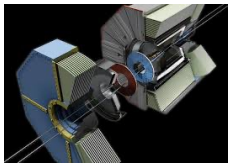
Probability and Statistics: some basics for HEP

Alan Schwartz
University of Cincinnati

US Belle II Summer School
Virginia Polytechnic Institute and State University
(Virginia Tech)
25 June 2025

- *definitions: probability, statistics, PDF, confidence interval, upper limit, p-value*
- *common PDFs: Gaussian, Poisson, Exponential*
- *the χ^2 distribution, cumulative χ^2 distribution, goodness-of-fit*
- *fitting data: method of maximum likelihood*
- *fitting data: method of least squares (χ^2 fit)*
- *some references*



Some definitions I

Probability *and* ***statistics*** *have different meanings:*

underlying probability A	➡	*observed result B*
what is the underlying probability A?	⬅	*observation B*

Example:

the probability of heads is 50% ➡ *flipping a coin 10 times gives 4 heads, 6 tails*

you observe 4 heads in 10 flips ➡ *best guess of probability of heads is 0.40*
(the "point estimate")

Blue *is probability,* ***green*** *is statistics. If you measure 4 heads in 10 flips, avoid saying "the most probable fraction of heads is 40%" – you should say the most* ***likely*** *fraction is 40%*

Probability Density Function (PDF):

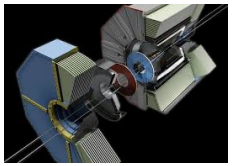
given a PDF $\mathcal{P}(x)$, the ***probability*** *of obtaining x in the interval (a,b) is the integral* $\int_a^b \mathcal{P}(x)\,dx$

Expectation value or Mean: $\mu \equiv \int_{-\infty}^{+\infty} x\,\mathcal{P}(x)\,dx$

Dispersion around the mean or Variance: $V \equiv \int_{-\infty}^{+\infty} (x-\mu)^2\,\mathcal{P}(x)\,dx$

Standard deviation: $\sigma \equiv \sqrt{V}$ or $\sigma^2 = V$

(for a Gaussian distribution, $\pm\sigma$ contains 68.3% of the area)



Some definitions II

90% confidence interval:

*Given an observation B, the **likelihood** that the true value of α lies in the interval (a,b) is 0.90.*

68.3% confidence interval:

*Given an observation B, the **likelihood** that the true value of α lies in the interval (a,b) is 0.683.*

Likelihood function:

The function $\mathcal{L}(x)$ that expresses the likelihood of α having a specific true value.

For a 90% confidence interval (a,b): $\int_a^b \mathcal{L}(x)\,dx = 0.90$

For a 68.3% confidence interval (a,b): $\int_a^b \mathcal{L}(x)\,dx = 0.683$

Error bars:

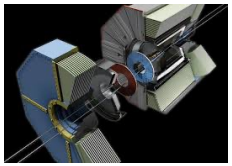
Suppose the most likely value of α is z, and the 68.3% confidence interval is (a,b). We quote both of these results together with the expression: $\alpha = z^{+(b-z)}_{-(z-a)}$

Example: if z = 5 and the 68.3% confidence interval is (1,8), then we write: $\alpha = 5^{+3}_{-4}$

90% confidence level upper limit ξ – Bayesian: $\int_{-\infty}^{\xi} \mathcal{L}(x)\,dx = 0.90$

90% confidence level upper limit ξ – frequentist: $\int_{-\infty}^{B} \mathcal{P}(x|\mu = \xi)\,dx = 0.10$

(for observation B)



Some definitions III

90% confidence level upper limit ξ – Bayesian: $\int_{-\infty}^{\xi} \mathcal{L}(x)\, dx = 0.90$

90% confidence level upper limit ξ – frequentist: $\int_{-\infty}^{B} \mathcal{P}(x|\mu = \xi)\, dx = 0.10$

(for observation B)

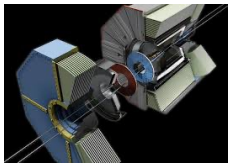
In words: *the frequentist upper limit is the value ξ which, if nature's true value, would give a probability of 10% of measuring a value B or smaller.*

How does one determine this? For simple situations, i.e., known background, one can calculate ξ using the Poisson PDF (see Neyman construction, Zech, Feldman-Cousins). For complicated situations, use toy MC.

