



# SOS School 2018

## Classical Interval Estimation and Hypothesis Testing

Tommaso Dorigo

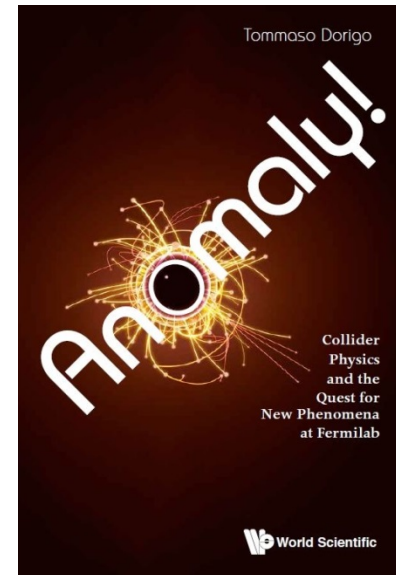
INFN Padova

La Londe Les Maures, May 28-29, 2018

# About Your Lecturer

- I am a INFN researcher, working in the CMS experiment at CERN since 2002
  - member of the CMS Statistics Committee, 2009- (and chair, 2012-2015)
  - Scientific coordinator of AMVA4NewPhysics network, 2015-
- Previously (1992-2010) I have worked in the CDF experiment at the Tevatron
- Besides research, I do physics outreach in a blog since 2005. The blog is now at [http://www.science20.com/quantum\\_diaries\\_survivor](http://www.science20.com/quantum_diaries_survivor)
- Ways to contact me:
  - Email: [tommaso.dorigo@gmail.com](mailto:tommaso.dorigo@gmail.com)
  - Skype: tonno923 (seldom online)
  - Twitter: dorigo
  - Phone: 3666995594
  - Office phone: 0039 – 049 967 7230

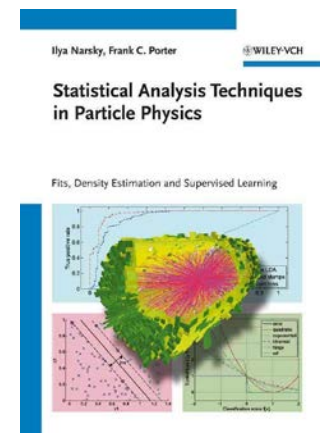
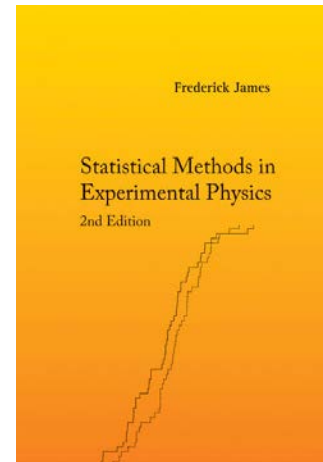
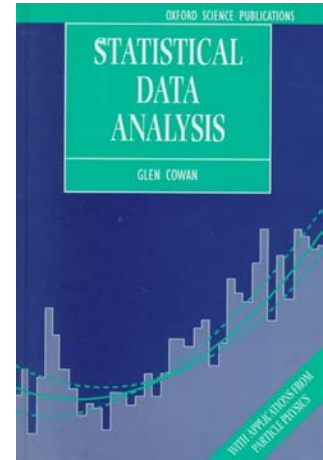
- I recently published a book on how HEP discoveries are made and not made – contains **discussions on how statistical inference is made in large particle physics experiments**



More info at the World Scientific page:  
<http://www.worldscientific.com/worldscibooks/10.1142/q0032>

# Better Advices for Books

- Glen Cowan, "*Statistical Data Analysis*", Oxford Science Publications 1998
  - Easy, clear, concise. Provides basic understanding on all the common topics, but lacks in-depth treatment of some advanced material important for HEP (e.g. MVA)
- F. James, "*Statistical Methods in Experimental Physics*", 2nd ed., World Scientific 2002
  - A serious handbook which contains advanced treatment of many important problems for HEP. Also not complete.
- I. Narsky, F. Porter, "*Statistical Analysis Techniques in Particle Physics*", Wiley 2014
  - A sharp focus on Multivariate Analysis techniques and their applications to HEP. Aimed at problem solving and extensive, although concise on any given topic.



# Practicalities

You can find the code used for some of the examples of these lectures in the links below

**Mind the underscores** →  
they are where you  
see a space in the name

These are simple ROOT  
macros – the code is  
ugly but hopefully easy  
to understand

Code for exercises in:

[http://www.pd.infn.it/%7Edorigo/Poisson\\_prob\\_fix.C](http://www.pd.infn.it/%7Edorigo/Poisson_prob_fix.C)

[http://www.pd.infn.it/%7Edorigo/Poisson\\_prob\\_fluct.C](http://www.pd.infn.it/%7Edorigo/Poisson_prob_fluct.C)

[http://www.pd.infn.it/%7Edorigo/FlipFlop\\_exercise.C](http://www.pd.infn.it/%7Edorigo/FlipFlop_exercise.C)

<http://www.pd.infn.it/%7Edorigo/FlipFlop.C>

<http://www.pd.infn.it/%7Edorigo/Coverage.C>

<http://www.pd.infn.ig/%7Edorigo/Die.C>

<http://www.pd.infn.ig/%7Edorigo/Die5.C>

[http://www.pd.infn.ig/%7Edorigo/Bootstrap\\_variance.C](http://www.pd.infn.ig/%7Edorigo/Bootstrap_variance.C)

A couple more practicalities:

- **text in green** shows proposed exercises
- **text in purple** indicates questions to you
- references[xx] are given in the text, listed at the end

**And don't forget to ask questions when I am not clear  
(surprisingly it does happen!)**

# Contents

Today:

- Classical interval estimation
- Derivation of upper and lower limits

Tomorrow:

- Hypothesis testing
- CLs and the Higgs search methodology

with examples and exercises scattered around.

Note: these slides are packed full with text. This has the purpose of making them easy to use offline - but this makes it hard for you to follow, especially if you take notes. So don't: there's everything you need already

# Statistics Matters!

- To be a good physicist, **one MUST understand Statistics:**
  - “*Our results were inconclusive, so we had to use Statistics*”  
**We are quite often in that situation in HEP !**
  - A good knowledge of Statistics allows you to make **optimal use** of your measurements, *obtaining more precise results than your colleagues*, other things being equal
  - It is **very easy to draw wrong inferences from your data**, if you lack some basic knowledge on Statistics (it is easy regardless!)
  - Foundational Statistics issues **play a role** in our measurements, because **different statistical approaches provide different results**
    - There is nothing wrong with this: the different results just **answer different questions**
    - The problem usually is, what is the question we should be asking ?  
→ Not always trivial to decide!
- We also as scientists have a **responsibility for the way we communicate our results**. Sloppy jargon, imprecise claims, probability-inversion statements are bad. **Who talks bad thinks bad !**

# What Is a Measurement ?

- When we (physicists) talk about the "measurement" of a physical quantity, **what do we actually mean ?**
  - I would say it is a procedure:
    - 1) use of a measuring device to extract observations (data) carrying information on the quantity
    - 2) analyse the data to extract the value of the quantity most consistent with the observations
    - 3) use some prescription to associate an uncertainty to the value found
- In Statistics, what one talks about is an "estimate" of the quantity, and the process involves two very distinct activities, called "point estimation" and "interval estimation", which roughly correspond to points 2) and 3) above
  - The two take different chapters in any Statistics book, for a good reason or two

# Point and Interval Estimation

- **Point estimation** can be awfully complicated, but it is almost always non-controversial
  - It works by defining an **estimator**, a function of the data which has good properties (no bias, small variance, consistency, efficiency...)
  - In making this choice, a careful **evaluation of what we know of the distribution of our data is CRUCIAL**
  - Two all-important estimators: the chisquare, the likelihood
  - But even more common and simple to remember as good examples of estimators are the *sample mean* and the *sample variance*
  - **PE is dealt with in J. Donini's lectures** – but I will introduce estimators and the MLE below.
- **Interval estimation** is more subtle – and it is what we really care about
  - provide the user with the range of values the parameter is likely to have
  - experimental design: minimize expected uncertainties on parameters of interest
  - BSM searches: "does it agree with the SM?" ← cannot answer with the estimate alone; the uncertainty without estimate is instead still useful !
- The core question we should always be asking ourselves is "**do my uncertainty bars cover at the stated confidence level ?**"



# A Parenthesis: Estimators

- Before we discuss interval estimation, coverage, and related topics, we need to introduce a few basic concepts we cannot do without
  - some of them are in J.D.'s lecture, but they might be covered only tomorrow – so this is my backup plan
  - If I repeat something it can only be beneficial
- The next few slides provide a few definitions we are going to use in the following:
  - expectation value, variance
  - estimators and some of their crucial properties
  - the MLE method
- We can skip whatever is trivial to you... But **stop me** if you need more explanation

# E[.]: the Mean

- The *probability density function* (pdf)  $f(\mathbf{x})$  of a random variable  $x$  is a normalized function which describes the probability to find  $x$  in a given range:

$$P(x, x+dx) = f(x)dx$$

- This is defined for continuous variables. For discrete ones, e.g.  $P(n|\mu) = e^{-\mu} \mu^n / n!$  ,  $P$  is a probability tout-court.

- The *expectation value* of the random variable  $x$  is then defined as

$$E[x] = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

← a function of the parameters of the model  $f$

- $E[x]$ , also called *population mean* , or simply **mean**, of  $x$ , thus depends on the distribution  $f(x)$ . Note that  $E[x]$  is not a function of  $x$ , but it is rather a fixed quantity dependent on the form of the PDF  $f(x)$ .
- The formulation of the expectation value is useful to define other properties of the PDF, as shown in the following.

# The Variance

- Of crucial importance to determine the property of a distribution is the “second central moment” of  $x$ ,

$$E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = V[x]$$

also called *variance*. The variance describes the "spread" of the PDF around its expectation value. It enjoys the property that

$$E[(x - E[x])^2] = E[x^2] - \mu^2,$$

as it is trivial to show.

- Also well-known is the *standard deviation*  $\sigma = \text{sqrt}(V[x])$ .

# Parameter Estimation: Definitions

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

$x$  is a random variable,  $\theta$  is a parameter. If you change  $\theta$ , you get a different PDF !

Suppose we have a **sample** of observed values:  $\vec{x} = (x_1, \dots, x_n)$

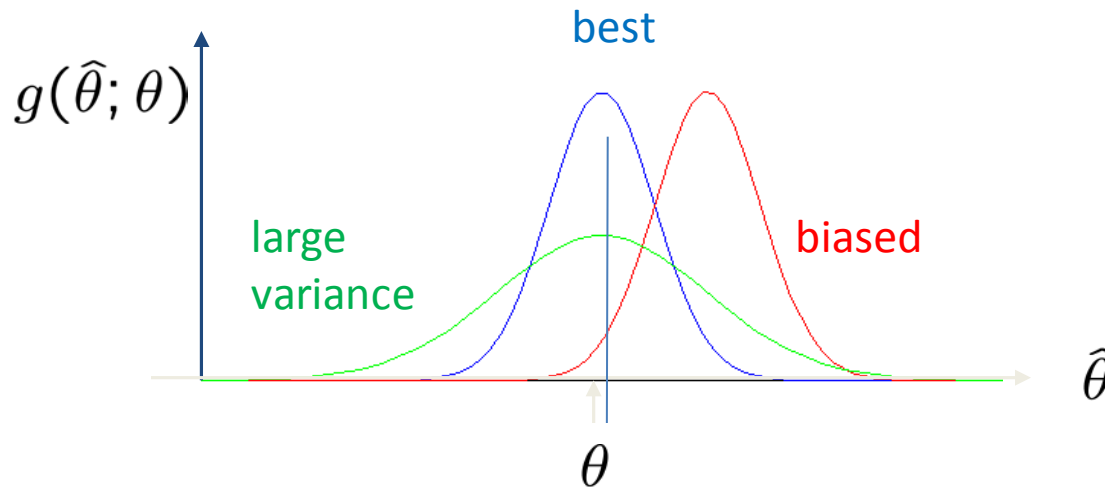
We often want to find some function of the data to **estimate** the parameter(s):

$$\hat{\theta}(\vec{x}) \quad \text{Note: the estimator gets written with a hat (or a *)}$$

Usually we say 'estimator' for the function of  $x_1, \dots, x_n$ ;  
'estimate' for the value of the estimator with a particular data set.

# Two Properties of Estimators

If we were to **repeat the entire measurement many times**, the estimates we get from each would follow a pdf:



We usually (not always!!!) want small (or zero) bias (systematic error):  $b = E[\hat{\theta}] - \theta$

this way, the average of repeated measurements should tend to the true value.

And we want a small variance (statistical error):  $V[\hat{\theta}]$  (will define better below)

**Note:** small bias & small variance are in general conflicting criteria. You know this from your experimental physics practice, but in Statistics this is a **rule**

# Estimators: a Few More Definitions

- Given a sample  $\{x_i\}$  of  $n$  observations of a random variable  $x$ , drawn from a pdf  $f(x)$ , one may construct a **statistic**: a function of  $\{x_i\}$  containing no unknown parameters. An **estimator** is a statistic used to estimate some property of a pdf. Using it on a set of data provides an **estimate** of the parameter.

- Estimators are **consistent** if they converge to the true value for large  $n$ .

- The expectation value of an estimator  $\hat{\theta}^*$  having a sampling distribution  $H(\hat{\theta}; \theta)$  is

$$E[\hat{\theta}(x)] = \int \hat{\theta} H(\hat{\theta}; \theta) d\theta$$

- Simple example of day-to-day estimators: the sample mean and the sample variance

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

Unbiased estimators of population mean and variance

- An estimator can be consistent even if biased: the average of an infinite replica of experiments with **finite  $n$**  will not in general converge to the true value, even if  $E[\hat{\theta}^*]$  will tend to  $\theta$  as  $n$  tends to infinity.

- Other properties of estimators (among which usually there are tradeoffs):

- **efficiency**: an efficient estimator (within some class) is the one with **minimum variance**
- **robustness**: the estimate is less dependent on the unknown true distribution  $f(x)$  for a more robust estimator (see example on OPERA later)

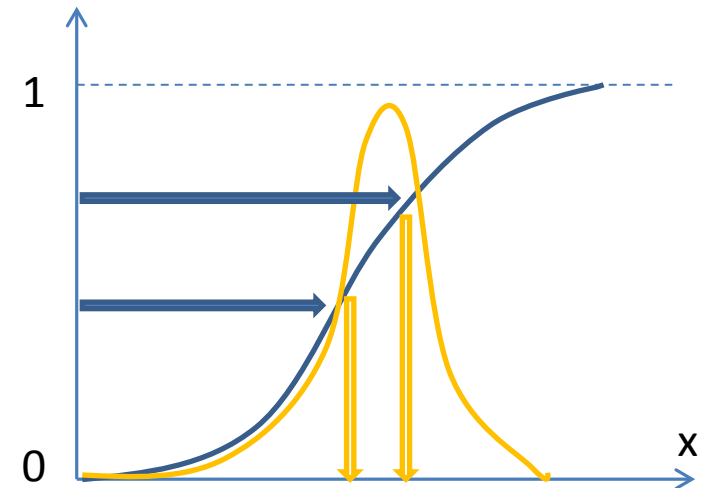
# A Final Digression: the (Toy) Monte Carlo Technique

- We often need to check the properties of our estimators in the specific conditions of our experiment – one example will be given later
  - For instance, we want to see if they are better than others, or if they depend on a tunable parameter we want to optimize it
- Often these details cannot be calculated algebraically, but we can use the Monte Carlo technique:
  - Simulate data with random generators
  - Repeat many times, each time extracting properties under study (optionally as a function of parameter to be optimized)
  - Study properties of estimators as  $f(\text{tunable parameters})$
- To generate pseudo-data one may rely on built-in functions in statistics packages (root, R, etc.)
  - We are spoiled by these built-in functions! We need to remember how to do the basic things by ourselves...
- One important part is to know how to generate data according to  $f(x)$  using a simple `rndm()` function. To do this one needs to find the cumulative  $F(x)$  and invert it. See next slide

How many of you know how to do that ?

# The General Idea

- You have a histogram, or a function,  $f(x)$ . You want to create pseudo-data that are distributed like it, to study other properties
- From that  $f(x)$  you can always derive the cumulative function  $F(x)$ :  
$$F(x) = \int_{-\infty}^x f(t)dt$$
- Then just throw a random number in  $[0,1]$
- Find the  $x$  where the cumulative function has that value
  - and you are done!





# E.g. How To Get Data Distributed as $f(x)=\exp(-x)$ ?

- First obtain  $F(x)$ , the cumulative function:
  - $F(x) = \int_0^x f(x')dx' = \int_0^x e^{-x} dx = 1 - e^{-x} = y$
- Next, invert it:
  - $x = -\log(1 - y)$
- Finally, account for the range of  $x$  you wish to generate, e.g.  $[0, x_{\max}]$ :
  - $x = -\log[1 - y(1 - e^{-x_{\max}})]$  (you multiply  $y$  by the integral of the pdf in the required range ( $<1$ ) to account for the restriction)
- Voila – if  $y$  is uniformly distributed in  $[0,1]$ ,  $x$  as computed above is distributed as  $f(x) = e^{-x}$  in  $[0, x_{\max}]$  !
  - Try it at home: derive recipe to get  $f(x)=x^2$

```
void Example_exp(int N=1000, int xmax=10.) {
```

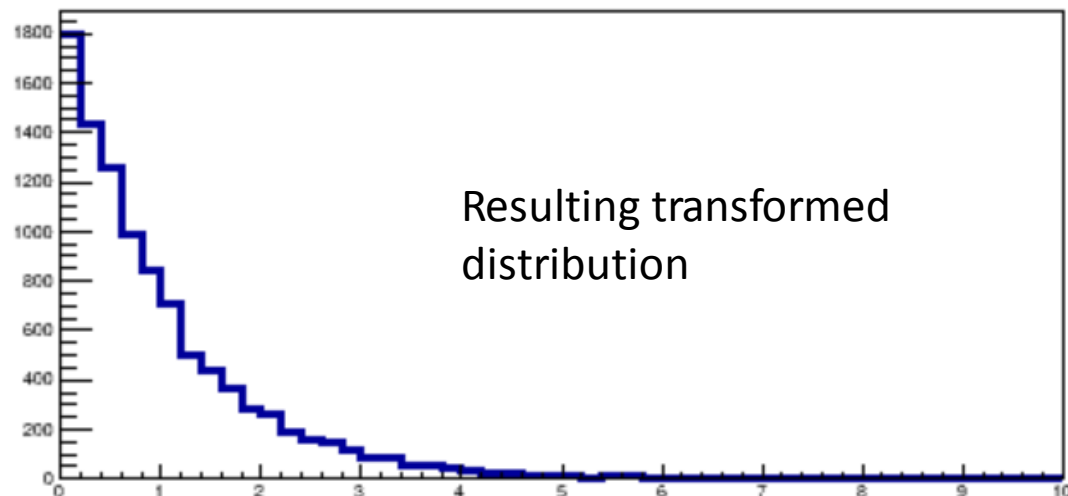
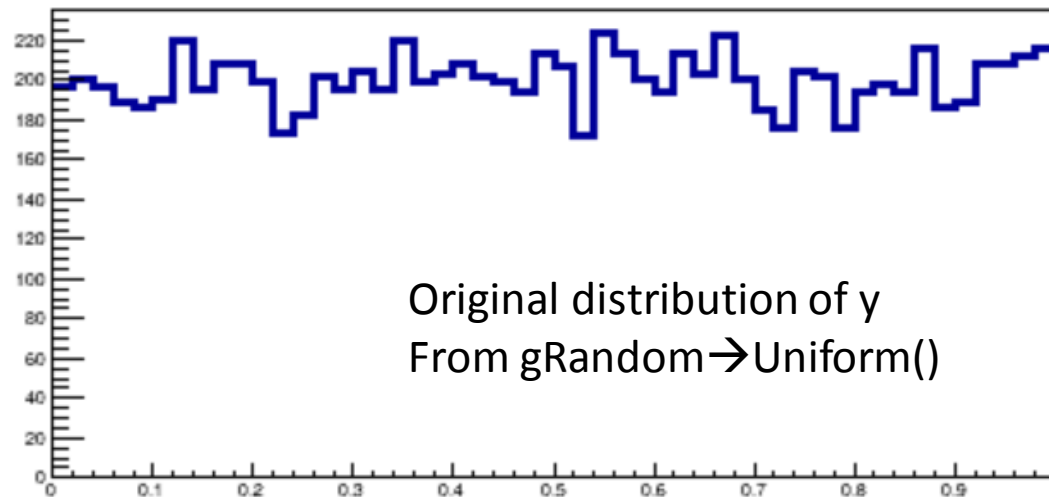
```
    // Change preset generator  
    // NB other versions of TRandom are flawed  
    // -----
```

```
    delete gRandom;  
    TRandom3 * myRNG = new TRandom3  
    gRandom = myRNG;
```

```
    int Nbins=50;  
    TH1D * Data0 = new TH1D ("Data0", "Data0", Nbins, 0, 10, 0);  
    TH1D * Data1 = new TH1D ("Data1", "Data1", Nbins, 0, 10, 0);
```

```
    for (int i=0; i<N; i++) {  
        double y = gRandom->Uniform(0,1);  
        Data0->Fill(y);  
        double x = -log(1-y*(1-exp(-1)));  
        Data1->Fill(x);  
    }
```

```
    TCanvas * C = new TCanvas ("C", "C", 1000, 1000);  
    C->Divide(1,2);  
    C->cd(1);  
    Data0->SetMinimum(0);  
    Data0->SetLineWidth(3);  
    Data0->Draw();  
    C->cd(2);  
    Data1->SetLineWidth(3);  
    Data1->Draw();
```



# The Method of Maximum Likelihood

- Take a pdf for a random variable  $x$ ,  $f(\mathbf{x}; \theta)$  which is analytically known, but for which the value of  $m$  parameters  $\theta$  is unknown. *The method of maximum likelihood allows us to estimate the parameters  $\theta$  if we have a set of data  $x_i$  distributed according to  $f$ .*
- The probability of our observed set  $\{x_i\}$  depends on the distribution of the pdf. Assuming that the measurements are independent, we have

$$p = \prod_{i=1}^n f(x_i; \theta) dx_i \quad \text{to find } x_i \text{ in } [x_i, x_i + dx_i]$$

- The likelihood function

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

is then a *function of the parameters  $\theta$*  only. It is written as the joint pdf of the  $x_i$ , but we *treat those as fixed*

- Using  $L(\theta)$  one can define “maximum likelihood estimators” for the parameters  $\theta$  as the *values which maximize the likelihood*, i.e. the solutions  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$  of the equation

$$\left( \frac{\partial L(\theta)}{\partial \theta_j} \right)_{\theta = \hat{\theta}} = 0 \quad \text{for } j = 1, \dots, m$$

Note: The ML requires **(and exploits!)** the *full knowledge* of the distributions

# Variance of the MLE

- In the simplest cases, i.e. when one has **unbiased estimates and Gaussian distributed data**, one can estimate the variance of the maximum likelihood estimate with the simple formula

$$\hat{\sigma}^2_{\theta=\theta_0} = \left( -\frac{\partial^2 \ln L}{\partial \theta^2} \right)_{\theta=\theta_0}^{-1}$$

This is also the default used by MIGRAD to return the uncertainty of a MLE from a fit.

However, note that this is **only a lower limit of the variance** in conditions when errors are not Gaussian and when the ML estimator is unbiased. A general formula called the **Rao-Cramer-Frechet inequality** gives this lower bound as

$$V[\hat{\theta}] \geq \left( 1 + \frac{\partial b}{\partial \theta} \right)^2 / E \left[ -\frac{\partial^2 \ln L}{\partial \theta^2} \right]$$

(b is the bias, E is the expectation value)

# Example: the Loaded Die

Imagine you want to test whether a die is loaded. Your **hypothesis** might be that the probabilities of the six occurrences are **not** equal, but rather that

$$\begin{aligned}P(1) &= 1/6 - t/2 \\P(2) &= P(3) = P(4) = P(5) = 1/6 - t/8 \\P(6) &= 1/6 + t\end{aligned}$$

Your data comes from **N=20 repeated throws** of the die, whereupon you get:

$$x_i = 1 : 3 \text{ trials}$$

$$x_i = 2..5 : 3 \text{ trials each}$$

$$x_i = 6 : 5 \text{ trials}$$

The likelihood is the product of probabilities, so to estimate t you write L as

$$-\log(L(t)) = -\sum_{i=1}^N \log(P(x_i, t)) = -3\log(1/6 - t/2) - 12\log(1/6 - t/8) - 5\log(1/6 + t)$$

Setting the derivative wrt t to zero of  $-\log L$  yields a quadratic equation:

$$360t^2 - 249t + 16 = 0$$

This has one solution in the allowed range for t, **[-1/6, 1/3]**:  $t=0.072$ . Its uncertainty can be obtained by the variance, computed as the inverse of the second derivative of the likelihood. This amounts to  $\pm 0.084$ . **The point estimate of the “load”, the MLE, is different from zero, but compatible with it. We conclude that the data cannot establish the presence of a bias.**

# Exercise With Root

Write a root macro that determines, using the likelihood of the previous slide, the value of the bias,  $t$ , and its uncertainty, given a random set of  $N$  (unbiased) die throws.

Directions:

- 1) Your macro will be called “Die.C” and it will have a function called “void Die(int N) {}”
- 2) Produce a set of  $N$  throws of the die by looping  $i=0\dots N-1$  and storing the result of  $(\text{int})(1+\text{gRandom}\rightarrow\text{Uniform}(0.,6.))$ ;
- 3) Call  $N_1$ =number of occurrence of 1;  $N_3$ =occurrences of 6;  $N_2$ =other results.
- 4) With paper and pencil, derive the coefficients of the quadratic equation in  $t$  for the likelihood maximum as a function of  $N_1, N_2, N_3$ .
- 5) Also derive the expression of  $-d^2\ln L/dt^2$  as a function of  $t$  and  $N_1, N_2, N_3$ .
- 6) Insert the obtained formulas in the code to compute  $t^*$  and its uncertainty  $\sigma(t^*)$ .
- 7) Print out the result of  $t$  in the allowed range  $[-1/6, 1/3]$  and its uncertainty. If there are two solutions in that interval, print the result away from the boundary.
- 8) How frequently do you get a result for  $t$  less than one standard deviation away from 0?

# Die.C

```
void Die(int N=100) {
    int res[100000];
    int n1=0, n2=0, n3=0;
    for (int i=0; i<N; i++) {
        res[i]=1+(int)gRandom->Uniform(0.,6.);
        if (res[i]==1) {
            n1++;
        } else if (res[i]<6) {
            n2++;
        } else {
            n3++;
        }
    }
    cout << endl << " Die throwing results:" << endl;
    cout << " n1 = " << n1;
    cout << " n2 = " << n2;
    cout << " n3 = " << n3 << endl << endl;
    // Quadratic equation for max of L: coefficients
    double a = 18*(n1+n2+n3);
    double b = -3*(7*n1+n2+10*n3);
    double c = -(4*n1+n2-8*n3);
    double rms, t1, t2;
    double discr=b*b-4*a*c;
    double tmin=-1./6., tmax=1./3., tstar=0;
    if (discr<0) {
        cout << " No solution for max likelihood" << endl;
    } else {
        t1 = (-b-sqrt(discr))/(2*a);
        t2 = (-b+sqrt(discr))/(2*a);
        if (t1>=tmin && t1<=tmax) {
            if (t2>=tmin && t2<=tmax) {
                // NNBB when n1=0 there is always one solution at t=0.33333
                if (t1-tmin>tmax-t2) {
                    cout << " Bias is estimated to be t = " << t1;
                    tstar=t1;
                }
            }
        }
    }
}
```



```

    } else {
        cout << " Bias is estimated to be t = " << t2;
        tstar=t2;
    }
} else {
    cout << " Bias is estimated to be t = " << t1;
    tstar=t1;
}
} else if (t2>=tmin && t2<=tmax) {
    cout << " Bias is estimated to be t = " << t2;
    tstar=t2;
}
// Determine error from inverse of second derivative of logL
double d2logl;
if (tstar>-1/6. && tstar<1/3.) {
    d2logl=9*n1/pow(1-3*tstar,2)+ 9*n2/pow(4-3*tstar,2)+ 36*n3/pow(1+6*tstar,2);
    rms = sqrt(1/d2logl);
}
cout << " +- " << rms << endl;
}
// Compute corresponding p-values
cout << endl;
cout << " This corresponds to the following probabilities:" << endl;
cout << " p(1) = " << 1./6.-tstar/2. << " +- " << rms/2. << endl;
cout << " p(x) = " << 1./6.+tstar/8. << " +- " << rms/8. << endl;
cout << " p(6) = " << 1./6.+tstar << " +- " << rms << endl << endl;
}

```



# Another Exercise: Solve With LS Method

- We just used the ML method to estimate the load on the die. But we could have also done it with the chisquared method

→ try it at home, we can look at the results tomorrow

Hints:

- write down the chisquare
- derive WRT the load  $t$
- set the derivative to zero, solve for  $t$
- find  $t_1, t_2$  such that  $\chi^2(t_1) = \chi^2(t_2) = \chi^2(t) + 1$

# Loaded Die: Least-Square Solution

- We just have to write a chisquare as a function of the data  $N_i=(3,3,3,3,3,5)$  and the load  $t$ :

$$\chi^2 = \sum_{i=1}^6 \frac{(N_i - e_i(t))^2}{\sigma_i^2}$$

where  $e_i(t)$  are the expected times that result "i" appears in 20 throws, i.e.  $e_i = 20 P(i)$  where, as before,

$$\begin{aligned} P(1) &= 1/6 - t/2 \\ P(2) &= P(3) = P(4) = P(5) = 1/6 - t/8 \\ P(6) &= 1/6 + t \end{aligned}$$

Note that we can use the information of  $N_2, N_3, N_4, N_5$  distributions if we wish – it just amounts to consider them as separate in the  $\chi^2$ .

Once we have the  $\chi^2(t)$ , we may compute its derivative w.r.t.  $t$ , and set it to zero, then solve for  $t \rightarrow$  this will yield our point estimate  $t^*$

The interval will be obtained by finding  $t_1, t_2$  such that

$$\chi^2(t_1) = \chi^2(t_2) = \chi^2(t^*) + 1$$

Results: ....

Comparing with the likelihood solution, we see that ... ?

Of the two ways to compute the chisquare the preferable one is ... ?

# Calculation

Inputs:  $N, n_1, n_x, n_6$  ( $x = \text{sum of } 2,3,4,5$ )

$$e_i(t) = N \cdot p(i,t) \rightarrow e_1(t) = N \cdot (1/6 - t/2); e_x(t) = 4 \cdot N \cdot (1/6 - t/8) = N \cdot (2/3 - t/2); e_6(t) = N \cdot (1/6 + t) \quad (\rightarrow e_{\text{tot}} = N)$$

$$S_1 = [n_1 - e_1(t)]^2 / n_1 = [n_1^2 - 2 \cdot n_1 \cdot N \cdot (1/6 - t/2) + N^2 \cdot (1/6 - t/2)^2] / n_1 =$$

$$n_1 - N/3 + N \cdot t + N^2 / (36 \cdot n_1) - N^2 \cdot t / (6 \cdot n_1) + N^2 \cdot t^2 / (4 \cdot n_1)$$

$$S_x = [n_x - e_x(t)]^2 / n_x = [n_x^2 - 2 \cdot n_x \cdot N \cdot (2/3 - t/2) + N^2 \cdot (2/3 - t/2)^2] / n_x =$$

$$n_x - 4 \cdot N/3 + N \cdot t + 4 \cdot N^2 / (9 \cdot n_x) - 2 \cdot N^2 \cdot t / (3 \cdot n_x) + N^2 \cdot t^2 / (4 \cdot n_x)$$

$$S_6 = [n_6 - e_6(t)]^2 / n_6 = [n_6^2 - 2 \cdot n_6 \cdot N \cdot (1/6 + t) + N^2 \cdot (1/6 + t)^2] / n_6 =$$

$$n_6 - N/3 - 2 \cdot N \cdot t + N^2 / (36 \cdot n_6) + N^2 \cdot t / (3 \cdot n_6) + N^2 \cdot t^2 / n_6$$

$$dS_1/dt = N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1)$$

$$dS_x/dt = N - 2 \cdot N^2 / (3 \cdot n_x) + N^2 \cdot t / (2 \cdot n_x)$$

$$dS_6/dt = -2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6$$

$$dS_1/dt + dS_x/dt + dS_6/dt = 0 \rightarrow$$

$$N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1) + N - 2 \cdot N^2 / (3 \cdot n_x) + N^2 \cdot t / (2 \cdot n_x) - 2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6 = 0$$

$$t \cdot N^2 \cdot [1 / (2 \cdot n_1) + 1 / (2 \cdot n_x) + 2 / n_6] - [N^2 \cdot (1 / (6 \cdot n_1) + 2 / (3 \cdot n_x) - 1 / (3 \cdot n_6))] = 0$$

$$t = [1 / (6 \cdot n_1) + 2 / (3 \cdot n_x) - 1 / (3 \cdot n_6)] / [1 / (2 \cdot n_1) + 1 / (2 \cdot n_x) + 2 / n_6] =$$

$$= (n_x \cdot n_6 + 4 \cdot n_1 \cdot n_6 - 2 \cdot n_1 \cdot n_x) / (6 \cdot n_1 \cdot n_x \cdot n_6) / (3 \cdot n_x \cdot n_6 + 3 \cdot n_1 \cdot n_6 + 12 \cdot n_1 \cdot n_x) / (6 \cdot n_1 \cdot n_x \cdot n_6) =$$

$$= (n_x \cdot n_6 + 4 \cdot n_1 \cdot n_6 - 2 \cdot n_1 \cdot n_x) / (3 \cdot n_x \cdot n_6 + 3 \cdot n_1 \cdot n_6 + 12 \cdot n_1 \cdot n_x)$$

# Calculation, using all results (2,3,4,5)

Inputs:  $N, n_1, n_x, n_6$  ( $x=2,3,4,5$ )

$$e_i(t) = N \cdot p(i,t) \rightarrow e_1(t) = N \cdot (1/6 - t/2); e_x(t) = N \cdot (1/6 - t/8); e_6(t) = N \cdot (1/6 + t) \quad (\rightarrow e_{\text{tot}} = N)$$

$$S_1 = [n_1 - e_1(t)]^2 / n_1 = [n_1^2 - 2 \cdot n_1 \cdot N \cdot (1/6 - t/2) + N^2 \cdot (1/6 - t/2)^2] / n_1 =$$

$$n_1 - N/3 + N \cdot t + N^2 / (36 \cdot n_1) - N^2 \cdot t / (6 \cdot n_1) + N^2 \cdot t^2 / (4 \cdot n_1)$$

$$S_x = [n_x - e_x(t)]^2 / n_x = [n_x^2 - 2 \cdot n_x \cdot N \cdot (1/6 - t/8) + N^2 \cdot (1/6 - t/8)^2] / n_x =$$

$$n_x - N/3 + N \cdot t/4 + N^2 / (36 \cdot n_x) - N^2 \cdot t / (24 \cdot n_x) + N^2 \cdot t^2 / (64 \cdot n_x)$$

$$S_6 = [n_6 - e_6(t)]^2 / n_6 = [n_6^2 - 2 \cdot n_6 \cdot N \cdot (1/6 + t) + N^2 \cdot (1/6 + t)^2] / n_6 =$$

$$n_6 - N/3 - 2 \cdot N \cdot t + N^2 / (36 \cdot n_6) + N^2 \cdot t / (3 \cdot n_6) + N^2 \cdot t^2 / n_6$$

$$dS_1/dt = N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1)$$

$$dS_x/dt = N/4 - N^2 / (24 \cdot n_x) + N^2 \cdot t / (32 \cdot n_x)$$

$$dS_6/dt = -2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6$$

$$dS_1/dt + dS_2/dt + dS_3/dt + dS_4/dt + dS_5/dt + dS_6/dt = 0 \rightarrow$$

$$N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1) + N/4 - N^2 / (24 \cdot n_2) + N^2 \cdot t / (32 \cdot n_2) + N/4 - N^2 / (24 \cdot n_3) + N^2 \cdot t / (32 \cdot n_3) + N/4 - N^2 / (24 \cdot n_4) + N^2 \cdot t / (32 \cdot n_4) + N/4 - N^2 / (24 \cdot n_5) + N^2 \cdot t / (32 \cdot n_5) - 2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6 = 0$$

$$t \cdot N^2 \cdot [1 / (2 \cdot n_1) + 1 / (32 \cdot n_2) + 1 / (32 \cdot n_3) + 1 / (32 \cdot n_4) + 1 / (32 \cdot n_5) + 2 / n_6] - [N^2 \cdot (1 / (6 \cdot n_1) + 1 / (24 \cdot n_2) + 1 / (24 \cdot n_3) + 1 / (24 \cdot n_4) + 1 / (24 \cdot n_5) - 1 / (3 \cdot n_6))] = 0$$

$$t = [1 / (6 \cdot n_1) + 1 / (24 \cdot n_2) + 1 / (24 \cdot n_3) + 1 / (24 \cdot n_4) + 1 / (24 \cdot n_5) - 1 / (3 \cdot n_6)] / [1 / (2 \cdot n_1) + 1 / (32 \cdot n_2) + 1 / (32 \cdot n_3) + 1 / (32 \cdot n_4) + 1 / (32 \cdot n_5) + 2 / n_6] =$$

$$= 4/3 \cdot [4/n_1 + 1/n_2 + 1/n_3 + 1/n_4 + 1/n_5 - 8/n_6] / [16/n_1 + 1/n_2 + 1/n_3 + 1/n_4 + 1/n_5 + 64/n_6]$$

Intermezzo: Area Preservation

*or*

Two Chisquared and a Likelihood

# Know the Properties of Thy Estimators

- Issues (and errors hard to trace) may arise in the simplest of calculations, if you do not know the properties of the tools you are working with.
- Take the simple problem of combining three measurements of the *same quantity*. Make these be counting rates, i.e. with Poisson uncertainties:

- $A_1 = 100$
- $A_2 = 90$
- $A_3 = 110$



If they aren't,  
don't combine!

