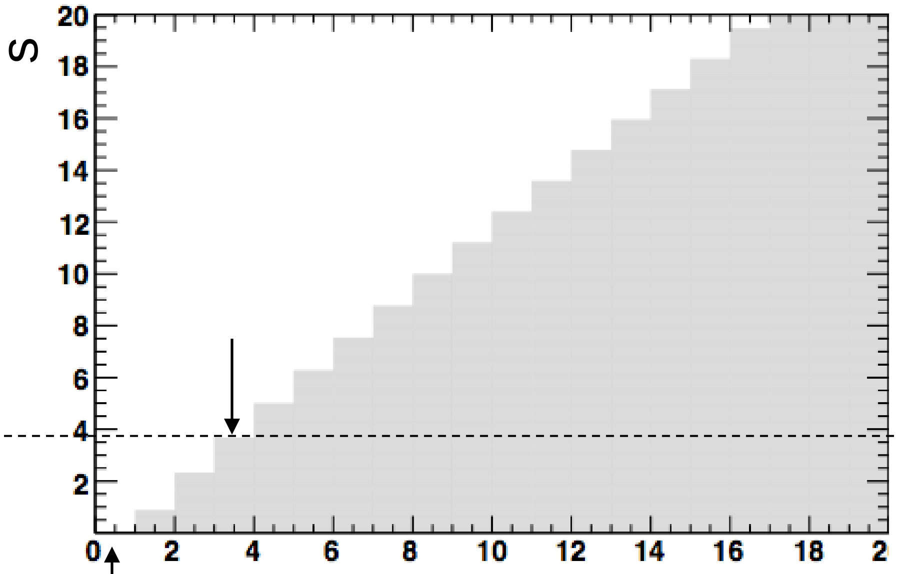


The contour yields the limit for a given observation

Another example in the presence of background (b=3)

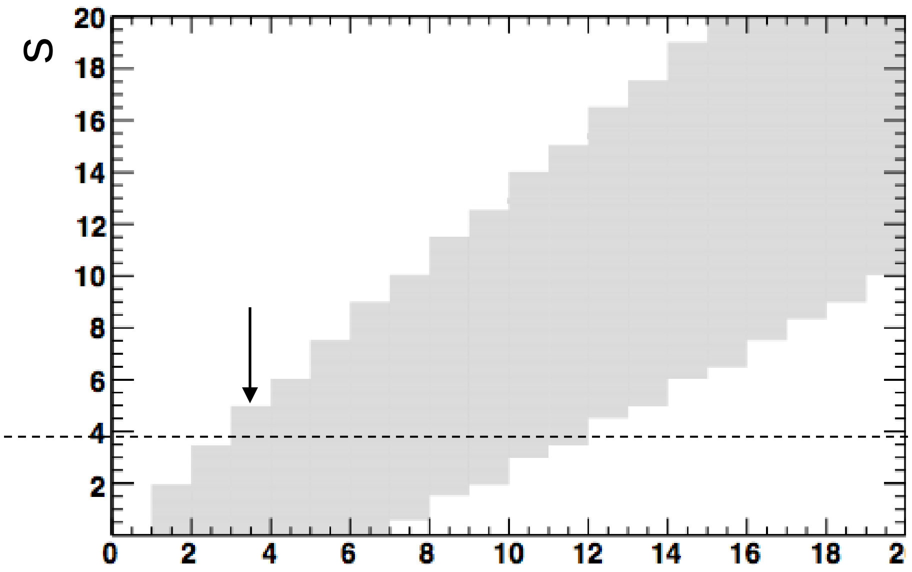
$$E = P(n \mid s) = e^{-(s+b)}\frac{(s+b)^n}{n!}$$

95% Confidence area (E = n)



See Neyman.C

n_{obs}



n_{obs}

Example of construction of the central confidence interval (as for the Poisson error)

Problem 1 : 0 signal is excluded (non sense) !

Problem 2 : When/how to switch to a central confidence interval ?