Coverage: *if the upper or lower limit holds for a statistical ensemble, we say "the limit provides coverage"* *(under-coverage: the limit is not true for an ensemble)*

p-value: $p = \int_{-\infty}^{B} \mathcal{P}(x|\mu = \xi)\, dx$ *(if very small, observation B is suspiciously low)*

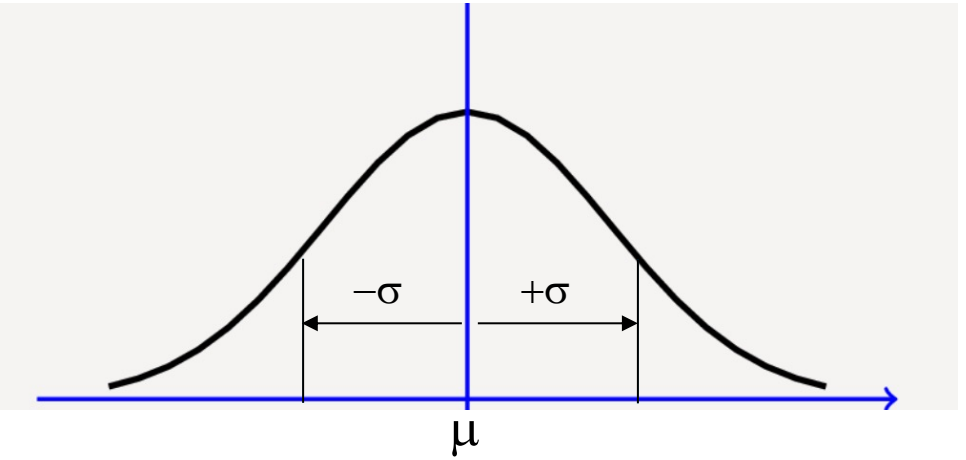
$p = \int_{B}^{\infty} \mathcal{P}(x|\mu = \xi)\, dx$ *(if very small, observation B is suspiciously high)*



Some common PDFs I

Gaussian distribution:

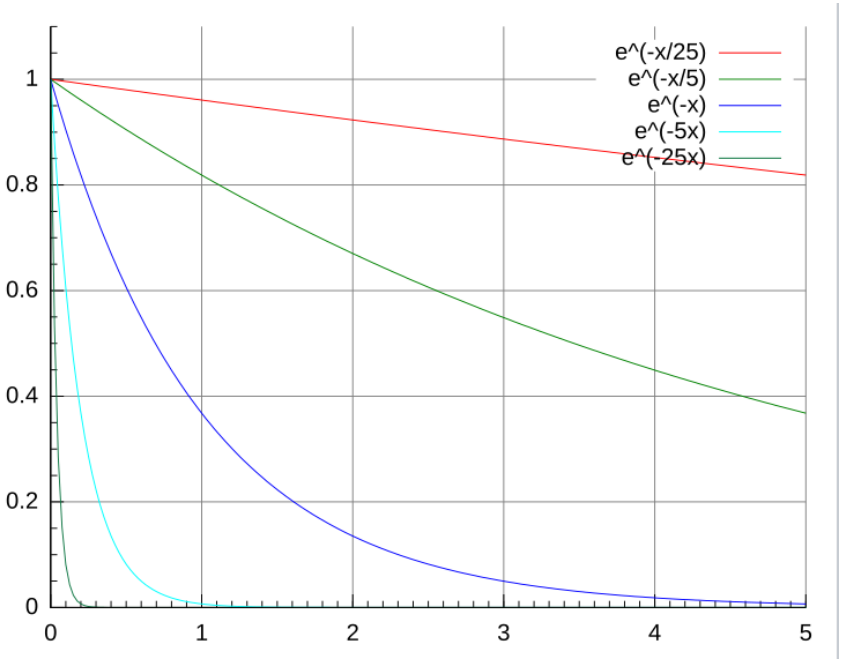
$$\mathcal{P}(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)}$$



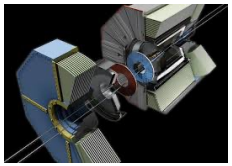
- *68.3% of the likelihood is within ±1σ*
- *"standard normal": μ = 0, σ = 1*
- *Gaussian distributions have a special importance due to the Central Limit Theorem: the sum of a large number of deviations about mean values is distributed according to a Gaussian distribution with mean $\mu_1 + \mu_2 + \mu_3 + \ldots$ and variance $V = V_1 + V_2 + V_3 \ldots$ [or $\sigma = \sqrt{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 \ldots)}$] regardless of what distributions the individual deviations follow*

Exponential function (e.g., particle lifetime):

$$\mathcal{P}(x) = \frac{1}{\tau}\, e^{-x/\tau}$$



- *mean value μ = τ , but most probable value is 0*
- *cumulative distribution is the same exponential*
- *has unusual property that shifting the distribution horizontally is equivalent to shifting it vertically*