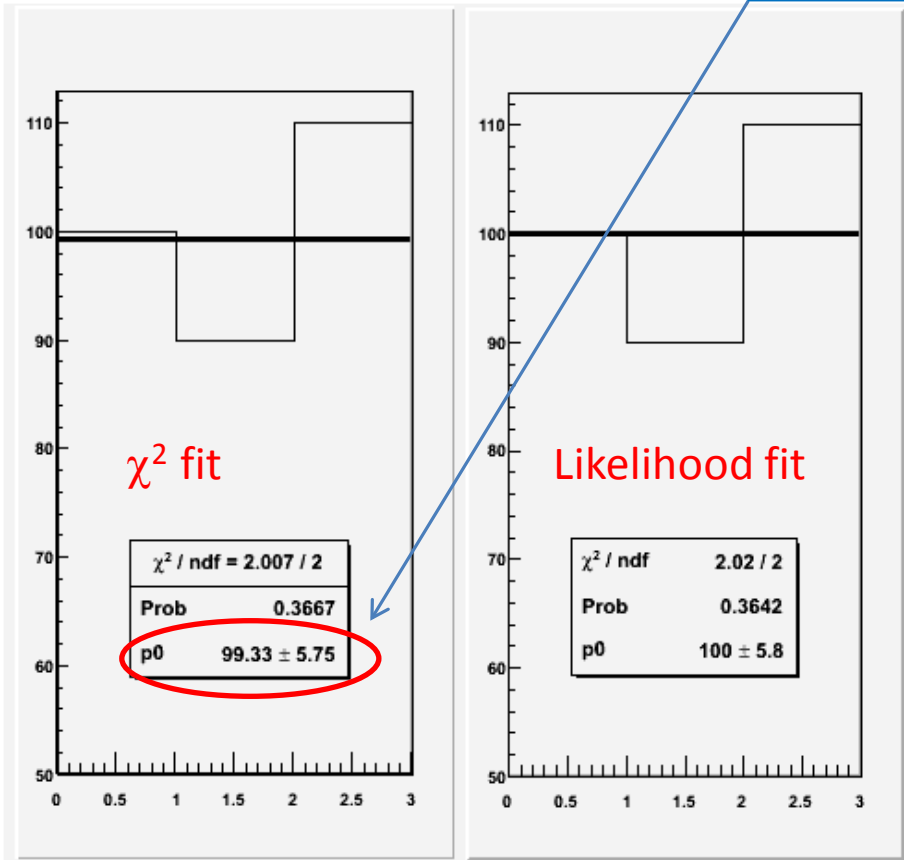
These measurements are **fully compatible with each other**, given that the estimates of their uncertainties are  $\sqrt{A_i} = \{10, 9.5, 10.5\}$  respectively. We may thus proceed to **average** them, obtaining  **$\langle A \rangle = 100.0 \pm 5.77$**

Now imagine, for the sake of argument, that we were on a lazy mood, and rather than do the math we **used a  $\chi^2$  fit** to evaluate  $\langle A \rangle$ .

Surely we would find the same answer as the simple average of the three numbers, right?

... Wrong!

the  $\chi^2$  fit does not “preserve the area” of the fitted histogram



## WTF is going on ??

Let us dig a little bit into this matter. This requires us to **study the detailed definition** of the test statistics we employ in our fits.

In general, a  $\chi^2$  statistic results from a **weighted sum of squares**; the *weights should be the inverse variances of the true values*.

Unfortunately, we do not know the latter!

# Two Chisquareds and a Likelihood

- The “standard” definition is called “Pearson’s  $\chi^2$ ”, which for Poisson data we write as

$$\chi_P^2 = \sum_{i=1}^k \frac{(N_i - n)^2}{n} \quad (\text{here } \mathbf{n} \text{ is the best fit value, } \mathbf{N}_i \text{ are the measurements})$$

- The other (AKA “modified”  $\chi^2$ ) is called “Neyman’s  $\chi^2$ ”:

$$\chi_N^2 = \sum_{i=1}^k \frac{(N_i - n)^2}{N_i}$$

- While  $\chi_P^2$  uses the best-fit variances at the denominator,  $\chi_N^2$  uses the individual **estimated variances**. Although both of these least-square estimators have asymptotically a  $\chi^2$  distribution, and display **optimal properties**, they use **approximated weights**.

The result is a pathology: neither definition preserves the area in a fit!

$\chi_P^2$  **overestimates the area**,  $\chi_N^2$  **underestimates it**. In other words, neither works to make a simple weighted average !

- The maximization of the Poisson maximum likelihood, 
$$L_P = \prod_{i=1}^k \frac{n^{N_i} e^{-n}}{N_i!}$$

instead preserves the area, and **obtains exactly the result of the simple average**.

[Proofs in the next slides.](#)



# Proofs – 1: Pearson's $\chi^2$

- Let us compute  $n$  from the minimum of  $\chi^2_P$ :

$$\chi_P^2 = \sum_{i=1}^k \frac{(N_i - n)^2}{n} \quad \text{note: a variable weight!}$$

$$0 = \frac{\partial \chi_P^2}{\partial n} = \sum_{i=1}^k \frac{2n(n - N_i) - (N_i - n)^2}{n^2}$$

$$0 = \sum_{i=1}^k (n^2 - N_i^2) = kn^2 - \sum_{i=1}^k N_i^2$$

$$\Rightarrow n = \sqrt{\frac{\sum_{i=1}^k N_i^2}{k}}$$

$n$  is found to be the *square root of the average of squares*, and is thus by force an **overestimate of the area!**

# 2 – Neyman's $\chi^2$

- If we minimize  $\chi^2_N$ ,

$$\chi^2_N = \sum_{i=1}^k \frac{(N_i - n)^2}{N_i} \leftarrow \text{again a variable weight}$$

we have:

$$0 = \frac{\partial \chi^2_N}{\partial n} = \sum_{i=1}^k \frac{2(N_i - n)}{N_i}$$

Just developing  
the fraction leads to

$$0 = \sum_{i=1}^k \left[ (N_i - n) \prod_{j=1, j \neq i}^k N_j \right] = \sum_{i=1}^k \left[ \prod_{j=1}^k N_j - n \prod_{j=1, j \neq i}^k N_j \right]$$

which implies that

$$\sum_{i=1}^k \prod_{j=1}^k N_j = n \sum_{i=1}^k \prod_{j=1, j \neq i}^k N_j$$

from which we finally get

$$\frac{1}{n} = \frac{\sum_{i=1}^k \prod_{j=1, j \neq i}^k N_j}{\sum_{i=1}^k \prod_{j=1}^k N_j} = \frac{1}{k} \sum_{i=1}^k \frac{1}{N_i}$$

the minimum is found for  $\mathbf{n}$  equal to the harmonic mean of the inputs – which is an **underestimate of the arithmetic mean!**

# 3 – The Poisson Likelihood $L_P$

- We minimize  $L_P$  by first taking its logarithm, and find:

$$L_P = \prod_{i=1}^k \frac{n^{N_i} e^{-n}}{N_i!}$$

$$\ln(L_P) = \sum_{i=1}^k (-n + N_i \ln n - \ln N_i!)$$

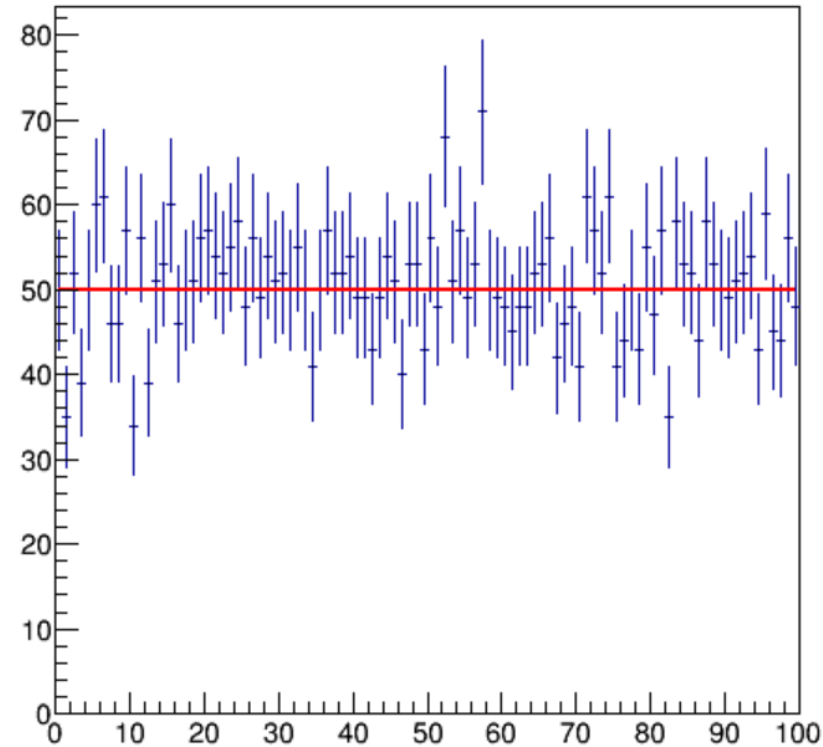
$$0 = \frac{\partial \ln(L_P)}{\partial n} = \sum_{i=1}^k \left( -1 + \frac{N_i}{n} \right) = -k + \frac{1}{n} \sum_{i=1}^k N_i$$

$$\Rightarrow n = \frac{\sum_{i=1}^k N_i}{k}$$

As predicted, the result for **n** is the arithmetic mean. Likelihood fitting preserves the area!

# Putting it together

- To check the behavior of the three fitting methods (remember: we are just considering them as ways to determine a **weighted average** here), we study a histogram with 100 bins
- Each bin is filled with  $N$  sampled from a  $\text{Poisson}(N|\mu)$
- We then fit the histogram to a constant by minimizing  $\chi^2_P$ ,  $\chi^2_N$ ,  $-2\ln(L_P)$  in turn
- We repeat many times, getting the average result for each fitting method
- We can then also study the ratio between the average result and the true  $\mu$  as a function of  $\mu$

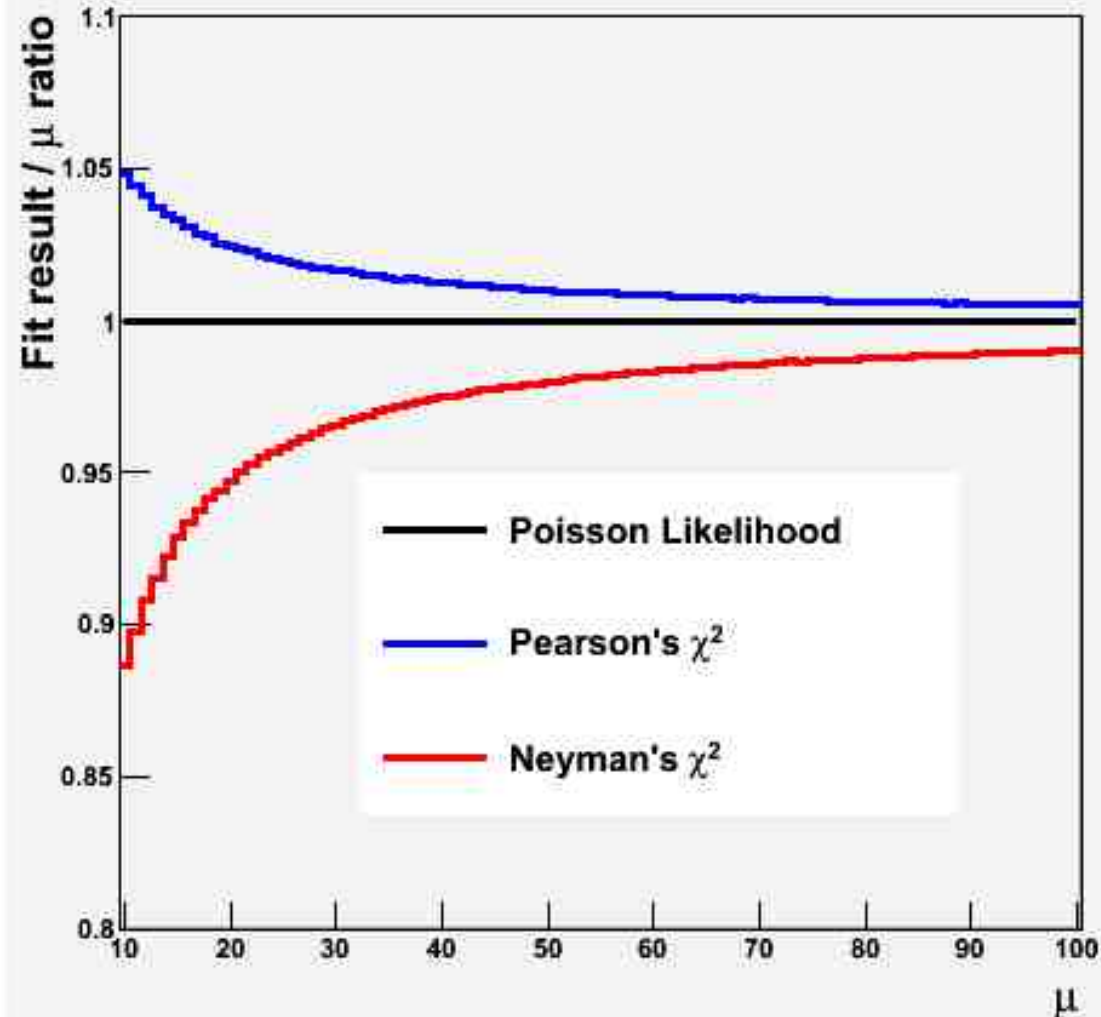


By the way, it's four lines of code:

```
TH1D * A = new TH1D("A", "", 100, 0., 100.);  
For (int i=1; i<101; i++) {  
A->SetBinContent(i,gRandom->Poisson(50.));}  
A->Fit("pol0"); // for Neyman's chi2
```

# Comparison vs $\mu$

Fit results with different  $\chi^2$



- One observes that **the convergence is slowest for Neyman's  $\chi^2$** , but the bias is significant also for  $\chi^2_p$ 
  - This result depends only marginally on the number of bins
- Keep that in mind when you fit a histogram!
- Standard ROOT fitting uses  $V=N_i \rightarrow$  Neyman's definition!

# Discussion

- What we are doing when we fit a constant through a set of  $k$  bin contents is to extract the common, unknown, true value  $\mu$  from which the entries were generated, by combining the  $k$  measurements

We have  $k$  Poisson measurement of this true value. Each equivalent measurement should have the same weight in the combination, because each is drawn from a Poisson of mean  $\mu$ , whose true variance is  $\mu$ .

But having no way to start with, we must use *estimates* of the variance as a (inverse) weight. So the  $\chi^2_N$  gives the different observations different weights  $1/N_i$ . Since negative fluctuations ( $N_i < \mu$ ) have larger weights, the result is downward biased!

What  $\chi^2_P$  does is different: it uses a common weight for all measurements, the fit result for the average,  $\mu^*$ . Since we minimize  $\chi^2_P$  to find  $\mu^*$ , larger denominators get preferred  $\rightarrow$  positive bias:  $\mu^* > \mu$ !

All methods have optimal asymptotic properties: consistency, minimum variance. However, **one seldom is in that regime**.  $\chi^2_P$  and  $\chi^2_N$  have problems when  $N_i$  is small. These drawbacks are solved by grouping bins, at the expense of *loss of information*.

$L_P$  does not have the approximations of the two sums of squares, and it has in general better properties. Cases when the use of a LL yields problems are rare. **Whenever possible, use a Likelihood!**

# Interval Estimation

# Confidence Level

- In classical statistics, the confidence level (CL) is a reference value chosen by the user
  - Most typical: CL=0.683 ("1-sigma")
  - Also quite used: CL=0.90, CL=0.95, CL=0.99
- The CL is used to define the **level of confidence** one wishes to have on the possible values of a quantity under study
  - One can alternatively set the **type-I error rate  $\alpha$** , as  
CL =  $1-\alpha$ .

What does one do with the CL? **One seeks to derive intervals** (uncertainty bars, or upper or lower limits) **that on average** (in a frequentist sense) **have the property of including the unknown, but fixed, true value of the quantity with a rate not smaller than CL.**

The notion of a CL stems from reasoning on the probability of getting data of some kind, under some hypothesis. To understand it, we need to discuss the **Neyman construction**.



# The Simplest Confidence Interval: the Standard Deviation

- The **standard deviation** is used in most simple applications as a *measure of the uncertainty of a point estimate*

– Sample standard deviation:

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

- For example: N **i.i.d.** observations  $\{x_i\}$  of random variable  $x$  with hypothesized pdf  $f(x;\theta)$ , with  **$\theta$  unknown**.  $X=\{x_i\}$  allows to compute the value that a suitable **estimator**  $\theta^*(\cdot)$  takes on  $X$ ,  **$\theta^*(X)$**
- Using an analytic method, or the RCF bound, or MC sampling techniques, one may usually cook up an estimate the standard deviation of  $\theta^*$ ,  **$\sigma_{\theta^*}^*$**
- The value  **$\theta^* \pm \sigma_{\theta^*}^*$**  is then reported. **What does this mean ?**
  - **Have a crack at it ! Spell out what it means to report that.**

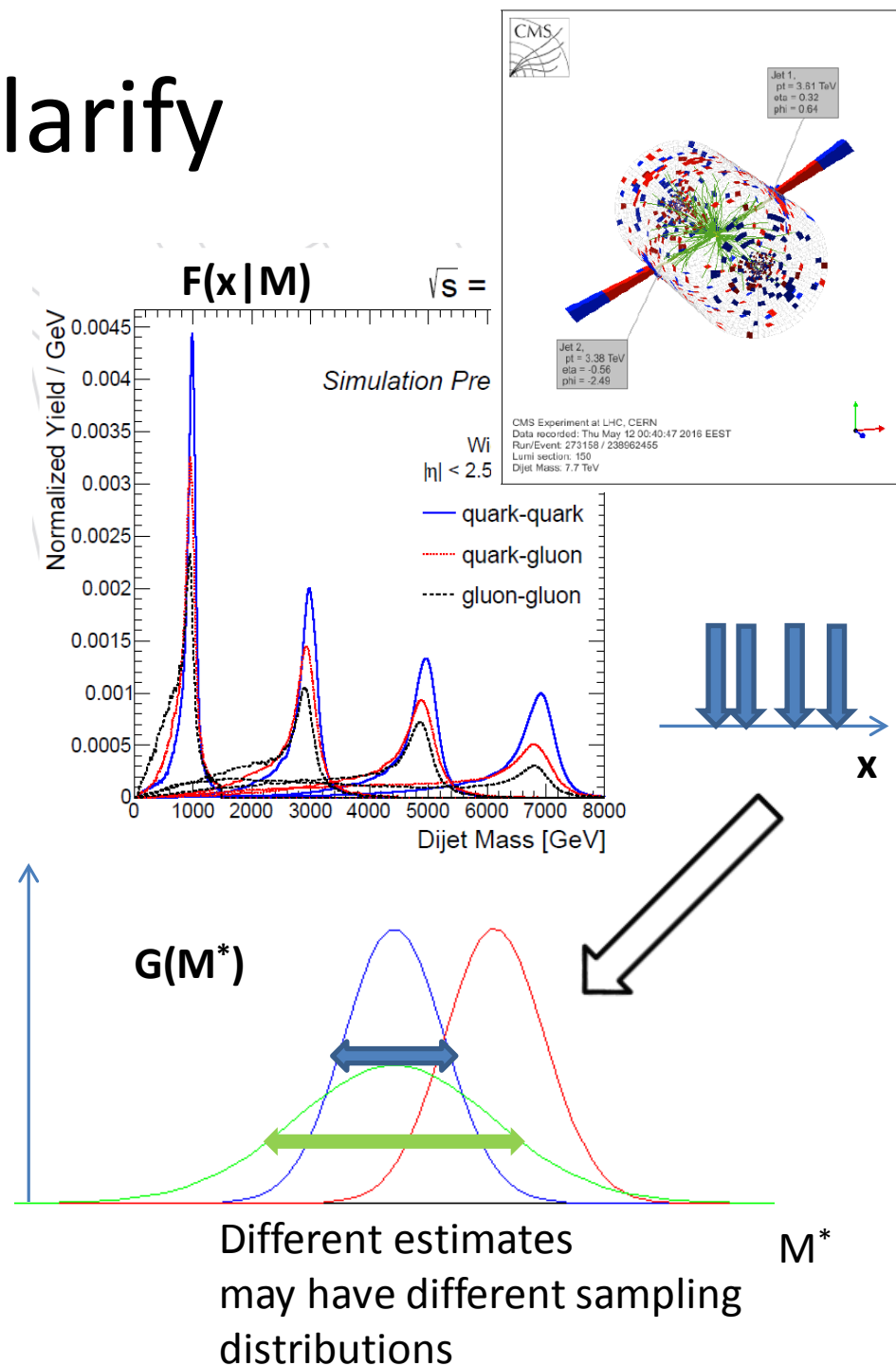
# $\theta^* \pm \sigma_{\theta^*}$ is Reported. What Does This Mean ?

- It means that in repeated estimates based on the same number  $N$  of observations of  $x$ ,  $\theta^*$  would distribute according to some pdf  $G(\theta^*)$  centered around a true value  $\theta$  with a true standard deviation  $\sigma_{\theta^*}$ , **respectively estimated** by  $\theta^*$  and  $\sigma_{\theta^*}$
- *In the large sample limit  $G()$  is a (multi-dimensional) Gaussian function*
- In most interesting cases for physics  $G()$  is not Gaussian, the large sample limit does not hold, 1-sigma intervals do not cover 68.3% of the time the true parameter, and we have better be careful in constructing intervals.
- But **we need to have a hunch of the pdf  $f(x;\theta)$**  to start with! (Or maybe not: when we can't, we assume it is itself Gaussian, and use the chisquare method.)

# One Example, to Clarify

- A strongly produced resonance of unknown mass in LHC data would result in events with two energetic jets. Let us assume we have a significant signal in our data,  $x$
- The PDF  $f(x;M)$  depends on  $M$ ; we may derive an estimate  $M^* \pm \sigma_{M^*}$  using  $x$  and some estimator – the easiest one being the sample mean
- Our interval estimation procedure returns intervals that hopefully fulfil the requirements on the confidence level chosen
- Still, there is **no guarantee** that the true value  $M$  is within the quoted interval around  $M^*$  !
- Yet in a frequentist sense our interval **covers it** 68.3% of the time
- The fact is that intervals constructed in a less than rigorous manner often FAIL to fulfil that requirement

**The question is how to construct confidence intervals that "work" in general**



# Neyman's Confidence Interval Recipe

Note: the recipe is designed to cover correctly. Thus, **one could not, on average, win money** by betting that the result of a measurement does not contain the true value, by using payoff odds corresponding to the stated type-I error rate (eg. 5%  $\rightarrow$  20:1)

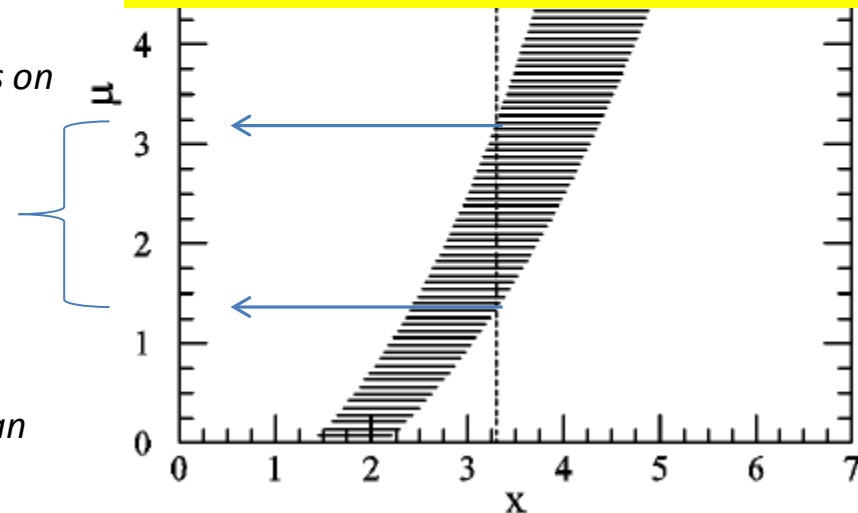
- Specify a model which provides the probability density function of a particular observable  $x$  being found, for each value of the unknown parameter of interest:  $p(x|\mu)$
- Also choose a Type-I error rate  $\alpha$  (e.g. 31.7%, or 5%), or the corresponding CL
- For each  $\mu$ , draw a horizontal acceptance interval  $[x_1, x_2]$  such that

$$p(x \in [x_1, x_2] | \mu) = 1 - \alpha.$$

**There are infinitely many ways of doing this:** it all depends on what you want from your data

- for upper limits, integrate the pdf from  $x$  to infinity
- for lower limits do the opposite
- might want to choose **central** intervals
- or **shortest** intervals?

- In general: **an ordering principle** is needed to well-define.
  - Upon performing an experiment, you measure  $x=x^*$ . You can then draw a vertical line through it.
- $\rightarrow$  The vertical **confidence interval**  $[\mu_1, \mu_2]$  (with Confidence Level C.L. =  $1 - \alpha$ ) is the union of all values of  $\mu$  for which the corresponding acceptance interval is intercepted by the vertical line.



# Important Notions on C. I.'s

**What is a vector ?** A vector is an element of a vector space (a set with certain properties).

Similarly, *a **confidence interval** is defined to be “an element of a confidence set”, the latter being a set of intervals defined to have the property of frequentist coverage under sampling!*

Let the unknown true value of  $\mu$  be  $\mu_t$ . In repeated experiments, the confidence intervals will have different endpoints  $[\mu_1, \mu_2]$ , depending on the random variable  $x$ .

*A fraction C.L. =  $1 - \alpha$  of intervals obtained by Neyman's construction will contain (“cover”) the fixed but unknown  $\mu_t$ :  $P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$ .*

It is important thus to realize two facts:

- 1) **the random variables in this equation are  $\mu_1$  and  $\mu_2$ , and not  $\mu_t$**
  - 2) **Coverage is a property of the set, not of an individual interval !** For a Frequentist, the interval either covers or does not cover the true value, regardless of  $\alpha$ .
- Classic **FALSE statement** you should avoid making:

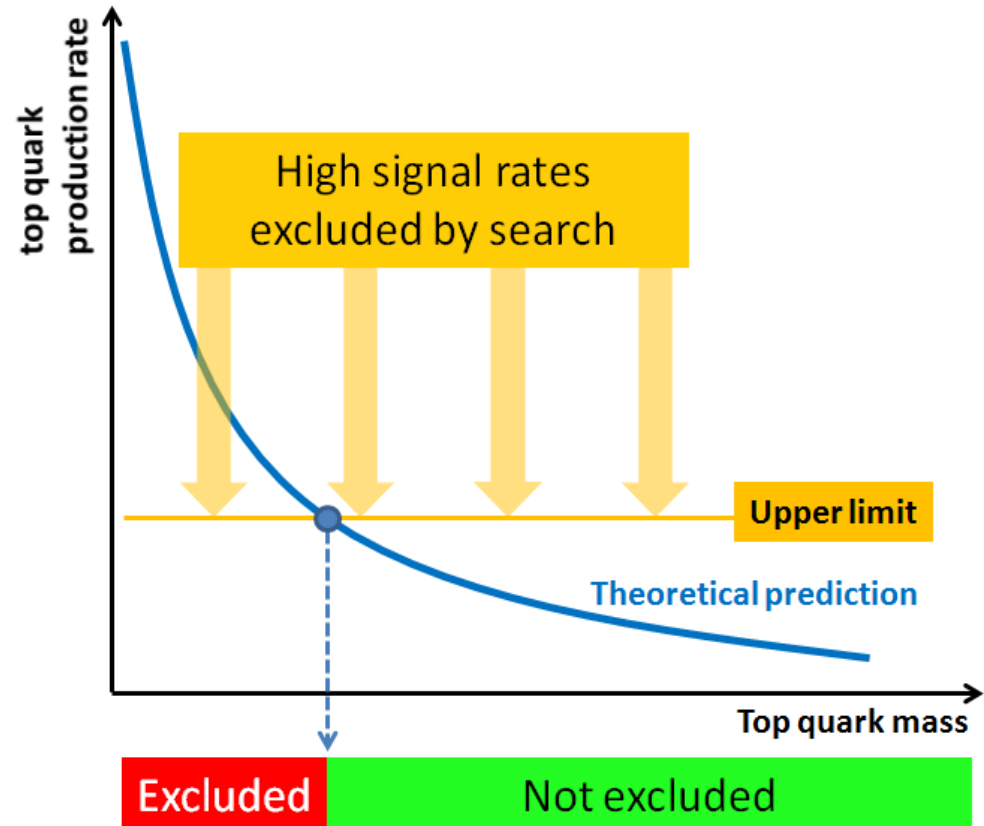
**“The probability that the true value is within  $\mu_1$  and  $\mu_2$  is 68%” !**

*The confidence interval instead does consist of those values of  $\mu$  for which the observed  $x$  is among the most probable (in sense specified by ordering principle)*

Also note: **“repeated sampling” does not require one to perform the same experiment all of the times** for the confidence interval to have the stated properties. Can even be different experiments and conditions! A big issue is what is the **relevant space** of experiments to consider.

# Upper Limits: How We Use Them

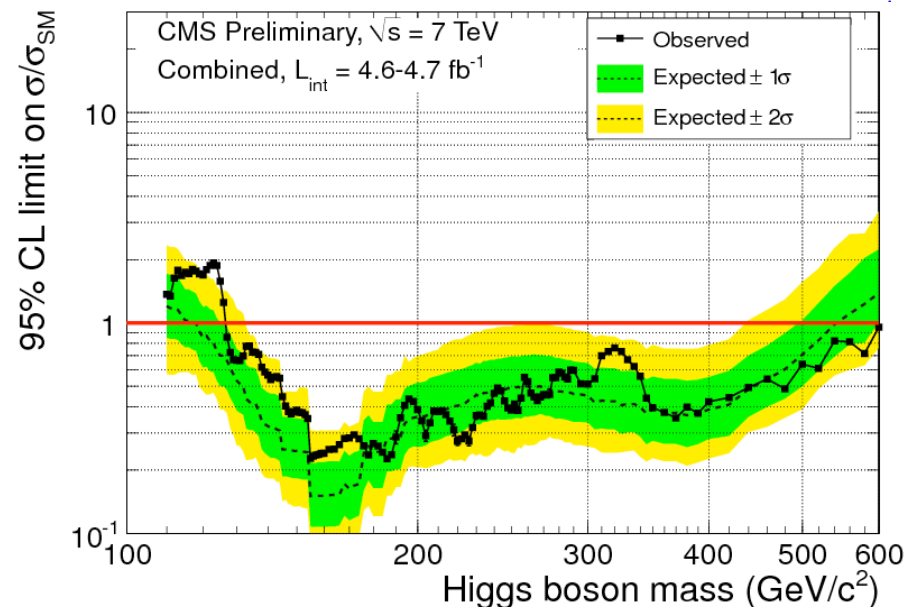
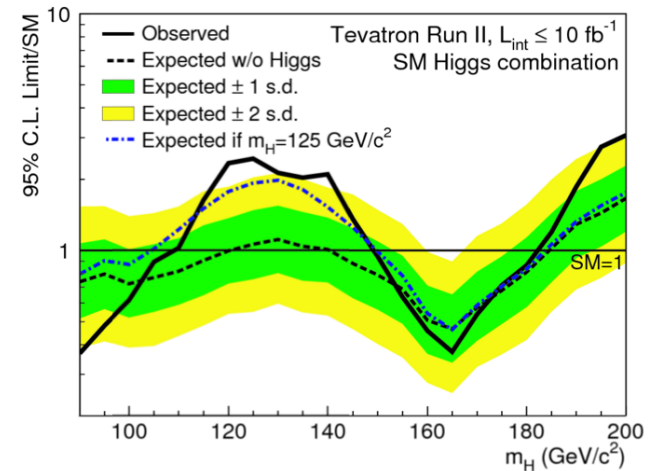
- If we do not see a signal we can exclude the new physics model  
→ (simple hypothesis test)
- More often we have a unknown parameter, and we exclude ranges of its value
  - Typically this is the mass of the particle
- We can e.g. derive **lower limits on the particle mass** from **upper limits** on the signal strength, by comparing those to a **theoretical model**



Luckily, the lower mass limit is useful information, worth a publication !

# The Problem Is Relevant in Fundamental Physics and Astrophysics !

- To give you the flavour of the relevance of the problem of setting correct upper limits, suffices to tell the story of the Higgs search
- For a long time (late 1990s) all we could say was where the particle could \*not\* be
- The competition (also for funding) centred on that information rather than the observation of the particle
- At the end of the seminar I will discuss the details of the method used.



