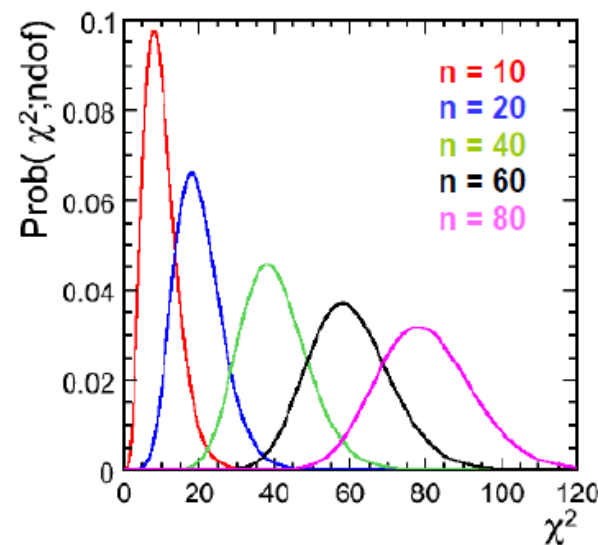
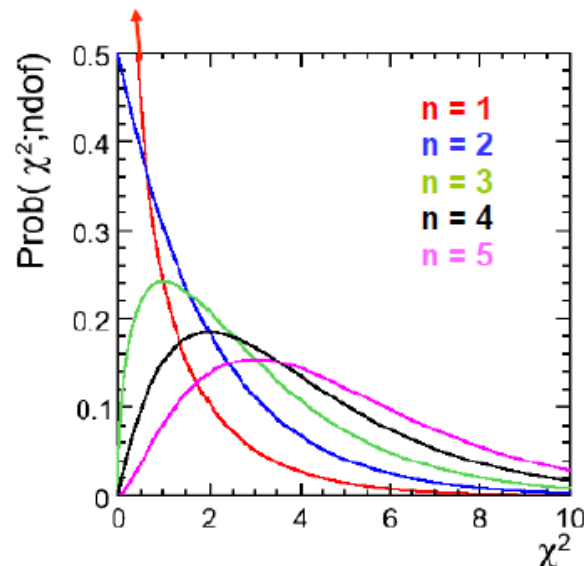


# Statistics and Data Analysis (HEP)

## Fitting and Hypothesis testing

- The  $\chi^2$  test
- Likelihood functions
- Fitting
- Binned maximum likelihood
- Unbinned maximum likelihood



Follow the course/slides from M. A. Thomson lectures at Cambridge University

- Lecture 1: Back to basics  
Introduction, Probability distribution functions, Binomial distributions, Poisson distribution
- Lecture 2: The Gaussian Limit  
The central limit theorem, Gaussian errors, Error propagation, Combination of measurements, Multi-dimensional Gaussian errors, Error Matrix
- Lecture 3: Fitting and Hypothesis Testing  
The  $\chi^2$  test, Likelihood functions, Fitting, Binned maximum likelihood, Unbinned maximum likelihood
- Lecture 4: The Dark Arts Part I  
Bayesian Inference, Credible Intervals
- Lecture 5: The Dark Arts II  
The Frequentist approach, Confidence Intervals, Limits near physical boundaries, Systematic Uncertainties .

# Introduction

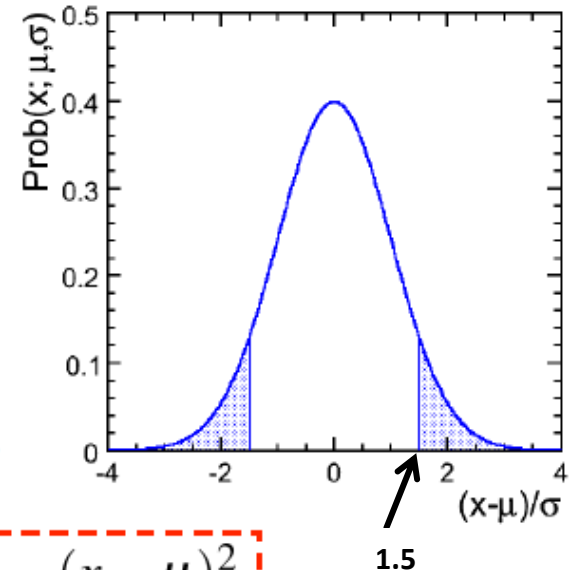
- ★ Given some **data** (event counts, distributions) and a **particular theoretical model**
  - are the data consistent with the model:
    - hypothesis testing
    - goodness of fit
  - in the context of the model, what are our best estimates of its parameters:
    - fitting
- ★ In both cases, need a measure of consistency of data with our model
- ★ Start with a discussion of  $\chi^2$

# The Chi-Squared Statistic

- ★ Suppose we measure a parameter,  $x \pm \sigma$ , which a theorist says should have the value  $\mu$
- ★ Within this simple model, we can write down the prior probability of obtaining the value  $x \pm \sigma$  given the prediction

$$P(\text{data}; \text{prediction}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

- ★ To express the consistency of the data, ask the question “if the model is correct what is the probability of obtaining result at least as far from the prediction as the observed value”
- ★ This is simply the fraction of the area under the Gaussian with  $|x - \mu| > |x_{obs} - \mu|$
- ★ e.g. if  $1.5\sigma$  from the prediction: **13%**
- ★ Only care about degree of consistency, not whether we are on the +ve or -ve side, so equivalently want the probability



$$P(\chi^2 > \chi_{obs}^2) \quad \text{where} \quad \chi^2 = \frac{(x - \mu)^2}{\sigma^2}$$

- ★ For Gaussian distributed variables,  $\chi^2$ , forms the basis of our consistency test

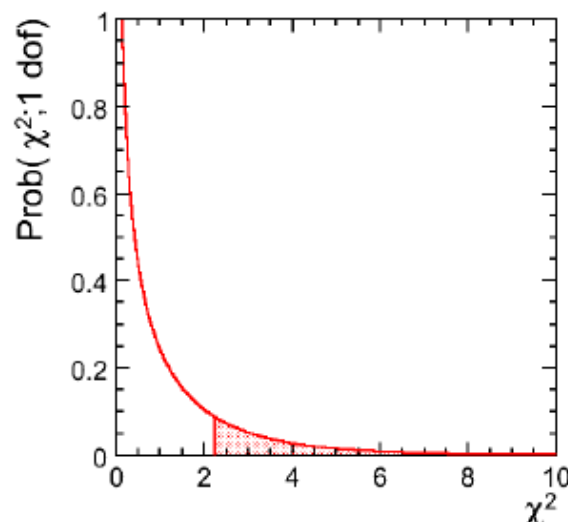
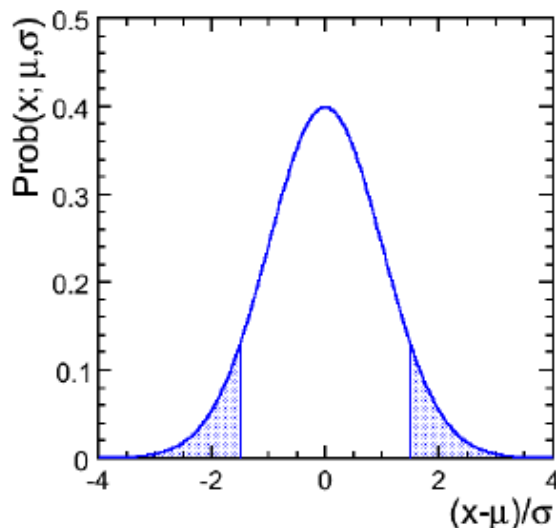
★ The probability distribution for,  $\chi^2$ , can be obtained easily from the ID distribution

$$P(\chi^2)d(\chi^2) = G(x)dx \quad \text{and} \quad \frac{d\chi^2}{dx} = 2 \frac{(x-\mu)}{\sigma^2} = 2 \frac{\sqrt{\chi^2}}{\sigma}$$

$$P(\chi^2) = 2 \times \frac{1}{\sqrt{2\pi}\sigma} \frac{\sigma}{2\sqrt{\chi^2}} \exp\left\{-\frac{\chi^2}{2}\right\}$$

Factor of two from  
+ve and -ve parts  
of Gaussian

$$P(\chi^2) = \frac{1}{\sqrt{2\pi}} (\chi^2)^{-\frac{1}{2}} \exp\left\{-\frac{\chi^2}{2}\right\}$$