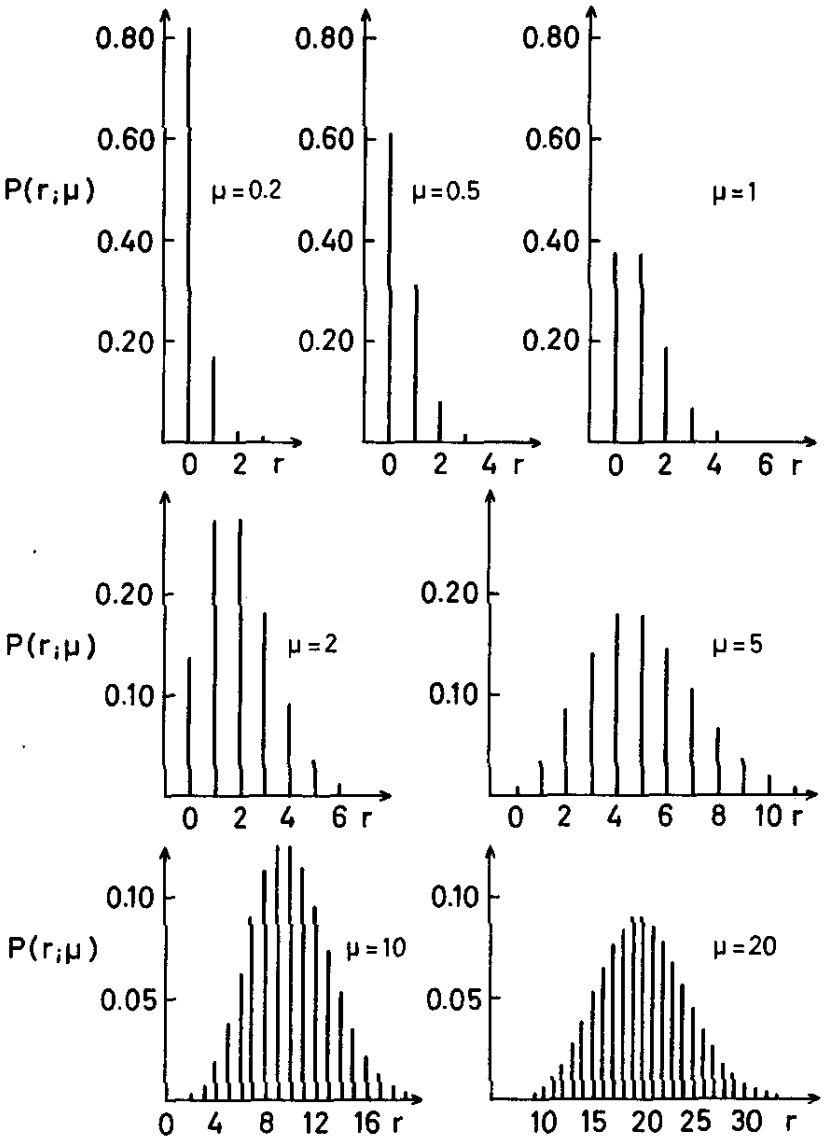


Some common PDFs II

Poisson distribution (discrete value):

$$\mathcal{P}(N;\mu) = \frac{1}{N!}\,\mu^N e^{-\mu}$$

- μ is a real number, but N is an integer
- most likely value for N is the integer nearest μ
- for μ large (>15), Poisson distribution → Gaussian distribution with mean μ and variance μ ($\sigma = \sqrt{\mu}$)

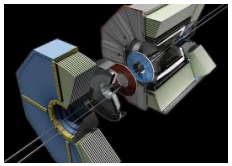


Poisson errors:

Suppose one observes N events; what is nature's mean value μ? (i.e., to calculate a branching fraction). Obviously, the best guess for μ is N; but it's also useful to give a confidence interval, i.e., $\mu \in (a,b)$ at 68.3% CL. The convention (frequentist):

- a is value of μ for which there is 15.87% probability of observing $\geq$ N events
- b is value of μ for which there is 15.87% probability of observing $\leq$ N events