# Coverage, or the Lack Thereof

Take a typical HEP graph: event counts in a mass histogram, with  $\sqrt{N}$  bars

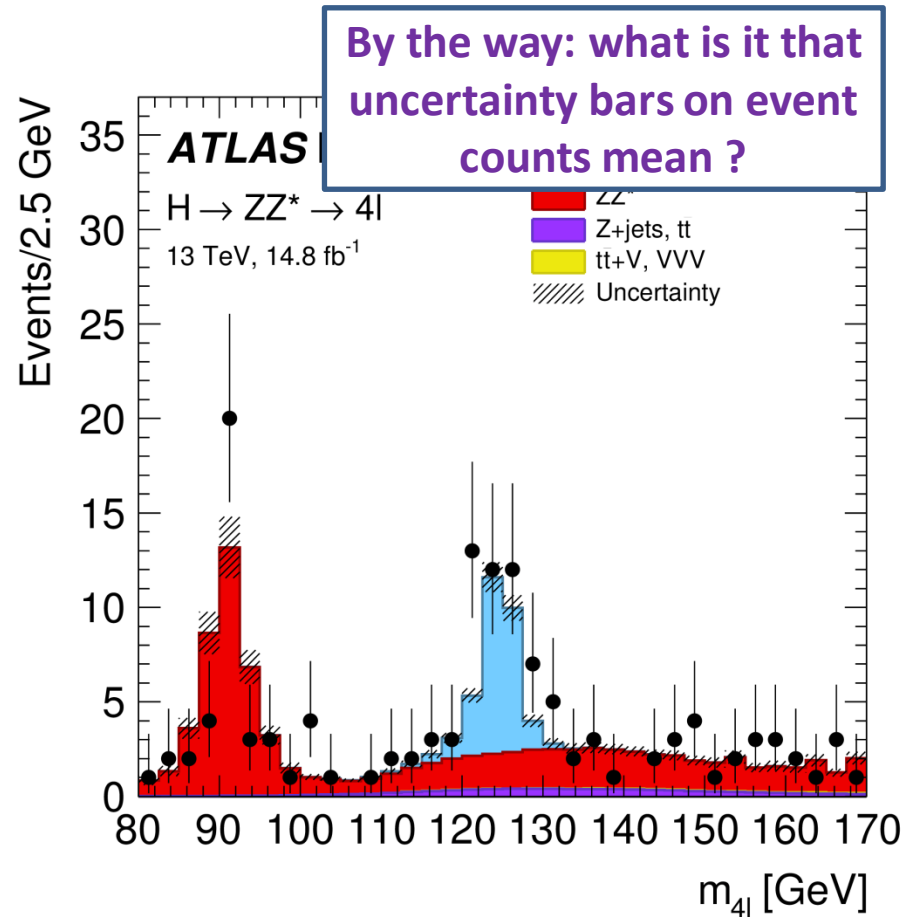
[Remember: for a Poisson,  $\mu = \sigma^2$ ]

What are those uncertainty bars supposed to mean? They report central intervals and nothing is said, so these should "cover" at 68.3%. Do they ?

Alas, usually **they don't**, as **the Gaussian approximation for the Poisson distribution breaks down for small N**

Suppose John claims  $x$  is in  $[a, b]$  with 68.3% confidence, but in fact the CL of the procedure is only 50%.

→ John is a liar ! He gave a misrepresentation of the information content of the measurement !



Of course, a solution exists: it was obtained in the fifites by Garwood, who used **Neyman's construction** for the Poisson distribution

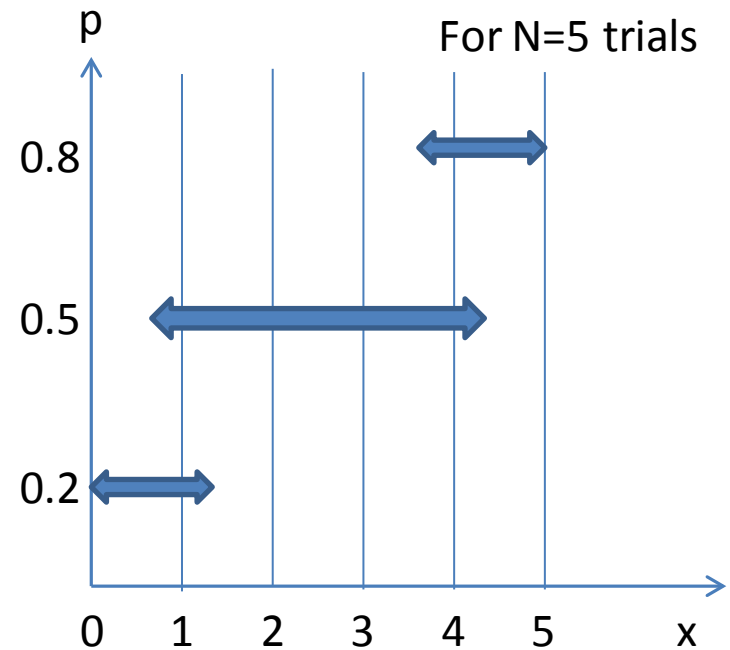


# On Undercoverage

- It is **BAD**. A frequentist shouldn't allow it.
- **E.g: if you state a limit or an interval at 95% CL and it turns out that, for the true value  $\mu$ , the coverage is actually 85%, you have significantly underestimated the uncertainty bars of your measurement – and your type-I error rate is 3-fold larger !!!**
- Undercoverage results from approximate expressions for the variance, or from other specific aspects of the problem
  - See example of likelihood of loaded die later
- Undercoverage can also results from **apparently innocuous procedures** in the derivation of our results, like
  - deciding whether to quote a limit or a confidence interval *a posteriori*
  - modifying details of analysis “because something does not look right” in your background estimate
  - Not publishing results that are controversial !

# Overcoverage

- *Coverage* is usually guaranteed by the frequentist Neyman construction. But this includes overcoverage.
- **Overcoverage:** sometimes the pdf  $p(x|\theta)$  is discrete  $\rightarrow$  it may not be possible to find exact boundary values  $x_1, x_2$  for each  $\theta$ ; one thus errs conservatively by including  $x$  values (according to one's ordering rule) until  $\sum_i p(x_i|\theta) > 1-\alpha$   
 $\rightarrow \theta_1$  and  $\theta_2$  will **overcover**



Let's make an example with the Binomial

$$F(N;r,p) = N! p^r(1-p)^{N-r}/[r!(N-r)!]$$

For  $N=5, p=0.5$ :

$$F(5;0,0.5)=0.5^5=0.031$$

$$F(5;1,0.5)=5*0.5^5=0.156$$

$$F(5;2,0.5)=10*0.5^5=0.313$$

$$F(5;3,0.5)=10*0.5^5=0.313$$

$$F(5;4,0.5)=5*0.5^5=0.156$$

$$F(5;5,0.5)=0.5^5=0.031$$

0.938

For  $N=5, p=0.8$ :

$$F(5;0,0.8)=0.2^5=0.0003$$

$$F(5;1,0.8)=5*0.2^4*0.8=0.0064$$

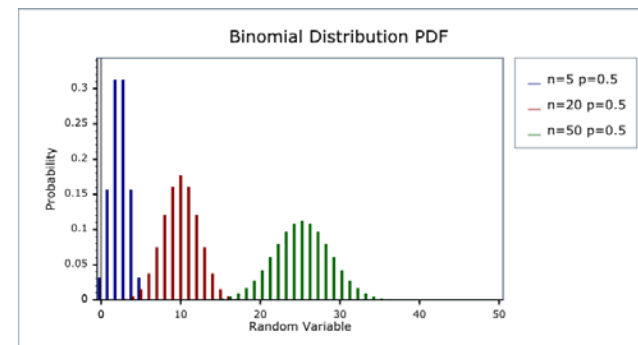
$$F(5;2,0.8)=10*0.2^3*0.8^2=0.0512$$

$$F(5;3,0.8)=10*0.2^2*0.8^3=0.2048$$

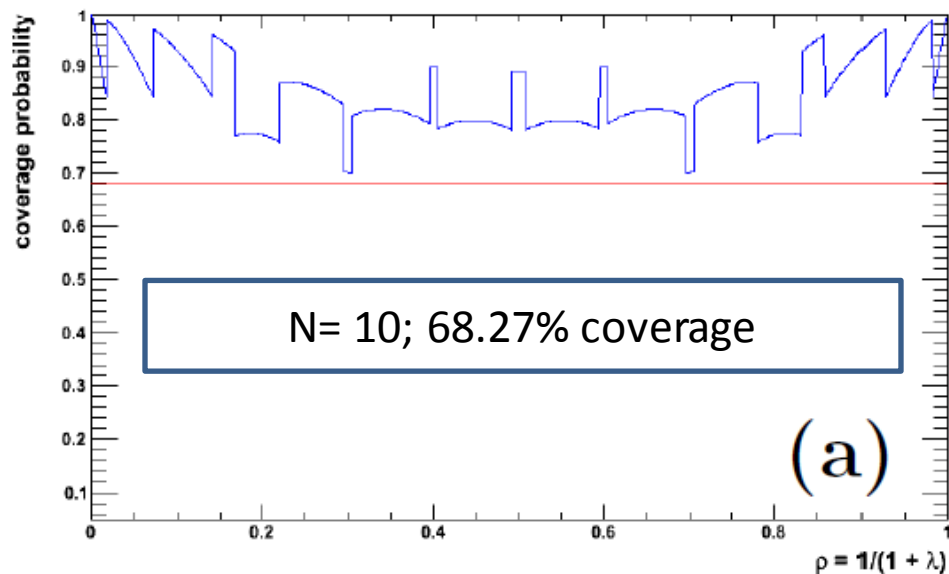
$$F(5;4,0.8)=5*0.2*0.8^4=0.4096$$

$$F(5;5,0.8)=0.8^5=0.3277$$

0.737



- The Binomial error bars for a small number of trials is indeed a complex problem!
- The (true) **variance is  $\sigma = \sqrt{\rho(1-\rho)/N}$**  , but its **ESTIMATE  $\sigma^* = \sqrt{\rho^*(1-\rho^*)/N}$**  (with  $\rho^* = \text{Successes}/N$ ) (so-called Wald interval) **fails** badly for small  $N$  and  $\rho^* \rightarrow 0,1$
- **Clopper-Pearson**: intervals obtained from Neyman's construction with a *central interval* ordering rule. **They overcover sizeably for some values of the trials/successes.**
- Lots of technology has been deployed to improve properties of binomial intervals



In HEP (and astro-HEP) the interest is related to the famous **on-off problem** (determine a expected background from a sideband)

# Wilson Score Interval for Binomial

[Cousins and Tucker, 0905.3831](#)

Already in 1927, Edwin Wilson [9] realized that since the rms depends on the unknown parameter  $\rho$ , the more appropriate way to invoke the Gaussian approximation was by consistently inverting the test using the rms of the null hypothesis for each value of  $\rho$ . For the lower endpoint, one uses the lowest value  $\rho_1$  such that  $\rho_1 + Z_{\alpha/2}\sqrt{\rho_1(1-\rho_1)/n_{\text{tot}}}$  contains  $\hat{\rho}$ . Analogously for the upper endpoint, one uses the largest value  $\rho_2$  such that  $\rho_2 - Z_{\alpha/2}\sqrt{\rho_2(1-\rho_2)/n_{\text{tot}}}$  contains  $\hat{\rho}$ . Letting  $T = (Z_{\alpha/2})^2/n_{\text{tot}}$ , this leads to a quadratic equation in  $\rho$  for the endpoints,  $(\rho - \hat{\rho})^2 = T\rho(1 - \rho)$ , with solutions

$$\rho = \frac{\hat{\rho} + T/2}{1 + T} \pm \frac{\sqrt{\hat{\rho}(1 - \hat{\rho})T + T^2/4}}{1 + T}. \quad (20)$$

These endpoints form the *Wilson score interval*; in spite of the fact that it is a non-iterative solution using nothing more than a square root, sadly it is commonly overlooked in favor of the Wald interval when a quick Gaussian estimate is desired.

$$Z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) = -\Phi^{-1}(\alpha/2)$$

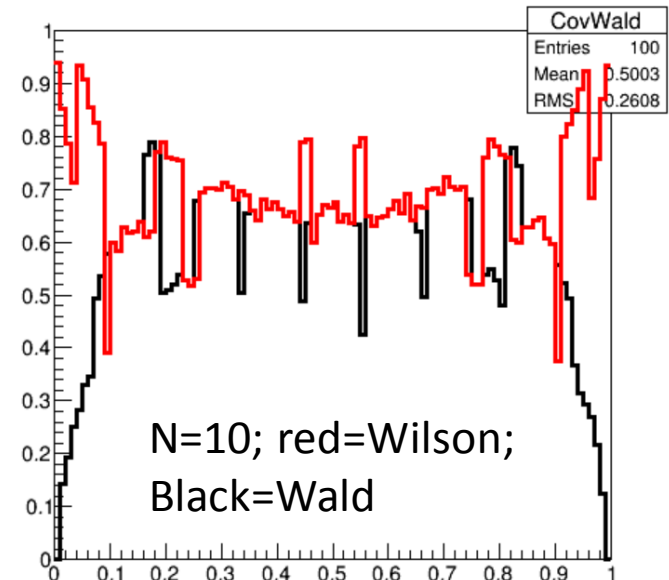
where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z \exp(-t^2/2) dt = \frac{1 + \text{erf}(Z/\sqrt{2})}{2},$$

so that

$$Z = \sqrt{2} \text{erf}^{-1}(1 - \alpha).$$

E.g.,  $Z_{\alpha/2} = 1$  for  $\alpha/2 = 0.159$ , and  $Z_{\alpha/2} = 1.64$  for  $\alpha/2 = 0.05$ .

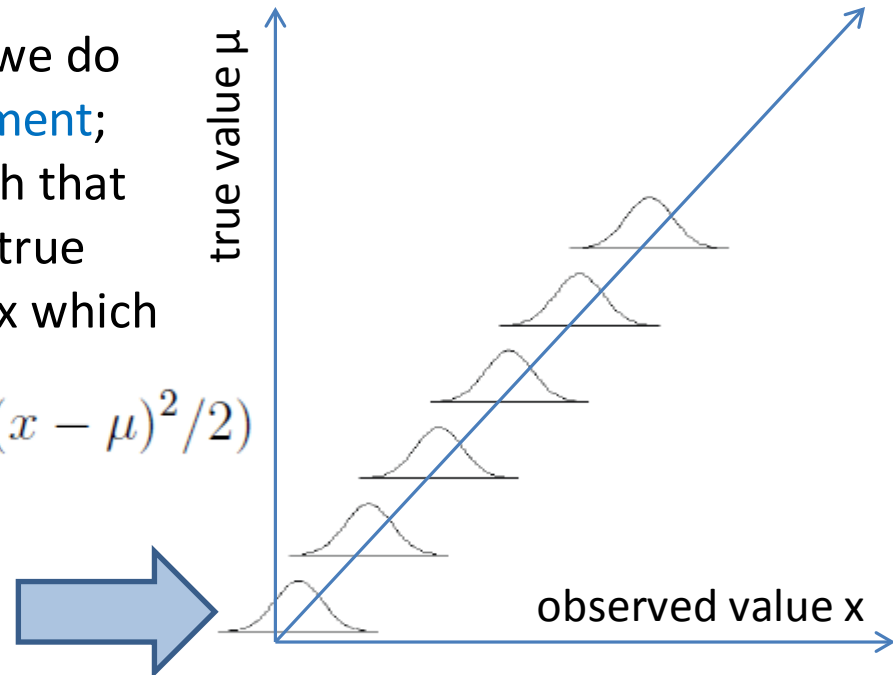


# Confidence Intervals and Flip-Flopping

- Here we want to understand a couple of issues that the Neyman construction can run into, for the very common case of the **measurement of a bounded parameter** and the derivation of upper limits on its value
- Typical observables falling in this category: cross section for a new phenomenon; or neutrino mass
- We take the simplifying assumption that we do a **unbiased Gaussian-resolution measurement**; we also renormalize measured values such that the variance is 1.0. In that case if  $\mu$  is the true value, our experiment will return a value  $x$  which is distributed as

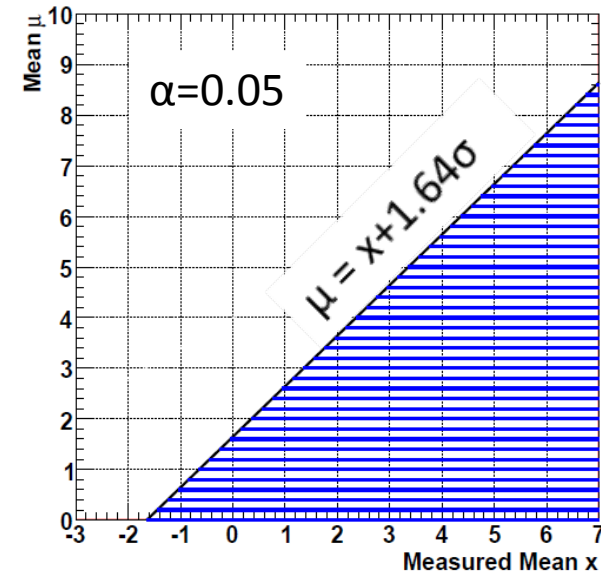
$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-(x - \mu)^2/2)$$

**Nota bene:  $x$  may assume negative values!**



# Neyman Construction for Bounded Parameter

- Gaussian measurement with known sigma ( $\sigma=1$  assumed in graph) of **bounded parameter**  $\mu \geq 0$
- Classical method for  $\alpha=0.05$  produces upper limit  $\mu < x + 1.64\sigma$  (or  $\mu < x + 1.28\sigma$  for  $\alpha=0.1$ )
- **for  $x < -1.64$  this results in the empty set!**
  - in violation of one of Neyman's own demands (confidence set does not contain empty sets)
- Also note:  $x \ll 0$  casts doubt on  $\sigma=1$  hypothesis  
→ rather than telling about value of  $\mu$  **the result could be viewed as a GoF test**



**Flip-flopping:** “since we observe no significant signal, we proceed to derive upper limits...”  
As a result, the upper limits undercover! (**Unified approach by Feldman and Cousins** solves the issue)

The attitude that one might take, upon measuring, say, a particle cross section which is negative (say if your backgrounds fluctuated up such that  $N_{\text{obs}} < B_{\text{exp}}$ ), is to **quote zero, and report an upper limit** which, in units of sigma, is

$$x^{\text{up}} = \text{sqrt}(2) * \text{ErfInverse}(1-2\alpha)$$

where  $\alpha$  is the desired confidence level.  $x^{\text{up}}$  is such that the integral of the Gaussian from minus infinity to  $x^{\text{up}}$  is  $1-\alpha$  (one-tailed test).

If, however, one finds  $x > D$ , where  $D$  is one's discovery threshold (say, 3-sigma or 5-sigma), one feels entitled to say one has "measured" a non-zero value of the parameter – a discovery of the Higgs, or a measurement of a non-zero neutrino mass. What the physicist will then report is rather an interval: to be consistent with the chosen test size  $\alpha$ , he will then quote central intervals which cover at the same level:  $x_{\text{meas}} \pm E(\alpha/2)$ , with  $E(\alpha) = \text{sqrt}(2) * \text{ErfInverse}(1-2*\alpha)$ .

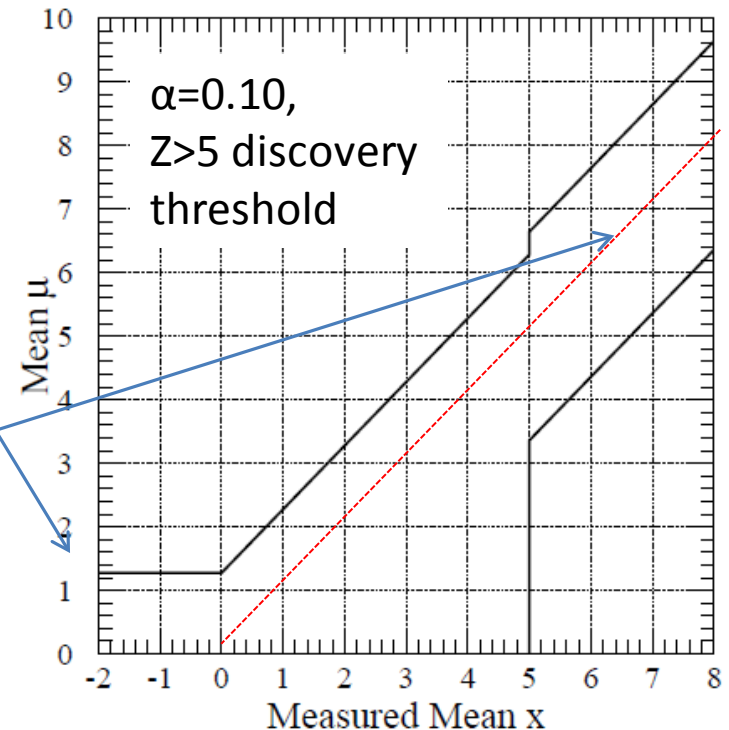
The confidence belt may then take the form shown on the graph on the right.

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{x^{\text{up}}}{\sqrt{2}}\right)$$

$$2\Phi(x) - 1 = \text{erf}\left(\frac{x^{\text{up}}}{\sqrt{2}}\right)$$

$$\frac{x^{\text{up}}}{\sqrt{2}} = \text{erfinv}(2\Phi(x) - 1)$$

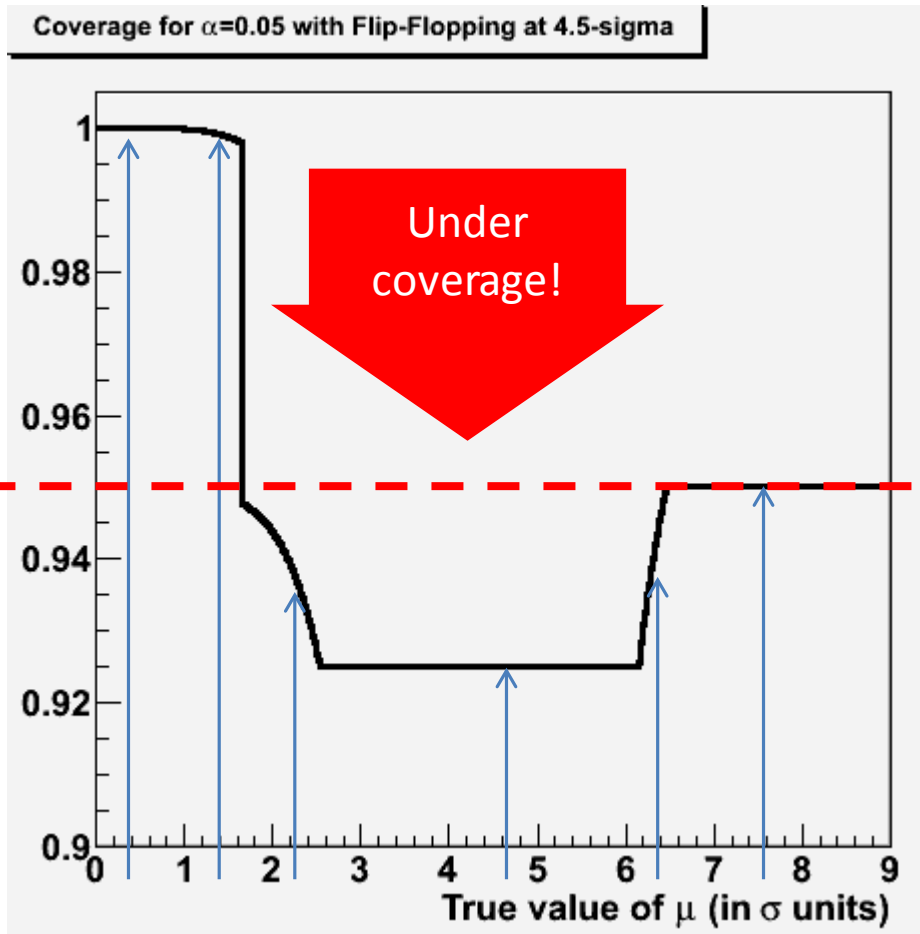
$$x^{\text{up}} = \sqrt{2} \text{erfinv}[2(1-\alpha) - 1] = \sqrt{2} \text{erfinv}(1-2\alpha)$$



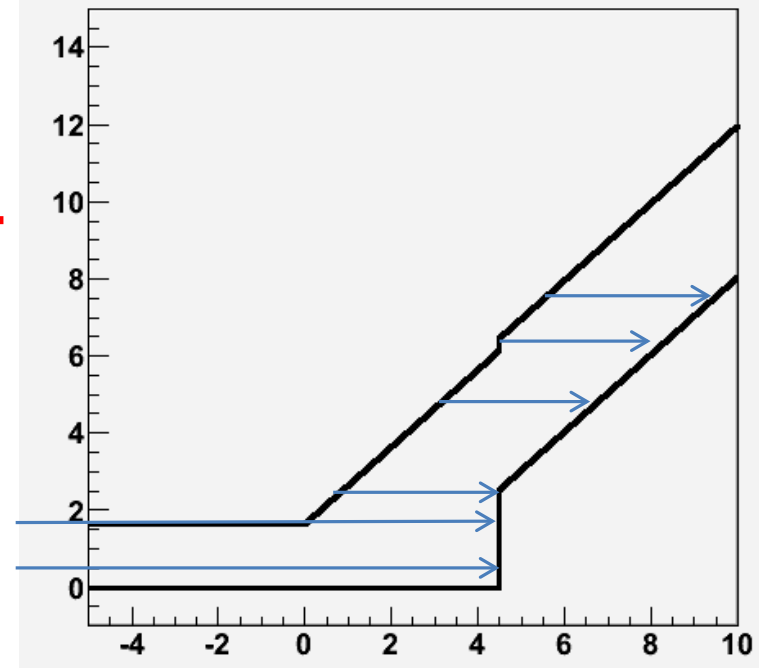
# Flip-Flopping Illustrated

- E.g.  $\alpha=0.05$ , Disc. Threshold = 4.5

The issue of Flip-Flopping and the empty set problem can be cured in the frequentist setting by the recipe advocated by G.Feldman and R.Cousins in 1998, based on a likelihood-ratio ordering of the acceptance intervals. The FC technique is widely used in HEP



Flip-flopping Confidence belt

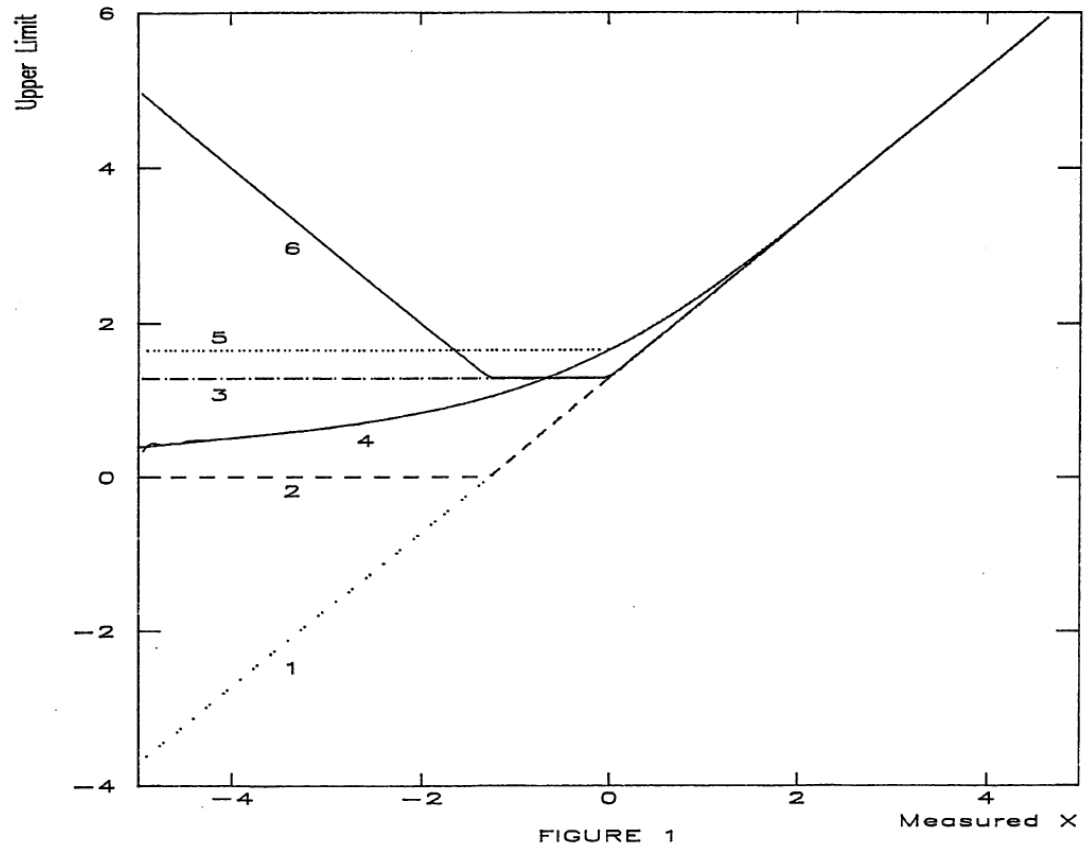




# Bounded $\mu$ Problem: Proposed Solutions

The graph illustrates various choices for confidence belts one can construct for the bounded parameter problem

The most principled among classical constructions is the one provided by **Feldman and Cousins** in 1998  
Bayesians have their own solution too



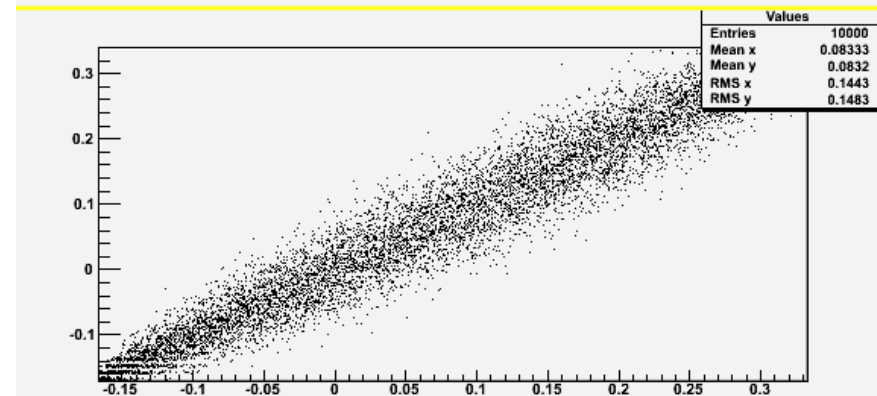
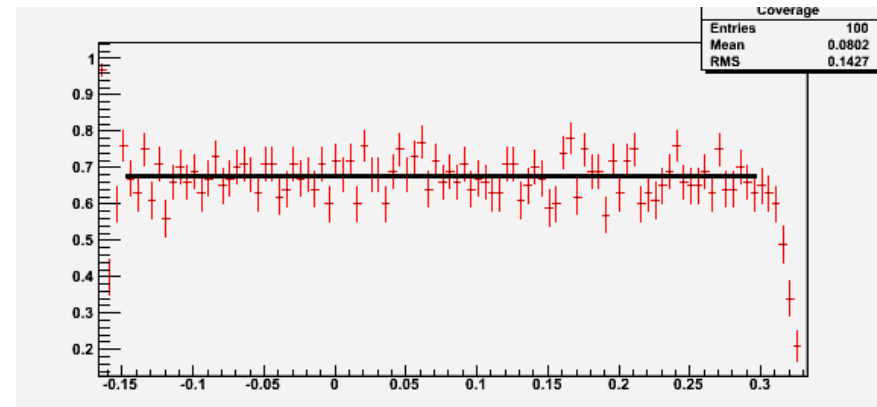
- (1) Neyman's recipe for 90% upper limits:  $\mu_{UL} = x + 1.28$ .
- (4) Bayesian solution: step-function prior
- (6) McFarlane's "loss of confidence"

# One Further Example of Coverage

- We can re-use the program "Die.C" You may modify it to compute the coverage of the likelihood intervals. → Die5.C

To do that, one must add a TH1D\* called "Coverage" and a cycle on the true parameter values, taking care of simulating the die throws correctly taking into account the bias  $t$ . Then one counts how often the likelihood has the true value within its interval, as a function of the true value.

By running it you will find that the coverage is only approximate for small number of throws, especially when your true value of the parameter  $t$  (the "increase in probability" of throws giving a 6) lies close to the boundaries  $-1/6, 1/3$ .



# Food for Thought: Relevant Subsets

Neyman's method applied to Gaussian measurement with known  $\sigma$  of a parameter with unknown **positive** mean  $\mu$  yields upper limits at 95% CL in the form  $\mu_{UL} = x + 1.64\sigma$ . **The procedure guarantees coverage, and yet...**

- Yet one can devise a betting strategy against it at 19:1 odds, using no more information than the observed  $x$ , and be guaranteed to win in the long run!
  - How? *Just choose a real constant  $k$ : bet that the interval does not cover when  $x < k$ , pass otherwise.*
  - For  $k < -1.64$  this wins EVERY bet! For larger  $k$ , advantage is smaller but is still  $> 0$ .

**Surely then, the procedure is not making the best inference on the data ?**

# Conditioning and Ancillary Statistics

In the bounded parameter problem, the flaw of being subject to winning bet strategies can be amended by adding a horizontal line or interval (such that any c.i. will contain that value of  $\mu$ ), but it **feels like a hack**

In other cases one can identify **ancillary statistics** and use them to **partition the space** into **relevant subsets**.

- “**Ancillary statistic**”:  $f(\text{data})$  yielding **information about the precision of the estimate** of the parameter of interest, but **no information about the parameter's value**.
- Most typical case in HEP: **branching fraction** measurement. With  $N_A, N_B$  event counts in two channels one finds that

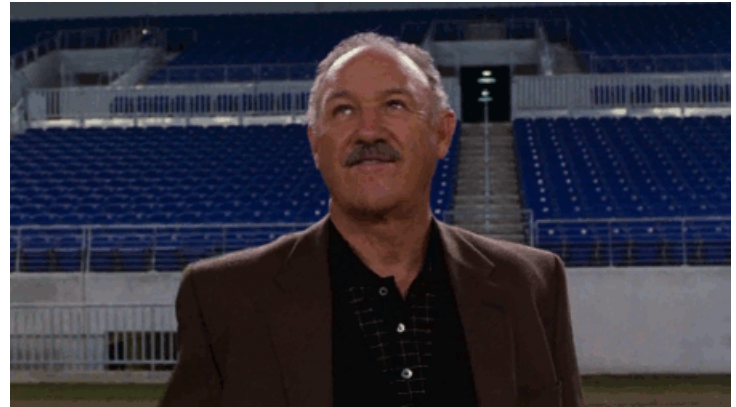
$$\begin{aligned} P(N_A, N_B) &= \text{Poisson}(N_A) \times \text{Poisson}(N_B) = \\ &= \text{Poisson}(N_A + N_B) \times \text{Binomial}(N_A | N_A + N_B) \end{aligned}$$

By using the second expression, one may **condition to having observed  $N_A + N_B$**  in total, and then **ignore the ancillary statistic  $N_A + N_B$** , since all the information on the BR is in the conditional binomial factor

→ by **restricting the sample space**, the problem is simplified.