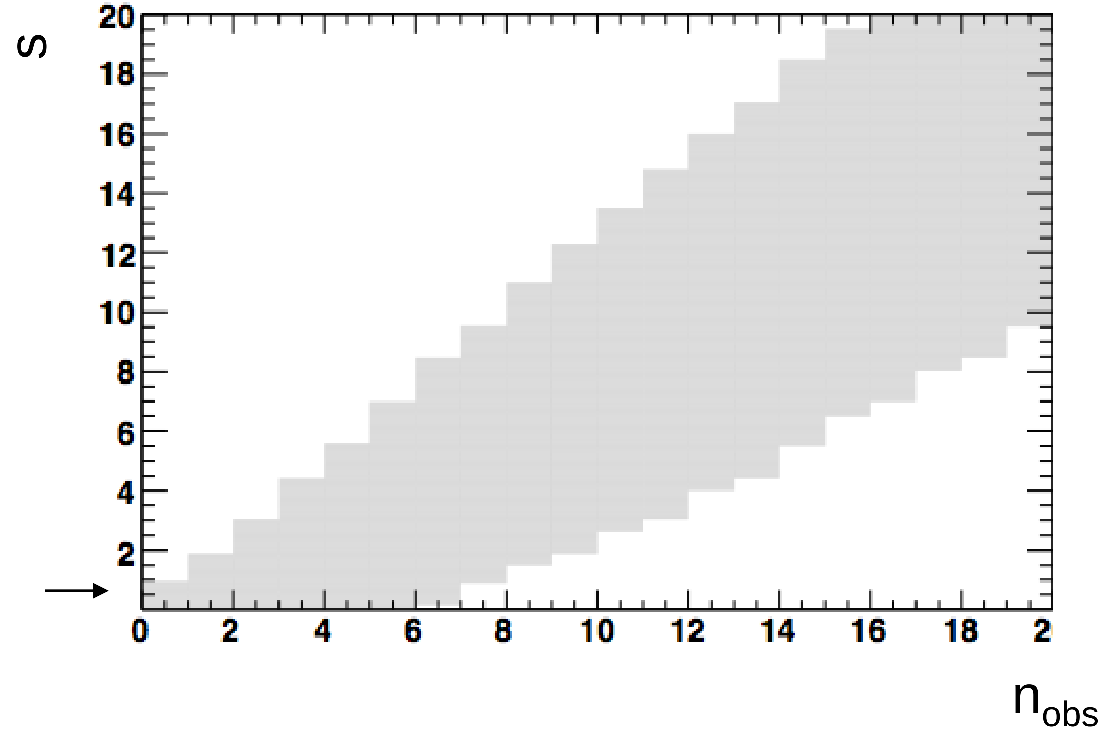
Flip-Flopping

Both problems are solved by the G. J. Feldman and R. D. Cousins ordering parameter :

$$P(n \mid s) \equiv \frac{P(n \mid s+b)}{P(n \mid \max(0, n-b))}$$

Best signal estimate in the data (maximizes P)

See Neyman.C



Note the small exclusion (it is in good agreement with the background fluctuation)

This method is related to a general Lemma (see next slide) and has inspired more advanced techniques in hypotheses testing…

The Neyman-Pearson Lemma

The underlying concept in ordering experiments is really to quantify the compatibility of the observation with the signal hypothesis (H1) …

The problem of testing Hypotheses was studied in the 30's by Jerzy Neyman and Egon Pearson…

They have shown that the ratio of likelihoods of an observation under the two hypotheses is the most powerful tool (or test-statistic or order parameter) to

$$E = \frac{P(H_1 \mid x)}{P(H_0 \mid x)}$$

The Profile Likelihood

A very useful tool to compute limits, observation or discovery sensitivties and treat systematics is the Profile Likelihood…

Let's again take the example of the H$\rightarrow$gg analysis at LHC *(in ATLAS)*

QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

We have a simple model for the background :

$$b(m,\theta) = \theta_1 e^{-\theta_2 m}$$

Relies only on two parameters

Assume a very simple model for the signal :

$$s(m,\mu) = \mu s \times Gauss(m)$$

The Gaussian is centered at 120 GeV/c^2 and a width of 1.4 GeV/c^2

The Profile Likelihood

The overall fit model is very simple :

$$L(\mu,\theta \mid data) = \prod_{i \in data} (s(m_i,\mu) + b(m_i,\theta))$$

This model relies essentially only on two types of parameters :

- The signal strength parameter (μ) It is essentially the signal normalization
- The nuisance parameters (θ) Background description in the *"side bands"*

$$\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu) \mid data)}{L(\hat{\mu},\hat{\theta} \mid data)}$$