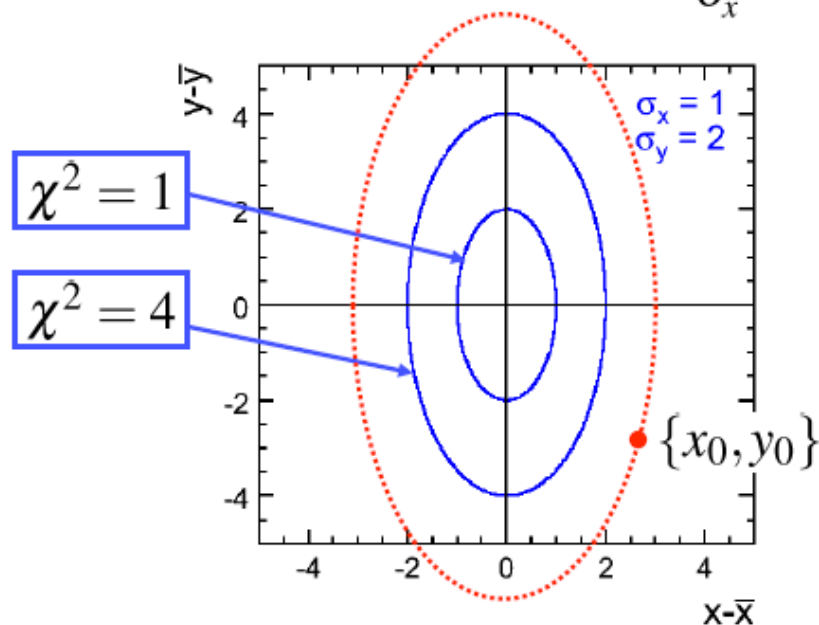
# The Chi-Squared Statistic in higher dimensions

- ★ So far, this isn't particularly useful...
- ★ But now extend this to two dimensions (ignoring correlations for the moment, although we now know how to include them)

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{1}{2} \left[ \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\}$$

- ★ Lines of equal probability are equivalent to lines of equal  $\chi^2$

$$\chi^2 = \chi_x^2 + \chi_y^2 = \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}$$



- ★ What if I measure  $\{x_0, y_0\}$
- ★ How consistent is this with expected values ?
- ★ ANSWER: the probability of obtaining smaller probability than observed
  - i.e. integrate 2D PDF over region where  $\chi^2 > \chi_{obs}^2$

★ It is worth seeing how this works mathematically...

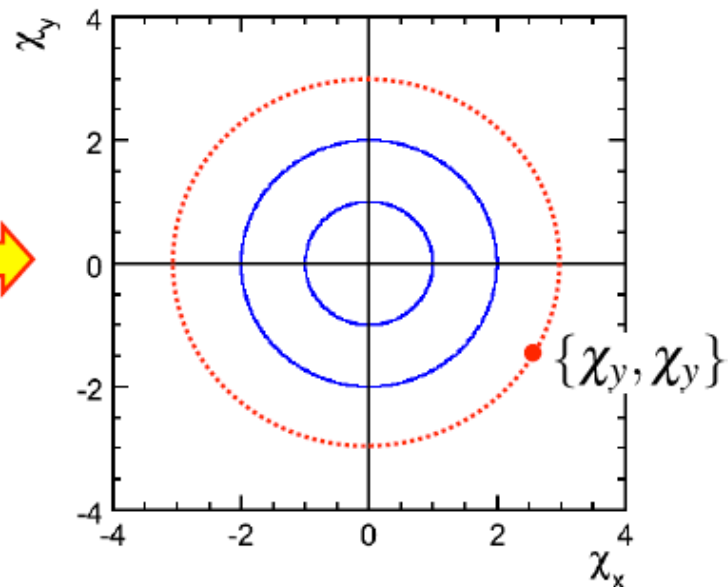
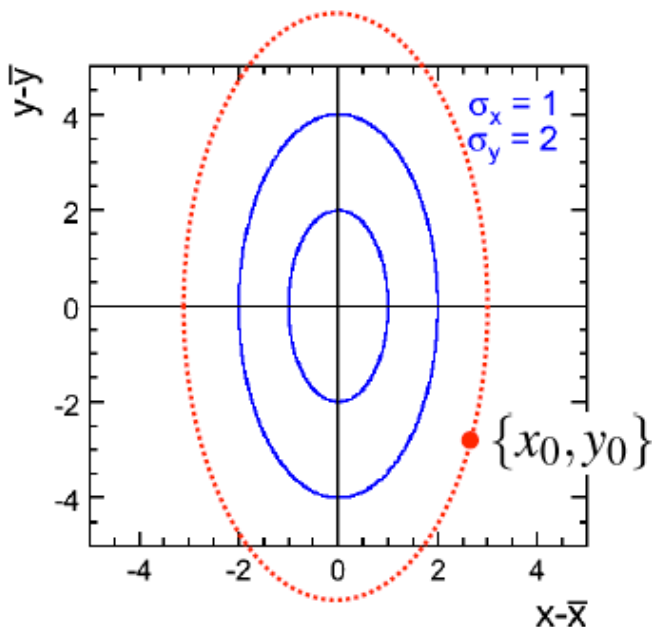
- First transform error ellipse into circular contours

$$\{x, y\} \Rightarrow \{\chi_x, \chi_y\} \quad \chi_x^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} \quad \chi_y^2 = \frac{(y - \mu_y)^2}{\sigma_y^2}$$

$$P(\chi_x, \chi_y) d\chi_x d\chi_y = P(x, y) dx dy$$

- Therefore

$$P(\chi_x, \chi_y) = \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} [\chi_x^2 + \chi_y^2] \right\}$$



- ★ Only interested in “radius” i.e.  $\chi^2$ 
  - so transform  $\{\chi_x, \chi_y\} \Rightarrow \{\chi, \phi\}$

with  $\delta\chi_x \delta\chi_y = \chi \delta\chi \delta\phi$

$$\begin{aligned} P(\chi, \phi) \delta\chi \delta\phi &= P(\chi_x, \chi_y) \delta\chi_x \delta\chi_y \\ &= \chi P(\chi_x, \chi_y) \delta\chi \delta\phi \end{aligned}$$

$$P(\chi) \delta\chi = 2\pi \chi P(\chi_x, \chi_y) \delta\chi$$

$$P(\chi) = \chi \exp\left(-\frac{\chi^2}{2}\right)$$

- ★ Therefore, probability distribution in chi-squared:

$$P(\chi^2; n=2) = \frac{1}{2} \exp\left(-\frac{\chi^2}{2}\right)$$

$$\delta(\chi^2) = 2\chi \delta\chi$$

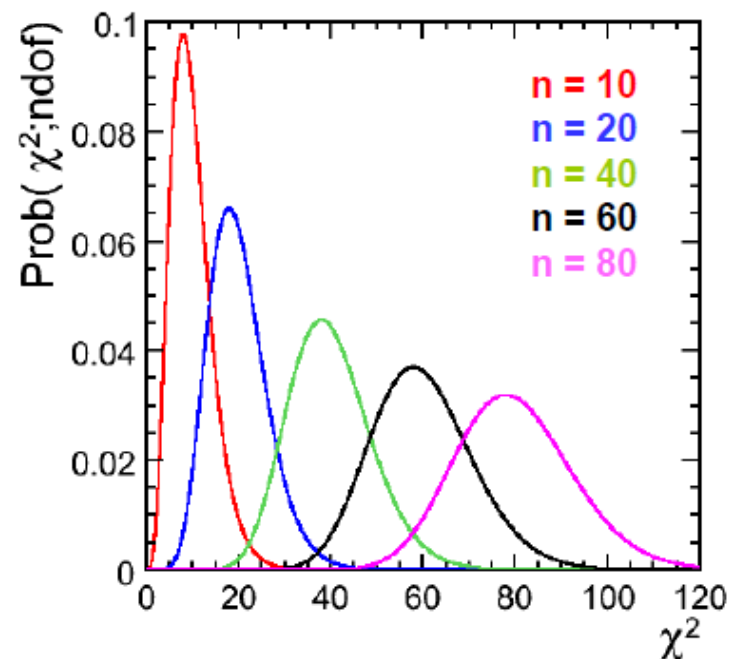
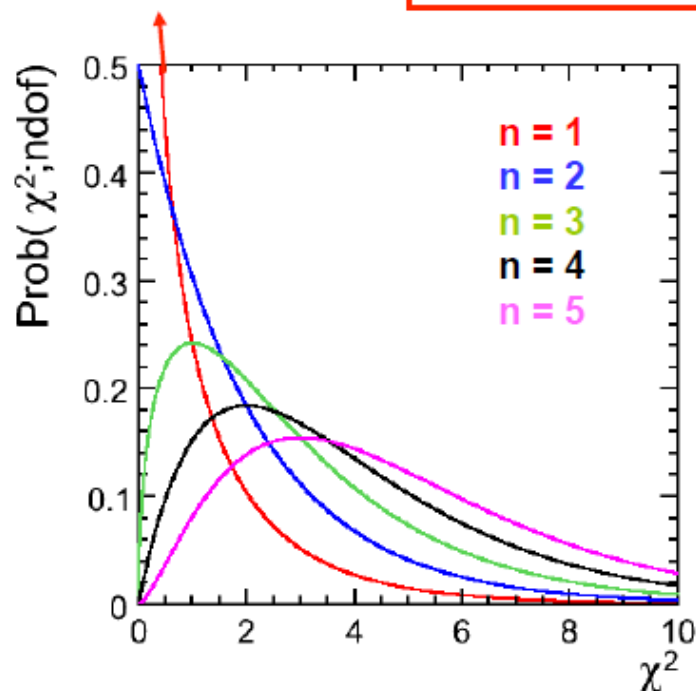
- ★ For two Gaussian distributed variables, we now have an expression for the chi-squared probability distribution !