# Cox Weighting Procedure

Things get even more intriguing in the famous example by [B. Cox\[2\]](#):

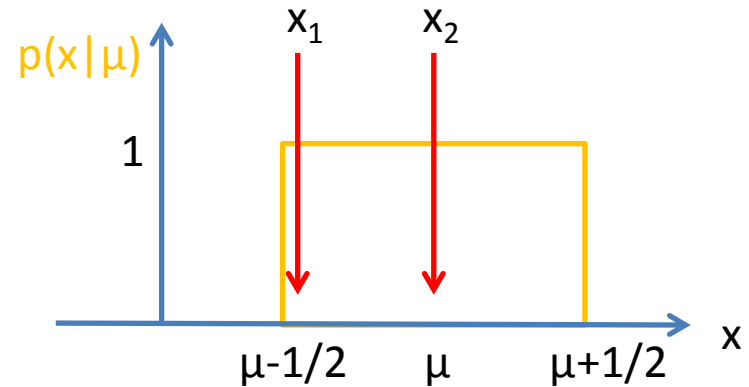


Flip a coin to decide whether to use a 10% scale (if you get tails) or a 1% scale (if you get heads) to measure an object's weight. [Which error do you quote for your measurement, upon getting heads ?](#)

Of course the knowledge of your device allows you to estimate that your precision is 1% - but a full NP construction (which is **unconditional** on the outcomes) would **require you to include the coin flipping in the procedure!**

# Locating the Box

- Another example:  
Find  $\mu$  using  $x_1, x_2$  sampled from  
 $p(x|\mu) = \text{Uniform}[\mu-1/2, \mu+1/2]$



Suppose e.g. that  $\mu=1$ , and take the two datasets,  
A: {0.99,1.01} ; B: {0.6,1.4}. **What would you prefer to measure?**

- NP procedures maximizing power in the unconditional space yield the same confidence interval for both data sets A and B; however, **B restricts the set of possible  $\mu$  to [0.9,1.1] while A only restricts it to [0.51,1.49] !**
- **There exists in fact an ancillary statistics  $|x_1-x_2|$  which carries no information on  $\mu$ , yet it can be used to divide the sample space in subsets where inference can be more or less powerful.**
- See **R. Cousins** for more discussion

# Relevant Subsets: Take-Away Bit

***Point made:*** *The quality of your inference depends on the breadth of the “whole space” you are considering. The more you can restrict it, the better (i.e. the more relevant) your inference becomes*

- Ancillary statistics are not easy to find, but they are quite useful!

→ Look for ancillary statistics in your everyday measurements!

# Properties of Estimators Relevant for Interval Estimation

- A *uniformly minimum variance unbiased estimator* (UMVU) for a parameter is the one which has the minimum variance possible, **for any value** of the unknown parameter it estimates.
- The form of the UMVU estimator depends on the distribution of the parameter!
- **Minimum variance bound:** it is given by the RCF inequality

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \left(E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right]\right)^{-1}$$

→ A unbiased estimator ( $b=0$ ) may have a variance as small as the inverse of the second derivative of the likelihood function, but not smaller.

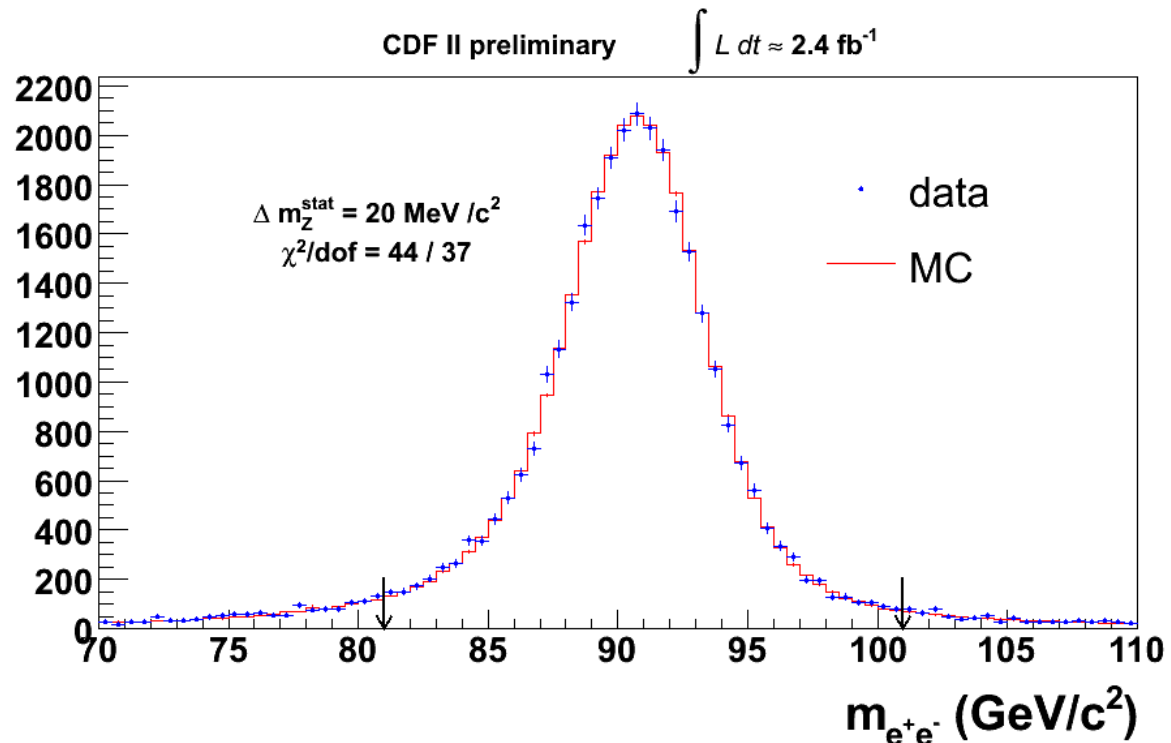
- Two related properties of estimators are **efficiency** and **robustness**.
  - **Efficiency:** the ratio of the variance to the *minimum variance bound*
  - Robustness:** more robust estimators are less dependent on deviations from the assumed underlying pdf
- Simple examples:
  - **Sample mean:** most used estimator for centre of a distribution - it is the UMVU estimator of the mean, if the distribution is Normal; however, for non-Gaussian distributions it may not be the best choice.
  - **Sample mid-range** (defined later): UMVU estimator of the mean of a *uniform distribution*
- Both sample mean and sample mid-range are efficient (asymptotically efficiency=1) for the quoted distribution (Gaussian and box, respectively). But for others, they are not. **Robust estimators have efficiency less dependent on distribution**



# A Robust Estimator: Trimmed Mean

- Often we have a sample of measurements, most of which are drawn from a narrow PDF, but we know that there is some "background" that follows a wider distribution
  - Simple example: a  $Z \rightarrow ee$  peak from collider data

- We might want narrow "signal"
- If we take the sample mean, we might not know the overall distribution and ruin the accuracy
  - Use a "Trimmed Mean" of the central quantiles
  - The estimate becomes



# Trimmed Mean Example: $Z \rightarrow ee$ Mass Distribution

Here we have 100  $Z \rightarrow ee$  candidates, and want a quick-and-dirty check of the energy scale in the EM calorimeter

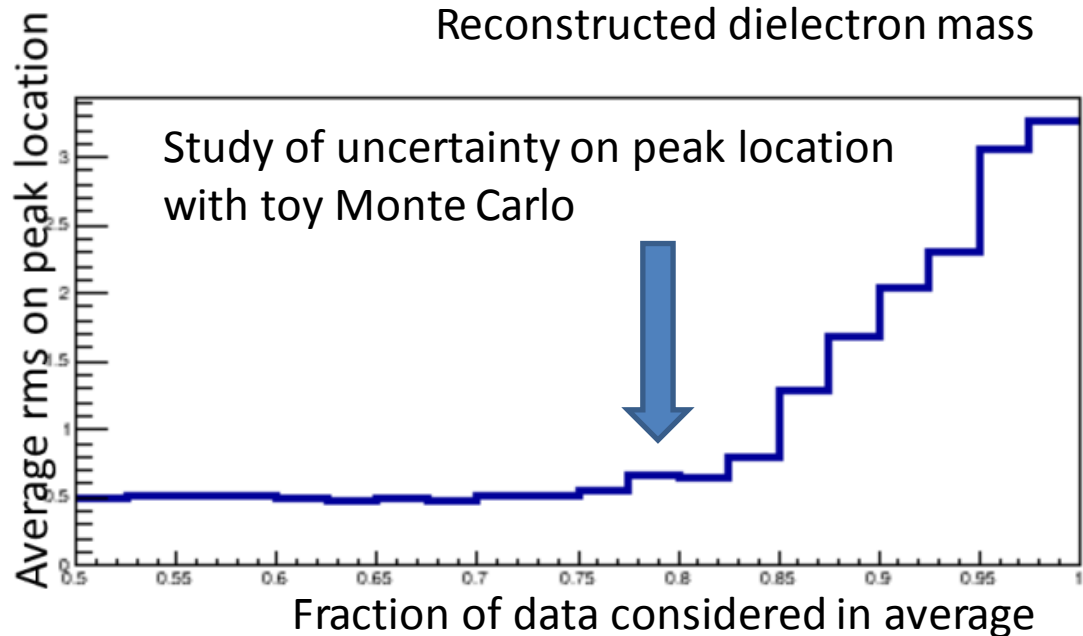
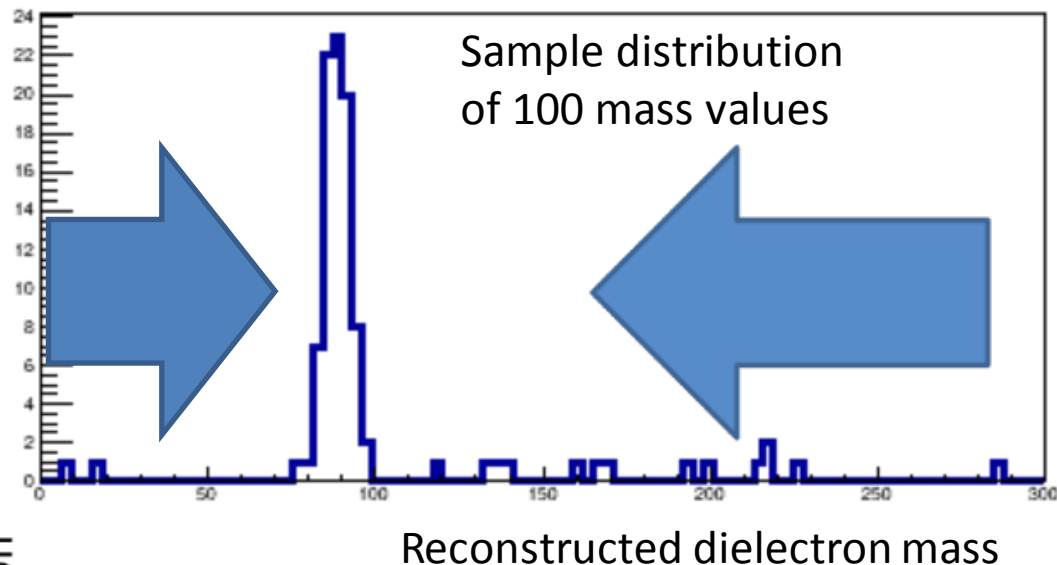
There is obviously some background, distributed at random values

To get the peak position we could fit the distribution, but a quicker way is to take the **trimmed mean**.

Average error of estimate indicates we should not average all data

In this case, for  $r < 0.8$  we are **insensitive to the background noise!**

*One obviously needs to find the proper working point for one's own problem  
→ Use toy MC technique!*



# Choosing Estimators: Another Example

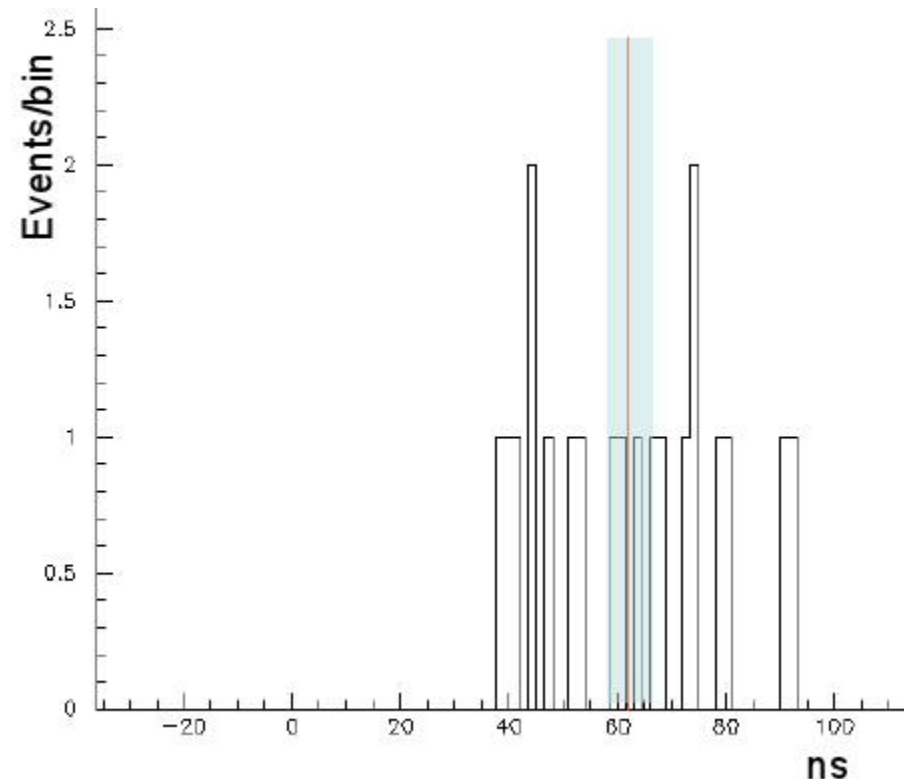
You are all familiar with the OPERA measurement of neutrino velocities

You may also have seen the graph below, which shows the distribution of  $\delta t$  (in nanoseconds) for individual neutrinos sent from narrow bunches at the end of October 2011

Because times are subject to random offset (jitter from GPS clock), you **might expect this to be a Box distribution**

OPERA quoted its best estimate of the  $\delta t$  as the **sample mean** of the measurements. **Would you have a better idea ?**

- This is **NOT the best choice** of estimator for the location of the center of a square distribution!
- OPERA quotes the following result:  
 **$\langle \delta t \rangle = 62.1 \pm 3.7$  ns**
- The **UMVU estimator for the Box is the mid-range**,  
 **$\delta t = (t_{\max} + t_{\min}) / 2$**
- You may understand why sample mid-range is better than sample mean: *once you pick the extrema, the rest of the data carries no information on the center!!!* It only adds noise to the estimate of the average!
- The larger N is, the larger the disadvantage of the sample mean.

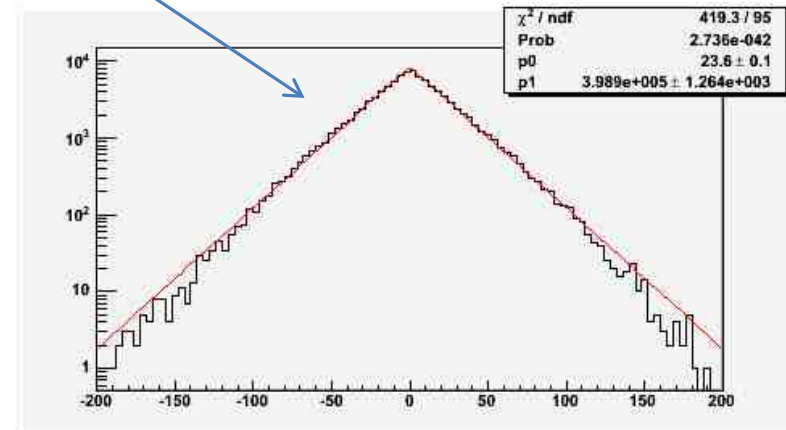
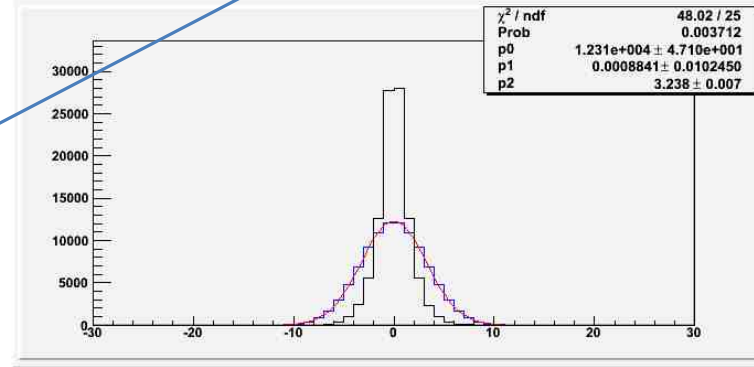
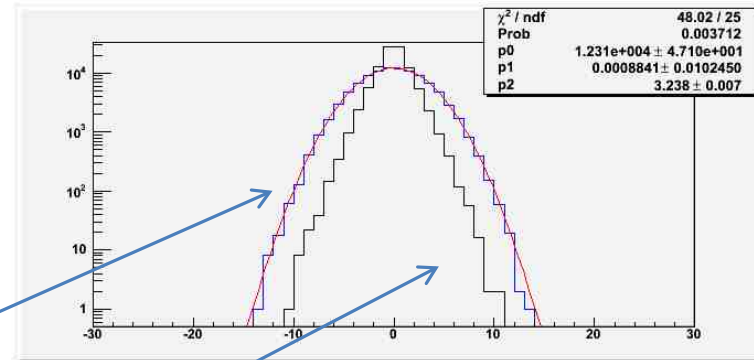


# Expected Uncertainty on Mid-Range and Average

- 100,000 n=20-entries histograms, with data distributed uniformly in [-25:25] ns
  - Average is asymptotically distributed as a Gaussian; for 20 events this is already a **good approximation**. Expected width is **3.24 ns**
  - Error on average consistent with Opera result
  - Mid-point has expected error of **1.66 ns**
  - if  $\delta t = (t_{\max} + t_{\min})/2$ , mid-point distribution  $P(n \delta t)$  is asymptotically a Laplace distribution; again 20 events are seen to already be **close to asymptotic behaviour** (but note departures at large values)
  - **If OPERA had used the mid-point, they would have halved their statistical uncertainty:**
  - $\langle \delta t \rangle = 62.1 \pm 3.7 \text{ ns} \rightarrow \langle \delta t \rangle = 65.2 \pm 1.7 \text{ ns}$

NB If you were asking yourselves what is a Laplace distribution:

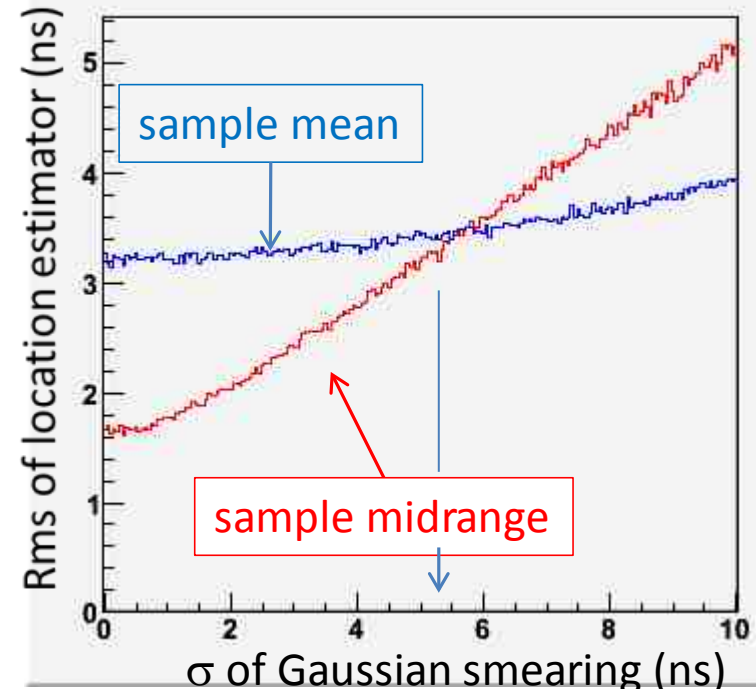
$$f(x) = 1/2b \exp(-|x-\mu|/b)$$



# However... The Devil Is in the Details

- Although the conclusions above are correct if the underlying pdf of the data is exactly a box distribution, **things change rapidly if we look at the real problem in more detail**
- Additional random smearings affect timing measurements:
  - the proton bunch has a peaked shape with 3ns FWHM
  - other effects contribute to smear randomly each timing measurement
- of course there may also be biases –fixed offsets due to imprecise corrections made to the delta t determination; these systematic uncertainties do not affect our conclusions, because they do not change the shape of the p.d.f
- **But the random smearings do affect our conclusions regarding the least variance estimator, since they change the p.d.f. !**

- One may assume that the smearings are Gaussian. The real p.d.f. from which the 20 timing measurements are drawn is then a convolution of a Gaussian with a Box distribution.
- Inserting that modification in the generation of toys one can study the effect. With 20-event samples, a Gaussian smearing with 6ns sigma is already enough to make the expected variance equal for the two estimators
- Timing smearings in Opera are likely  $O(6\text{ns}) \rightarrow$  **They did well in using the sample mean**



# Choice of Estimators: Take-Away Bit

- **Point made:** the intrinsic properties of estimators are not enough to choose them: the **problem at hand** [defined by the **pdf, e.g.  $p(x|\mu)$ , and the amount of data**] must be considered carefully when deciding how to perform a point and interval estimate
  - for point estimates, **bias** is usually a concern
  - But **variance** is equally important
  - In fact what one should minimize is the **Minimal Squared Error  $MSE = b^2 + \sigma^2$** , which is the expectation value of the squared difference between true and estimated value.
- To determine the UMVU (or a good substitute) is sometimes easy, sometimes hard. **A toy MC analysis can often be quite useful to understand what is optimal**, as analytical calculations are not always feasible

# Intermezzo: On The Weighted Average

# Weighted Average: the Basics

- Suppose we need to **combine two different, independent measurements** with variances  $\sigma_1, \sigma_2$  of the same physical quantity  $x_0$ :

– we denote them with

$$x_1(x_0, \sigma_1), x_2(x_0, \sigma_2)$$

← the PDFs are  $G(x_0, \sigma_i)$

- Let us combine them linearly to get the result with the smallest possible variance,

$$x = cx_1 + dx_2$$

→ What are  $c, d$  such that  $\sigma_x$  is smallest ?

Let us try this simple exercise

**Answer:** we first of all note that  $d=1-c$  if we want  $\langle x \rangle = x_0$  (*reason with expectation values to convince yourself of this*). Then, we express the variance of  $x$  in terms of the variance of  $x_1$  and  $x_2$

$$x = cx_1 + (1-c)x_2$$

$\sigma_x^2 = c^2\sigma_1^2 + (1-c)^2\sigma_2^2$ , and **find  $c$  which minimizes the expression**. This yields:

$$x = \frac{x_1 / \sigma_1^2 + x_2 / \sigma_2^2}{1 / \sigma_1^2 + 1 / \sigma_2^2}$$

$$\sigma_x^2 = \frac{1}{1 / \sigma_1^2 + 1 / \sigma_2^2}$$

The generalization of these formulas to  $N$  measurements is trivial



# Linearization and Correlation

In the method of LS *the linear approximation in the covariance (Taylor series expansion to first order) may lead to strange results*

Let us consider the LS minimization of a combination of two measurements of the same physical quantity  $k$ , for which the covariance terms be all known.

In the **first case** let there be a **common offset error**  $\sigma_c$ . We may combine the two measurements  $x_1, x_2$  with LS by computing the inverse of the covariance matrix:

$$V = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 + (\sigma_1^2 + \sigma_2^2) \sigma_c^2} \begin{pmatrix} \sigma_2^2 + \sigma_c^2 & -\sigma_c^2 \\ -\sigma_c^2 & \sigma_1^2 + \sigma_c^2 \end{pmatrix}$$
$$\chi^2 = \frac{(x_1 - k)^2 (\sigma_2^2 + \sigma_c^2) + (x_2 - k)^2 (\sigma_1^2 + \sigma_c^2) - 2(x_1 - k)(x_2 - k) \sigma_c^2}{\sigma_1^2 \sigma_2^2 + (\sigma_1^2 + \sigma_2^2) \sigma_c^2}$$

The minimization of the above expression leads to the following expressions for the best estimate of  $k$  and its standard deviation:

$$\hat{k} = \frac{x_1 \sigma_2^2 + x_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

The best fit value does not depend on  $\sigma_c$ , and corresponds to the weighted average of the results when the individual variances  $\sigma_1^2$  and  $\sigma_2^2$  are used.

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \sigma_c^2$$

This result is what we expected, and all is good here.

# Normalization Error: *Hic Sunt Leones*

In the **second case** we take two measurements of k having a **common scale error**.

$$V = \begin{pmatrix} \sigma_1^2 + x_1^2 \sigma_f^2 & x_1 x_2 \sigma_f^2 \\ x_1 x_2 \sigma_f^2 & \sigma_2^2 + x_2^2 \sigma_f^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 + (x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2) \sigma_f^2} \begin{pmatrix} \sigma_2^2 + x_2^2 \sigma_f^2 & -x_1 x_2 \sigma_f^2 \\ -x_1 x_2 \sigma_f^2 & \sigma_1^2 + x_1^2 \sigma_f^2 \end{pmatrix}$$

$$\chi^2 = \frac{(x_1 - k)^2 (\sigma_2^2 + x_2^2 \sigma_f^2) + (x_2 - k)^2 (\sigma_1^2 + x_1^2 \sigma_f^2) - 2(x_1 - k)(x_2 - k)x_1 x_2 \sigma_f^2}{\sigma_1^2 \sigma_2^2 + (x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2) \sigma_f^2}$$

This time the minimization produces these results for the best estimate and its variance:

Try this at home to see how it works!

$$\hat{k} = \frac{x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2 \sigma_f^2}$$

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2 \sigma_2^2 + (x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2) \sigma_f^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2 \sigma_f^2}$$

Before we discuss these formulas, let us test them on a simple case:

$$x_1 = 10 \pm 0.5,$$

$$x_2 = 11 \pm 0.5,$$

$$\sigma_f = 20\%$$

This yields the following disturbing result:

$$k = 8.90 \pm 2.92 !$$

What is going on ???

# Shedding Some Light on the Disturbing Result

- The fact that we get a result outside the range of inputs requires investigation.
- Rewrite the result by dividing it by the weighted average result obtained ignoring the scale correlation:



$$\hat{k} = \frac{x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2 \sigma_f^2}$$

$$\bar{x} = \frac{x_1^2 \sigma_2^2 + x_2^2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

$$\Rightarrow \frac{\hat{k}}{\bar{x}} = \frac{1}{1 + \frac{(x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2} \sigma_f^2}$$

If the two measurements differ, their squared difference divided by the sum of the individual variances plays a role in the denominator. In that case **the LS fit “squeezes the scale” by an amount allowed by  $\sigma_f$  in order to minimize the  $\chi^2$ .**

This is because *the LS expression uses only first derivatives of the covariance*: the individual variances  $\sigma_1$ ,  $\sigma_2$  do not get rescaled when the normalization factor is lowered, but the points get closer.

This may be seen as a shortcoming of the linear approximation of the covariance, but it might also be viewed as a *careless definition of the covariance matrix itself* instead (see next slide) !

- In fact, let us try again. We had defined earlier the covariance matrix as

$$V = \begin{pmatrix} \sigma_1^2 + x_1^2 \sigma_f^2 & x_1 x_2 \sigma_f^2 \\ x_1 x_2 \sigma_f^2 & \sigma_2^2 + x_2^2 \sigma_f^2 \end{pmatrix}$$

- The expression above contains the estimates of the true value, not the true value itself. We have learned to **beware** of this earlier... What happens if we instead try using the following ?

$$V = \begin{pmatrix} \sigma_1^2 + k^2 \sigma_f^2 & k^2 \sigma_f^2 \\ k^2 \sigma_f^2 & \sigma_2^2 + k^2 \sigma_f^2 \end{pmatrix}$$

The minimization of the resulting  $\chi^2$ ,

$$\chi^2 = \frac{(x_1 - k)^2 (\sigma_2^2 + k^2 \sigma_f^2) + (x_2 - k)^2 (\sigma_1^2 + k^2 \sigma_f^2) - 2(x_1 - k)(x_2 - k)k^2 \sigma_f^2}{\sigma_1^2 \sigma_2^2 + (\sigma_1^2 + \sigma_2^2)k^2 \sigma_f^2}$$

produces as result the weighted average

$$\hat{k} = \frac{x_1 \sigma_2^2 + x_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

- The same would be obtained by maximizing the likelihood

$$L = \exp \left[ -\frac{(x_1 - k)^2}{2(\sigma_1^2 + x_1^2 \sigma_f^2)} \right] \exp \left[ -\frac{(x_2 - k)^2}{2(\sigma_2^2 + x_2^2 \sigma_f^2)} \right]$$

or even minimizing the  $\chi^2$  defined as

$$\chi^2 = \frac{(fx_1 - k)^2}{(f\sigma_1)^2} + \frac{(fx_2 - k)^2}{(f\sigma_2)^2} + \frac{(f-1)^2}{\sigma_f^2}$$

Note that the latter corresponds to “averaging first, dealing with the scale later”.

# When Do Results Outside Bounds Make Sense ?

- Let us now go back to the general case of taking the average of two correlated measurements, when the correlation terms are expressed in the general form :

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- The LS estimators provide the following result for the weighted average [Cowan 1998]:

$$\hat{x} = wx_1 + (1-w)x_2 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} x_1 + \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} x_2$$

whose (inverse) variance is

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2} \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2} \right) = \frac{1}{\sigma_1^2} + \frac{1}{1-\rho^2} \left( \frac{\rho}{\sigma_1} - \frac{1}{\sigma_2} \right)^2$$

From the above we see that once we take a measurement of  $x$  of variance  $\sigma_1^2$ , a second measurement of the same quantity will reduce the variance of the average unless  $\rho = \sigma_1/\sigma_2$ .

But what happens if  $\rho > \sigma_1/\sigma_2$  ? In that case the weight  $w$  gets negative, and the average goes outside the “psychological” bound  $[x_1, x_2]$ .

The reason for this behaviour is that with a large positive correlation the two results are likely to lie on the same side of the true value! On which side they are predicted to be by the LS minimization depends on which result has the smallest variance.

# Exercise

- Suppose you have a measurement  $x_1$  of a physical quantity  $x$ , with a variance  $\sigma_1^2=1.0$ . You are offered to improve the knowledge of  $x$  by performing a second measurement  $x_2$  with variance  $\sigma_2=4.0$  and taking the weighted average of the two. You can choose to do this with two different methods. The first method offers a result with a 50% correlation with  $x_1$ ; the second offers a correlation of 75%.
- Which one should you choose and why ?

# How Can That Be ?

It seems a paradox, but it is not. Again, the reason why we cannot digest the fact that the best estimate of the true value  $\mu$  be outside of the range of the two measurements is our incapability of understanding intuitively the mechanism of large correlation between our measurements.

- **John:** “I took a measurement, got  $x_1$ . I now am going to take a second measurement  $x_2$  which has a larger variance than the first. Do you mean to say I will more likely get  $x_2 > x_1$  if  $\mu < x_1$ , and  $x_2 < x_1$  if  $\mu > x_1$  ??”

**Jane:** “That is correct. Your second measurement ‘goes along’ with the first, because your experimental conditions made the two highly correlated and  $x_1$  is more precise.”

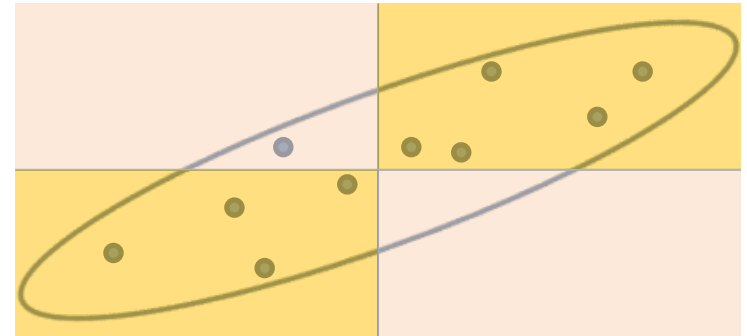
**John:** “But that means my second measurement is **utterly useless!**”

**Jane:** “Wrong. It will in general **reduce the combined variance**. Except for the very special case of  $\rho = \sigma_1 / \sigma_2$ , the weighted average will converge to the true  $\mu$ . **LS estimators are consistent !!**”.

# Jane vs John, Round 1

**John:** “I still can’t figure out how on earth the average of two numbers can be outside of their range. It just fights with my common sense.”

**Jane:** “You need to think in probabilistic terms. Look at this error ellipse: it is thin and tilted (high correlation, large difference in variances).”



**John:** “Okay, so ?”

**Jane:** “Please, would you pick a few points at random within the ellipse?”

**John:** “Done. Now what ?”

**Jane:** “Now please tell me whether they are mostly on the same side (orange rectangles) or on different sides (pink rectangles) of the true value.”

**John:** “Ah! Sure, all but one are on orange areas”.

**Jane:** “That’s because their correlation makes them likely to “go along” with one another.”



# Round 2: a Geometric Construction

**Jane:** “And I can actually make it even easier for you. Take a two-dimensional plane, draw axes, draw the bisector: the latter represents the possible values of  $\mu$ . Now draw the error ellipse around a point of the diagonal. Any point, we’ll move it later.”

**John:** “Done. Now what ?”

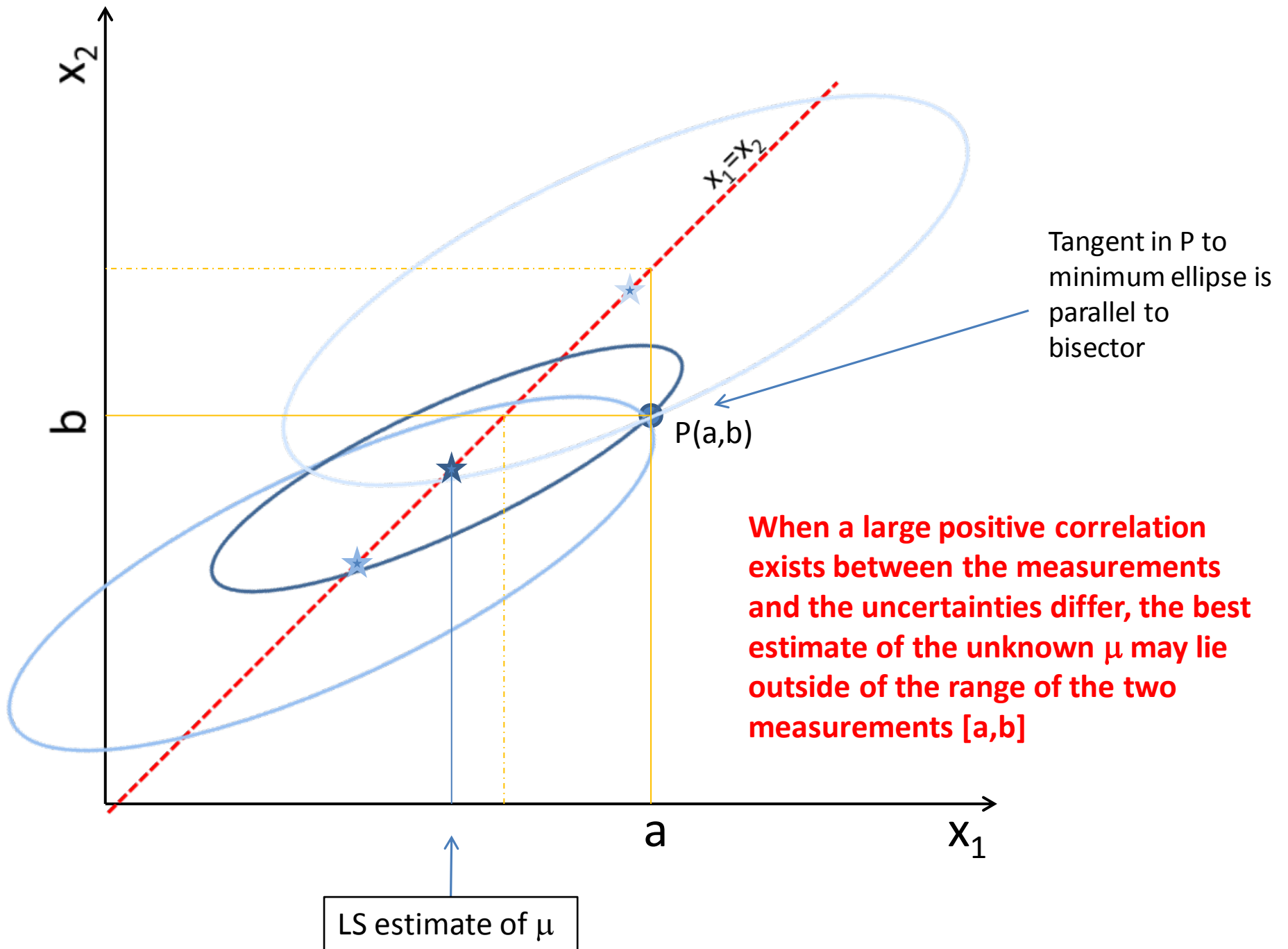
**Jane:** “Now enter your measurements  $x=a$ ,  $y=b$ . That corresponds to picking a point  $P(a,b)$  in the plane. Suppose you got  $a>b$ : you are on the lower right triangle of the plane. To find the best estimate of  $\mu$ , move the ellipse by keeping its center along the diagonal, and try to scale it also, such that you intercept the measurement point  $P$ .”