$\longleftarrow$ Test of a given signal hypothesis μ

$\longleftarrow$ Best fit of the data

Prescription similar to the Feldman Cousins

Usually work with the estimator : $q_\mu = -2\ln(\lambda(\mu))$ Because …

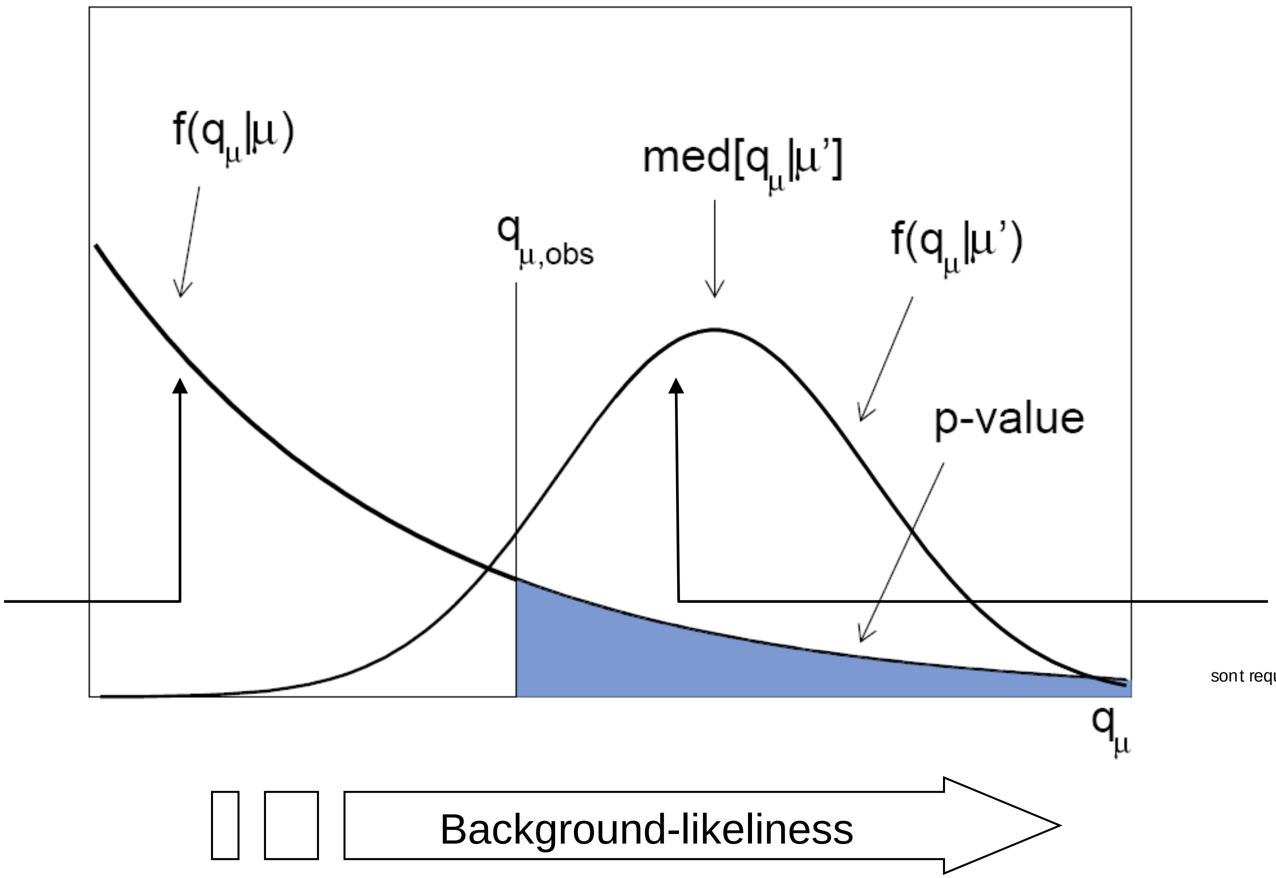
Wilks' Theorem

Under the H_μ Signal hypothesis the PL is distributed as a χ^2 with 1 d.o.f. !

(v.i.z a well know analytical function)

To estimate the overall statistical behavior, toy MC full experiments are simulated and fitted !

QuickTime™ et un décompresseur sont requis pour visionner cette image.