Problem: Show that:

$$P(\chi^2; n=3) = \frac{1}{2\pi} (\chi^2)^{\frac{1}{2}} \exp\left(-\frac{\chi^2}{2}\right) \quad \text{and}$$

$$P(\chi^2; n) \propto (\chi^2)^{\frac{(n-2)}{2}} \exp\left(-\frac{\chi^2}{2}\right)$$

$$P(\chi^2; n) \propto (\chi^2)^{\frac{(n-2)}{2}} \exp\left(-\frac{\chi^2}{2}\right)$$



★ For any number of variables (degrees of freedom) can now obtain

$$P(\chi^2 > \chi_{obs}^2) = \int_{\chi_{obs}^2}^{\infty} P(\chi^2, n) d\chi^2$$

★ Already done for you, e.g. tables or more convenient `TMATH::Prob( $\chi^2, n$ )`

# Properties of chi-squared

★ For n degrees of freedom

$$\langle \chi^2 \rangle = n$$

Proof:

$$\begin{aligned} \langle \chi^2 \rangle &= \left\langle \sum_{i=1}^n \chi_i^2 \right\rangle \\ &= \sum_{i=1}^n \left\langle \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right\rangle \\ &= \sum_{i=1}^n 1 = n \end{aligned}$$

$$\begin{aligned} \text{or } \langle \chi^2 \rangle &= \frac{\int \chi^2 P(\chi^2; n) d(\chi^2)}{\int P(\chi^2; n) d(\chi^2)} \\ &= \frac{\int \chi^2 (\chi^2)^{\frac{(n-2)}{2}} e^{-\frac{\chi^2}{2}} 2\chi d\chi}{\int (\chi^2)^{\frac{(n-2)}{2}} e^{-\frac{\chi^2}{2}} 2\chi d\chi} \\ &= \frac{\int x^{n+1} e^{-\frac{x^2}{2}} dx}{\int x^{n-1} e^{-\frac{x^2}{2}} dx} \\ &= n \end{aligned}$$

★ For n degrees of freedom

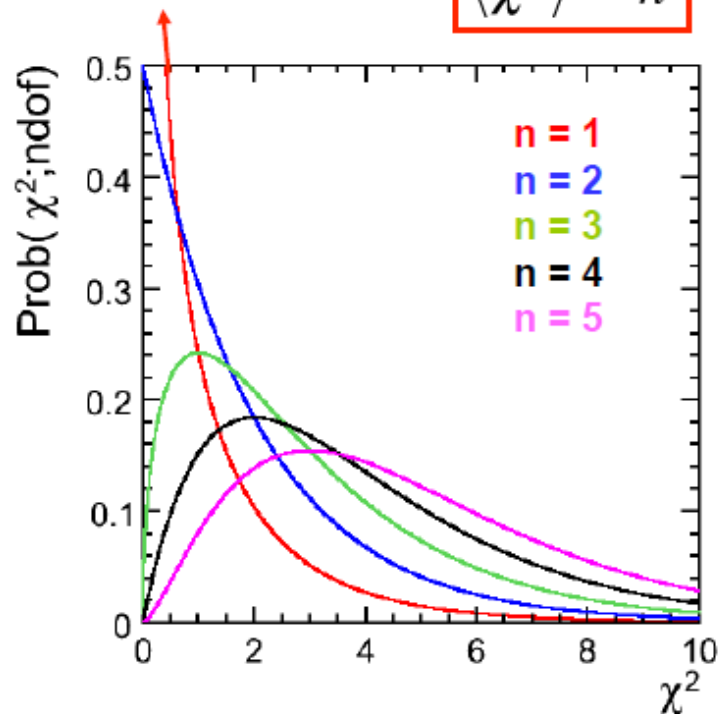
$$Var(\chi^2) = 2n$$

Proof:

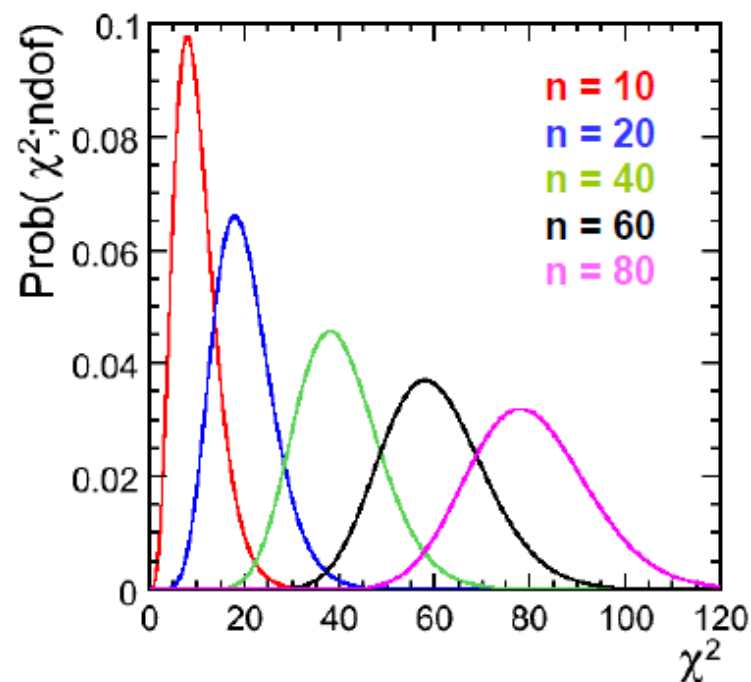
$$\begin{aligned} Var(\chi^2) &= \langle \chi^2 \chi^2 \rangle - \langle \chi^2 \rangle^2 \\ &= n \langle \chi_i^4 \rangle + n(n-1) - n^2 \\ &= 3n + n^2 - n - n^2 \\ &= 2n \end{aligned}$$

$$\begin{aligned} \langle \chi^4 \rangle &= \frac{\int x^{n+3} e^{-\frac{x^2}{2}} dx}{\int x^{n-1} e^{-\frac{x^2}{2}} dx} \\ &= I_{n+3} / I_{n-1} \\ &= (n+2)n \\ Var(\chi^2) &= \langle \chi^4 \rangle - \langle \chi^2 \rangle^2 \\ &= n^2 + 2n - n^2 = 2n \end{aligned}$$

$$\langle \chi^2 \rangle = n$$



$$\text{Var}(\chi^2) = 2n$$



★ For large  $n$ ; the distribution tends to a Gaussian

(large  $\sim 40$ )

Useful for quick estimates...

# Example

- ★ Suppose we have an absolute prediction for a distribution, e.g. a differential cross section and we can account perfectly for experimental effects (efficiency, background, bin-bin migration)
- ★ Measure number of events in bins of  $\cos\theta$ ,  $n_i$ , and compare to prediction  $\mu_i$
- ★ If prediction is correct, expect the observed number of events in a given bin to follow a Poisson distribution of mean  $\mu_i$
- ★ If the expectations in all bins are sufficiently large, the Poisson distribution can be approximated as a Gaussian with **mean**  $\mu_i$  and **variance**  $\mu_i$
- ★ In this limit can use chi-squared for consistency with hypothesis

$$\chi_i^2 = \frac{(n_i - \mu_i)^2}{\mu_i}$$

Expected fluctuations  
around mean

- ★ For **N** bins, have **N** independent (approximately) Gaussian distributed variables
- ★ Overall consistency of data with prediction assessed using

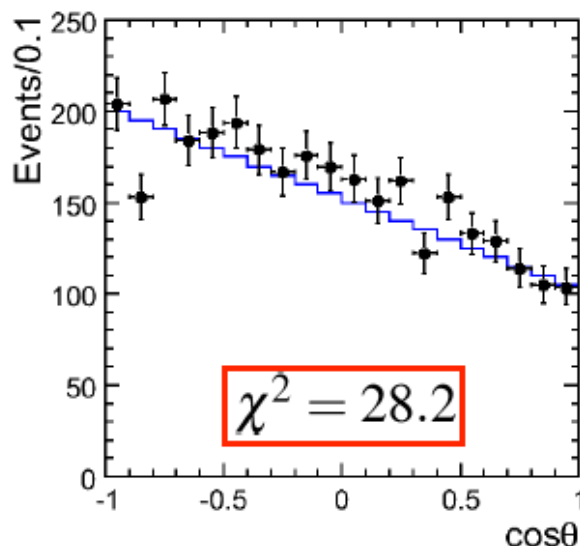
$$\chi^2 = \sum_i \chi_i^2$$

- ★ If hypothesis (prediction) is correct expect

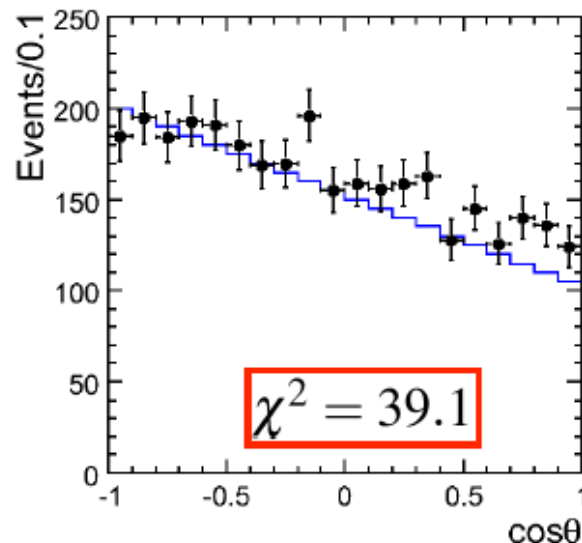
$$\langle \chi^2 \rangle = n$$

$$\text{Var}(\chi^2) = 2n$$

$$\langle \chi^2 \rangle = 20$$



$$\sigma_{\chi^2} = \sqrt{40} = 6.3$$



▪ **Quick estimates:**

$$\approx 8.2/6.3 = 1.3\sigma$$

$$\approx 19.1/6.3 = 3.0\sigma$$

▪ But  $N=20$ , not very large, these estimates only give an indication of the agreement.