**John:** “But there’s an infinity of ellipses that fulfil that requirement”.

**Jane:** “That’s correct. But **we are only interested in the smallest ellipse!** Its center will give us the best estimate of  $\mu$ , given  $(a,b)$ , the ratio of their variances, and their correlation.”

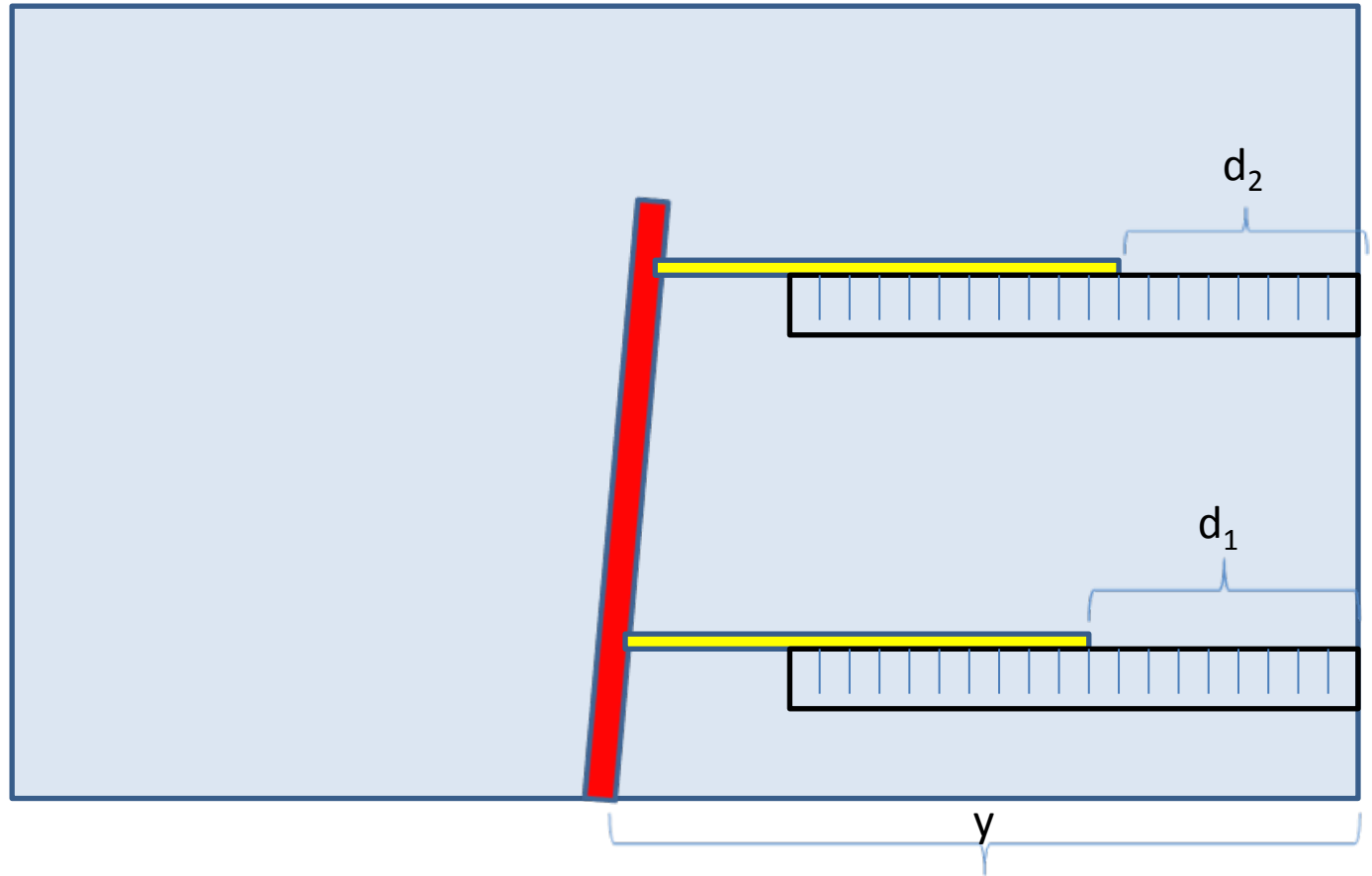
**John:** “Oooh! Now I see it! It is bound to be outside of the interval!”

**Jane:** “Well, that is not true: **it is outside of the interval only because the ellipse you have drawn is thin and its angle with the diagonal is significant.** In general, the result depends on how correlated the measurements are (how thin is the ellipse) as well as on how different the variances are (how big is the angle of its major axis with the diagonal). Note also that in order for the “result outside bounds” to occur, the correlation must be positive!



# Trivia – Try It at Home

Here is a simple arrangement with which you can test whether or not a significant correlation between two measurements causes the effect we have been discussing.



Here we measure  $y$  with a ruler shorter than  $y$ , by taking  $d_1$  and  $d_2$  and using the yellow stick as an offset. The arrangement is such that we set the yellow stick from the edge of the red bar, and the red bar may have an angle error WRT the orthogonal to  $y$ . The non-zero angle causes a correlation between the two measurements  $d_1$  and  $d_2$ . It turns out that  $y_1 = d_1 + a$  and  $y_2 = d_2 + a$  ( $a$  being the length of the yellow stick) will be on the same side of the true value of  $y$ , if the angle error is larger than the other uncertainties in the measurements.

# When Chi-By-Eye Fails !

Which of the PDF (parton distribution functions!) models shown in the graph is a best fit to the data:

CTEQ4M (horizontal line at 0.0) or MRST (dotted curve) ?

**You cannot tell by eye!!!**

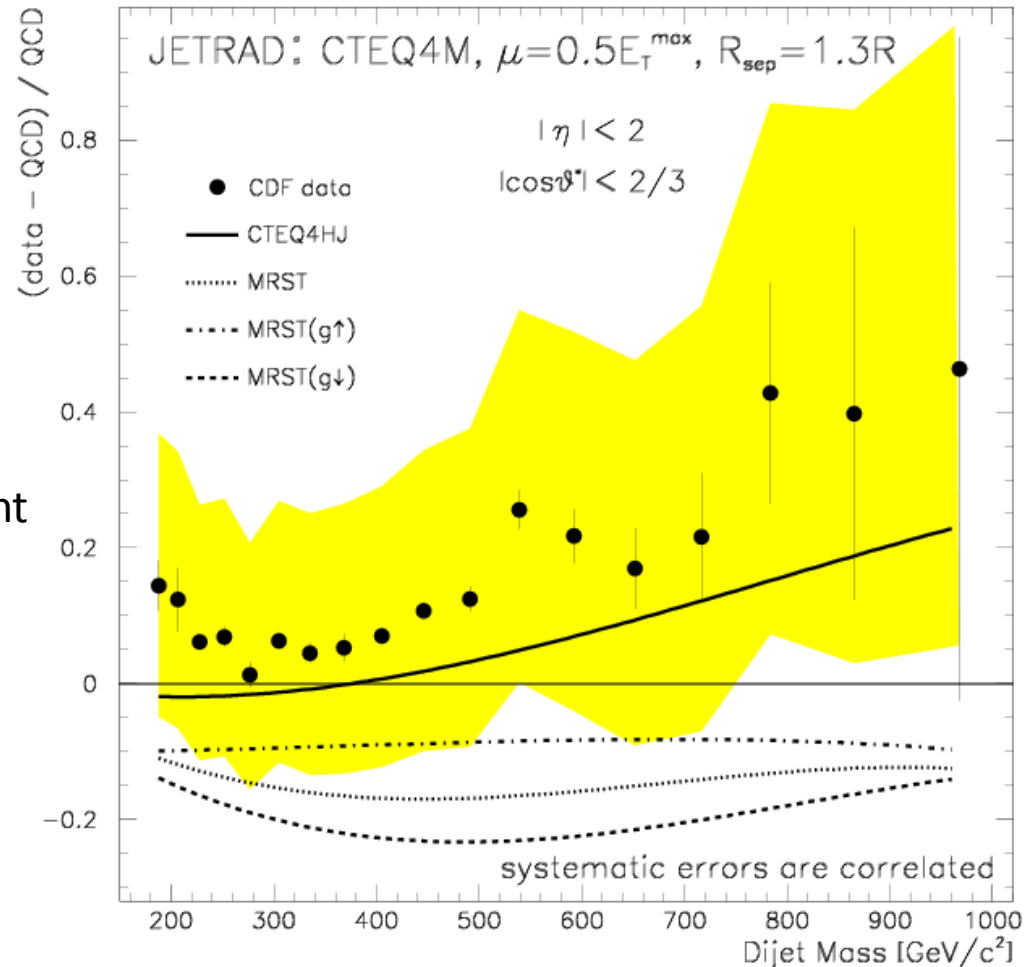
The presence of large correlations makes the normalization much less important than the shape.

$p\text{-value}(\chi^2 \text{ CTEQ4M}) = 1.1\text{E-}4,$   
 $p\text{-value}(\chi^2 \text{ MRST}) = 3.2\text{E-}3 :$

The MRST fit has a 30 times higher p-value than the CTEQ4M fit !

Take-home lessons:

- Be careful with LS fits in the presence of large common systematics!
- Do not trust your eye when data points carry significant bin-to-bin correlations!



Source: 1998 CDF measurement of the differential dijet mass cross section using 85/pb of Run I data, F. Abe et al., The CDF Collaboration, Phys. Rev. Lett. 77, 438 (1996)

# Drawing Home a Few Lessons

- If I managed to thoroughly confuse you, I have reached my goal!  
There are a number of lessons to take home from this:
  - Even the simplest problems can be easily mishandled if we do not pay a lot of attention...
  - **Correlations may produce surprising results.** The average of highly-correlated measurements is an especially dangerous case, because a small error in the covariance leads to large errors in the point estimate.
  - Knowing the PDF your data are drawn from is CRUCIAL (but you then have to use that information correctly!)
  - **Statistics is hard! Pay attention to it if you want to get correct results !**

# Hypothesis Testing and GOF

- A few basic definitions
- Statistical significance: what is it ?
- The Jeffrey-Lindley Paradox
- Some examples

# Hypothesis Testing: Generalities

We are often concerned with **proving or disproving a theory**, or comparing and **choosing between different hypotheses**.

In general this is a different problem than that of estimating a parameter, but the two are tightly connected.

If nothing is known a priori about a parameter, naturally one uses the data to **estimate** it; if however theory predictions exist, the problem is better formulated as a **test of hypothesis**.

Within the idea of hypothesis testing one must also consider **goodness-of-fit tests**: **in that case there is only one hypothesis** to test (e.g. a particular value of a parameter as opposed to any other value), so some of the possible techniques are not applicable

A hypothesis is **simple** if it is completely specified; otherwise (e.g. if depending on the unknown value of a parameter) it is called **composite**.



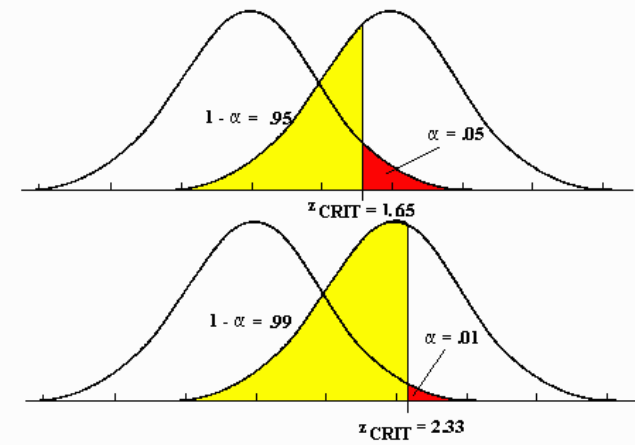
# Nuts and Bolts of Hypothesis Testing

- $H_0$ : null hypothesis
- $H_1$ : alternate hypothesis
- Three main parameters in the game:
  - $\alpha$ : **type-I error rate**; probability that  $H_0$  is true although you accept the alternative hypothesis
  - $\beta$ : **type-II error rate**; probability that you fail to claim a discovery (accept  $H_0$ ) when in fact  $H_1$  is true
  - $\theta$ , parameter of interest (describes a continuous hypothesis, for which  $H_0$  is a particular value). E.g.  $\theta=0$  might be a zero cross section for a new particle
- Common for  $H_0$  to be **nested** in  $H_1$

Can compare different methods by plotting the test statistic for  $H_0$  and  $H_1$  and look at  $\alpha$  vs  $\beta$

- Usually there is a tradeoff between  $\alpha$  and  $\beta$ ; often a **subjective decision, involving cost** of the two different errors.
- Tests may be more powerful in specific regions of an interval

In classical hypothesis testing, **test of  $\theta=0$  equates to asking whether 0 is in the confidence interval** (HT  $\leftrightarrow$  Interval estimation)

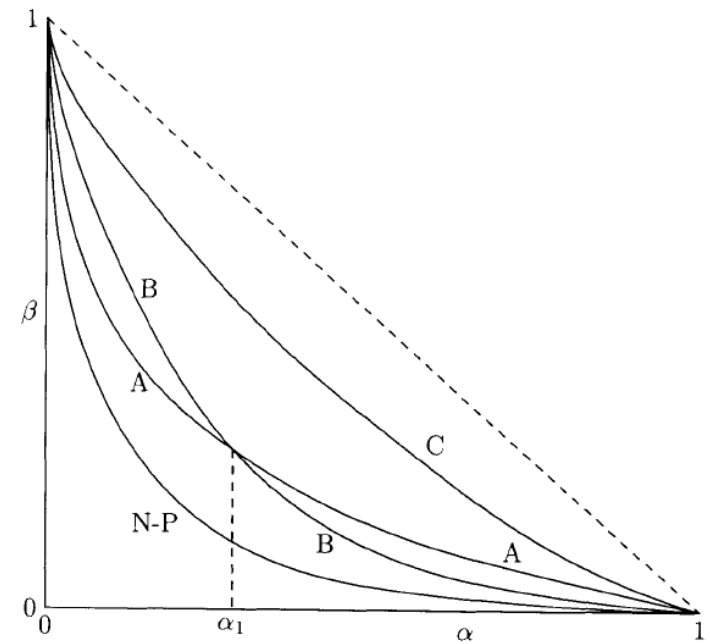


Above, a smaller  $\alpha$  is paid with a larger type-II error rate (yellow area)  
 $\rightarrow$  smaller power  $1-\beta$

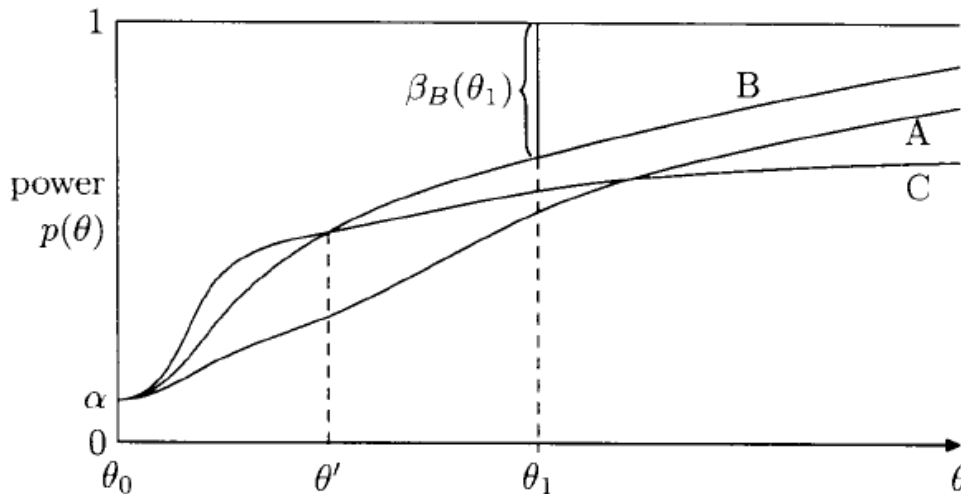


# Alpha vs Beta and Power Graphs

- Very general framework of classification
- **Choice of  $\alpha$  and  $\beta$  is conflicting**: where to stay in the curve provided by your analysis method highly depends on habits in your field
- What makes a difference is the **test statistic**: note how the N-P likelihood-ratio test outperforms others in the figure – reason is N-P lemma (see below)



As data size increases, power curve becomes closer to step function



The power of a test usually also depends on the parameter of interest: different methods may have better performance in different parameter space points

UMP (**uniformly most powerful**): has the highest power for any  $\theta$

Fig. 10.3. Power functions of tests A, B, and C at significance level  $\alpha$ . Of these three tests, B is the best for  $\theta > \theta'$ . For smaller values of  $\theta$ , C is better.

# Statistical Significance: What It Is

Statistical significance reports the probability that an experiment obtains data **at least as discrepant as** those actually observed, under a given "null hypothesis"  $H_0$

- In physics  $H_0$  *usually describes the currently accepted and established theory*
- Given **data X** and a **test statistic T** (a function of X), one may obtain a **p**-value as the **probability of obtaining a value of T at least as extreme as the one observed**, if  $H_0$  is true.

**p** can then be converted into the corresponding number of "sigma," *i.e.* standard deviation units from a Gaussian mean. This is done by finding **x** such that **the integral from x to infinity** of a unit Gaussian equals **p**:

$$\frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{t^2}{2}} dt = p$$

According to the above recipe, a **15.9%** probability is a one-standard-deviation effect; a **0.135%** probability is a three-standard-deviation effect; and a **0.0000285%** probability corresponds to five standard deviations - "**five sigma**" in jargon.

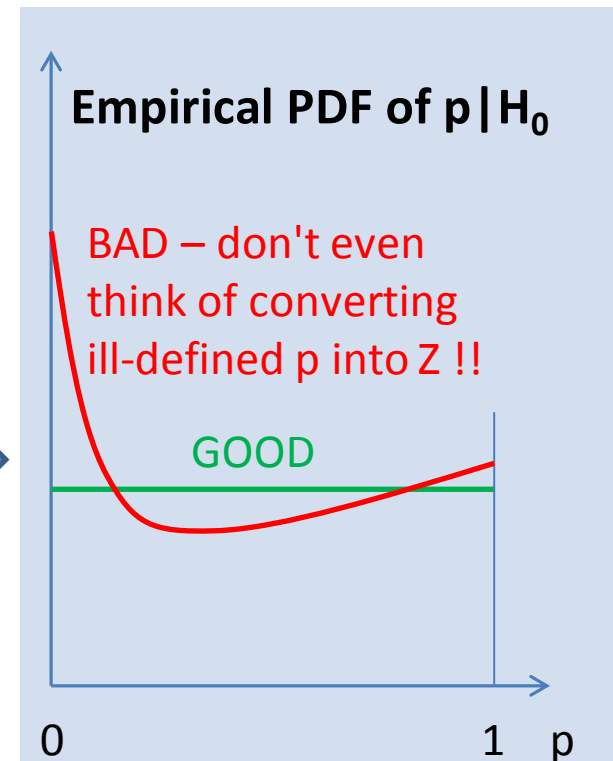
# Notes

The convention is to use a “one-tailed” Gaussian: we do not care about departures of  $x$  from the mean in the *un-interesting direction*

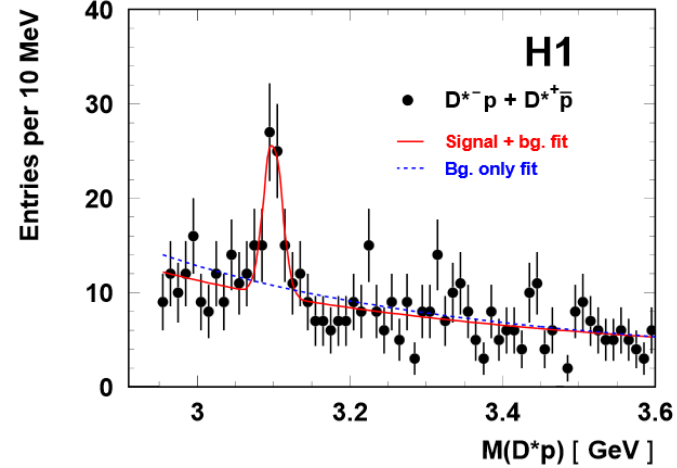
The conversion of  $p$  into  $\sigma$  is independent of experimental detail. Using  $N\sigma$  rather than  $p$  is just a **shortcut, nothing more** !

In particular, using “sigma” units does in no way mean we are operating some kind of Gaussian approximation anywhere in the problem

The whole construction rests on a proper definition of the  $p$ -value. Any shortcoming of the properties of  $p$  (e.g. a tiny non-flatness of its PDF under the null hypothesis) totally invalidates the meaning of the derived  $N\sigma$



# An Important Ingredient: Wilks' Theorem

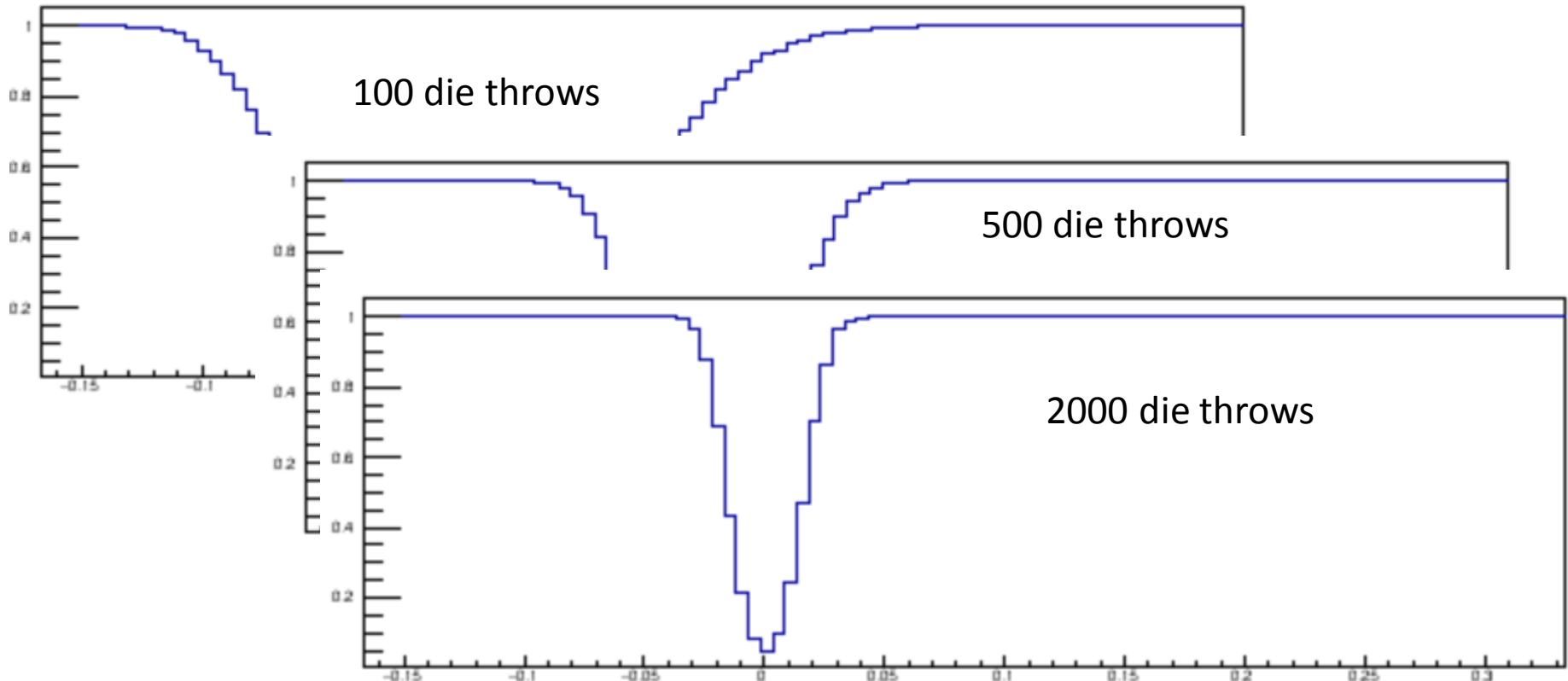


- An almost ubiquitous method to derive a significance from a likelihood fit is the one of invoking **Wilks' theorem**
  - that is, some invoke it although they are not aware they are doing it !
- One has a likelihood under the null hypothesis,  $L_0$  (say, a background-only fit), and a likelihood for an alternative,  $L_1$  (a signal+background fit)
- One takes  $-2(\ln L_1 - \ln L_0) = -2\Delta(\ln L)$  and interprets it as a chisquare
- $P(\chi^2)$  can then be obtained, and from it a Z-value
  - But people regularly forget that this is only applicable when the two hypotheses are connected by  $H_0$  being a particular case of  $H_1$  (fixing of one parameter): they must be **nested models**.
  - In most cases this is not so: we routinely test a  $H_1$  where one of the parameters is not present in  $H_0$  (mass  $m$  for  $\sigma=0$ ).

Fortunately, often even when the regularity conditions demanded by the theorem are not met, the asymptotic properties of  $\Delta \ln L$  are good enough

# Power of the Die Load Test

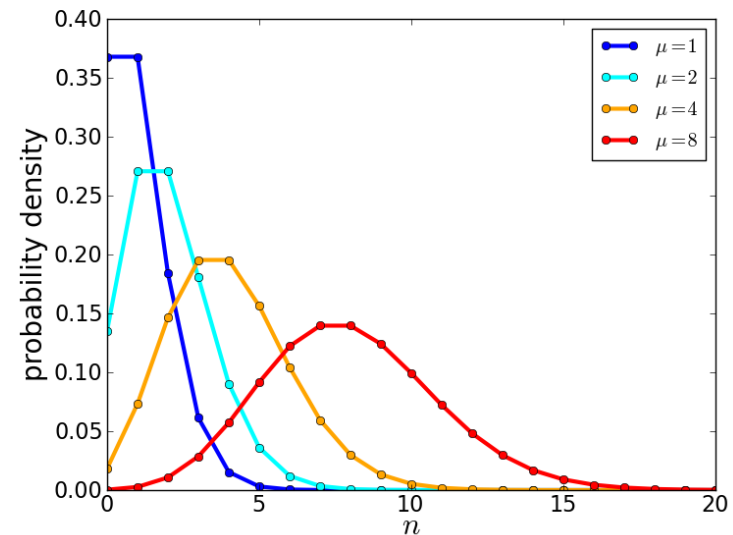
- We can revisit the macro Die5.C, which studies the hypothesis that there is a load in the die, and study the power of the test (is  $t=0$  in the critical region?) as the data size increases



# The Poisson distribution

- We all know what the Poisson distribution is:

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



- The expectation value of a Poisson variable with mean  $\mu$  is  $\mathbf{E(n) = \mu}$
- its variance is  $V(n) = \mu$

The Poisson is a discrete distribution. It describes the probability of getting exactly  $n$  events in a given time, if these occur independently and randomly at constant rate (in that given time)  $\mu$

BEWARE

Other fun facts:

- it is a limiting case of the Binomial [ $P(n) = \binom{N}{n} p^n (1-p)^{N-n}$ ] for  $p \rightarrow 0$ , in the limit of large  $N$
- it converges to the Normal for large  $\mu$

# The Compound Poisson Distribution

- Less known is the **compound Poisson distribution**, which **describes the sum of N Poisson variables, all of mean  $\mu$ , when N is also a Poisson variable of mean  $\lambda$ :**

$$P(n; \mu, \lambda) = \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^n e^{-N\mu}}{n!} \frac{\lambda^N e^{-\lambda}}{N!} \right]$$

- Obviously the expectation value is  $E(\mathbf{n}) = \lambda\mu$
- The variance is  $V(\mathbf{n}) = \lambda\mu(1+\mu)$
- One seldom has to do with this distribution in practice. Yet it is necessary for a physicist to know it exists, and to recognize it is different from the simple Poisson distribution.

**Why ? Should you really care ?**

Let me ask before we continue: **how many of you knew about the existence of the compound Poisson distribution?**

# An Example of the Compound Poisson: Bootstrapping

- Bootstrapping: creating new samples from a dataset by fishing events at random, with replacement
- The idea of bootstrapping is that **inference on properties of a unknown distribution from which we have a sample of data can be obtained by inference on resampled sets**
- Example (from Wikipedia): assume we are interested in the average height of people worldwide. We only measure the heights of  $N$  (say 10000) individuals. From that single sample, only one estimate of the mean can be obtained. In order to reason about the population, we need some sense of the variability of the mean that we have computed.
  - By resampling with replacement we may construct many (say 1000) sets of size  $N$ , and study the distribution of the mean (or of the variance, or whatever statistic we are interested in)



# The PDF of Bootstrapped Sets


Most common situation: you have a histogram of events in the original dataset, such that each bin content has **Poisson** properties.

What are the statistical properties of the bin entries in the bootstrapped histograms ?

- Quantitatively: **if the expectation value of a bin's content is  $\mu$ , what is the associated variance  $\sigma^2$  ?**

As you might have correctly guessed, the variance in the number of entries in each bin is **larger than  $\sigma^2 = \mu$**  as the Poisson distribution would imply.

$$P(n; \mu, \lambda) = \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^n e^{-N\mu}}{n!} \frac{\lambda^N e^{-\lambda}}{N!} \right]$$

$$V(n) = \lambda\mu(1+\mu)$$


**EXERCISE: write a program that tests this.**

# Bootstrap\_variance.C

```
void Bootstrap_variance (double Ndata=10000, int Nrep=100, double
fracBoot=1.0) {
    // Ndata = Expectation value of number of events in original histogram
    // Nrep = Number of Bootstrap replicas drawn
    // fracBoot = fraction of Ndata drawn in Bootstrapped sets
    double NdataB=Ndata*fracBoot;
    const int Nbins = 100; // We fix the number of bins to 100
    if (Ndata>100000) {
        cout << "Too much data per sample, reduce to <100000. Exiting.." <<
endl;
        return;
    }
    // Repeat many times to get average of variance over replicas
    double sumvar =0;
    double Average_content=0;
    for (int i=0; i<Nrep; i++) {
        double data[100000];
        int thisdata = gRandom->Poisson(Ndata);
        for (int j=0; j<thisdata; j++) { // Generate histogram data
            double x = gRandom->Uniform(0.,(double)Nbins);
            data[j]= x;
        }
        // Create Bootstrap sample
        double Bdata[100000];
        thisdata = gRandom->Poisson(NdataB);
        Average_content+=thisdata;
```

```
for (int j=0; j<thisdata; j++) {
    int index=(int)gRandom->Uniform(0.,Ndata);
    if (index==Ndata) index=Ndata-1;
    Bdata[j]=data[index];
}
// Study statistical properties of Bdata in each bin by computing
the bin-by-bin variances
    int Contents[Nbins];
    double sum=0;
    double sum2=0;
    for (int k=0; k<Nbins; k++) {
        Contents[k]=0;
        for (int j=0; j<thisdata; j++) {
            if (Bdata[j]>=(double)k && Bdata[j]<(double)k+1.) {
                Contents[k]++;
            }
        }
        sum+= Contents[k];
        sum2+= Contents[k]*Contents[k];
    }
    double var = sum2/Nbins-pow(sum/Nbins,2);
    sumvar +=var;
}
Average_content = Average_content/Nrep;
double Average_variance = sumvar/Nrep;
cout << endl;
cout << " Average variance in bootstrapped sets is " <<
Average_variance << endl;
cout << " Expectation for compound Poisson is " <<
NdataB/Nbins*(1+Average_content/Ndata) << endl;
cout << " Variance for a Poisson distribution is " <<
NdataB/Nbins << endl;
}
```

# Example: 100 Bins With $\mu=20$

- We generate bootstrapped replicas of 2000 events each sampled from the same data, with 100 bins
  - We can then measure the variance within each bin and compare to Poisson and Compound Poisson expectations
  - We vary the fraction of resampling from 0.2 to 0.8 to see the effect on the actual variance of multiple entries in the same bin

```
root [5] Bootstrap_variance(10000,1000,0.2);  
  
Aver. variance of bin contents in bootstrapped sets is 23.7969  
Expectation for compound Poisson is 23.9975  
Variance for a Poisson distribution is 20  
  
root [6] Bootstrap_variance(5000,1000,0.4);  
  
Aver. variance of bin contents in bootstrapped sets is 27.7172  
Expectation for compound Poisson is 28  
Variance for a Poisson distribution is 20  
  
root [7] Bootstrap_variance(2500,1000,0.8);  
  
Aver. variance of bin contents in bootstrapped sets is 35.6591  
Expectation for compound Poisson is 35.9953  
Variance for a Poisson distribution is 20
```

# Take-Away Bits

- 1) Bootstrapping is powerful, but be careful with the handling of resulting uncertainty estimates!
- 2) The compound Poisson is more common than you'd think
- 3) Knowing the properties of the PDF you sample from is crucial
  - this is a common theme of these lessons

C. B. A. McCusker and I. Cairns

Cornell-Sydney University Astronomy Center, Physics Department, The University of Sydney, Sydney, Australia

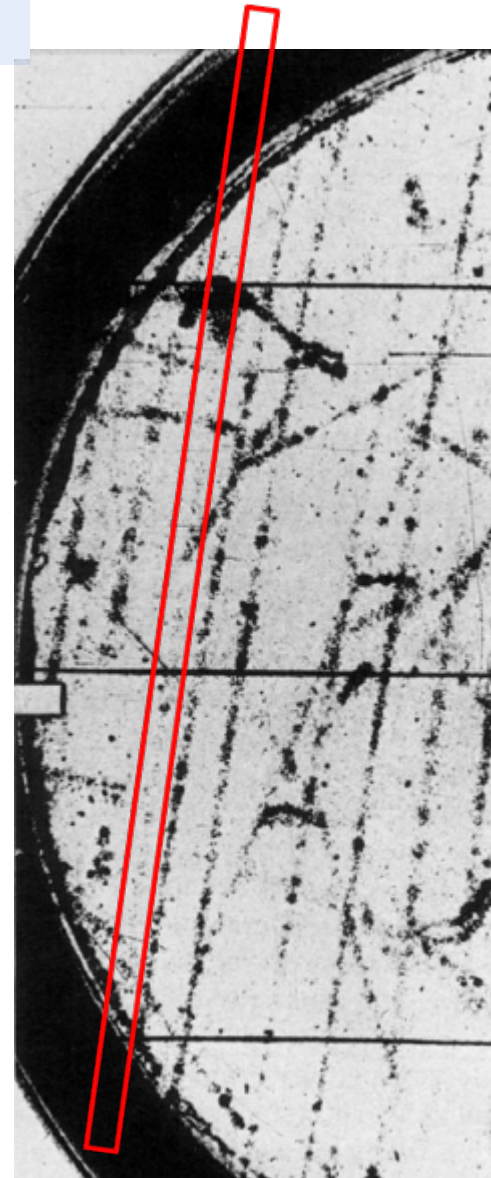
(Received 3 September 1969)

In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

In 1968 the gentlemen named in the above clip observed four tracks in a Wilson chamber whose apparent ionization was compatible with the one expected for particles of charge  $2/3e$ . Successively, they published a paper where they showed a track which could not be anything but a fractionary charge particle! In fact, it produced **110 counted droplets** per unit path length against an expectation of **229** (from the **55,000 observed tracks**).