Values a, b obtained from tables or online calculators



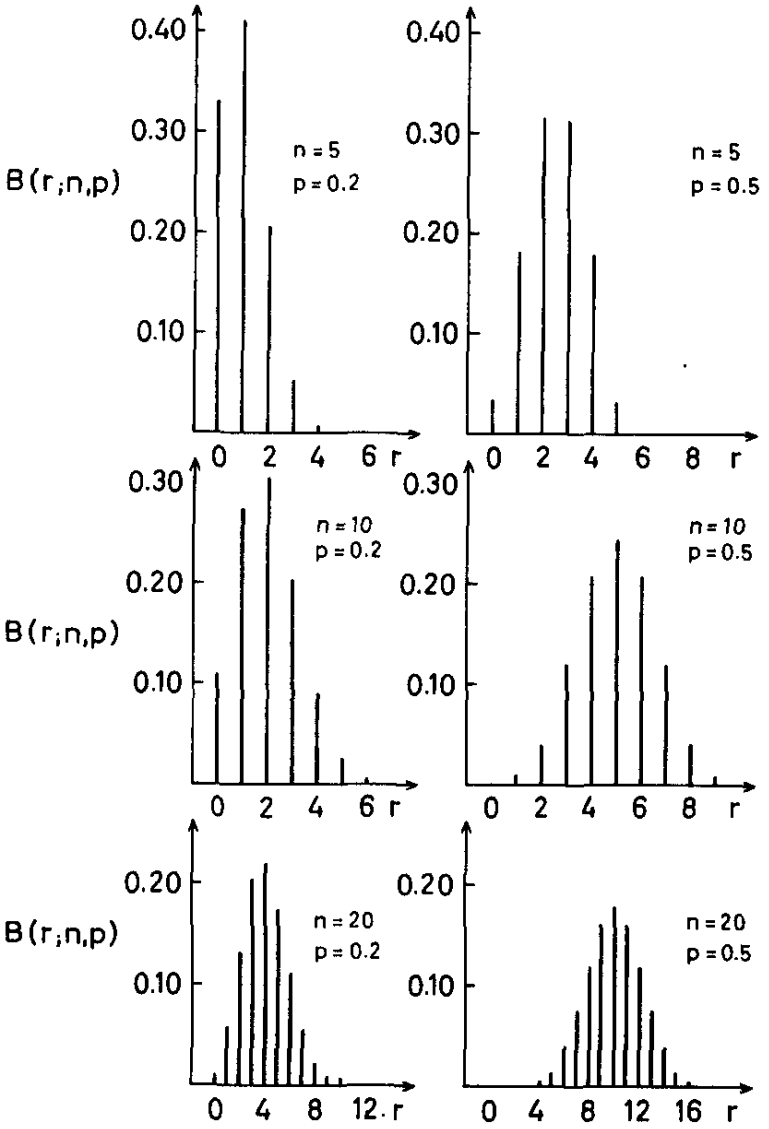
Some common PDFs III

Binomial distribution (discrete value):

$$\mathcal{P}(m;N,p) = \left(\begin{array}{c} N \\ m \end{array} \right) p^m q^{N-m} \qquad (p+q=1)$$

binomial coefficient: $\left(\begin{array}{c} N \\ m \end{array} \right) \equiv \frac{N!}{m!\,(N-m)!}$

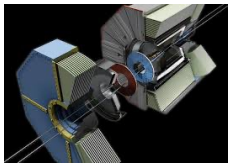
- *p, q are probabilities of success and failure, i.e., p + q = 1 ("heads" and "tails")*
- *N, m are integers: m is the subset of N with sucess ("heads")*
- *Variance of this distribution is Npq*



Binomial errors:

Suppose one observes m successes out of N trials (i.e., an efficiency ε = m/N); what is nature's probability of success p? Obviously, the best guess for p is m/N; but it's also useful to give a confidence interval, i.e., p $\in$(a,b) at 68.3% CL. The convention:

- *$\pm$ the standard deviation = $\pm\sqrt{(Npq)}$ is quoted as the 68.3% confidence interval, since it contains 68.3% of the probability. Note: probability $\rightarrow$ likelihood is usually forbidden but here can be shown to provide coverage*



Some common PDFs IV

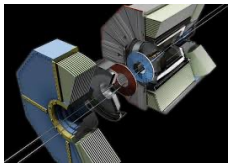
Relation between Poisson distribution and Binomial distribution:

Suppose you have a sample of N particles, f of which decay "forwards," b = N – f of which decay "backwards." Let p= probability of forward decay, q = 1-p = probability of backwards decay. Since f is a subset of N, the distribution of forward decays follows a binomial distribution. If we now release the requirement of fixed N, i.e., N follows a Poisson distribution with mean μ, then $\mathcal{P}$ becomes a joint probability as follows:

$$\begin{aligned}
\mathcal{P}(f;N,p) &= \frac{N!}{f!\,b!}\,p^f q^b \qquad (\text{note: } f+b=N;\ p+q=1) \\
&\rightarrow \frac{N!}{f!\,b!}\,p^f q^b \left(\frac{1}{N!}\,\mu^N e^{-\mu}\right) \\
&= \left(\frac{1}{f!}\,(p\mu)^f e^{-p\mu}\right)\left(\frac{1}{b!}\,(q\mu)^b e^{-q\mu}\right)
\end{aligned}$$

The last expression is the product of two Poisson distributions: one for forward decays with mean $p\mu$, and one for backwards decays with mean $q\mu$. This illustrates the points:

- *binomial distribution + release fixed N $\rightarrow$ Poisson or*
- *Poisson from a fixed number of trials $\rightarrow$ binomial*



Other important distributions

The χ^2 distribution:

Suppose we sum n random variables distributed with mean μ and variance σ^2:

$$\chi^2 \equiv \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

This variable will be distributed as:

$$f(\chi^2; n) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \left(\chi^2\right)^{n/2-1} e^{-\chi^2/2}$$