▪ Correct numbers (integral of expected  $\chi^2$  distribution) TMath::Prob( $\chi^2$ ,20)

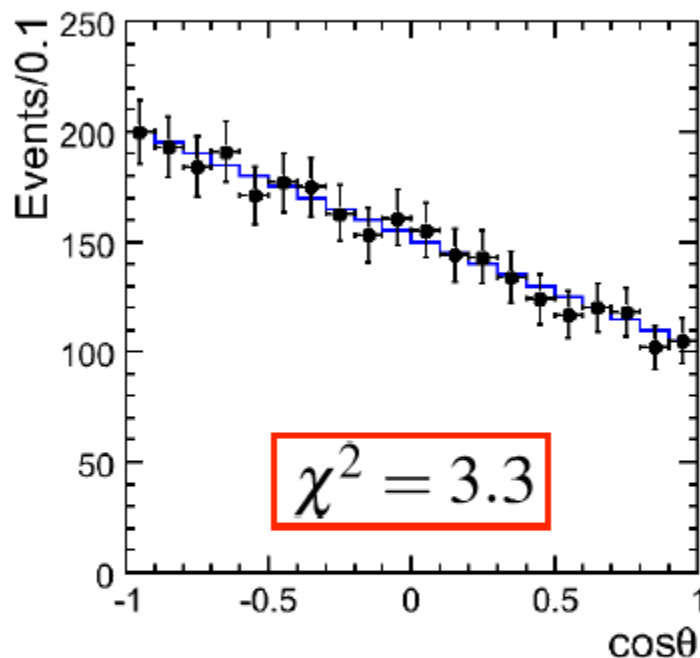
$$P(\chi^2 > 28.2; N = 20) = 10.4\%$$

Perfectly consistent: 10 % probability  
of getting a  $\chi^2$  worse than observed  
value by chance

$$P(\chi^2 > 39.1; N = 20) = 0.6\%$$

A bit dodgy: only 1 % probability  
of getting a  $\chi^2$  worse than observed  
value by chance

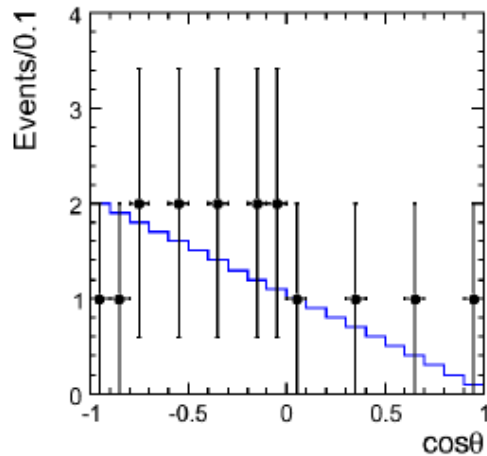
- What about a very small value of chi-squared, e.g.



- Expected  $\langle\chi^2\rangle = 20$
- Observed value is much smaller
- $P(\chi^2 > 3.3; N = 20) = 99.999\%$
- $P(\chi^2 < 3.3; N = 20) = 0.00001$
- Conclude 1/1000000 chance of getting such a small value: highly suspicious...
- What could be wrong:
  - Errors not estimated correctly
  - ...

# Log-Likelihood

- ★ The chi-squared-n distribution is just a re-expression of a Gaussian for N-variables
- ★ If the expected numbers of events are low, have to base consistency of data and prediction on Poisson statistics



- For the  $i$ th bin:

$$P_i(n_i; \mu_i) = \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!}$$

- Therefore the joint probability of obtaining **exactly** the observed  $\{n_i\}$ , i.e. the likelihood  $L$

$$L = \prod_i P_i = \prod_i \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!}$$

- Convenient to take the **natural logarithm** (hence log-likelihood)

$$\ln L = \sum_i \ln \left( \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \right)$$

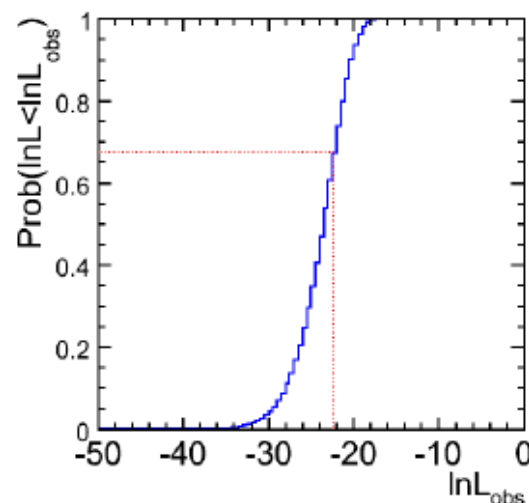
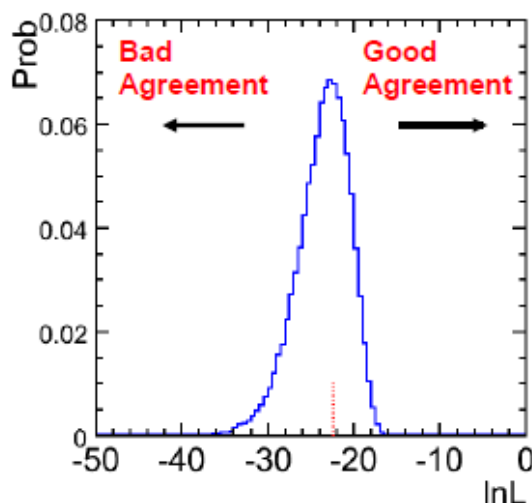
Poisson  
distributed  
variables

- ★ The likelihood is often **very small**. It is the probability of obtaining **exactly** the observed numbers of events in each bin

- For above distribution  $L = 2 \times 10^{-10}$  ( $\ln L = -22.4$ )

- ★ What constitutes a good value of log-likelihood ?
  - i.e. for the distribution shown, does  $\ln L = -22.4$  imply good agreement ?
- ★ There is no simple answer to this question
- ★ Unlike for the chi-squared distribution, there is **no general analytic form**
- ★ One practical way of assessing the consistency, is to generate many “toy MC” distributions according to expected distributions

```
Float_t LogLikelihood = 1.0;
TRandom2* r = new TRandom2();
for( Int_t i=1; i<= nbins; i++){
  Float_t expected = hist->GetBinContent(i);
  Int_t nObs = r->Poisson(expected);
  Double_t prob = TMath::Poisson(nObs,expected);
  LogLikelihood += log(prob);
}
```



- ★ Hence have obtain expected  $\ln L$  distribution for particular problem

# Relationship between chi-squared and likelihood

★ For Gaussian distributed variables

$$L(\{x_i\}) \propto e^{-\frac{\chi^2}{2}} \propto \prod_i \exp \left\{ -\frac{1}{2} \left[ \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right] \right\}$$

$$-\ln L = \frac{\chi^2}{2} + k$$

$$\chi^2 = -2 \ln L + \kappa$$

Chi-squared  $\times \frac{1}{2}$  is  $\ln L$  for Gaussian distributed variables

# Chi-Squared Fitting

★ Given some data and a prediction which depends on a set of parameters, we want to determine our best estimate for the parameters.

★ Construct the probability:  $P(\text{data}; \{x_i\})$

★ Best estimate is the set of parameters that maximises  $P(\text{data}; \{x_i\})$

## Simple Example:

- Two measurements of a quantity,  $x$ , using different methods

$$x = x_1 \pm \sigma_1 = 5.1 \pm 0.5 \quad x = x_2 \pm \sigma_2 = 6.0 \pm 0.3$$

(assume independent **Gaussian** errors)

- What is our best estimate of the true value of  $x$  ?

$$P(\text{data}; x) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[ \frac{(x_1 - x)^2}{\sigma_1^2} + \frac{(x_2 - x)^2}{\sigma_2^2} \right] \right\} = A e^{-\frac{\chi^2}{2}}$$

$$-\ln P = \frac{\chi^2}{2}$$

- Maximum probability corresponds to **minimum chi-squared**
- For **Gaussian** distributed variables: **fitting**  $\Rightarrow$  **minimising chi-squared**
- Here require

$$\frac{d\chi^2}{dx} = 0$$

$$\frac{d\chi^2}{dx} = 0 \quad \Rightarrow \quad -2 \frac{(x_1 - \bar{x})}{\sigma_1^2} - 2 \frac{(x_2 - \bar{x})}{\sigma_2^2} = 0$$