What is the probability to observe such a phenomenon ?  
We compute it in the following slide.

Note that if you are strong in nuclear physics and thermodynamics, **you may know that a scattering interaction produces on average about four droplets**. The scattering and the droplet formation are **independent Poisson processes**. However, if your knowledge of Statistics is poor, this observation does not allow you to reach the right conclusion. **What is the difference, after all, between a Poisson process and the combination of two ?**



# Significance of the Observation

Case A: **single Poisson process**, with  $\mu=229$ :

$$P(n \leq 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \approx 1.6 \times 10^{-18}$$

Since they observed 55,000 tracks, seeing at least one track with  $P = 1.6 \times 10^{-18}$  has a chance of occurring of  $1-(1-P)^{55000}$ , or about  **$10^{-13}$**

Case B: **compound Poisson process**, with  $\lambda\mu=229$ ,  $\mu=4$ :

One should rather compute

$$P'(n \leq 110) = \sum_{i=0}^{110} \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^i e^{-N\mu}}{i!} \frac{\lambda^N e^{-\lambda}}{N!} \right] \approx 4.7 \times 10^{-5}$$

from which one gets that the probability of seeing at least one such track is rather  $1-(1-P')^{55000}$ , or **92.5%. Ooops!**

**Bottomline:**

**You may know your detector and the underlying physics as well as you know your \*\*\*, but only your knowledge of basic Statistics prevents you from being fooled !**

# Going Bayesian: The Jeffreys-Lindley Paradox

So what happens if one tries to move to Bayesian territory ?

Consider a null hypothesis,  $H_0$ , on which we base a strong belief. In physics we do believe in our “point null” – a theory valid for a specific value  $\theta_0$  of a parameter  $\theta$  (say the photon mass being 0); in other sciences a true “point null” hardly exists

Comparing a point null  $\theta=\theta_0$  to an alternative which has a continuous support for  $\theta$ , we need to suitably encode this in a prior belief. Bayesians use a “probability mass” at  $\theta=\theta_0$  for  $H_0$ .

The use of probability masses to encode priors for a **simple-vs-composite test** throws a monkey wrench in the Bayesian paradigm, as it can be proven that **no matter how large and precise is the data, Bayesian inference strongly depends on the scale over which the prior is non-null** – that is, on the **prior belief** of the experimenter.

The **Jeffreys-Lindley paradox**[16] arises as frequentists and Bayesians draw **opposite conclusions on some data when comparing a point null to a composite alternative**. This fact bears relevance to the kind of tests we are discussing, so let us give it a look.

# The Paradox

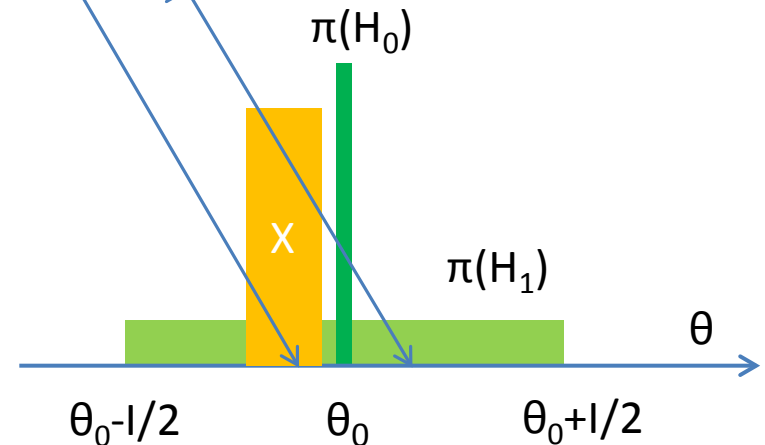
Take  $X_1 \dots X_n$  i.i.d. as  $X_i | \theta \sim N(\theta, \sigma^2)$ , and a prior belief on  $\theta$  constituted by a mixture of a **point mass  $p$  at  $\theta_0$**  and  **$(1-p)$  uniformly distributed in  $[\theta_0 - I/2, \theta_0 + I/2]$** .

In classical hypothesis testing the “critical values” of the sample mean delimiting the rejection region of  $H_0: \theta = \theta_0$  in favor of  $H_1: \theta \neq \theta_0$  at significance level  $\alpha$  are

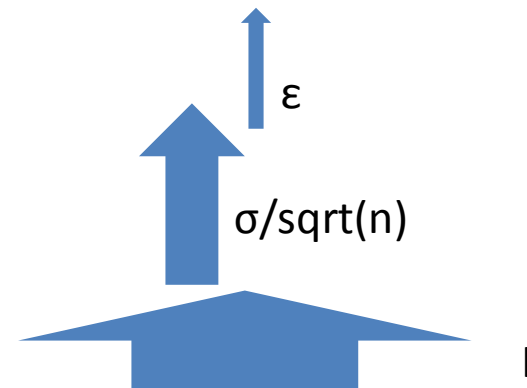
$$\bar{X} = \theta_0 \pm (\sigma / \sqrt{n}) z_{\alpha/2}$$

where  $z_{\alpha/2}$  is the significance corresponding to test size  $\alpha$  for a two-tailed normal distribution

The **paradox** is that **the posterior probability that  $H_0$  is true, conditional on seeing data in the critical region** (i.e. ones which exclude  $H_0$  in a classical  $\alpha$ -sized test) **approaches 1 (not  $\alpha$ , NB!) as the sample size becomes arbitrarily large.**



As evidenced by **R. Cousins[17]**, the paradox arises if there are three independent scales in the problem,  $\epsilon \ll \sigma / \sqrt{n} \ll I$ , i.e. the width of the point mass, the measurement uncertainty, and the scale  $I$  of the prior for the alternative hypothesis



**Common situation in HEP!!**



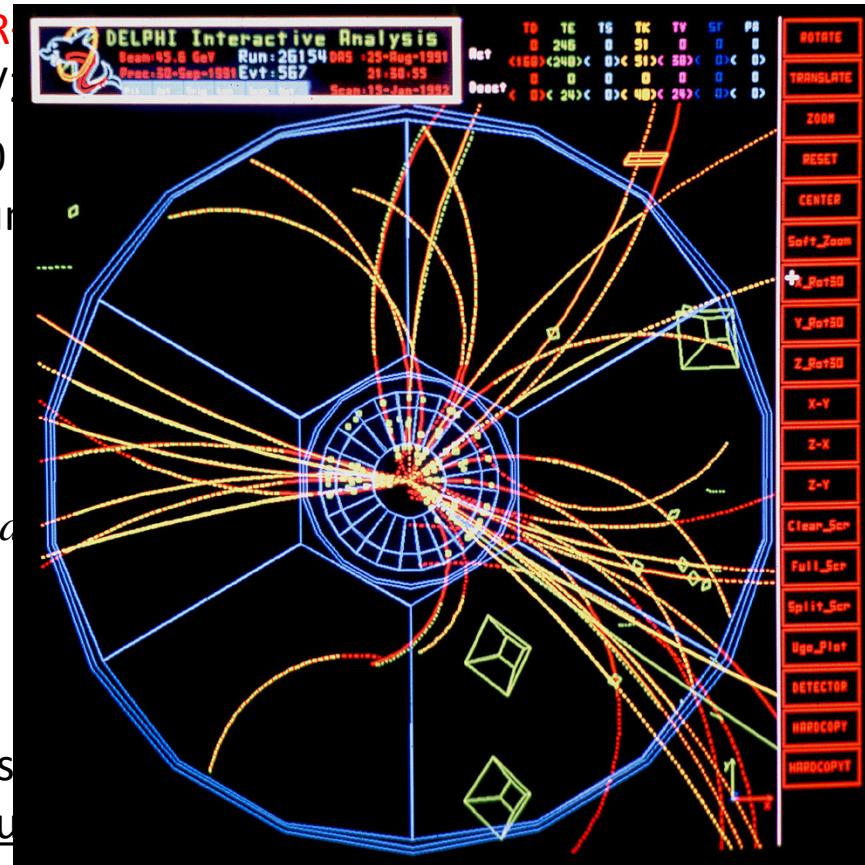
# JLP Example: Charge Bias of a Tracker

Imagine you want to investigate whether your detector has a bias in reconstructing positive versus negative curvature, say at a lepton collider ( $e^+e^-$ ). You take a unbiased set of collisions, and count positives and negatives in a set of  $n=1,000,000$ .

- You get  $n^+=498,800$ ,  $n^-=501,200$ . You want to test the hypothesis that the fraction of positive tracks, say, is  $R=0.5$  with a size  $\alpha=0.05$ .
- Bayesians will **need a prior  $\pi(R)$** : a typical choice would be to **assign equal probability to the chance that  $R=0.5$  and to it being different ( $R \neq 0.5$ )** and a uniform distribution of the remaining  $p=1/2$ .
- We are in high-statistics regime and away from 0 and 1, so we can use the normal distribution **for the Binomial**. The probability to observe a number of positives  $x$  is written, with  $x=n^+/n$ , as  $N(x, \sigma)$  with  $\sigma^2=x(1-x)/n$ . The **posterior probability** that  $R=0.5$  is then

$$P(R = \frac{1}{2} | x, n) \approx \frac{1}{2} \frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} / \left[ \frac{1}{2} \frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} + \frac{1}{2} \int_0^1 \frac{e^{-\frac{(x-R)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} dR \right]$$

from which a Bayesian concludes that there is a bias and actually the data strongly supports the null hypothesis



# JLP Charge Bias: Frequentist Solution

Frequentists calculate how often a result “**at least as extreme**” as the one observed arises by chance, if the underlying distribution is  $N(R, \sigma)$  with  $R=1/2$  and  $\sigma^2=x(1-x)/n$

One then has

$$P(x \leq 0.4988 \mid R = \frac{1}{2}) = \int_0^{0.4988} \frac{e^{-\frac{(t-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} dt = 0.008197$$
$$\Rightarrow P'(x \mid R = \frac{1}{2}) = 2 * P = 0.01639$$

(we multiplied by two since we would be just as surprised to observe an excess of positives as a deficit).

From this, frequentists conclude that the tracker is biased, since there is a less-than 5% probability,  $P' < \alpha$ , that a result as the one observed could arise by chance!

A frequentist thus draws the **opposite conclusion** of a Bayesian from the same (large body of) data !

# Notes on the JL Paradox

- The paradox has been used by Bayesians to criticize the way inference is drawn by frequentists:
  - Jeffreys: “*What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*” [18]
- Still, the Bayesian approach offers no effective substitute to the p-value
  - **Bayes factors**, which describe by how much prior odds are modified by the data, do not factor out the subjectivity of the prior when the JLP applies: even asymptotically, they retain a dependence on the scale of the prior of  $H_1$ .
- In JLP debates, Bayesians have argued that “the precise null” is never true.
  - However, we do believe our point nulls in HEP and astro-HEP!!  
(mass of photon==0; total electric charge of a system==0)

There is a large body of literature on the subject. The issue is an active research topic and is **not resolved**.

→ The trouble of picking  $\alpha$  in classical hypothesis testing is not automatically solved by moving to Bayesian territory.

# The Neyman-Pearson Lemma

- For **simple** hypothesis testing there is a recipe to find the **most powerful test**. It is based on the likelihood ratio.
- Take data  $X = \{X_1 \dots X_N\}$  and two hypotheses depending on the values of a discrete parameter:  $H_0 = \{\theta = \theta_0\}$  vs  $H_1 = \{\theta = \theta_1\}$ .  
If we write the expressions of size  $\alpha$  and power  $1 - \beta$  we have

$$\int_{w_\alpha} f_N(X | \theta_0) dX = \alpha$$

$$1 - \beta = \int_{w_\alpha} f_N(X | \theta_1) dX$$

The problem is then to find the **critical region**  $w_\alpha$  such that  $1 - \beta$  is maximized, given  $\alpha$ .  
We rewrite the expression for power as

$$1 - \beta = \int_{w_\alpha} \frac{f_N(X | \theta_1)}{f_N(X | \theta_0)} f_N(X | \theta_0) dX$$

which is an expectation value:

$$= E_{w_\alpha} \left[ \frac{f_N(X | \theta_1)}{f_N(X | \theta_0)} \mid \theta = \theta_0 \right]$$

This is maximized if we accept in  $w_\alpha$  all the values for which  $l_N(X, \theta_0, \theta_1) = \frac{f_N(X | \theta_1)}{f_N(X | \theta_0)} \geq c_\alpha$

So one chooses  $H_1$  if  $l_N(X, \theta_0, \theta_1) > c_\alpha$   
and  $H_0$  if instead  $l_N(X, \theta_0, \theta_1) \leq c_\alpha$

In order for this to work, hypotheses must be **simple**. The test above is called **Neyman-Pearson test**, and a test with such properties is the **most powerful**.

# Goodness-of-Fit Tests

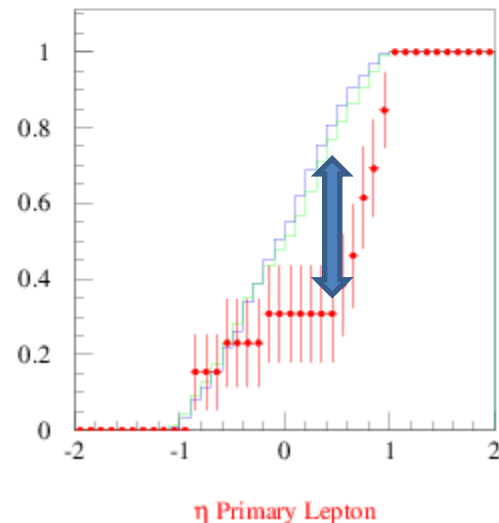
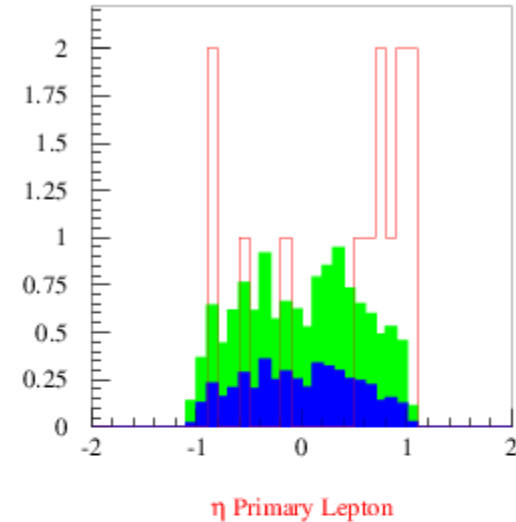
- If  $H_0$  is specified but the alternative  $H_1$  is not, then only the Type I error rate  $\alpha$  can be calculated, since **the Type II error rate  $\beta$  depends on having specified a particular  $H_1$ .**  
In this case the test is called a **test for *goodness-of-fit (to  $H_0$ )***.
- Hence the question “**Which g.o.f. test is best?**” is ill-posed, since the **power** depends on the alternative hypothesis, which is not given.
- In spite of the popularity of tests which give a statistic which one may directly connect with the size  $\alpha$  (in particular  $\chi^2$  and Kolomogorov tests), their ability to discriminate against variations with respect to  $H_0$  may be poor, i.e. they may have small power  $(1-\beta)$  against relevant alternative hypotheses
  - $\chi^2$  throws away information (sign, ordering)
  - Kolmogorov –Smirnov test only sensitive to biases, not to shape variations, and has terrible performance on tails (we'll see it in a minute)

# The Kolmogorov Test: an Example

- CDF, circa 2000: 13 weird events identified in a subset of sample used to extract top quark cross section
  - contain a “superjet”: a jet with a b-quark tag also containing a soft-lepton tag
  - expected 4.4 +/-0.6 events from background sources
  - $P(\geq 13 | 4.4 \pm 0.6) = 0.001$
  - Kinematic characteristics found in stark disagreement with expectation from SM sources
- Have no alternative model to compare → try a Goodness-of-Fit test
- Kolmogorov-Smirnov test: compare cumulative distributions of data and model  $f(x)$ ; find largest difference

$$d_{KS} = \text{Max}_{x \in [a,b]} \left| \int_a^x \text{data}(t) dt - \int_a^x f(t) dt \right|$$

Value of  $d_{KS}$  can then be used to extract a p-value, given data size.



# On Tail Probabilities: Choosing the Region of Interest

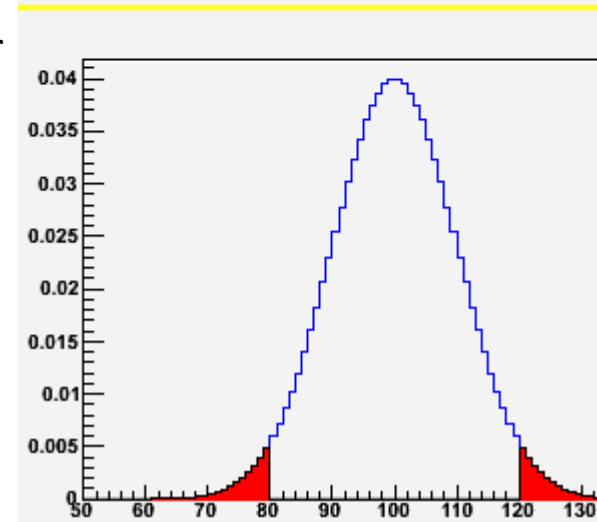
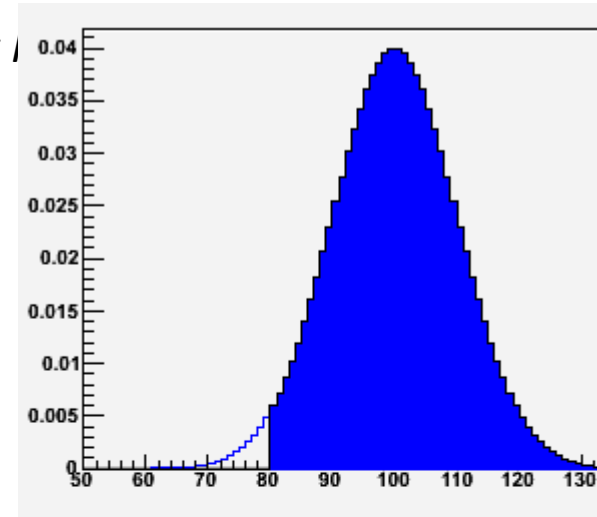
- Feynman's example:

*"Upon walking here this morning, the strangest thing ever happened to me. A car passed by, and I could read the plate: **JKZ 0533**. How weird is that ??! The probability that I saw such a combination of letters and numbers (assuming they are all used in this country) is one in  $10000 \cdot 26^3$ , or one in eighty-eight millions!"*

Correct... The paradox arises from not having defined beforehand the region of interest!

- A more common one: you have a counting experiment where background is predicted to be 100 events. You observe 80 events. How rare is that ?

- **Ill-posed question** ! Depends, to say the least, on whether you are interested only in excesses or in absolute departures!
- In the first case the **region of interest is  $N \geq x$** , which, for  $x=80$ , corresponds to a fractional area  $p = 0.977$ .
- In the second case, the **region of interest is  $|N-100| \geq |x-100|$**  which for  $x=80$  has an integral  $p = 0.0455$ .
- And one might imagine other ways to answer – a no-brainer being  $p = e^{-100} 100^{80}/80!$



# Intermezzo: Combination of p-Values

- Suppose you have several p-values, derived from different, independent tests. You may ask yourself several questions with them.
  - What is the probability that the smallest of them is as small as the one I got ?
  - What is the probability that the largest one is as small as the largest I observed ?
  - What is the probability that the product is as small as the one I can compute with these N values ?
- Please note! Your inference on the data at hand **strongly** depends on what test you perform, for a given set of data. In other words, **you cannot choose which test to run only upon seeing the data...**
- Suppose anyway you believe that each p-value tells something about the null hypothesis you are testing, so you do not want to discard any of them. Then one possibility is to **use the product of the N values**. The formula providing the cumulative distribution of the density of  $x = \prod x_i$  can be derived by induction (see [Roe 1992], p.129) and is

$$F_N(x) = x \sum_{j=0}^{N-1} \frac{1}{j!} |\log^j(x)|$$

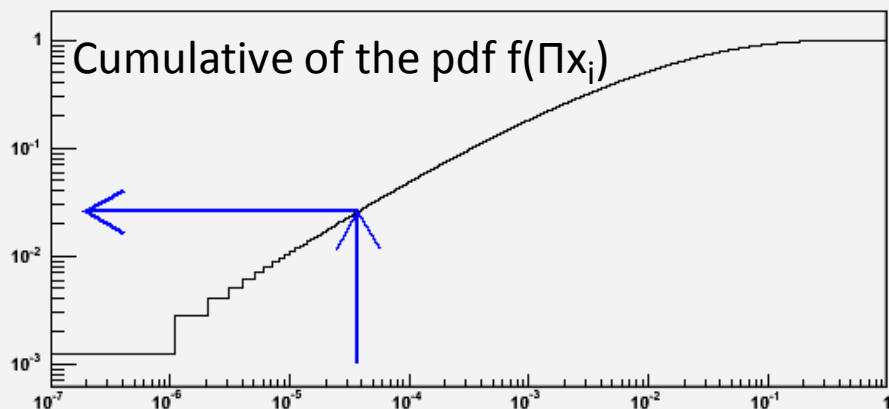
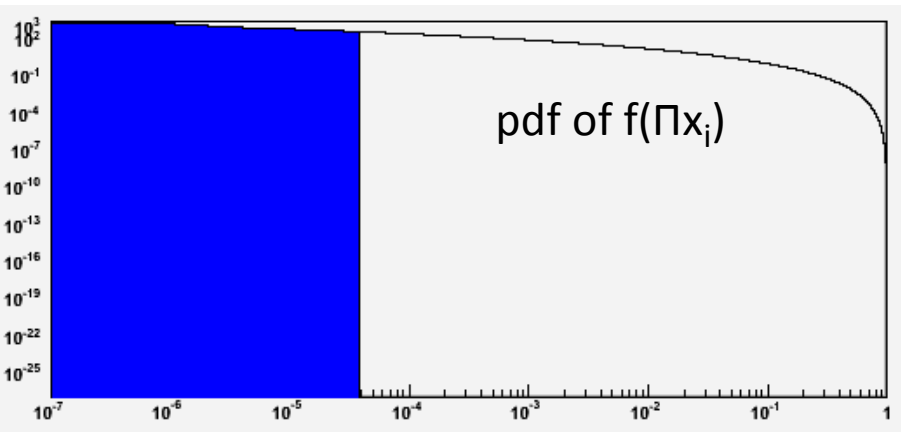
This accounts for the speed with which the product of N numbers in [0,1] tends to zero as N grows.



# Some Examples on the Product of Probabilities

To start let us take five *really uniformly* distributed p-values,  $x_1=0.1$ ,  $x_2=0.3$ ,  $x_3=0.5$ ,  $x_4=0.7$ ,  $x_5=0.9$ . Their product is 0.00945, and with the formula just seen we get  $P(0.00945)=0.5017$ . As expected.

- And what if instead  $x_1=0.00001$ ,  $x_2=0.3$ ,  $x_3=0.5$ ,  $x_4=0.7$ ,  $x_5=0.9$ ? The result is  $P(9.45 \cdot 10^{-7})=0.00123$ , which is rather large: one might think that the chance of getting one in five numbers as small as  $10^{-5}$  must occur only a few times in  $10^5$ . But **we are testing the product**, not the smallest of the five numbers !



And if now we let  $x_1=0.05$ ,  $x_2=0.10$ ,  $x_3=0.15$ ,  $x_4=0.20$ ,  $x_5=0.25$ , the test for the product yields  $P(3.75 \cdot 10^{-5})=0.0258$  (see picture on the right).

Also not a compelling rejection of the null... Compare with what you would get if you had asked “*what is the chance that five numbers are all smaller than 0.25 ?*”, whose answer is  $(0.25)^5=0.00098$ . This demonstrates that **the a-posteriori choice of the test is to be avoided !**

# Global P From Set of p-Values

- Authors of CDF “superjet” analysis tested a “complete set” of kinematical quantities; then computed global P of set of KS p-values using formula of combining p-values (assumed sampled from a Uniform distribution):

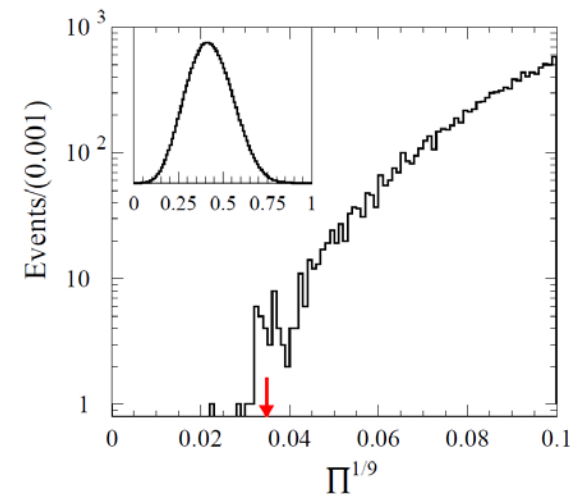
$$F_N(x) = x \sum_{j=0}^{N-1} \frac{1}{j!} |\log^j(x)|$$

→ **>6-sigma result!**

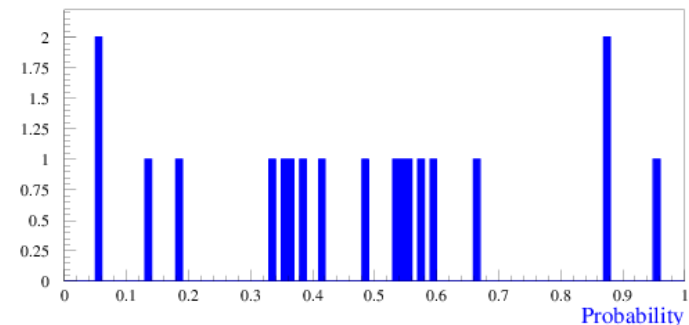
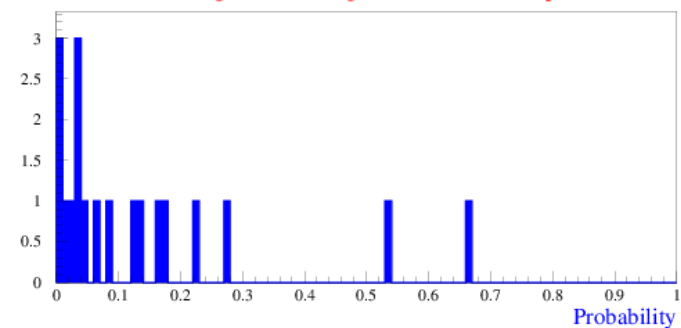
... But in absence of an alternative model

(really hard to cook given the weird

Variable	Events with a superjet <i>P</i> (%)	Complementary sample <i>P</i> (%)
$E_T^l$	2.6	70.9
$\eta^l$	0.10	72.7
$E_T^{suj}$	11.1	43.0
$\eta^{suj}$	15.2	73.4
$E_T^b$	6.7	8.6
$\eta^b$	6.8	80.0
$E_T^{l+b+suj}$	2.5	18.8
$y^{l+b+suj}$	13.8	7.8
$\delta\phi^{l,b+suj}$	1.0	77.9



Kolmogorov test - Signal and Control samples



# GoF Tests with Max Likelihood

- The maximum likelihood is a powerful method to estimate parameters, but **no measure of GoF is given**, because the expected value of L at maximum is not known
- The distribution of  $L_{\max}$  can be studied with toy MC  $\rightarrow$  one derives a p-value that a value as small as the one observed in the data arises, under the given assumptions
- Alternatively, one can bin the data, obtaining estimated mean values of entries per bin **from the ML fit**:

$$\hat{v}_i = n_{tot} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \hat{\theta}) dx$$

Then one can derive a  $\chi^2_L$  statistic using the ratio of likelihoods

and computing

$$\chi^2 = -2 \log \lambda$$

$$\lambda = \frac{L(n | \hat{v})}{L(n | n)}$$

since in this case the latter **follows a  $\chi^2$  distribution**.

The quantity  $\lambda(v) = L(n | v) / L(n | n)$  differs from the likelihood function by a normalization factor, and can thus be used for both parameter estimation and Goodness of Fit.

# Systematic Uncertainties

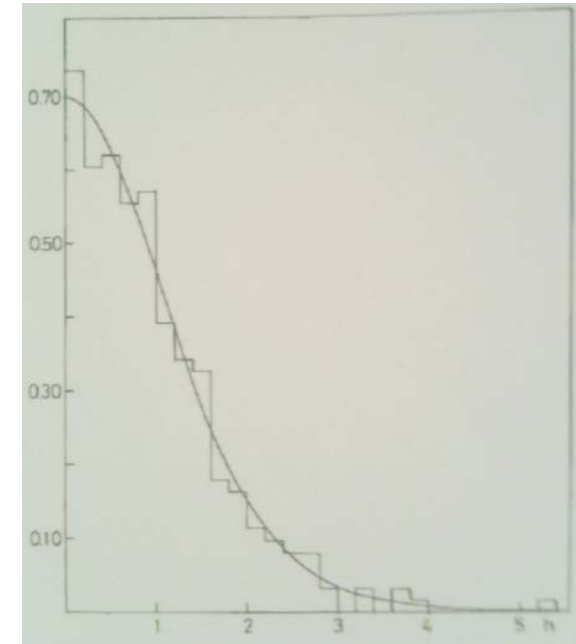
# A Study of Residuals

A study of the residuals of particle properties in the RPP in 1975 revealed that they were **not Gaussian**. Matts Roos et al. [20] considered residuals in kaon and hyperon mean life and mass measurements, and concluded that these **seem to all have a similar shape, well described by a Student distribution  $S_{10}(h/1.11)$** :

$$S_{10}\left(\frac{x}{1.11}\right) = \frac{315}{256\sqrt{10}} \left(1 + \frac{x^2}{12.1}\right)^{-5.5}$$

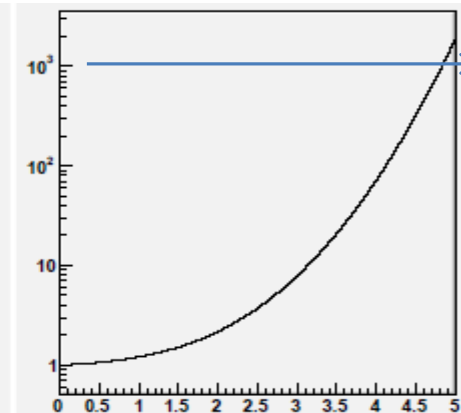
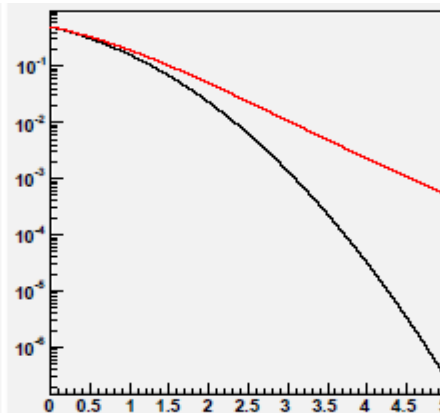
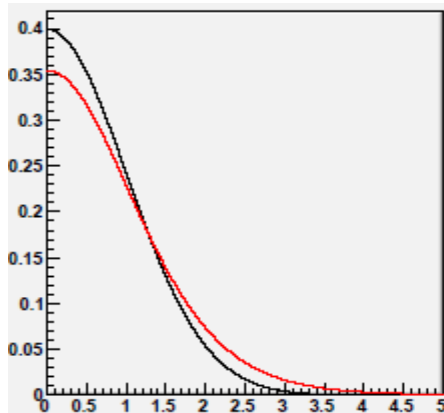
Of course, one cannot extrapolate to 5-sigma the behaviour observed by Roos and collaborators in the bulk of the distribution; however, one may consider this as evidence that **the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component**

The distribution of residuals of 306 measurements in [20]



Black: a unit Gaussian;  
red: the  $S_{10}(x/1.11)$  function

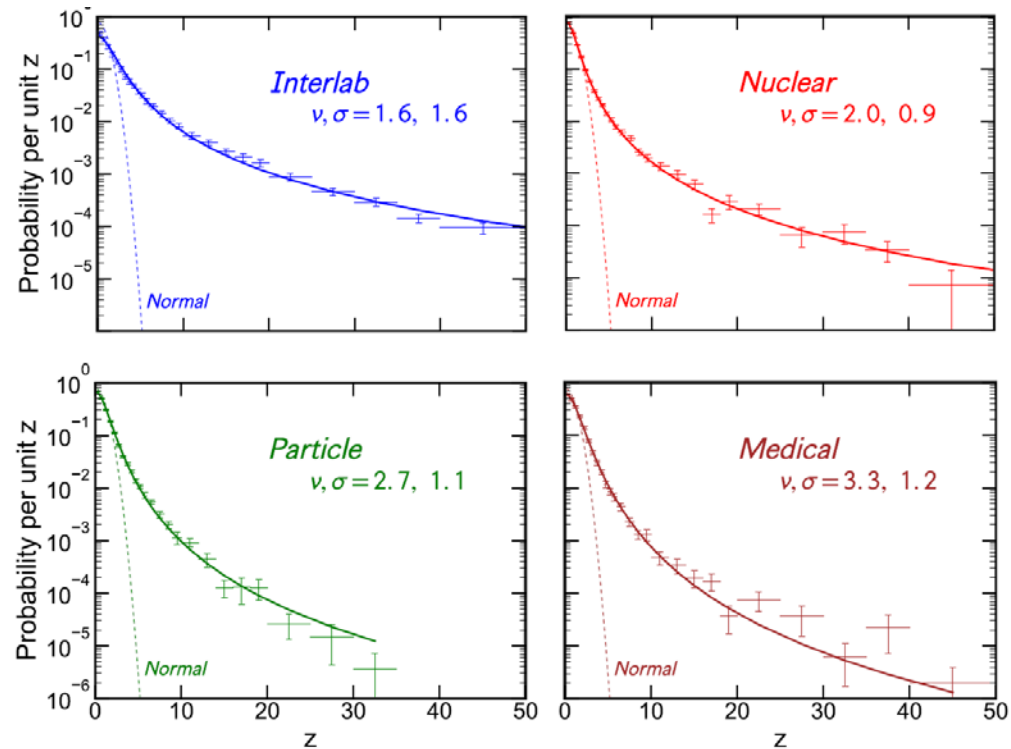
Left: 1-integral distributions of the two functions.  
Right: ratio of the 1-integral values as a function of  $z$



x1000!