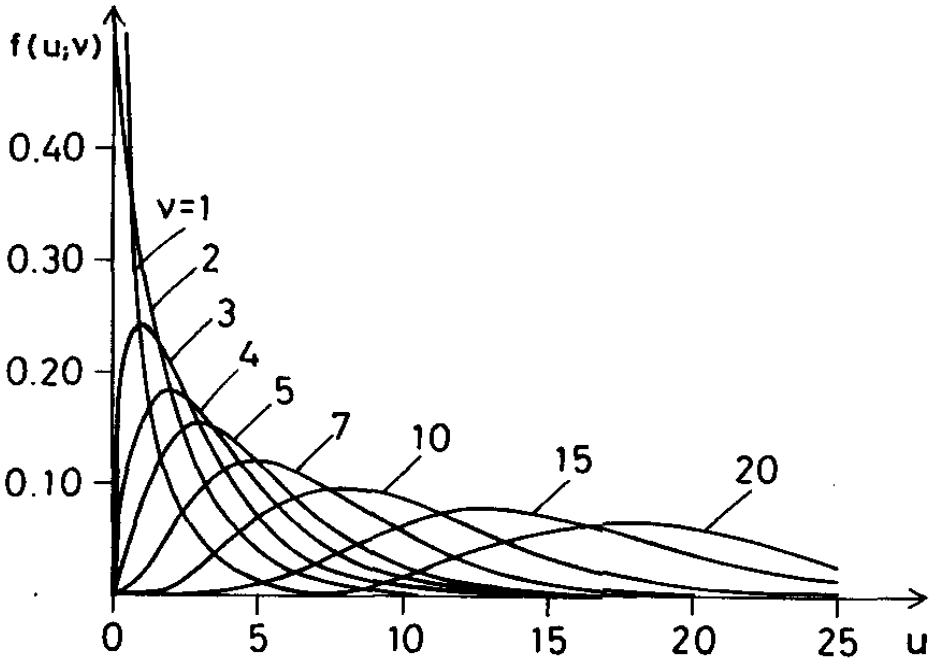
This distribution is referred to as "the χ^2 distribution for n degrees of freedom"

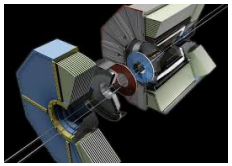
- *If μ is unknown, which is often the case, one constructs the variable*

$$\chi^2 \equiv \sum_{i=1}^{n} \left(\frac{x_i - \langle x \rangle}{\sigma} \right)^2$$

which is distributed according to a χ^2 distribution with (n-1) degrees of freedom.

- *For n large, χ^2 distribution $\rightarrow$ Gaussian distribution with mean n and V = 2n*
  [or $\sigma = \sqrt{(2n)}$]





Other important distributions

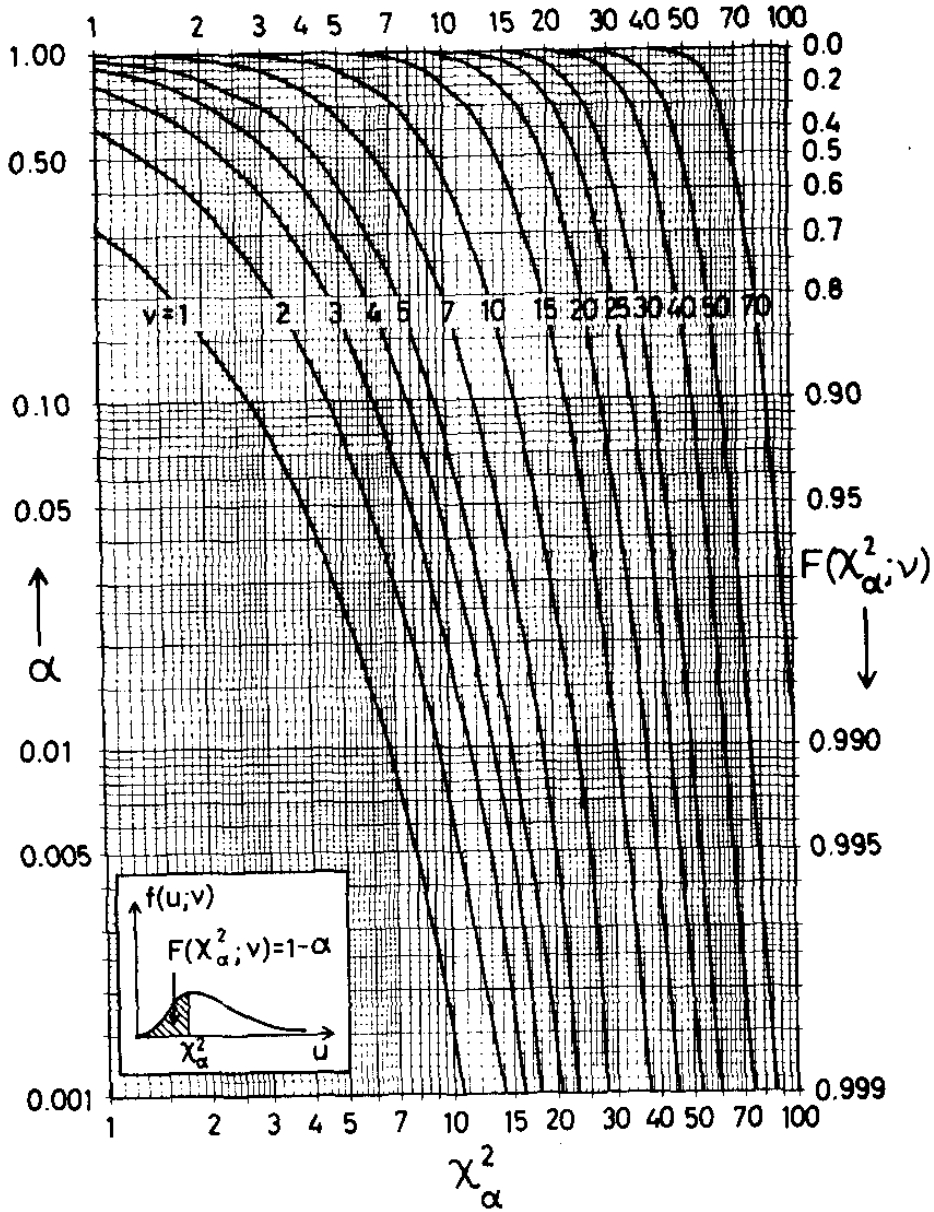
The cumulative χ^2 distribution:

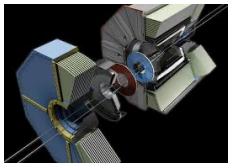
Suppose you have a set of observables x_i and want to see if they are consistent with a mean μ and variance σ^2. You calculate χ^2 as

$$\chi^2 \equiv \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

and see if the value is consistent with that expected from the χ^2 distribution. I.e., if the value is very high or very low, the guessed μ and σ are probably wrong. To make a quantitative statement of how likely the obtained χ^2 value is, one needs to know the probability content of the χ^2 distribution, i.e., the fractional area under parts of it. This is called the cumulative χ^2 distribution. It is often presented as a plot, but there are standard routines and online calculators for calculating it.

- *From right-most plot, χ^2 =7 for n=2 has α = 0.03 ⇒ 3% of ensemble is above this (⇒rare). For n =7, α = 0.40 ⇒ 40% of ensemble is above this (⇒not rare)*

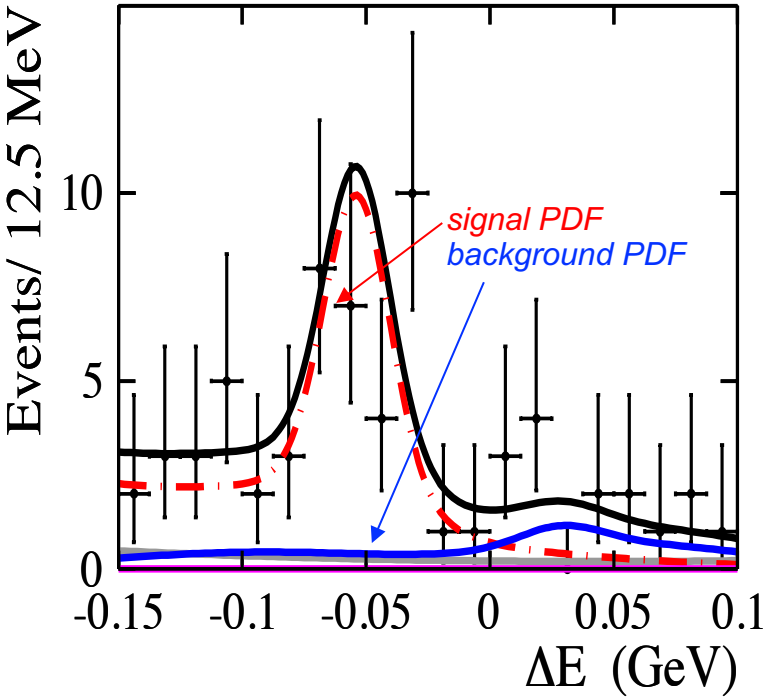




Fitting data

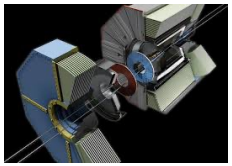
We often fit data to a PDF hypothesis, to determine signal and background yields. Three variations of this:

- *all parameters of PDF are fixed (e.g., from MC), and the resulting χ^2 ("goodness-of-fit") tells us if the assumed PDF is likely or unlikely, i.e., describes the data well.*
- *parameters of PDF are fixed to MC but adjusted ("calibrated") to reduce data/MC differences for a control sample; χ^2 tells us if this adjusted PDF describes the data*
- *analytic form (shape) of PDF is taken from MC, but parameters are floated in the fit. Fitter determines "best" values of the parameters, χ^2 tells us if this fitted PDF describes the data well. How is this fit performed?*