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

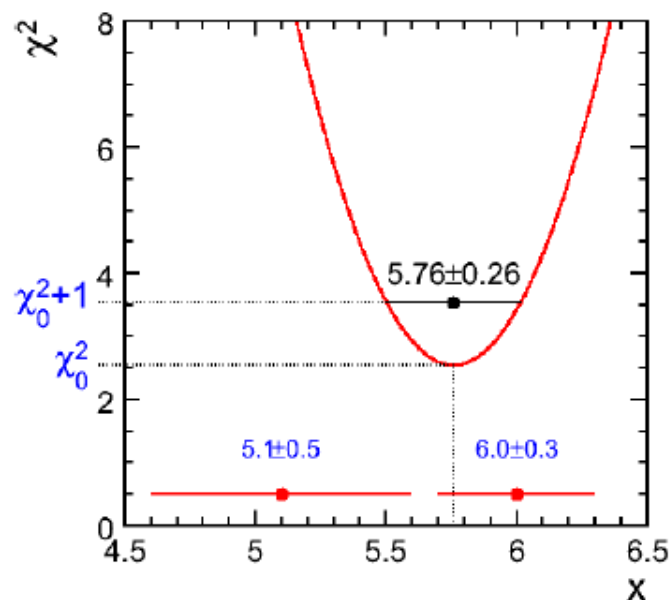
- Which is the formula we found previously for averaging two measurements
- **Note:** chi-squared is a quadratic function of the parameter  $x$
- Taylor expansion about minimum with

$$\frac{d^2\chi^2}{dx^2} = 2 \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \quad \text{and} \quad \frac{d^3\chi^2}{dx^3} = 0$$

gives  $\chi^2(x - \bar{x}) = \chi_0^2 + \frac{1}{2!} \frac{d^2\chi^2}{dx^2} (x - \bar{x})^2$

$$P(x) \propto e^{-\frac{\chi^2}{2}} = e^{-\frac{\chi_0^2}{2}} e^{-\frac{(x - \bar{x})^2}{2\sigma_x^2}}$$

therefore  $\sigma_x^2 = \left( \frac{1}{2} \left[ \frac{d^2\chi^2}{dx^2} \right]_{x=\bar{x}} \right)^{-1}$



- Started with two measurements of a quantity,  $x$

$$x = x_1 \pm \sigma_1 = 5.1 \pm 0.5 \quad x = x_2 \pm \sigma_2 = 6.0 \pm 0.3$$

- Best estimate of the true value of  $x$ , is that which maximises the likelihood, i.e. minimises chi-squared, giving a chi-squared at the minimum of:

$$\chi_0^2 = \left[ \frac{(x_1 - \bar{x})^2}{\sigma_1^2} + \frac{(x_2 - \bar{x})^2}{\sigma_2^2} \right]$$

- What is the probability that our data are consistent with a common mean?  
i.e. how do we interpret this chi-squared minimum

- Expanding and substituting for  $\bar{x}$

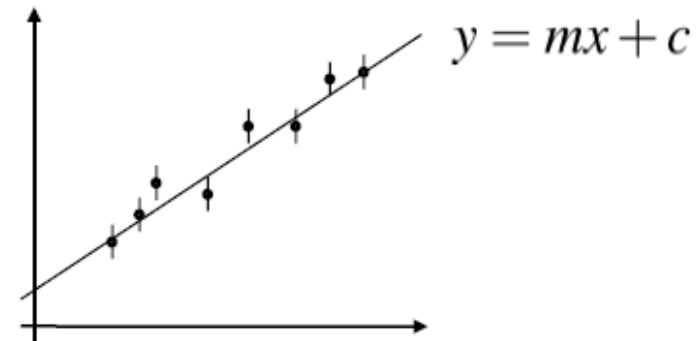
$$\begin{aligned} \chi_0^2 &= \frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} - \bar{x}^2 \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \\ &= \frac{(x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{(\Delta x)^2}{\sigma_{\Delta x}^2} \end{aligned} \quad \text{with} \quad \begin{cases} \Delta x = x_1 - x_2 \\ \sigma_{\Delta x}^2 = \sigma_1^2 + \sigma_2^2 \end{cases}$$

- Here the **minimum chi-squared corresponds** / distributed as chi-squared for a single Gaussian measurement, i.e. 1 degree of freedom

In general, the fitted **minimum chi-squared** is distributed as chi-squared for (number of measurements – number of fitted parameters) degrees of freedom

# Example: Straight line fitting

- ★ Given a set of points  $\{x_i, y_i \pm \sigma_i\}$  find the best fit straight line and the uncertainties



- ★ First define the **chi-squared**:  $\chi^2 = \sum_{i=0}^n \frac{(y_i - mx_i - c)^2}{\sigma_i^2}$

- ★ Minimise chi-squared with respect to the 2 parameters describing the model

$$\frac{\partial \chi^2}{\partial m} = \sum_{i=0}^n -2x_i \frac{(y_i - mx_i - c)}{\sigma_i^2} = 0 \quad \frac{\partial \chi^2}{\partial c} = \sum_{i=0}^n -2 \frac{(y_i - mx_i - c)}{\sigma_i^2} = 0$$

$$\Rightarrow \bar{m} \sum_{i=0}^n \frac{x_i^2}{\sigma_i^2} + \bar{c} \sum_{i=0}^n \frac{x_i}{\sigma_i^2} = \sum_{i=0}^n \frac{x_i y_i}{\sigma_i^2} \quad \bar{m} \sum_{i=0}^n \frac{x_i}{\sigma_i^2} + \bar{c} \sum_{i=0}^n \frac{1}{\sigma_i^2} = \sum_{i=0}^n \frac{y_i}{\sigma_i^2}$$

$$\Rightarrow \begin{pmatrix} s_{x^2} & s_x \\ s_x & s \end{pmatrix} \begin{pmatrix} \bar{m} \\ \bar{c} \end{pmatrix} = \begin{pmatrix} s_{xy} \\ s_y \end{pmatrix}$$

Where the sums are represented by

$$s_x = \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \quad \text{etc.}$$

- ★ For the **errors** first make Taylor expansion around the minimum **chi-squared**:

$$\chi^2(m, c) = \chi^2(\bar{m}, \bar{c}) + \frac{1}{2!} \frac{\partial^2 \chi^2}{\partial m^2} (m - \bar{m})^2 + \frac{1}{2!} \frac{\partial^2 \chi^2}{\partial m^2} (c - \bar{c})^2 + 2 \frac{1}{2!} \frac{\partial^2 \chi^2}{\partial m \partial c} (m - \bar{m})(c - \bar{c})^2$$

(since the function is quadratic there are no other terms)

which gives an elliptical contours

- ★ In terms of the **inverse error matrix**

$$\chi^2 = \mathbf{x}^T \mathbf{M}^{-1} \mathbf{x} \quad \text{with} \quad \mathbf{x} = \begin{pmatrix} m - \bar{m} \\ c - \bar{c} \end{pmatrix}$$

- ★ Hence

$$\mathbf{M}^{-1} = \begin{pmatrix} \frac{1}{2} \frac{\partial^2 \chi^2}{\partial m^2} & \frac{1}{2} \frac{\partial^2 \chi^2}{\partial m \partial c} \\ \frac{1}{2} \frac{\partial^2 \chi^2}{\partial m \partial c} & \frac{1}{2} \frac{\partial^2 \chi^2}{\partial c^2} \end{pmatrix} \quad (\mathbf{M}^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_i \partial a_j}$$

- ★ Here

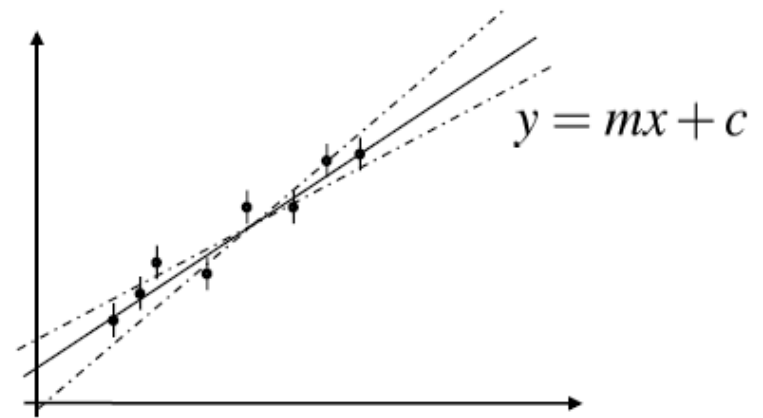
$$\mathbf{M}^{-1} = \begin{pmatrix} s_{x^2} & s_x \\ s_x & s \end{pmatrix}$$

$$\mathbf{M} = \frac{1}{s_{x^2}s - s_x^2} \begin{pmatrix} s & -s_x \\ -s_x & s_{x^2} \end{pmatrix} = \begin{pmatrix} \sigma_m^2 & \rho \sigma_m \sigma_c \\ \rho \sigma_m \sigma_c & \sigma_c^2 \end{pmatrix}$$

- ★ Giving

$$\rho = \frac{-s_x}{(ss_{x^2})^{\frac{1}{2}}} = - \frac{\sum \frac{x_i}{\sigma_i^2}}{\left( \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} \right)^{\frac{1}{2}}}$$

$$\rho = \frac{-s_x}{(ss_{x^2})^{\frac{1}{2}}} = - \frac{\sum \frac{x_i}{\sigma_i^2}}{\left(\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2}\right)^{\frac{1}{2}}}$$