# A Bigger, Meaner Study of Residuals

- David Bailey (U. Toronto) recently published an **article[15]** where use of large datasets is made (all of RPP, Cochrane medical and health database, Table of Radionuclides)
- 41,000 measurements of 3200 quantities studied
- The methodology is similar to that of Roos et al., but some shortcuts are made, and data input automation prevents more vetting (e.g. correlations not properly accounted for)



**Results are quite striking - we seem to have ubiquitous Student-t distributions in our Z values, with large tails – almost Cauchy-like.**

# Treatment of Systematic Uncertainties

Statisticians call these *nuisance parameters*

Any measurement is affected by them: the turning of an observation into a measurement requires **assumptions about parameters** and other quantities whose exact value is not perfectly known.

These parameters are typically correlated with the quantity being measured

→ their uncertainty affects the main measurement

E.g. going from an event count to a cross section requires knowing a number of additional inputs:  $N_b$ ,  $L$ ,  $\epsilon_{\text{sel}}$ ,  $\epsilon_{\text{trig}}$  ...

**The uncertainty of each of these has to be accounted for**

**Error propagation is the standard tool, but typically analytical solutions are inapplicable.**

# The Nuisance of Dealing with Nuisances

- Inclusion of effect of nuisances in interval estimation and hypothesis testing introduces complications. Each method has a recipe, but not universal nor always applicable
  - **Bayesian treatment:** one constructs the multi-dimensional prior pdf  $p(\theta)\prod_i p(\lambda_i)$  including all the parameters  $\lambda_i$ , multiplies by  $p(X_0 | \theta, \lambda)$ , and integrates all of the nuisances out, remaining with  $p(\theta | X_0)$
  - **Classical frequentist treatment:** scan the space of nuisance parameters; for each point do Neyman construction, obtaining multi-dimensional confidence region; project on parameter of interest
  - **Likelihood ratio:** for each value of the parameter of interest  $\theta^*$ , one finds the value of nuisances that globally maximizes the likelihood, and the corresponding  $L(\theta^*)$ . The set of such likelihoods is called the **profile likelihood**.



# Issues with the Three Methods

- Each “method” has problems:
  - Bayesian techniques: involve multi-Dimensional priors
  - Classical intervals: afflicted by overcoverage issues and intractability;
  - Likelihood intervals: usually suffer from undercoverage

We will not discuss them here further, but note that this is a topic at the forefront of research, for which no general recipe is valid.

- Often used are “hybrid” methods for integrating nuisance parameters out
  - for instance, treat nuisance parameters in a Bayesian way while treating the parameter of interest in a frequentist way, or “profile away” the nuisance parameters and then use any method.
  - Also possible is using Bayesian techniques and then evaluate their coverage properties.

# Inclusion of Nuisances in the Model

With data  $x$ , and knowing the pdf  $P(x|\theta)$ , you want to estimate a parameter  $\theta$ . However, the model is imperfect  $\rightarrow$  you can improve it by adding nuisances that affect it:  $p(x|\theta,\lambda)$ .

The inclusion of nuisances changes the problem and decreases the power of your inference.

To reduce the impact of nuisance parameters one may constrain their values by means of control or calibration measurements that produced some other data  $y$ .

If the measurements  $y$  are statistically independent from  $x$  and are described by a model  $P(y|\lambda)$ , you can then write a joint likelihood:  $L(\theta,\lambda) = p(x|\theta,\lambda) * P(y|\lambda)$

When using Monte Carlo to simulate the experiment, be sure to include the variation of both datasets!

# The Profile Likelihood Method

The PL method is best described in connection to an **hypothesis test**.

If one wants to test a hypothesis (e.g.  $H_0: \theta=0$ ), one needs to define a **critical region** where  $H_0$  is disproven.

In presence of nuisances  $\lambda$ ,  $H_0$  must be disproven for all values of the nuisances  $\rightarrow$  one tries to define a test statistic  $q_\theta$  whose pdf  $f(q_\theta, \theta)$  that is independent on  $\lambda$ . A good approximation to this is

$$\lambda_p(\theta) = \frac{L(\theta, \hat{\nu}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

double hat: ML value of nuisance that maximizes L for the specified  $\theta$   
Denominator has absolute max of L

Using Wilks' theorem one can show that  **$-2\log(\lambda_p)$  distributes like a chisquared and is independent on the nuisance parameters**. One can thus do HT and properly define a critical region

We will see more application of this technique when we discuss the Higgs search

# Poisson Probabilities

Exercise: write a root macro that inputs expected background counts  $B$  (with no error) and observed events  $N$ , and computes the probability of observing at least  $N$ , and the corresponding number of sigma  $Z$  for a Gaussian one-tailed test.

The p-value calculation should be straightforward: just sum from 0 to  $N-1$  the values of the Poisson (computing the factorial as you go along in the cycle), and derive  $p$  as 1-sum.

Deriving the number of sigmas that  $p$  corresponds to requires the inverse error function,  $\text{ErfInverse}(x)$  as

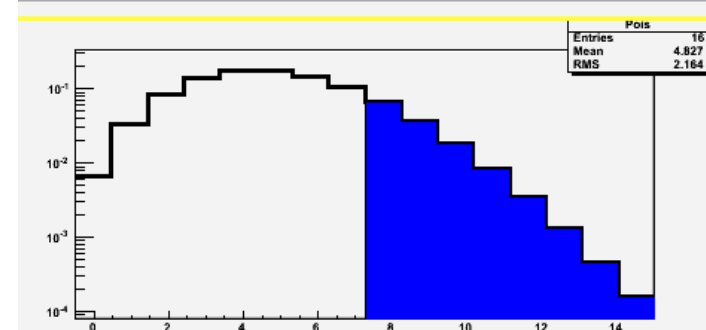
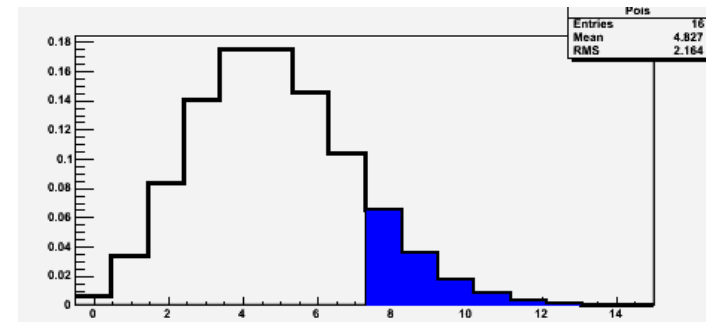
$$Z = \text{sqrt}(2) * \text{ErfInverse}(1-2p)$$

(it should be available as `TMath::ErfInverse(double)`)

You can also fill two distributions, one with the Poisson( $B$ ), the other with **only the bins  $\geq N$  filled** (and with `SetFillColor(kBlue)` or something) and plot them overlaid, to get something like the graph on the right (top: linear y scale; bottom: log y scale)

RECALL:

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



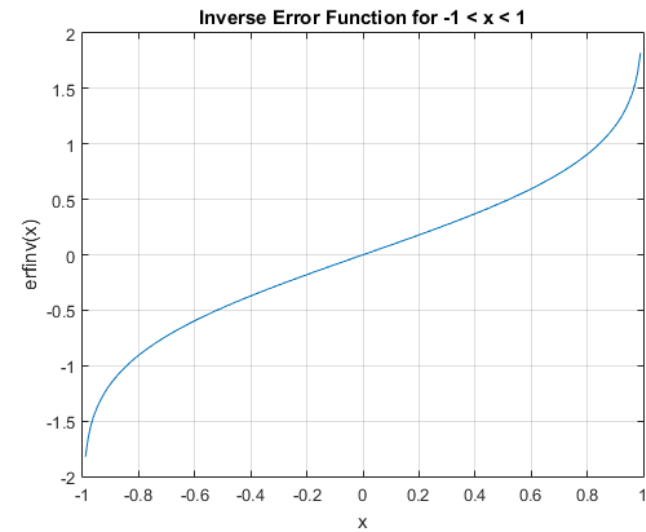
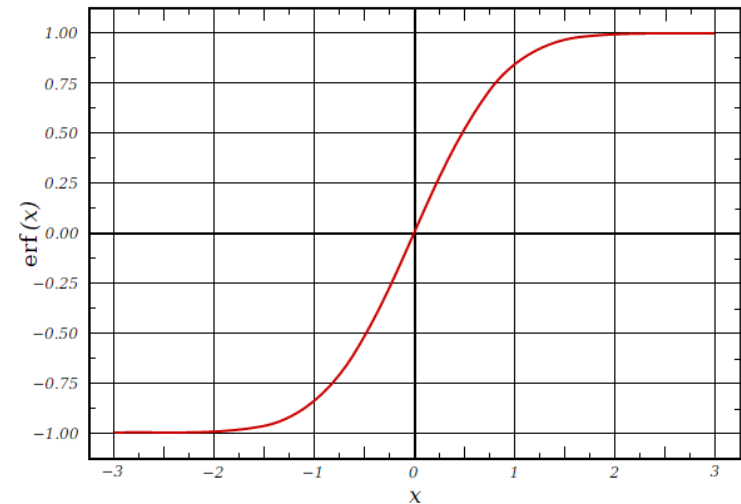
# Parenthesis – Erf and ErfInverse

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

- The error function and its inverse are useful tools in statistical calculations – you will encounter them frequently.
- The Erf can be used to obtain the integral of a Gaussian as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x}{\sqrt{2}} \right) \right]$$

The erfinverse function is used to convert alpha values into number of sigmas. We will see examples of that later on.



# One Possible Implementation

```
// Macro that computes p-value and Z-value
// of N observed vs B predicted Poisson counts
// -----
void Poisson_prob_fix (double B, double N) {

    int maxN = N*3/2; // extension of x axis
    if (N<20) maxN=2*N;
    TH1D * Pois = new TH1D ("Pois", "", maxN, -0.5, maxN,
0.5);
    TH1D * PoisGt = new TH1D ("PoisGt", "", maxN, -0.5,
maxN-0.5); // we also fill a "highlighted" portion

    double sum=0.;
    double fact=1.;
    for (int i=0; i<maxN; i++) {
        if (i>1) fact*=i; // calculate factorial
        poisson = exp(-B)*pow(B,i)/fact;
        if (i<N) sum+= poisson; // calculate 1-tail integral
        Pois->SetBinContent(i+1,poisson);
        if (i>=N) PoisGt->SetBinContent(i+1,poisson);
    }
    double P=1-sum; // get probability of >=N counts
    double Z = sqrt(2) * TMath::ErfInverse(1-2*P);
```

```
cout << "P of observing N=" << N << " or more events
if B=" << B << " : P=" << 1-sum << endl;
    cout << "This corresponds to " << Z << " sigma for a
Gaussian one-tailed test." << endl;

    Pois->SetLineWidth(3);
    PoisGt->SetFillColor(kBlue);
    TCanvas* T = new TCanvas ("T","Poisson
distribution", 500, 500);
// Plot the stuff
    T->Divide(1,2);
    T->cd(1);
    Pois->Draw();
    PoisGt->Draw("SAME");
    T->cd(2);
    T->GetPad(2)->SetLogy();
    Pois->Draw();
    PoisGt->Draw("SAME");
}
```

# Adding a Nuisance

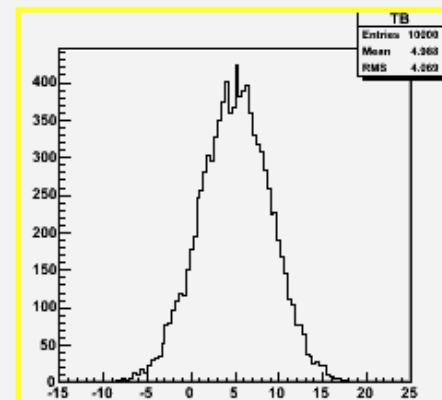
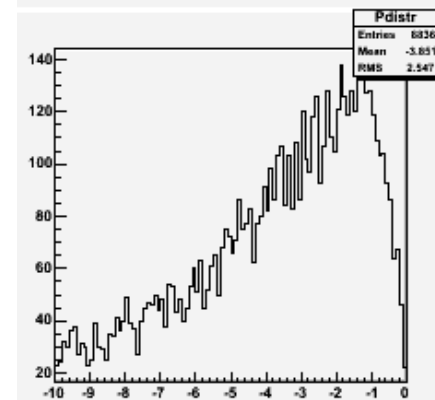
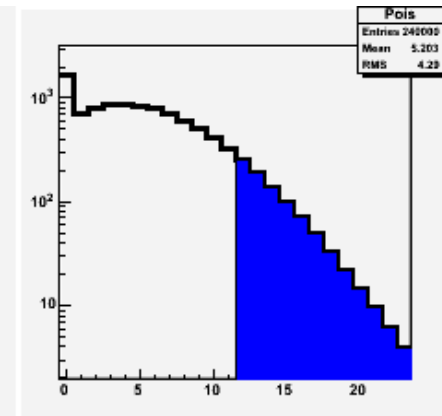
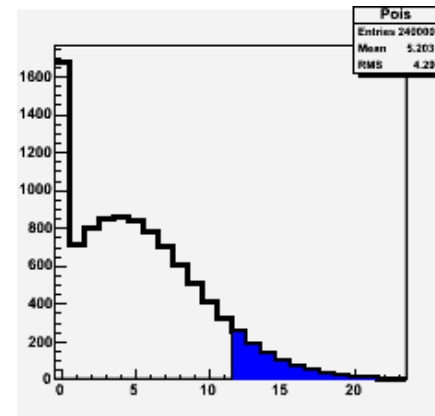
- Let us assume now that  $B'$  is not fixed, but known to some accuracy  $\sigma_B$ . We want to add that functionality to our macro. We can start with a Gaussian uncertainty.

You just have to throw a random number  $B=G(B',\sigma_B)$  to set  $B$ , and collect a large number (say 10k) of p-values as before, then take the average of them.

Upon testing it, you will discover that you need to enforce that  $B$  be non-negative.

What we do with the negative  $B$  determines the result we get, so we have to be careful, and ask ourselves what exactly do we mean when we say, e.g., “ $B=2.0\pm 1.0$ ”

Example below:  $B=5\pm 4$ ,  $N=12$



# A Possible Implementation

```
void Poisson_prob_fluct (double B, double SB, double N) {
    double Niter=10000;
    int maxN = N*3/2;
    if (N<20) maxN=2*N;
    TH1D * Pois = new TH1D ("Pois", "", maxN, -0.5, maxN-0.5);
    TH1D * PoisGt = new TH1D ("PoisGt", "", maxN, -0.5, maxN-0.5);
    // We throw a random Gaussian smearing SB to B, compute P,
    // and iterate Niter times; we then study the distribution
    // of p-values, extracting the average
    double Psum=0;
    TH1D * Pdistr = new TH1D ("Pdistr", "", 100, -10., 0.);
    TH1D * TB = new TH1D ("TB", "", 100, B-5*SB, B+5*SB);
    cout << "Start of cycle" << endl;
    for (int iter=0; iter<Niter; iter++) {
        // Extract B from G(B,SB)
        double thisB = gRandom->Gaus(B,SB);
        TB->Fill(thisB); // We keep track of the pdf of the background
        if (thisB<=0) thisB=0.; // Note this – what if we had rethrown it ?
        double sum=0.;
        double fact=1.;
        for (int i=0; i<maxN; i++) {
            if (i>1) fact*=i;
            double poisson = exp(-thisB)*pow(thisB,i)/fact;
            if (i<N) sum+= poisson;
            Pois->Fill((double)i,poisson);
            if (i>=N) PoisGt->Fill((double)i,poisson);
        }
        double thisP=1-sum;
        if (thisP>0) Pdistr->Fill(log(thisP));
        Psum+=thisP;
    }
    double P = Psum/Niter; // we use the average for our inference here
    double Z = sqrt(2) * ErfInverse(1-2*P);
    cout << "Expected P of observing N=" << N << " or more events if
        B="
        << B << "+-" << SB << " : P= " << P << endl;
    cout << "This corresponds to " << Z << " sigma for a Gaussian one-
        tailed test." << endl;

    // Plot the stuff
    Pois->SetLineWidth(3);
    PoisGt->SetFillColor(kBlue);
    TCanvas* T = new TCanvas ("T", "Poisson distribution", 500, 500);
    T->Divide(2,2);
    T->cd(1);
    Pois->DrawClone();
    PoisGt->DrawClone("SAME");
    T->cd(2);
    T->GetPad(2)->SetLogy();
    Pois->DrawClone();
    PoisGt->DrawClone("SAME");
    T->cd(3);
    Pdistr->DrawClone();
    T->cd(4);
    TB->Draw();
}
```



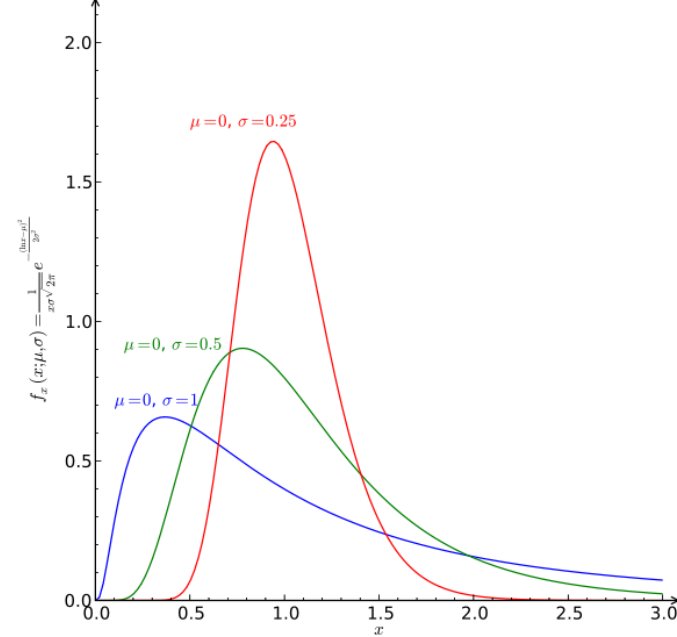
# Homework Assignment: Change to Log-Normal

Substitute the `gRandom->Gaus()` call such that you get a `B` distributed with a log-Normal pdf, **being careful to plug in the variance you really want**, and check what difference it makes.

It should be intuitive that the `LogNormal()` is the correct nuisance to use in many common situations. It correspond to saying “**I know `B` to within a factor of 2**”. Or think at a luminosity uncertainty...

This follows from the fact that while the Gaussian is the limit of the sum of many small random contributions, the limit of a product of small factors is a log-normal.

To get a logN quickly, just throw  $y = G(\mu, \sigma)$ ; then  $x = \exp(y)$  is what you need. However, note that with the ansatz “know `B` to within a certain factor”, we want the **median  $\exp(\mu)$**  to represent our central value, **not the mean  $e^{(\mu + \sigma^2/2)}$** ! So we **set  $\mu = \log(B)$** . To know what to set sigma to, we need to consider our ansatz:  $\sigma = \sigma_B / B$  corresponds to it.



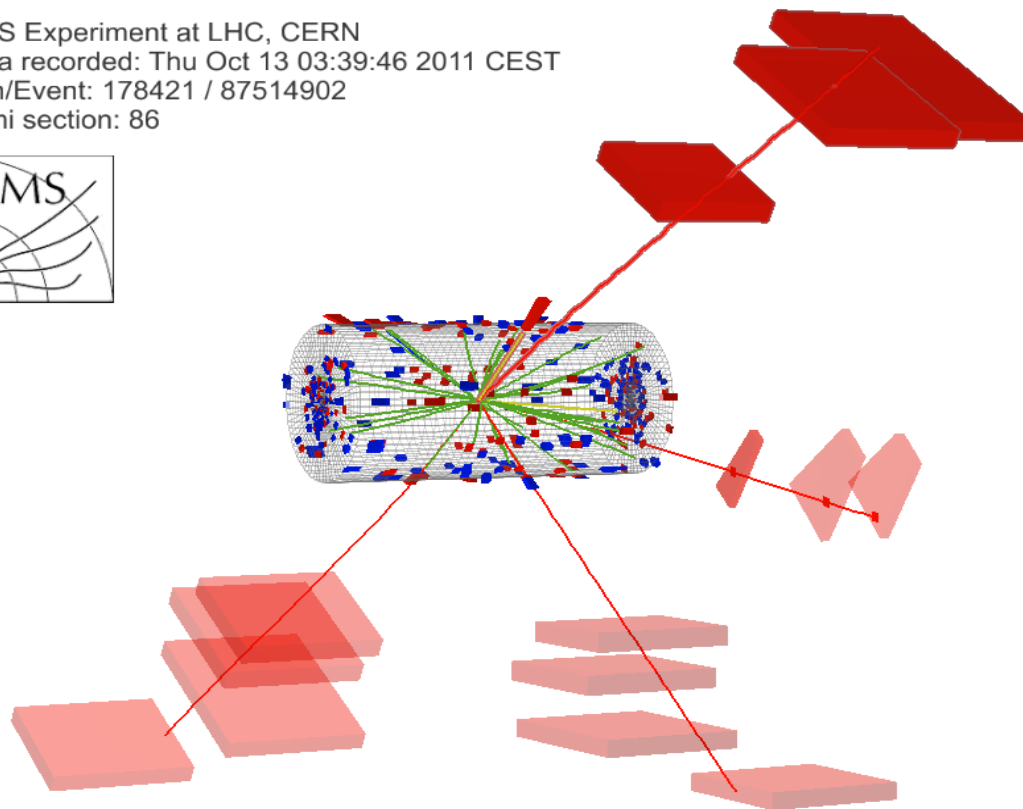
In the web area you find a version of `Poisson_prob_fluct.C` that does this

$$f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0$$

$$\begin{aligned} E[X] &= e^{\mu + \frac{1}{2}\sigma^2} \\ \text{Var}[X] &= (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \end{aligned}$$

# The Higgs Boson Search at the LHC

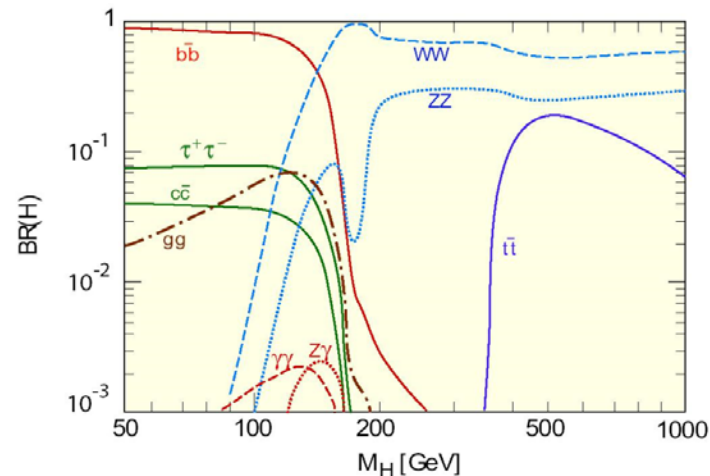
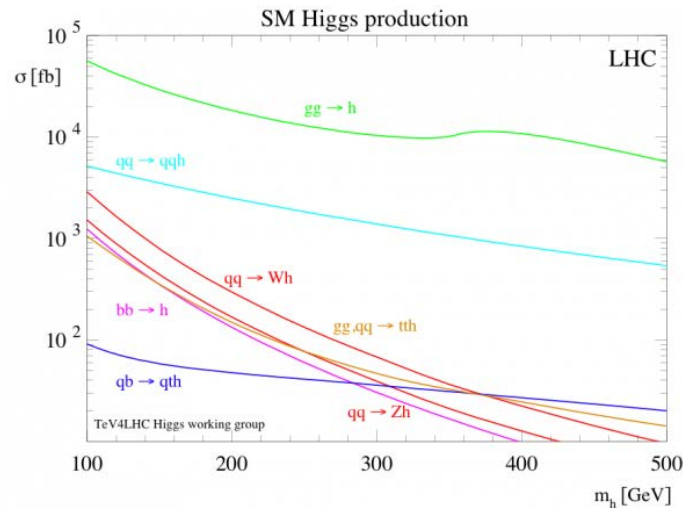
CMS Experiment at LHC, CERN  
Data recorded: Thu Oct 13 03:39:46 2011 CEST  
Run/Event: 178421 / 87514902  
Lumi section: 86



# Higgs Searches at LHC

- The Higgs boson has been sought for by ATLAS and CMS in all the main production processes and in a number of different final states, resulting from the varied production and decay modes:

- $qq \rightarrow Hqq$
- $gg \rightarrow H$
- $qq^{(\prime)} \rightarrow VH$
  
- $H \rightarrow ZZ$
- $H \rightarrow WW$
- $H \rightarrow gg$
- $H \rightarrow tt$
- $H \rightarrow bb$



- The importance of the goal brought together some of the best minds of CMS and ATLAS, to **define and refine the procedures to combine the above many different search channels, most of which have marginal sensitivity by themselves**

# Method

- The method used to set upper limits on the Higgs boson cross section uses the **CL<sub>s</sub> criterion** and the test statistic is a **profile log-likelihood ratio**. Dozens of nuisance parameters, with either 0% or 100% correlations, are considered
- Results have been produced as a combined upper limit on the “strength modifier”  $\mu = \sigma / \sigma_{SM}$ , as well as a “best fit value” for  $\mu$ , and a combined p-value of the null hypothesis. All of these are produced as a function of the unknown Higgs boson mass.
- **The technology is an advanced topic.** We can give a peek at the main points, including the construction of the CL<sub>s</sub> statistics and the treatment of nuisances, to understand the main architecture

# Nuts and Bolts of Higgs Combination

The recipe must be explained in steps. **The first one is of course the one of writing down extensively the likelihood function!**

- 1) One writes a global likelihood function, whose parameter of interest is the strength modifier  $\mu$ . If  $s$  and  $b$  denote signal and background, and  $\theta$  is a vector of systematic uncertainties, one can generically write for a single channel:

$$\mathcal{L}(\text{data} | \mu, \theta) = \text{Poisson}(\text{data} | \mu \cdot s(\theta) + b(\theta)) \cdot p(\tilde{\theta} | \theta)$$

Note that  $\theta$  has a “prior” coming from a hypothetical auxiliary measurement.

In the LHC combination of Higgs searches, nuisances are treated in a frequentist way by taking for them the likelihood which would have produced as posterior, given a flat prior, the PDF one believes the nuisance is distributed from. This differs from the Tevatron and LEP Higgs searches.

In L one may combine many different search channels where a counting experiment is performed as the product of their Poisson factors:

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-\mu s_i - b_i}$$

or from a unbinned likelihood over  $k$  events, factors such as:

$$k^{-1} \prod_i (\mu S f_s(x_i) + B f_b(x_i)) \cdot e^{-(\mu S + B)}$$

2) One then constructs a profile likelihood test statistic  $q_\mu$  as 
$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}$$

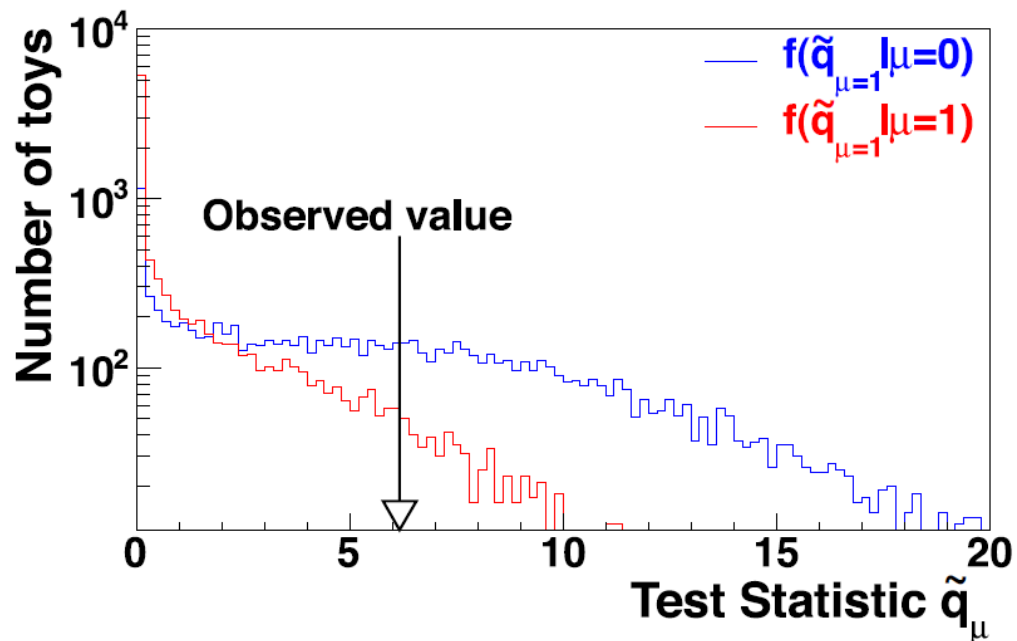
Note that the denominator has L computed with the values of  $\mu^\wedge$  and  $\theta^\wedge$  that globally maximize it, while the numerator has  $\theta = \theta^\wedge_\mu$  computed as the conditional maximum likelihood estimate, given  $\mu$ .

**A constraint is posed on the MLE  $\mu^\wedge$  to be confined in  $0 \leq \mu^\wedge \leq \mu$ :** this avoids negative solutions for the cross section, and ensures that best-fit values *above* the signal hypothesis  $\mu$  are not counted as evidence against it.

The above definition of a test statistic for  $CL_s$  in Higgs analyses differs from earlier instantiations

- LEP: no profiling of nuisances
- Tevatron:  $\mu=0$  in L at denominator

- 3) ML values  $\theta_\mu^\wedge$  for  $H_1$  and  $\theta_0^\wedge$  for  $H_0$  are then computed, given the data and  $\mu=0$  (bgr-only) and  $\mu>0$
- 4) Pseudo-data is then generated for the two hypotheses, **using the above ML estimates of the nuisance parameters.** With the data, one constructs the pdf of the test statistic given a signal of strength  $\mu$  ( $H_1$ ) and  $\mu=0$  ( $H_0$ ). This way has good coverage properties.



5) With the pseudo-data one can then compute the integrals defining p-values for the two hypotheses. For the signal plus background hypothesis  $H_1$  one has

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | \text{signal+background}) = \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu | \mu, \hat{\theta}_\mu^{obs}) d\tilde{q}_\mu$$

and for the null, background-only  $H_0$  one has

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | \text{background-only}) = \int_{\tilde{q}_0^{obs}}^{\infty} f(\tilde{q}_\mu | 0, \hat{\theta}_0^{obs}) d\tilde{q}_\mu$$

6) Finally one can compute the value called  $CL_s$  as

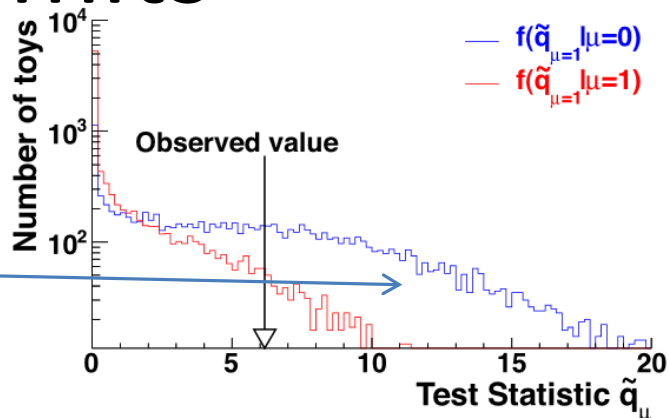
$$CL_s = p_\mu / (1 - p_b)$$

$CL_s$  is thus a “modified” p-value, in the sense that it describes how likely it is that the value of test statistic is observed under the alternative hypothesis **by also accounting for how likely the null is**: the drawing incorrect inferences based on extreme values of  $p_\mu$  is “damped”, and cases when one has no real discriminating power, approaching the limit  $f(q|\mu)=f(q|0)$ , are prevented from allowing to exclude the alternate hypothesis.

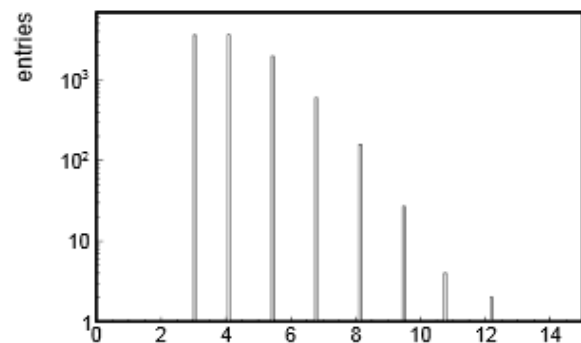
7) We can then **exclude  $H_1$  when  $CL_s < \alpha$** , the (defined in advance !) *size* of the test. In the case of Higgs searches, **all mass hypotheses  $H_1(M)$  for which  $CL_s < 0.05$  are said to be excluded** (one would rather call them “disfavoured”...)

# Derivation of Expected Limits

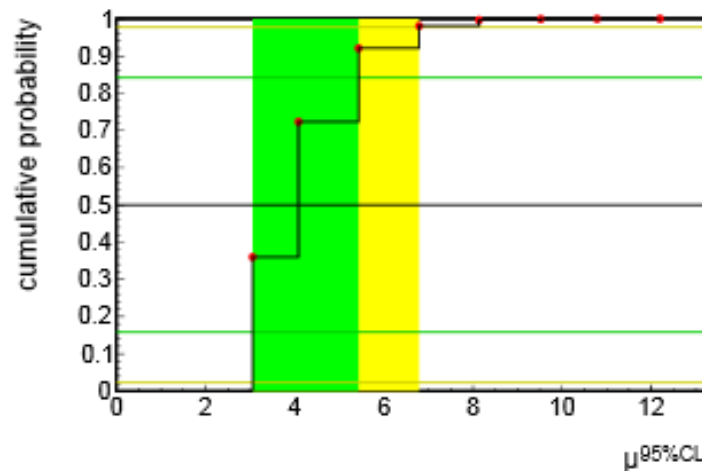
One starts with the **background-only hypothesis  $\mu=0$** , and determines a distribution of possible outcomes of the experiment with toys, obtaining the CLs test statistic distribution for each investigated Higgs mass point



From CLs one obtains the PDF of upper limits  $\mu^{UL}$  on  $\mu$  for each  $M_h$ . [E.g. on the right we assumed  $b=1$  and  $s=0$  for  $\mu=0$ , whereas  $\mu=1$  would produce  $\langle s \rangle = 1$ ]



Then one computes **the cumulative PDF of  $\mu^{UL}$**



Finally, one can derive the median and the intervals for  $\mu$  which correspond to 2.3%, 15.9%, 50%, 84.1%, 97.7% quantiles. These define the “expected-limit bands” and their center.