Two common methods:

- *Method of maximum likelihood* — *binned ML fit*, *unbinned ML fit*, *extended unbinned ML fit*
- *Method of least squares ("χ^2 fit")*



Method of maximum likelihood I

Suppose we have a set of n measurements x_1, x_2, x_3...x_n and a set of parameters θ_j we want to determine. We construct a likelihood function $\mathcal{L}$ as:

$$\mathcal{L}(\vec{x};\vec{\theta}) = \prod_{i=1}^{n} \mathcal{P}(x_i|\vec{\theta})$$

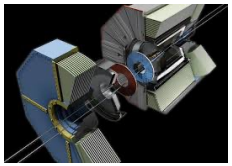
where $P(x_i|\vec{\theta})$ is the probability of measuring x_i given parameters $\vec{\theta}$. Thus, $\mathcal{L}$ is the joint conditional probability of measuring x_1, x_2, x_3...x_n for a fixed set of θ_j. The most likely values of θ_j are those that maximize $\mathcal{L}$. Computationally, as $\mathcal{L}$ can be quite small, we usually maximize the logarithm of $\mathcal{L}$:

$$\ln \mathcal{L}(\vec{x};\vec{\theta}) = \sum_{i=1}^{n} \ln \mathcal{P}(x_i|\vec{\theta})$$

(sum, instead of product)

$$\Rightarrow \quad \frac{\partial \ln \mathcal{L}}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \ln \mathcal{P}}{\partial \theta_j} = 0$$

$$\text{for} \quad \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_j^2} = \sum_{i=1}^{n} \frac{\partial^2 \ln \mathcal{P}}{\partial \theta_j^2} < 0$$



Method of maximum likelihood II

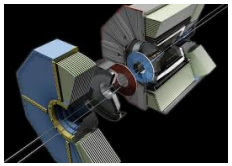
Example: fit for lifetime τ

$$
\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial \tau} &= \frac{\partial}{\partial \tau} \sum_{i=1}^{n} \ln \left(\frac{1}{\tau} e^{-t_i/\tau} \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \tau} \ln \left(\frac{1}{\tau} e^{-t_i/\tau} \right) \\
&= \sum_{i=1}^{n} \tau \, e^{t_i/\tau} \left[-\frac{e^{-t_i/\tau}}{\tau^2} + \frac{e^{-t_i/\tau}}{\tau} \left(\frac{t_i}{\tau^2} \right) \right] \\
&= \sum_{i=1}^{n} \left[-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right] \\
&= -\frac{n}{\tau} + \frac{1}{\tau^2} \sum_{i=1}^{n} t_i = 0 \\
\Rightarrow \quad n &= \frac{1}{\tau} \sum_{i=1}^{n} t_i \\
\Rightarrow \quad \tau &= \frac{1}{n} \sum_{i=1}^{n} t_i = \langle t \rangle
\end{aligned}
$$

so most-likely value is mean value

$$
\begin{aligned}
\left. \frac{\partial^2 \ln \mathcal{L}}{\partial \tau^2} \right|_{\tau = \langle t \rangle} &= \left. \frac{\partial}{\partial \tau} \left(-\frac{n}{\tau} + \frac{1}{\tau^2} \sum_{i=1}^{n} t_i \right) \right|_{\tau = \langle t \rangle} \\
&= \left. \left(\frac{n}{\tau^2} - \frac{2}{\tau^3} \sum_{i=1}^{n} t_i \right) \right|_{\tau = \langle t \rangle} \\
&= \frac{n}{\langle t \rangle^2} - \frac{2}{\langle t \rangle^3} \sum_{i=1}^{n} t_i \\
&= \frac{n}{\langle t \rangle^2} - \frac{2n}{\langle t \rangle^2} = -\frac{n}{\langle t \rangle^2}
\end{aligned}
$$

= negative, as desired for a maximum



Method of maximum likelihood III

What is the uncertainty on the best fit value(s) $\hat{\theta}$?

The uncertainty (68.3% confidence interval) is taken to be the square root of the variance, $\sqrt{V}$. How to calculate the variance?