- ★ Suppose we want to calculate the error on  $y$  as a function of  $x$  based on our fit:

$$\begin{aligned}\sigma_y^2 &= (x, 1) \begin{pmatrix} \sigma_m^2 & \rho \sigma_m \sigma_c \\ \rho \sigma_m \sigma_c & \sigma_c^2 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} \\ &= \sigma_m^2 x^2 + 2\rho x \sigma_m \sigma_c + \sigma_c^2\end{aligned}$$

$\frac{\partial y}{\partial m}$   
 $\frac{\partial y}{\partial c}$

- ★ It is worth noting that the correlation coefficient is proportional to  $\sum \frac{x_i}{\sigma_i^2}$ , hence one could fit after making the transformation  $x' = x - \bar{x}$  such that  $\sum \frac{x'_i}{\sigma_i^2} = 0$

and the uncertainties on the new intercept and gradient become **uncorrelated**

# Binned Maximum Likelihood Fits

- ★ So far only considered chi-squared fitting (i.e. assumes Gaussian errors)
- ★ In particle physics often dealing with low numbers of events and need to account for the Poisson nature of the data
- ★ In general can write down the joint probability

$$P(\text{data}; \text{parameters}) \equiv P(\{x_i\}; \{a_i\})$$

- e.g. if we predict a distribution should follow a first order polynomial

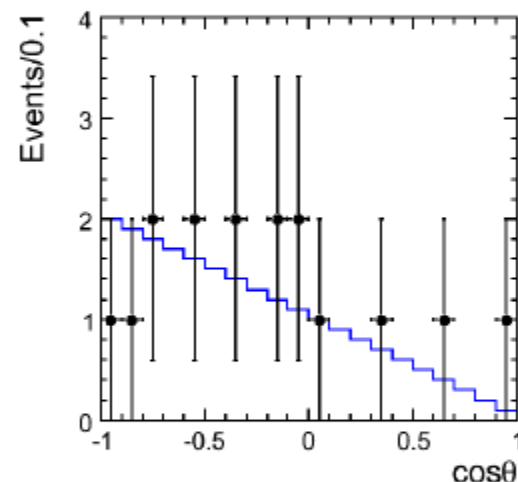
$$\mu_i = a_1 \cos \theta_i + a_0$$

- and measure events in bins of  $\cos \theta_i$
- define likelihood based on Poisson statistics

$$L = \prod_i P_i = \prod_i \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!}$$

- best estimates of parameters, defined by maximum likelihood or, equivalently, the minimum of -log-likelihood

$$\mathcal{L} = -\ln L = \sum_i -\ln \left( \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \right) = \sum_i \ln(n_i!) + \mu_i - n_i \ln \mu_i$$



- ★ Maximising the likelihood corresponds to solving the set of linear equations

$$\frac{\partial \mathcal{L}}{\partial a_i} = 0$$

- ★ To estimate the errors on the parameters, expand  $-\ln L$ , about its minimum

$$\mathcal{L} = \mathcal{L}(\bar{a}_i) + \frac{1}{2!} \sum_i (a_i - \bar{a}_i)^2 \frac{\partial^2 \mathcal{L}}{\partial a_i^2} + \frac{1}{2!} \sum_{i \neq j} (a_i - \bar{a}_i)(a_j - \bar{a}_j) \frac{\partial^2 \mathcal{L}}{\partial a_i \partial a_j} + O((a - \bar{a})^3)$$

- ★ Unlike the case of Gaussian errors, the

$$\frac{\partial^3 \mathcal{L}}{\partial a^3} \neq 0$$

- hence the resulting likelihood surface will not have a quadratic form

- ★ For the moment, restrict the discussion to a single variable

$$-\ln L = \mathcal{L} = \mathcal{L}(\bar{a}) + \frac{1}{2!} (a - \bar{a})^2 \frac{\partial^2 \mathcal{L}}{\partial a^2} + O((a - \bar{a})^3)$$

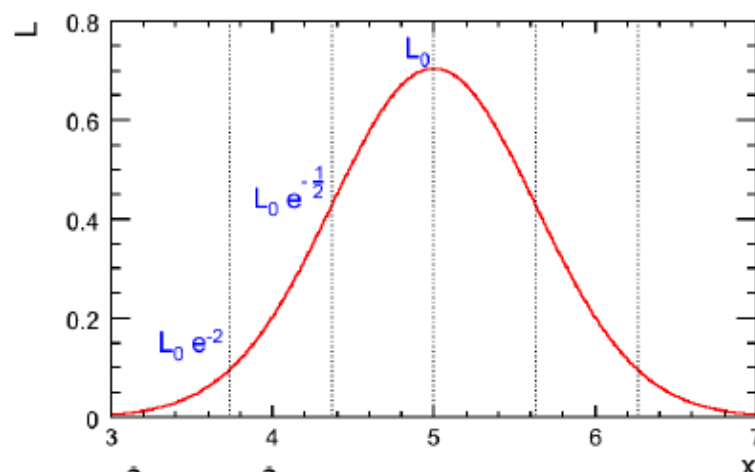
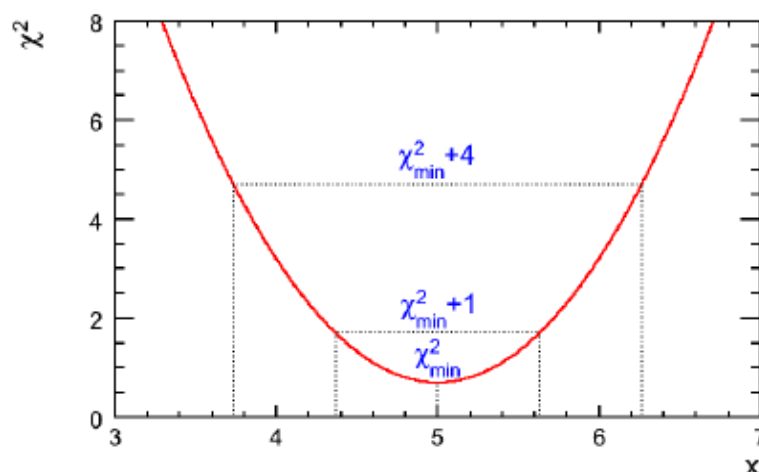
$$L = e^{-\mathcal{L}} = e^{-\mathcal{L}(\bar{a})} \times \underbrace{e^{-\frac{1}{2!} (a - \bar{a})^2 \frac{\partial^2 \mathcal{L}}{\partial a^2}}}_{\text{Gaussian}} \times \underbrace{e^{-O((a - \bar{a})^3)}}_{\text{Higher Order}}$$

- ★ If resulting likelihood distribution is “sufficiently Gaussian” could assign an estimate of the error on our measurement as:

$$\sigma^2 = \left( \frac{\partial^2 \mathcal{L}}{\partial a^2} \right)^{-1}$$

- ★ This is **OK in the Gaussian limit**, but in general it is not very useful
- ★ Usually adopt a Gaussian inspired procedure...
- ★ For Gaussian distributed variables, have a parabolic chi-squared curve which gives a Gaussian likelihood

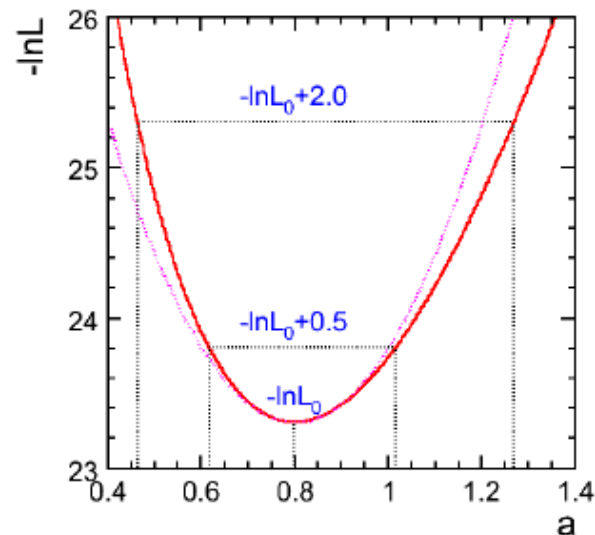
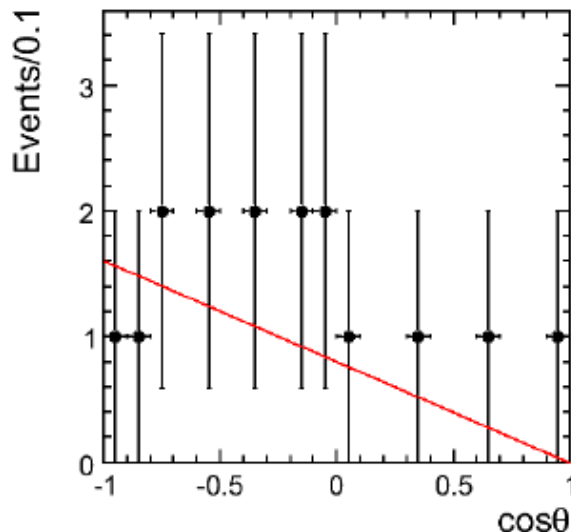
$$L = L_0 e^{-\frac{\chi^2}{2}} \quad \text{with} \quad L_0 = e^{-\frac{\chi_{min}^2}{2}}$$