# Quantifying the Significance of a Signal in the Higgs Search

- To test for the significance of an excess of events, given a  $M_h$  hypothesis, one uses the bgr-only hypothesis and constructs a modified version of the  $q$  test statistic:

$$q_0 = -2 \ln \frac{\mathcal{L}(\text{data}|0, \hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})} \quad \text{and } \hat{\mu} \geq 0.$$

- This time we are testing any  $\mu > 0$  versus the  $H_0$  hypothesis. One builds the distribution  $f(q_0 | 0, \theta_0^{\text{obs}})$  by generating pseudo-data, and derives a p-value corresponding to a given observation as

$$p_0 = P(q_0 \geq q_0^{\text{obs}}) = \int_{q_0^{\text{obs}}}^{\infty} f(q_0 | 0, \hat{\theta}_0^{\text{obs}}) dq_0.$$

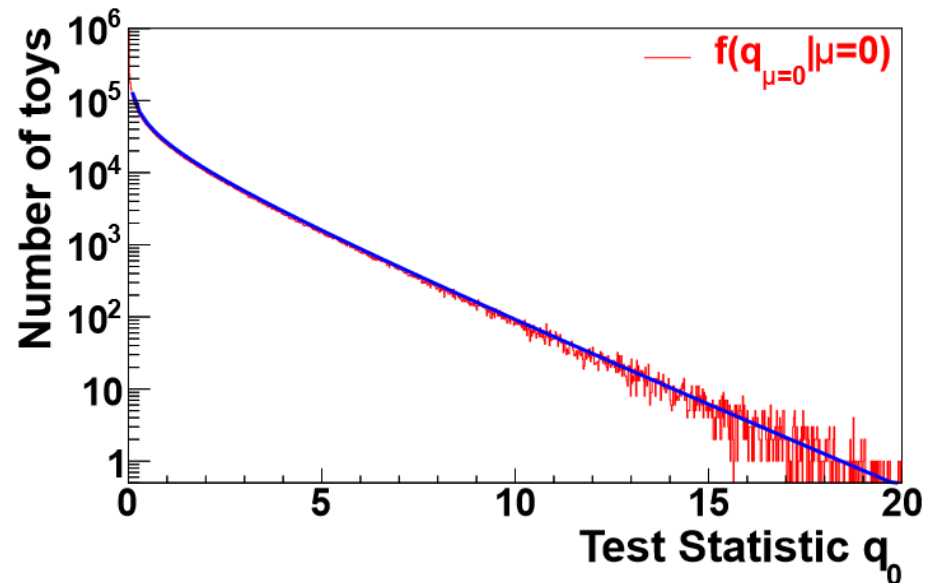
- One then converts  $p$  into  $Z$  using the relation

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = \frac{1}{2} P_{\chi^2_1}(Z^2)$$

where  $p_{\chi^2}$  is the survival function for the 1-dof chisquared.

- Often it is impractical to generate large datasets given the complexity of the search (dozens of search channels and sub-channels, correlated among each other). One then relies on a very good asymptotic approximation:
- The derived p-value and the corresponding Z value are “local”: they correspond to the specific hypothesis that has been tested (a specific  $M_h$ ) as  $q_0$  also depends on  $M_h$  (the search changes as  $M_h$  varies)
- When dealing with many searches, one needs to get a global p-value and significance, i.e. evaluate a trials factor. How to do it in complex situations is explained in the next slide.

$$p^{estimate} = \frac{1}{2} \left[ 1 - \text{erf} \left( \sqrt{q_0^{obs}/2} \right) \right]$$



# Trials Factors in the Higgs Search

When dealing with complex cases (Higgs combination), a study comes to help.

Wilks' theorem does not apply, and the complication of combining many different search channels makes the option of throwing huge number of toys impractical

Fortunately it has been shown how the trials factor can be counted in. First of all one defines a test statistic encompassing all possible Higgs mass values:

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

This is the maximum of the test statistic defined above for the bgr-only, across the many tests performed at the various possible masses of the Higgs boson.

One can use an asymptotic “regularity” of the distribution of the above  $q$  to get a global p-value by using a technique derived by Gross and Vitells [Vitells 2010].

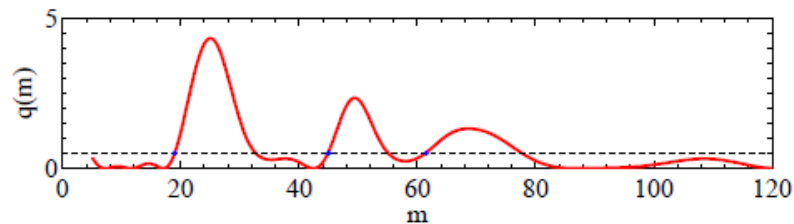
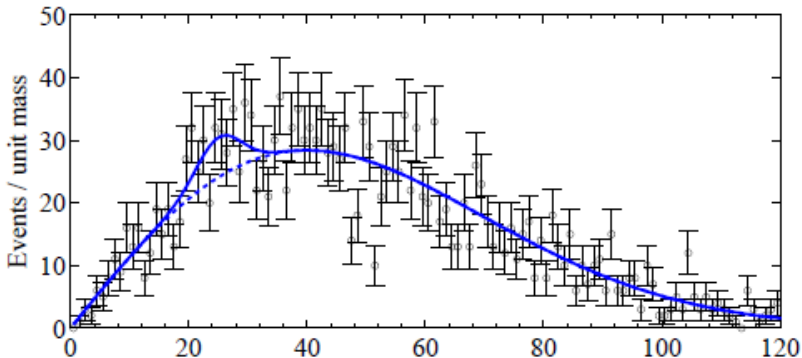
# Local Minima and Upcrossings

One counts the **number of “upcrossings” of the distribution of the test statistic**, as a function of mass. Its wiggling tells you how many independent places you have been searching in.

The number of local minima in the fit to a distribution is closely connected to the freedom of the fit to pick signal-like fluctuations in the investigated range

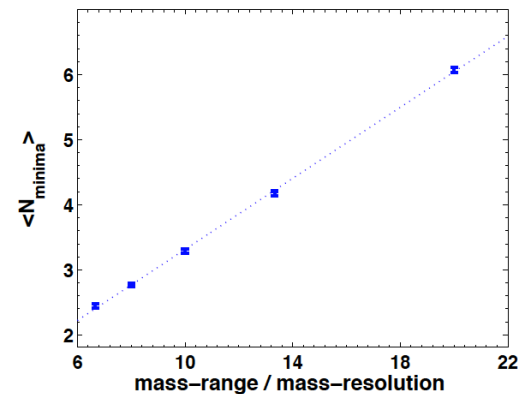
The number of times that the test statistic (below, the likelihood ratio between  $H_1$  and  $H_0$ ) crosses some reference point is a measure of the trials factor. One estimates the global p-value with the number  $N_0$  of upcrossings from a minimal value of the  $q_0$  test statistics (for which  $p=p_0$ ) by the formula

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$$



The number of upcrossings can be best estimated using the data themselves at a low value of significance, as it has been shown that the dependence on  $Z$  is a simple negative exponential:

$$\langle N_u \rangle = \langle N_{u_0} \rangle e^{-(u-u_0)/2}$$



# Trial Factors Example

- Imagine that you scan the Higgs mass and find a maximum  $q_0$  of 9, which according to

$$p^{estimate} = \frac{1}{2} \left[ 1 - \text{erf} \left( \sqrt{q_0^{obs}/2} \right) \right]$$

corresponds to a local p-value of 0.13% and a local Z-value of  $3\sigma$ , the latter computed using

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx = \frac{1}{2} P_{\chi_1^2}(Z^2)$$

- You then look at the distribution of  $q_0$  as a function of  $M_h$  and count the number of upcrossings at a level  $u_0=1$  (where the significance is  $Z=1$  as per above formulas) finding that there are 8 of them. You can then get  $\langle N_u \rangle$  for  $u=9$  using

$$\langle N_u \rangle = \langle N_{u_0} \rangle e^{-(u-u_0)/2}$$

which gives  $\langle N_u \rangle = 0.1465$

- The global p-value can be then computed as  $p_{glob} = 0.1465 + 0.0013$  using the formula below. One concludes that the trial factor is about 100 in this case.

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$$

# Conclusions

- **Statistics is NOT trivial.** Not even in the simplest applications!
- A understanding of the different methods to derive results (eg. for upper limits) is crucial to make sense of the often conflicting results one obtains even in simple problems
  - The key in HEP is to try and derive results with different methods –if they do not agree, we get wary of the results, plus we learn something
- Making the right choices for what method to use is an expert-only decision, so...

**You** should become an **expert in Statistics**, if you want to be a good particle physicist (or even if you want to make money in the financial market)
- The slide of this course are nothing but an appetizer. To really learn the techniques, you must **put them to work**

# References In Random Order

- F. James, *Statistical Methods in Experimental Physics* (II<sup>nd</sup> ed.), World Scientific (2006)
- G. Cowan, *Statistical Data Analysis*, Clarendon Press (1998)
- [R. Cousins, HCPSS lectures \(2009\)](#)
- G. D'Agostini, *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN Yellow Report 99/03 (1999)
- A. Stuart, K. Ord, S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A, 6<sup>th</sup> edition (1999)
- D. Cox, *Principles of Statistical Inference*, Cambridge UP (2006)
- B. P. Roe, *Probability and Statistics in Experimental Physics*, Springer-Verlag (1992)
- [R. Cousins and J. Tucker, 0905.3831 \(2009\)](#)
- [R. Cousins, Arxiv:1109.2023 \(2011\)](#)
- [R. Cousins, "Why Isn't Every Physicist a Bayesian?", Am. J. Phys. 63, n.5, pp. 398-410 \(1995\)](#)
- [E. Gross, "Look Elsewhere Effect", Banff \(2010\) \(see p.19\)](#)
- [E. Gross and O. Vitells, "Trials factors for the look elsewhere effects in High-Energy Physics", Eur.Phys.J.C70:525-530 \(2010\)](#)
- [ATLAS and CMS Collaborations, ATLAS-CONF-2011-157 \(2011\); CMS PAS HIG-11-023 \(2011\)](#)
- [ATLAS Collaboration, CMS Collaboration, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in summer 2011", ATL-PHYS-PUB-2011-818, CMS NOTE-2011/005 \(2011\).](#)
- G. Feldman and R. D. Cousins, "A Unified Approach to the Classical Statistical Analysis of Small Signals", Phys. Rev. D 57 (1998) 3873.
- D. Cox, "Some Problems Connected with Statistical Inference", Ann. Math. Stat. 29 (1958) no. 2, 357-372.
- R. D. Cousins, "Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter", [arXiv:1109.2023 \(2011\)](#).
- E. Gross and O. Vitells, "Trials factors for the Look-Elsewhere Effect in High-Energy Physics", Eur. Phys. J. C70 (2010) 525-530.
- M. Roos, M. Hietanen, and M. Luoma, "A new procedure for averaging particle properties", Phys.Fenn. 10 (1975) 21.
- D. Bailey, "Not Normal: the uncertainties of scientific measurements", ArXiv:1612.00778 (2016).
- D.V. Lindley, "A statistical paradox", *Biometrika*, 44 (1957) 187-192.
- R. D. Cousins, "The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics", arxiv:1310.3791v4 (2014).
- H. Jeffreys, "Theory of Probability", 3<sup>rd</sup> edition Oxford University Press, Oxford, 385.
- C. McCusker, I. Cairns, PRL 23, 658 (1969)

# Backup Material



# Loaded Die: Least-Square Solution

- We just have to write a chisquare as a function of the data  $N_i=(3,3,3,3,3,5)$  and the load  $t$ :

$$\chi^2 = \sum_{i=1}^6 \frac{(N_i - e_i(t))^2}{\sigma_i^2}$$

where  $e_i(t)$  are the expected times that result "i" appears in 20 throws, i.e.  $e_i = 20 P(i)$  where, as before,

$$\begin{aligned} P(1) &= 1/6 - t/2 \\ P(2) &= P(3) = P(4) = P(5) = 1/6 - t/8 \\ P(6) &= 1/6 + t \end{aligned}$$

Note that we can use the information of  $N_2, N_3, N_4, N_5$  distributions if we wish – it just amounts to consider them as separate in the  $\chi^2$ .

Once we have the  $\chi^2(t)$ , we may compute its derivative w.r.t.  $t$ , and set it to zero, then solve for  $t \rightarrow$  this will yield our point estimate  $t^*$

The interval will be obtained by finding  $t_1, t_2$  such that

$$\chi^2(t_1) = \chi^2(t_2) = \chi^2(t^*) + 1$$

Results: ....

Comparing with the likelihood solution, we see that ... ?

Of the two ways to compute the chisquare the preferable one is ... ?

# Calculation

Inputs:  $N, n_1, n_x, n_6$  ( $x = \text{sum of } 2,3,4,5$ )

$$e_i(t) = N \cdot p(i,t) \rightarrow e_1(t) = N \cdot (1/6 - t/2); e_x(t) = 4 \cdot N \cdot (1/6 - t/8) = N \cdot (2/3 - t/2); e_6(t) = N \cdot (1/6 + t) \quad (\rightarrow e_{\text{tot}} = N)$$

$$S_1 = [n_1 - e_1(t)]^2 / n_1 = [n_1^2 - 2 \cdot n_1 \cdot N \cdot (1/6 - t/2) + N^2 \cdot (1/6 - t/2)^2] / n_1 =$$

$$n_1 - N/3 + N \cdot t + N^2 / (36 \cdot n_1) - N^2 \cdot t / (6 \cdot n_1) + N^2 \cdot t^2 / (4 \cdot n_1)$$

$$S_x = [n_x - e_x(t)]^2 / n_x = [n_x^2 - 2 \cdot n_x \cdot N \cdot (2/3 - t/2) + N^2 \cdot (2/3 - t/2)^2] / n_x =$$

$$n_x - 4 \cdot N/3 + N \cdot t + 4 \cdot N^2 / (9 \cdot n_x) - 2 \cdot N^2 \cdot t / (3 \cdot n_x) + N^2 \cdot t^2 / (4 \cdot n_x)$$

$$S_6 = [n_6 - e_6(t)]^2 / n_6 = [n_6^2 - 2 \cdot n_6 \cdot N \cdot (1/6 + t) + N^2 \cdot (1/6 + t)^2] / n_6 =$$

$$n_6 - N/3 - 2 \cdot N \cdot t + N^2 / (36 \cdot n_6) + N^2 \cdot t / (3 \cdot n_6) + N^2 \cdot t^2 / n_6$$

$$dS_1/dt = N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1)$$

$$dS_x/dt = N - 2 \cdot N^2 / (3 \cdot n_x) + N^2 \cdot t / (2 \cdot n_x)$$

$$dS_6/dt = -2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6$$

$$dS_1/dt + dS_x/dt + dS_6/dt = 0 \rightarrow$$

$$N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1) + N - 2 \cdot N^2 / (3 \cdot n_x) + N^2 \cdot t / (2 \cdot n_x) - 2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6 = 0$$

$$t \cdot N^2 \cdot [1 / (2 \cdot n_1) + 1 / (2 \cdot n_x) + 2 / n_6] - [N^2 \cdot (1 / (6 \cdot n_1) + 2 / (3 \cdot n_x) - 1 / (3 \cdot n_6))] = 0$$

$$t = [1 / (6 \cdot n_1) + 2 / (3 \cdot n_x) - 1 / (3 \cdot n_6)] / [1 / (2 \cdot n_1) + 1 / (2 \cdot n_x) + 2 / n_6] =$$

$$= (n_x \cdot n_6 + 4 \cdot n_1 \cdot n_6 - 2 \cdot n_1 \cdot n_x) / (6 \cdot n_1 \cdot n_x \cdot n_6) / (3 \cdot n_x \cdot n_6 + 3 \cdot n_1 \cdot n_6 + 12 \cdot n_1 \cdot n_x) / (6 \cdot n_1 \cdot n_x \cdot n_6) =$$

$$= (n_x \cdot n_6 + 4 \cdot n_1 \cdot n_6 - 2 \cdot n_1 \cdot n_x) / (3 \cdot n_x \cdot n_6 + 3 \cdot n_1 \cdot n_6 + 12 \cdot n_1 \cdot n_x)$$

# Calculation, using all results (2,3,4,5)

Inputs:  $N, n_1, n_x, n_6$  ( $x=2,3,4,5$ )

$$e_i(t) = N \cdot p(i,t) \rightarrow e_1(t) = N \cdot (1/6 - t/2); e_x(t) = N \cdot (1/6 - t/8); e_6(t) = N \cdot (1/6 + t) \quad (\rightarrow e_{\text{tot}} = N)$$

$$S_1 = [n_1 - e_1(t)]^2 / n_1 = [n_1^2 - 2 \cdot n_1 \cdot N \cdot (1/6 - t/2) + N^2 \cdot (1/6 - t/2)^2] / n_1 =$$

$$n_1 - N/3 + N \cdot t + N^2 / (36 \cdot n_1) - N^2 \cdot t / (6 \cdot n_1) + N^2 \cdot t^2 / (4 \cdot n_1)$$

$$S_x = [n_x - e_x(t)]^2 / n_x = [n_x^2 - 2 \cdot n_x \cdot N \cdot (1/6 - t/8) + N^2 \cdot (1/6 - t/8)^2] / n_x =$$

$$n_x - N/3 + N \cdot t/4 + N^2 / (36 \cdot n_x) - N^2 \cdot t / (24 \cdot n_x) + N^2 \cdot t^2 / (64 \cdot n_x)$$

$$S_6 = [n_6 - e_6(t)]^2 / n_6 = [n_6^2 - 2 \cdot n_6 \cdot N \cdot (1/6 + t) + N^2 \cdot (1/6 + t)^2] / n_6 =$$

$$n_6 - N/3 - 2 \cdot N \cdot t + N^2 / (36 \cdot n_6) + N^2 \cdot t / (3 \cdot n_6) + N^2 \cdot t^2 / n_6$$

$$dS_1/dt = N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1)$$

$$dS_x/dt = N/4 - N^2 / (24 \cdot n_x) + N^2 \cdot t / (32 \cdot n_x)$$

$$dS_6/dt = -2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6$$

$$dS_1/dt + dS_2/dt + dS_3/dt + dS_4/dt + dS_5/dt + dS_6/dt = 0 \rightarrow$$

$$N - N^2 / (6 \cdot n_1) + N^2 \cdot t / (2 \cdot n_1) + N/4 - N^2 / (24 \cdot n_2) + N^2 \cdot t / (32 \cdot n_2) + N/4 - N^2 / (24 \cdot n_3) + N^2 \cdot t / (32 \cdot n_3) + N/4 - N^2 / (24 \cdot n_4) + N^2 \cdot t / (32 \cdot n_4) + N/4 - N^2 / (24 \cdot n_5) + N^2 \cdot t / (32 \cdot n_5) - 2 \cdot N + N^2 / (3 \cdot n_6) + 2 \cdot N^2 \cdot t / n_6 = 0$$

$$t \cdot N^2 \cdot [1 / (2 \cdot n_1) + 1 / (32 \cdot n_2) + 1 / (32 \cdot n_3) + 1 / (32 \cdot n_4) + 1 / (32 \cdot n_5) + 2 / n_6] - [N^2 \cdot (1 / (6 \cdot n_1) + 1 / (24 \cdot n_2) + 1 / (24 \cdot n_3) + 1 / (24 \cdot n_4) + 1 / (24 \cdot n_5) - 1 / (3 \cdot n_6))] = 0$$

$$t = [1 / (6 \cdot n_1) + 1 / (24 \cdot n_2) + 1 / (24 \cdot n_3) + 1 / (24 \cdot n_4) + 1 / (24 \cdot n_5) - 1 / (3 \cdot n_6)] / [1 / (2 \cdot n_1) + 1 / (32 \cdot n_2) + 1 / (32 \cdot n_3) + 1 / (32 \cdot n_4) + 1 / (32 \cdot n_5) + 2 / n_6] =$$

$$= 4/3 \cdot [4/n_1 + 1/n_2 + 1/n_3 + 1/n_4 + 1/n_5 - 8/n_6] / [16/n_1 + 1/n_2 + 1/n_3 + 1/n_4 + 1/n_5 + 64/n_6]$$

# Coverage of Flip-Flopping Experiment

- We want to write a routine that determines the true coverage of the procedure discussed above for a Gaussian measurement of a bounded parameter:
  - $x_{\text{meas}} < 0 \rightarrow$  quote size- $\alpha$  upper limit as if  $x_{\text{meas}} = 0$ ,  $x^{\text{up}} = \text{sqrt}(2) * \text{ErfInverse}(1-2\alpha)$
  - $0 \leq x_{\text{meas}} < D \rightarrow$  quote size- $\alpha$  upper limit,  $x^{\text{up}} = \text{sqrt}(2) * \text{ErfInverse}(1-2\alpha) + x_{\text{meas}}$
  - $x_{\text{meas}} \geq D \rightarrow$  quote central value  $\pm \alpha/2$  error bars,  $x_{\text{meas}} \pm \text{sqrt}(2) * \text{ErfInverse}(1-\alpha)$

## Guidelines:

1. insert proper includes (we want to compile it or it'll be too slow)
2. header: pass through it alpha, D, and N\_pexp
3. define useful variables and histogram containing coverage values
4. loop on x\_true values from 0 to 10 in 0.1 steps  $\rightarrow i=0 \dots < 100$  steps,  $x_{\text{true}} = 0.05 + 0.1 * i$
5. for each x\_true:
  1. zero a counter C
  2. loop many times (eg. N\_pexp, defined in header)
  3. throw  $x_{\text{meas}} = \text{gRandom} \rightarrow \text{Gaus}(x_{\text{true}}, 1.)$
  4. derive x\_down and x\_up depending on x\_meas:
    1. if  $x_{\text{meas}} < 0$  then  $x_{\text{down}} = 0$  and  $x_{\text{up}} = \text{sqrt}(2) * \text{ErfInverse}(1-2 * \alpha)$
    2. if  $0 \leq x_{\text{meas}} < D$  then  $x_{\text{down}} = 0$  and  $x_{\text{up}} = x_{\text{meas}} + \text{sqrt}(2) * \text{EI}(1-2 * \alpha)$
    3. if  $x_{\text{meas}} \geq D$  then  $x_{\text{down,up}} = x_{\text{meas}} \pm \text{sqrt}(2) * \text{EI}(1-\alpha)$
  5. if x\_true is in  $[x_{\text{down}}, x_{\text{up}}]$  C++
6. fill histogram of coverage at x\_true with C/N\_pexp
7. plot and enjoy

# Coverage of Flip-flopping measurement

```
void FlipFlop (double alpha=0.05, double D=4.5, double Npexp=1000) {
```

```
    double x_true;
    double x_meas;
    double sigma = 1;
    double x_down;
    double x_up;
    double covers=0.;
    double El1 = sqrt(2)*TMath::ErfInverse(1-alpha);
    double El2a=sqrt(2)*TMath::ErfInverse(1-2*alpha);
```

```
    TH1D * Coverage_vs_xtrue = new TH1D("Coverage_vs_xtrue", "Coverage vs x_true", 100, 0., 10.);
    TH1D * BeltUp = new TH1D ("BeltUp", "Flip-flopping Confidence belt", 15000, -5.,10.);
    TH1D * BeltDo = new TH1D ("BeltDo", "Flip-flopping Confidence belt", 15000, -5.,10.);
```

```
    cout << "Critical values:" << endl;
    cout << "For xmeas < 0 : 0 < xtrue < " << El2a*sigma << endl;
    cout << "For 0<xmeas<" << D << " : 0 < xtrue < xmeas+"
        << El2a*sigma << endl;
    cout << "For xmeas>=D : xmeas-" << El1*sigma << " < xtrue < xmeas+"
        << El1*sigma << endl;
    cout << endl;
    for (int ix=0; ix<100; ix++) {
```

```
        x_true = 0.05 + 0.1*ix;
        covers=0;
        for (int pexp=0; pexp<Npexp; pexp++) {
```

```
            // A Gaussian measurement with uncertainty sigma
            x_meas = gRandom->Gaus(x_true,sigma);
```

```
            if (x_meas<D) { // Not significantly different from zero, will report upper limit
                x_down = 0;
                x_up = El2a*sigma;
                if (x_meas>0) x_up = x_meas + x_up;
            } else { // will report an interval
                x_down = x_meas-El1*sigma;
                x_up = x_meas+El1*sigma;
            }
        }
    }
```

```
        // compute coverage
        if (x_true>=x_down && x_true<x_up) covers++;
    }

    Coverage_vs_xtrue->Fill(x_true,covers/Npexp);
}
```

```
    // Belt plot
    for (inti=0; i<15000;i++) {
        x_meas = -4.9995 + i*0.001;
        if (x_meas<0) {
            BeltUp->Fill(x_meas,El2a);
            BeltDo->Fill(x_meas,0.);
        } else if (x_meas<D) {
            BeltUp->Fill(x_meas,x_meas+El2a);
            BeltDo->Fill(x_meas,0.);
        } else {
            BeltUp->Fill(x_meas,x_meas+El1);
            BeltDo->Fill(x_meas,x_meas-El1);
        }
    }
```

```
    gStyle->SetOptStat(0);
```

```
    TCanvas * W2 = new TCanvas ("W2", "Coverage of flip-flopping NP construction", 500, 500);
    W2->cd();
    Coverage_vs_xtrue->SetLineWidth(3);
    Coverage_vs_xtrue->Draw();
```

```
    TCanvas * W = new TCanvas ("W", "Confidence belt", 500, 500);
    W->cd();
    BeltUp->SetMinimum(-1);
    BeltUp->SetMaximum(15);
    BeltUp->SetLineWidth(3);
    BeltDo->SetLineWidth(3);
    BeltUp->Draw();
    BeltDo->Draw("SAME");
}
```

# Coverage.C

(add at the top the #include commands needed to compile it)

```
void Coverage (double alpha, double disc_threshold=5.) {
// Only valid for the following:
// -----
if (disc_threshold-sqrt(2)*ErfInverse(1.-2*alpha/2.)<
    sqrt(2)*ErfInverse(1.-2*alpha)) {
    cout << "Too low discovery threshold, code not suitable. " << endl;
    cout << "Try a larger threshold" << endl;
    return;
}
char title[100];
int idisc_threshold=disc_threshold;
int fracdisctresh =10*(disc_threshold-idisc_threshold);
if (alpha>=0.1) {
    sprintf (title, "Coverage for #alpha=0.%d with Flip-Flopping at %d.%d-sigma",
(int)(10.*alpha),idisc_threshold, fracdisctresh);
} else {
    sprintf (title, "Coverage for #alpha=0.0%d with Flip-Flopping at %d.%d-
sigma", (int)(100.*alpha),idisc_threshold, fracdisctresh);
}
TH1D * Cov = new TH1D ("Cov", title, 1000, 0., 2.*disc_threshold);
Cov->SetXTitle("True value of #mu (in #sigma units)");

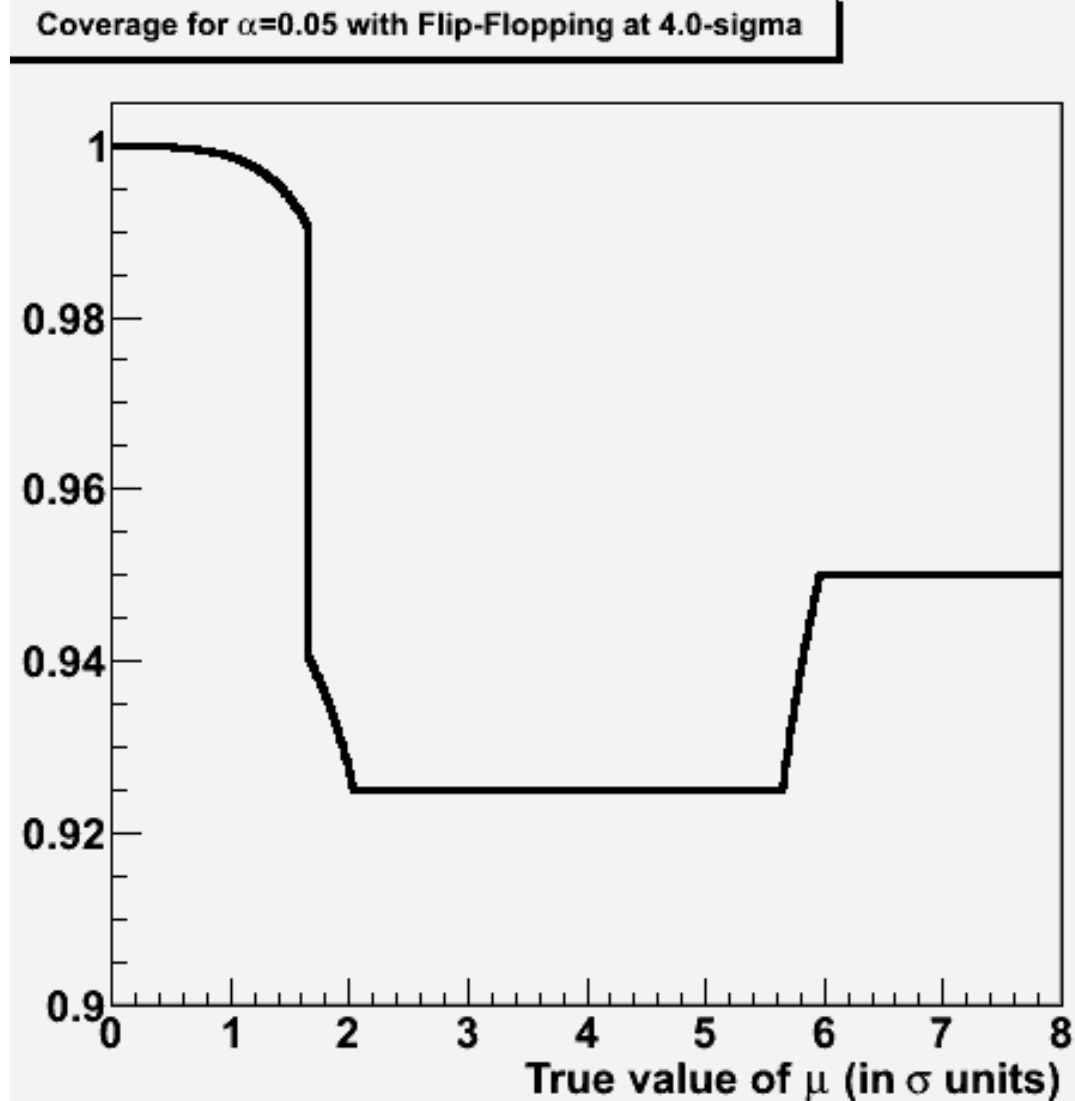
// Int Gaus-1:+1 sigma is TMath::Erf(1./sqrt(2.))
// To get 90% percentile (1.28): sqrt(2)*ErfInverse(1.-2*0.1)
// To get 95% percentile (1.64): sqrt(2)*ErfInverse(1.-2*0.05)
double cov;
for (int i=0; i<1000; i++) {
    double mu = (double)i/(1000./(2*disc_threshold))+
        0.5*(2*disc_threshold/1000);
```

```
if (mu<sqrt(2)*ErfInverse(1.-2*alpha)) { // 1.28, so mu within upper 90% CL
    cov = 0.5*(1+TMath::Erf((disc_threshold-mu)/sqrt(2.)));
} else if (mu< disc_threshold-sqrt(2)*ErfInverse(1.-2*alpha/2.)) { // <3.36
    cov = 1.-alpha-0.5*(1.-TMath::Erf((disc_threshold-mu)/sqrt(2.)));
} else if (mu<disc_threshold+
    sqrt(2)*ErfInverse(1.-2*alpha)) { // 6.28
    cov = 1.-1.5*alpha;
} else if (mu<disc_threshold+sqrt(2)*ErfInverse(1.-2*alpha/2.)) { // 6.64 {
    cov = 1.-alpha/2.-0.5*(1+TMath::Erf((disc_threshold-mu)/sqrt(2.)));
} else { cov = 1.-alpha; }
Cov->Fill(mu,cov);
}
char filename[40];
if (alpha>=0.1) {
    sprintf(filename,"Coverage_alpha_0.%d_obs_at_%d_sigma.eps",
(int)(10.*alpha),idisc_threshold);
} else {
    sprintf(filename,"Coverage_alpha_0.0%d_obs_at_%d_sigma.eps",
(int)(100.*alpha),idisc_threshold);
}
TCanvas * C = new TCanvas ("C","Coverage", 500,500);
C->cd();
Cov->SetMinimum(1.-2*alpha);
Cov->SetLineWidth(3);
Cov->Draw();
C->Print(filename);
// Now plot confidence belt
```

Here is e.g. the exact calculation of coverage for flip-flopping at 4-sigma and a test size  $\alpha=0.05$

Can get it by running:

```
root> .L Coverage.C+;  
root> Coverage(0.05,4.);
```



# Maximum Likelihood

- Take a pdf for a random variable  $x$ ,  $f(\mathbf{x}; \theta)$  which is analytically known, but for which the value of  $m$  parameters  $\theta$  is not. The *method of maximum likelihood* allows us to estimate the parameters  $\theta$  if we have a set of data  $x_i$  distributed according to  $f$ .

- The probability of our observed set  $\{\mathbf{x}_i\}$  depends on the distribution of the pdf. If the measurements are independent, we have

$$p = \prod_{i=1}^n f(x_i; \theta) dx_i \quad \text{to find } x_i \text{ in } [x_i, x_i + dx_i[$$

- The likelihood function

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

is then a **function of the parameters  $\theta$**  only. It is written as the joint pdf of the  $x_i$ , but *we treat those as fixed*. L is not a pdf! NOTA BENE! **The integral under L is MEANINGLESS.**

- Using  $L(\theta)$  one can define “maximum likelihood estimators” for the parameters  $\theta$  as the values which maximize the likelihood, i.e. the solutions  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$  of the equation

$$\left( \frac{\partial L(\theta)}{\partial \theta_j} \right)_{\theta = \hat{\theta}} = 0 \quad \text{for } j=1 \dots m$$

Note: The ML requires (**and exploits!**) the *full knowledge* of the distributions



# Maximum Likelihood for Gaussian pdf

- Let us take  $n$  measurements of a random variable distributed according to a Gaussian PDF with  $\mu$ ,  $\sigma$  unknown parameters. We want to use our data  $\{x_i\}$  to estimate the Gaussian parameters with the ML method.
- The log-likelihood is

$$\log L(\mu, \sigma^2) = \sum_{i=1}^n f(x_i; \mu, \sigma^2) = \sum_{i=1}^n \left( \log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \log \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

- The MLE of  $\mu$  is the value for which  $d \ln L / d\mu = 0$ :

$$\frac{d \ln L}{d\mu} = \sum_{i=1}^n \frac{(-2\mu - 2x_i)}{2\sigma^2}$$

$$0 = \sum_{i=1}^n (-2\mu - 2x_i)$$

$$\rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

So we see that **the ML estimator of the Gaussian mean is the sample mean.**

We can easily prove that **the sample mean is a unbiased estimator of the Gaussian  $\mu$** , since its expectation value is

$$\begin{aligned}
 E[\hat{\mu}] &= \int \dots \int \hat{\mu}(x_1 \dots x_n) F(x_1 \dots x_n; \mu) dx_1 \dots dx_n \\
 &= \int \dots \int \frac{1}{n} \sum_i x_i \left[ \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j - \mu)^2}{2\sigma^2}} \right] dx_1 \dots dx_n \\
 &= \frac{1}{n} \sum_{i=1}^n \int x_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} dx_i \prod_{j=1(\neq i)}^n \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j - \mu)^2}{2\sigma^2}} dx_j \\
 &= \frac{1}{n} \sum_{i=1}^n \mu = \mu
 \end{aligned}$$

The same is **not true** of the ML estimate of  $\sigma^2$ ,

$$\begin{aligned}
 \frac{d \ln L}{d\sigma^2} &= \sum_{i=1}^n \left( -\frac{1}{2\sigma^2} + \frac{1}{\sigma^4} \frac{(x_i - \mu)^2}{2} \right) \\
 0 &= \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} \\
 \rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2
 \end{aligned}$$

since one can find as above that  $E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$

**The bias vanishes for large n.** Note that a unbiased estimator of the Gaussian  $\sigma$  exists: it is the **sample variance**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

which is a unbiased estimator of the variance for any pdf. But it is not the ML one.