For one fitted parameter, it can be shown that

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

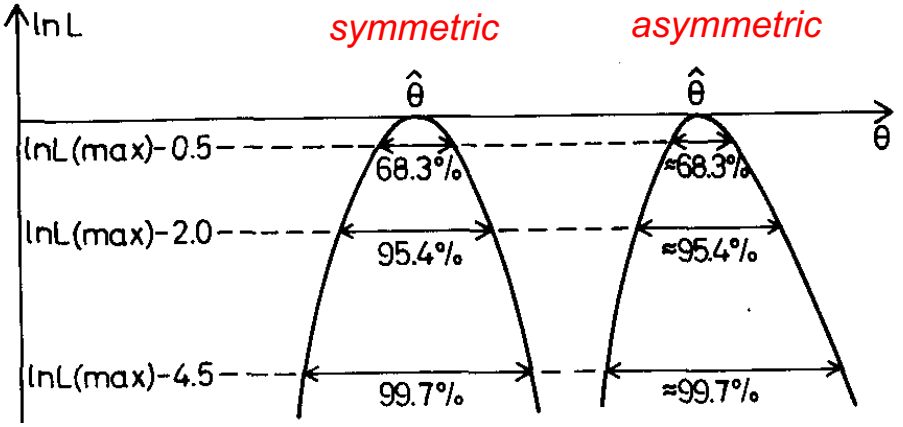
For a multi-parameter fit, it can be shown that

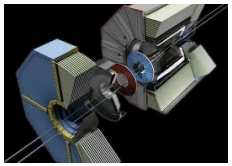
$$\left(V^{-1}\right)_{ij}(\hat{\theta}) = -\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \, \partial \theta_j}\right|_{\theta=\hat{\theta}}$$

In practice, e.g., MINUIT, the derivatives $\partial\mathcal{L}/\partial\theta$, $\partial^2\mathcal{L}/\partial\theta^2$, $\partial^2\mathcal{L}/(\partial\theta_i\,\partial\theta_j$ at $\theta=\hat{\theta}$ are evaluated numerically. MINOS calculates them to much higher precision than MIGRAD.

For a large number of observables x_i, the likelihood function $\mathcal{L} \rightarrow$ Gaussian, or ln $\mathcal{L} \rightarrow$ parabolic. The likelihood content of $\mathcal{L}$, used to obtain confidence intervals for the fitted parameters, can then be calculated analytically. The results are as shown:

(in practice, $-\ln\mathcal{L}$ is used, and confidence intervals are obtained from rise above the minimum)





Method of least squares (χ^2 fit)

Suppose we have a set of n measurements x_1, x_2, $x_3 \ldots x_n$ with uncertainties σ_1, σ_2, $\sigma_3 \ldots \sigma_n$, and a theoretical model predicts values f_1, f_2, $f_3 \ldots f_n$ for these observables. The model depends on a set of parameters θ_j that we want to determine. For example: if x_i were the bin contents of a helicity distribution, θ_1 would be the branching fraction of the decay mode, and θ_2 would be the fraction of longitudinal polarization.

The most-likely values of θ_j would be those that minimize the χ^2 statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - f_i)^2}{\sigma_i^2}$$

If the y_i are correlated, then the most-likely values of θ_j minimize

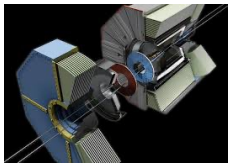
$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - f_i) \left(V^{-1}\right)_{ij} (x_j - f_j)$$

where V is the covariance matrix for the x_i measurements.

Note that if the measured values x_i are normally distributed about mean values μ_i with standard deviations σ_i, then the least-squares method and maximum-likelihood method are identical:

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-(x_i - f_i)^2/(2\sigma_i^2)} \\
&= \frac{1}{\sqrt{2\pi}(\sigma_1 \sigma_2 \sigma_3 \ldots \sigma_n)} e^{-\sum_{i=1}^{n} (x_i - f_i)^2/(2\sigma_i^2)} \\
\Rightarrow \ln \mathcal{L} &= -\sum_{i=1}^{n} \frac{(x_i - f_i)^2}{2\sigma_i^2} + C \\
&= -\frac{\chi^2}{2} + C
\end{aligned}$$

Since χ^2 is positive-definite, max(ln $\mathcal{L}$) corresponds to min(χ^2)



References

There are many excellent books on probability and statistics.

My favorites:

Frodesen, Skjeggestad, and Tofte, Probability and Statistics in Particle Physics (Columbia University Press, 1979). out of print

F. James, Statistical Methods in Experimental Physics, 2nd Edition (World Scientific, 2006). This is the update of the First Edition by Eadie, Drijard, James, Roos, and Sadoulet (North-Holland, 1971).

F. Reif, Chapter 1 of Fundamentals of Statistical and Thermal Physics (Waveland Press, 1965).

Other highly regarded books:

L. Lyons, Statistics for Nuclear and Particle Physicists (Cambridge University Press, 1989).

R. J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Wiley, 1993).

Additional recommended reading:

Particle Data Group (G. Cowan), "Review on Probability," "Review on Statistics," PRD 110, 030001 (2024).

G. Feldman and R. Cousins, "Unified Approach to the Classical Statistical Analysis of Small Signals," PRD 57, 3873 (1998).

Rolke, Lopez, Conrad, and James, "Limits and Confidence Intervals in the Presence of Nuisance Parameters," NIM A 551, 493 (2005).