- ★ “1 sigma errors” defined the points where  $\chi^2 \rightarrow \chi^2 + 1$
- ★ Or, **EQUIVALENTLY**, where the Relative Likelihood compared to the maximum likelihood decreases by  $e^{-1/2}$

# Example

e.g. assume a form  $\mu_i = a(1 - \cos \theta_i)$



★ Not unreasonable to use estimate of “1 sigma” uncertainty as the values:

$$-\ln L \rightarrow -\ln L + 0.5$$

Similarly for the “2 sigma” uncertainty:

$$-\ln L \rightarrow -\ln L + 2$$

★ At these points, have probabilities of  $e^{-1/2}$  and  $e^{-2}$  relative to maximum

★ BUT: no guarantee that 68 % of PDF lies within  $\pm 1\sigma$

Interpretation requires some care – a dark art (see later)

# Likelihood ratio

★ Safest way is to generate toy MC experiments, perform the fit, and thus obtain the **expected lnL distribution**

★ However, there is an invaluable trick

- For Poisson errors, we minimised the function

$$\mathcal{L} = -\ln L = -\sum_i \ln \left( \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \right) = \sum_i \ln(n_i!) + \mu_i - n_i \ln \mu_i$$

- Free to add a constant to this – doesn't affect the result
- Here **the data are fixed** and we vary the expectation
- Add the lnL of observing  $n_i$  given an expectation of  $n_i$

$$\begin{aligned} \mathcal{L} \rightarrow \mathcal{L} &= -\sum_i \ln \left( \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \right) + \boxed{\sum_i \ln \left( \frac{n_i^{n_i} e^{-n_i}}{n_i!} \right)} \equiv -\ln \left[ \frac{L(n_i; \mu_i)}{L(n_i; n_i)} \right] \\ &= \sum_i \ln n_i! + \mu_i - n_i \ln \mu_i - \ln n_i! - n_i + n_i \ln n_i \\ &= \sum_i \mu_i - n_i + n_i \ln \frac{n_i}{\mu_i} \\ &= \sum_i \frac{\mu_i^2 - \mu_i n_i}{\mu_i} + n_i \ln \left( 1 + \frac{n_i - \mu_i}{\mu_i} \right) \end{aligned}$$

Likelihood ratio

★ In the limit where the  $\mu_i$  are “not too small, in region of best fit  $\frac{n_i - \mu_i}{\mu_i}$  is small

$$\begin{aligned}
 \mathcal{L} &= \sum_i \frac{\mu_i^2 - \mu_i n_i}{\mu_i} + n_i \ln \left[ 1 + \left( \frac{n_i - \mu_i}{\mu_i} \right) \right] \\
 &= \sum_i \frac{\mu_i^2 - \mu_i n_i}{\mu_i} + n_i \left( \frac{n_i - \mu_i}{\mu_i} \right) - \frac{n_i}{2} \left( \frac{n_i - \mu_i}{\mu_i} \right)^2 + O \left\{ n_i \left( \frac{n_i - \mu_i}{\mu_i} \right)^3 \right\} \\
 &\approx \sum_i \frac{\mu_i^2 - \mu_i n_i + n_i^2 - \mu_i n_i}{\mu_i} - \frac{n_i}{2} \left( \frac{n_i - \mu_i}{\mu_i} \right)^2 \\
 &= \sum_i \frac{(n_i - \mu_i)^2}{\mu_i} - \frac{n_i}{2} \left( \frac{n_i - \mu_i}{\mu_i} \right)^2 \\
 &= \sum_i \frac{(n_i - \mu_i)^2}{\mu_i} \left( 1 - \frac{n_i}{2\mu_i} \right) \quad \text{with } \langle n_i \rangle = \mu_i \\
 &\approx \frac{1}{2} \sum_i \frac{(n_i - \mu_i)^2}{\mu_i} \quad \text{For Poisson distributed variables } \sigma_i^2 = \mu_i \\
 \mathcal{L} &= -\ln \lambda = \frac{\chi^2}{2} \quad \text{with } \lambda = \frac{L(n_i; \mu_i)}{L(n_i; n_i)}
 \end{aligned}$$

★ Hence  $-2 \ln \lambda$  is distributed as  $\chi^2$  in the limit of “large” n

- ★ This is a very useful trick.
- ★ When fitting a histogram with Poisson errors

**ALWAYS**

- Perform a **maximum likelihood**, not a chi-squared fit
- Use the likelihood ratio

$$-2 \ln \lambda = \sum_i \mu_i - n_i + n_i \ln \frac{n_i}{\mu_i}$$

**NOTE**

- Best fit parameters determined by

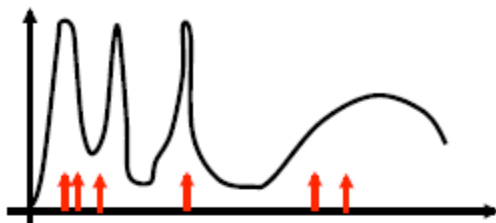
$$\frac{\partial[-2 \ln \lambda]}{\partial a_i} = \sum_i \left(1 - \frac{n_i}{\mu_i}\right) \frac{\partial \mu_i}{\partial a_i} = 0 \quad \mu_i(\{a_i\})$$

- At the best fit point  $-2 \ln \lambda$  tends to a chi-squared distribution for ***n-m*** d.o.f.

# Unbinned Maximum Likelihood Fits

- ★ For some applications, binning data results in a loss of precision

- e.g. sparse data and a rapidly varying prediction



- ★ In other cases there is simply no need to bin data: **unbinned maximum likelihood**

**NOTE:** this is a “shape-only” fit: normalisation doesn’t enter

- ★ Suppose we can construct the predicted **PDF** for the data as a function of the parameter of interest

- e.g. make N measurements of decay time,  $\{t_i\}$ , and want to estimate lifetime  $\tau$

- Write down **NORMALISED PDF**

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

- We can now write down the likelihood of obtaining our set of data

$$L(\{t_i\}) = \prod_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}}$$

- Obtain lifetime by maximising likelihood (or equivalently  $\ln L$ )

$$\begin{aligned}\ln L &= -\sum_i \frac{t_i}{\tau} + \ln \tau \\ \frac{\partial \ln L}{\partial \tau} &= \frac{\sum_i t_i}{\tau^2} - \frac{N}{\tau} \\ \Rightarrow \tau &= \frac{1}{N} \sum_i t_i\end{aligned}$$

**i.e. the expected, but not entirely obvious result**

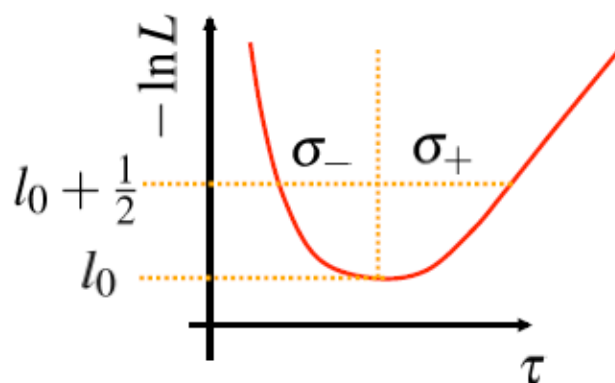
- **For the error estimate take the second derivative**

$$\begin{aligned}\frac{\partial^2 \ln L}{\partial \tau^2} &= -\frac{2}{\tau^3} \sum_i t_i + \frac{N}{\tau^2} \\ &= -\frac{N}{\tau^2} \\ \sigma_\tau &= \left( -\frac{\partial^2 \ln L}{\partial \tau^2} \right)^{-1} \\ &= \frac{\tau}{N^{\frac{1}{2}}}\end{aligned}$$

- **Since we now know what we are doing, it is immediately obvious that the error is not symmetric**

$$\frac{\partial^3 \ln L}{\partial \tau^3} \neq 0$$

- **The likelihood function is**



- **Usual to quote with asymmetric errors, e.g.**

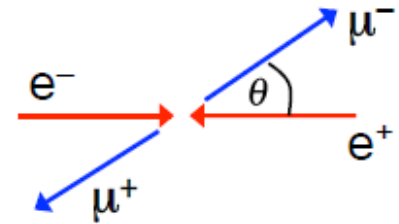
$$\tau = 1.0_{-0.4}^{+0.6}$$

# Example

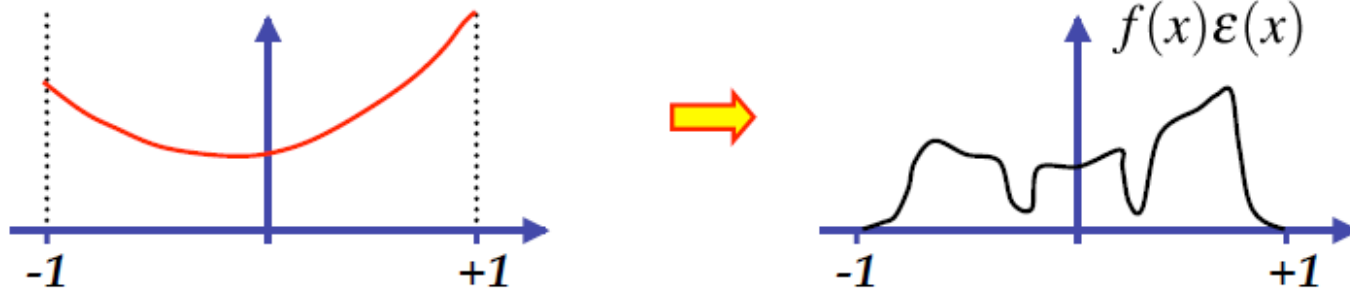
## ★ Forward-Backward asymmetry at LEP

- Expected angular distribution of form

$$f(x) \propto (1 + x^2 + \frac{3}{4}Ax) \quad x = \cos \theta$$



- But measured angular distribution depends on efficiency  $\varepsilon(x)$



- However  $\varepsilon(x)$  is not known (at least not precisely)
- But it is known to be symmetric
- Sufficient to construct an unbinned maximum likelihood fit
- First write down the PDF in terms of the unknown efficiency

$$P(x) \propto f(x)\varepsilon(x) \propto \varepsilon(x)(1 + x^2 + \frac{3}{4}Ax)$$