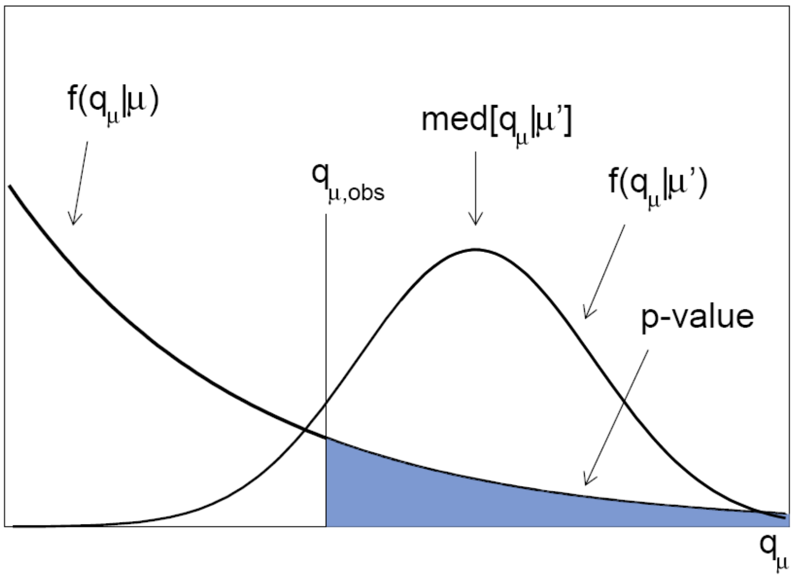
QuickTime™ et un décompresseur sont requis pour visionner cette image.

Signal-plus-background Toy experiments

Background only Toy experiments (μ'=0)

95% CL Limits

The observed 95% CL upper limit on μ is obtained by varying μ until the p value :



$$1 - CL_{s+b} = p = \int_{q_{obs}}^{+\infty} f(q_\mu \,|\, \mu)\, dq_\mu = 5\%$$

Analytically simple

This means in other words that if there is a signal with strength μ, the false exclusion probability is 5%.

The 95% CL exclusion sensitivity is obtained by varying μ until the p value :

$$p = \int_{med(\underbrace{q_\mu|0}_{\text{Background only experiments}})}^{+\infty} f(q_\mu \,|\, \mu)\, dq_\mu = 5\%$$

Exclusion Results

Performing this analysis for several mass hypotheses and using CL_{s+b} the exclusion has the same problem as the simple Poisson exclusion with background…

No-Signal (H_0) and presence of Signal (H_1)…

i.e. a signal of 0 can be excluded with a fluctuation of the background

décompresseur
sont requis pour visionner cette image.

QuickTime™ et un
décompresseur

We thus apply the (conservative) "modified frequentist" approach that requires :

$$CL_s = CL_{s+b} / CL_b = 5\% \quad \text{where} \quad CL_b = \int_{q_{obs}}^{+\infty} f(q_\mu \mid 0)\, dq_\mu$$

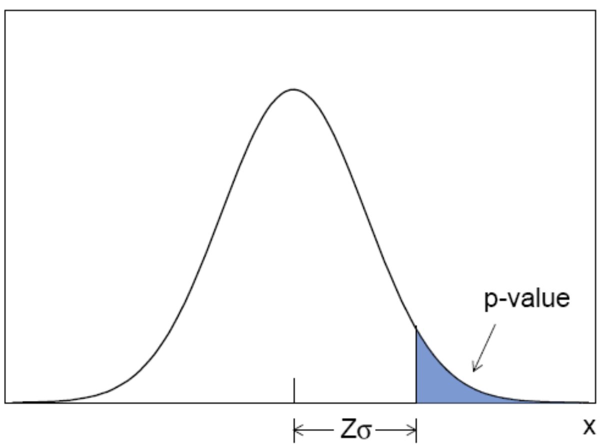
Observation and Discovery

The method is essentially the same, only the estimator changes…we now use q_0

In this case the $f(q_0|0)$ will be distributed as a χ^2 with 1 d.o.f. (Wilks' theorem)

$$p = \int_{q_{obs}}^{+\infty} f(q_0 \mid 0) dq_0$$

- To claim an observation (3 σ) : the conventional p-value required is 1.35 10^{-3}
- To claim an observation (5 σ) : the conventional p-value required is 2.87 10^{-7}



Corresponds to the "one sided" convention

This means in other words that in absence of signal, the false discovery probability is p.

« a probability of 1 in 10 000 000 is *almost* impossible to estimate »

R. P. Feynman

(What do you care what other people think?)

Conclusion

We went through an overview of the fundamental concepts of statistics for HEP

If possible take some time to play with the Root-Macros for hands-on experience

You should now be able to understand the following plot !

QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

There is a lot more for you/us to learn about statistical techniques

In particular concerning the treatment of systematics

So be patient and take some time to understand the techniques step by step…
… and follow Laplace's advice about statistics !