- For unbinned maximum likelihood fit **PDF must be normalised**

$$\int_{-1}^{+1} f(x)\varepsilon(x)dx = 1$$

- Since  $\varepsilon(x)$  is symmetric normalisation gives

$$\int_{-1}^{+1} (1+x^2)\varepsilon(x)dx = 1 \quad \text{i.e. independent of } A$$

- Hence the normalised PDF is of the form

$$P(x) = \kappa \varepsilon(x) (1 + x^2 + \frac{3}{4}Ax)$$

- For the N observed values  $\{x_i\}$  the log-likelihood function becomes:

$$\ln L = \sum_i \ln \kappa + \ln \varepsilon(x_i) + \ln \left( 1 + x_i^2 + \frac{3}{4}Ax_i \right)$$

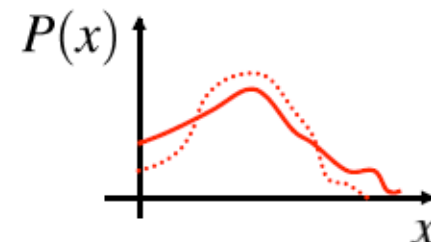
- For a maximum 
$$\frac{\partial \ln L}{\partial A} = \frac{3}{4} \sum_i \frac{x_i}{1 + x_i^2 + \frac{3}{4}Ax_i} = 0$$

- Which can be solved (preferably minimised by MINUIT) despite the fact we don't know the precise form of the PDF

# Extended Maximum Likelihood Fits

- ★ **Unbinned maximum likelihood** uses only shape information
- ★ In the **extended maximum likelihood fit** include normalisation
- ★ Suppose you observe  $n$  events with  $x_i$  and expect a total of  $\mu$  with **PDF** which is a function of some parameter you wish to measure  $P(x)$

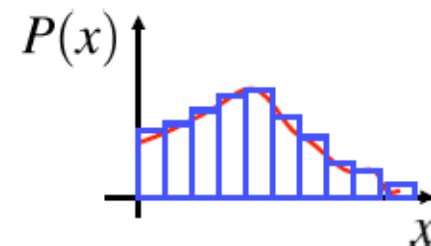
$$L(\{x_i\}) = \underbrace{\frac{e^{-\mu} \mu^n}{n!}}_{\text{Poisson}} \times \underbrace{\prod_i P(x_i)}_{\text{Unbinned ML}}$$



$$\ln L = -\mu + n \ln \mu - \ln n! + \sum_i \ln P(x_i)$$

- ★ Just for fun... suppose our **PDF is binned** with an expectation of  $\mu_j/\mu$  in each bin

$$\begin{aligned} \ln L &= -\mu + n \ln \mu - \ln n! + \sum_i \ln \frac{\mu_j}{\mu} \\ &= -\mu + n \ln \mu - \ln n! - \sum_i \ln \mu + \sum_i \ln \mu_j \\ &= -\mu - \ln n! + \sum_i \ln \mu_j \end{aligned}$$



- if there are  $n_j$  events observed in each bin

$$\ln L = -\mu - \ln \mu! + \sum_j n_j \ln \mu_j$$

$$\ln L = -\mu - \ln n! + \sum_j n_j \ln \mu_j$$

$$= \sum_j \mu_j - \ln n! + \sum_j n_j \ln \mu_j$$

$$= \sum_j \ln \left( \frac{e^{-\mu_j} \mu_j^{n_j}}{n_j!} \right) - \ln \frac{n!}{\sum_j n_j!}$$

For a given data set  
this is a constant

Binned Poisson  
Likelihood

Can you see where  
this comes from ?

Hence our previous expression for “**Binned Maximum likelihood**”  
is just an **Extended Maximum Likelihood** fit with a binned PDF

# Fitting Summary

- ★ Have covered main fitting/goodness of fit issues:
  - definition of chi-squared
  - chi-squared fitting
  - definition of likelihood functions and relation to chi-squared
  - likelihood fitting techniques
- ★ Next we will consider more carefully the **interpretation**

# Appendix: The Error Function

- ★ For a single variable the **Chi-Squared Probability**

$$P(\chi^2 < \chi_0^2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-x_-}^{+x_+} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \quad x_{\pm} = \mu \pm \chi_0 \sigma$$

- ★ Change variable  $\chi = \frac{x-\mu}{\sigma}$

$$P(\chi^2 < \chi_0^2) = \frac{1}{\sqrt{2\pi}} \int_{-\chi_0}^{+\chi_0} \exp\left\{-\frac{\chi^2}{2}\right\} d\chi$$

- ★ Change variable again  $t^2 = \chi^2/2$  and integrate over positive values only

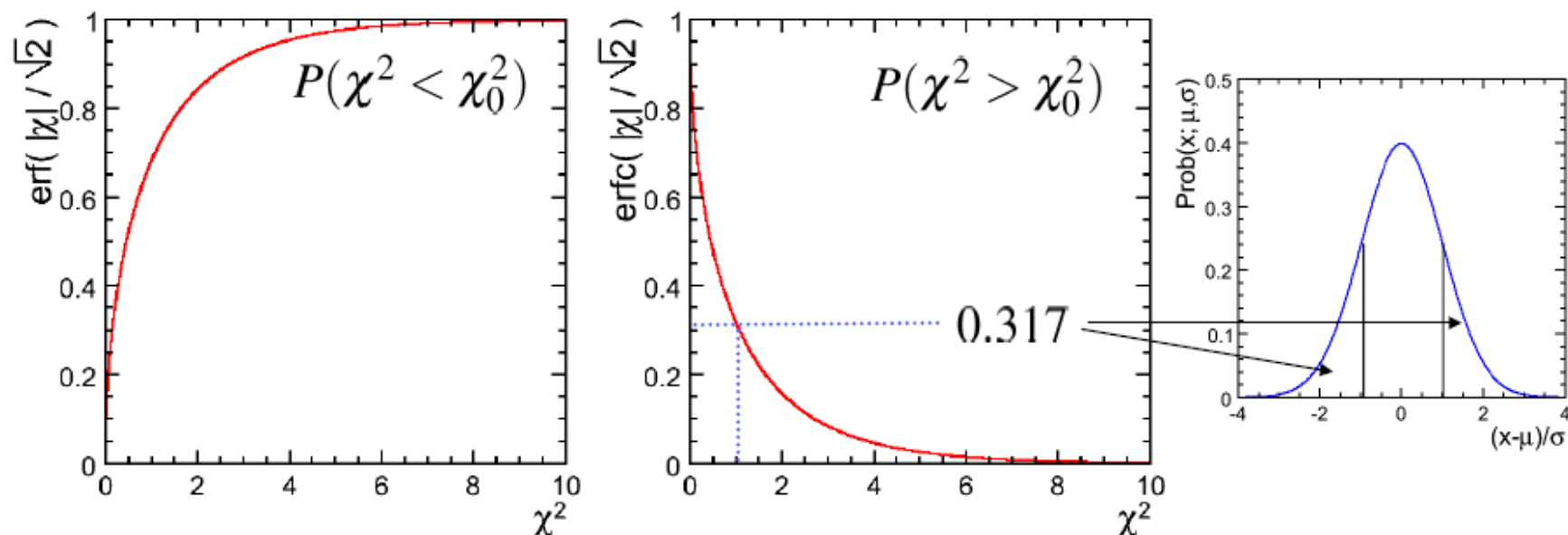
$$\begin{aligned} P(\chi^2 < \chi_0^2) &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\chi_0} e^{-\frac{\chi^2}{2}} d\chi \\ &= \frac{2}{\sqrt{\pi}} \int_0^{\chi_0/\sqrt{2}} e^{-t^2} dt \\ &= \text{erf}\left(\frac{\chi}{\sqrt{2}}\right) \end{aligned}$$

Error function

- ★ The probability of obtaining a value of  $\chi^2 > \chi_0^2$  by chance is:

$$P(\chi^2 > \chi_0^2) = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\chi_0/\sqrt{2}} e^{-t^2} dt = \text{erfc}\left(\frac{\chi}{\sqrt{2}}\right)$$

Complement of the Error Function



- ★ This is nothing more than a different way of expressing a 1D Gaussian distribution (or more correctly its two-sided integral)