Statistical data analysis

András LÁSZLÓ
laszlo.andras@wigner.hu
Wigner Research Centre for Physics

5 November 2021

Outline

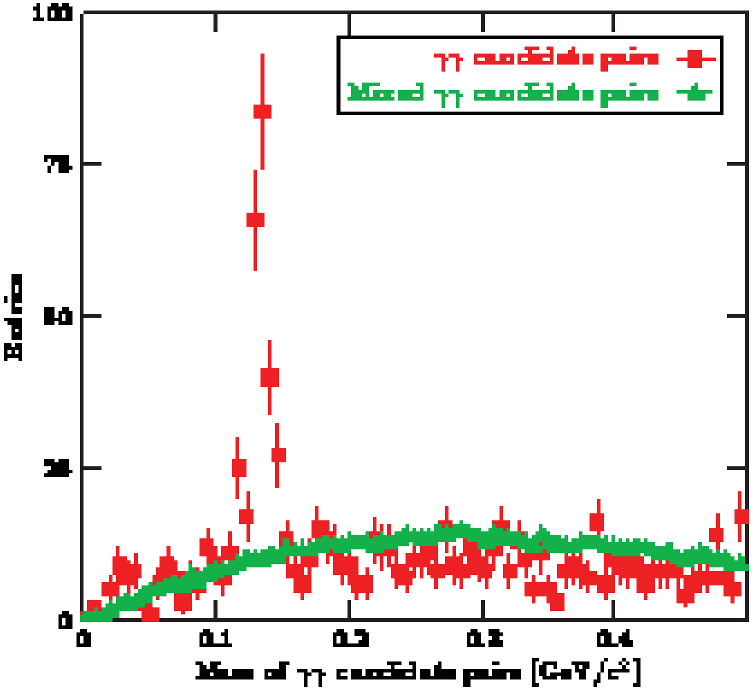
- Introduction
- Helper computer program packages
- Important identities and distributions
- Experimental estimation of density functions
- How it works: (pseudo)random number generation
- How it works: generating important distributions
- Statistical error propagation in linearized approximation
- Fitting to density functions (large statistics)
- Fitting to density functions (low statistics)
- Fine comparison of distributions, statistical tests
- Effect of non-ideal detector response, unfolding
- Systematic errors

Introduction

In experimental particle physics we often study distribution of various observables to draw physics conclusions.

For handling such problematics, we need a little probability theory / statistics. This is a whole science, we will only go through some often occuring techniques.

Example: in $\pi^0 \to \gamma\gamma$ decay we measure the momenta of the γ-s, and assuming that they are from π^0 decay, we try to reconstruct true mass spectrum of π^0-s.



But this is only example. One can do a great number of such things.

The topic has extended literature. A few sources:

- Bronstejn, Szemengyajev: Handbook for mathematics (statistics chapter, fundamental)
- Bohm, Zech: Introduction to statistics and data analysis for physicists (statistics, new)
- Feller: Introduction to probability theory (theoretical probability theory, a classic)
- etc etc (there are a number of books on market, each of them with different emphasis)
- Particle Physics Booklet (pocket book, to be used as handbook)
- Press, Flannery, Teukolsky, Vetterling: Numerical Recipes (numerical algorithms, classic)

Helper computer program packages

One can choose from a variety of tools. Some common examples:

- GNUPlot + AWK scripts (often used, efficient for simple problems)
- ROOT (a CERN and HEP standard, it is rather lousy, but lot of features)
- Python + MathPlotLib (begins to be standard, tidy, lot of features)
- BLOP-Plot (on sourceforge.net, non-standard, small community, tidy, developable)
- Often one has to write own libraries in C or C++ (specific tools for specific problems)

Typically an analysis is divided into 3 steps:

1. Data production: from raw measurement data one aims to reconstruct the physics parameters of a collision event (reconstruction), or the same on simulated data (simulation + reconstruction). This is normally done with a large number of jobs sent to big computer clusters, and with the official software of the experiment.
2. Data analysis: we analyze with statistical methods the data of the big DST-s ("Data Summary Tape"-s), produced in the above process. Typically done with large number of jobs sent to big computing clusters, with our own custom analysis software.
3. Data visualization and interpretation (plotting, fitting etc). Typically we perform this on our own laptop on small data produced above (on distributions etc), with our own software.

In this lecture, we mainly address the problematics occuring in steps 2 and 3.

Important identities and distributions

Let x and y be two probability variable, with their values in $\mathbb{R}^n$, having independent probability density functions (pdfs) f and g. That is: let the common pdf of (x, y) be $(x, y) \mapsto f(x)\, g(y)$. Then, the pdf of the probability variable $z := x + y$ will be $f \star g$, which is the convolution:

$$(f \star g)(z) := \int f(z - x) g(x) \mathrm{d}x.$$

Easily follows:

- If x and y has finite expectation value, then $x + y$ has also, and $\mu(x + y) = \mu(x) + \mu(y)$ holds. (Expectation value is additive.)
- If x and y has finite sigma-square (variance), then $x + y$ has also, and $\sigma(x + y)^2 = \sigma(x)^2 + \sigma(y)^2$ holds in such case. (Sigma-square is additive in such case.)

If we have n copies $x_1, \ldots, x_n$ from the same probability variable, with the same pdf f, then their arithmetic mean $z := \frac{1}{n} \sum_{k=1}^{n} x_k$ shall be a probability variable with pdf $z \mapsto n\, f^{\star(n)}(n\, z)$ (scaled n-fold convolution).

From the additivity of μ and σ^2 follows:

- The expectation value of arithmetic mean $m(x) := \frac{1}{n}\sum_{k=1}^{n} x_k$ is $\mu(x)$, i.e. expectation value can be estimated from n samples.
- The expectation value of empirical variance $s^2(x) := \frac{1}{n-1}\sum_{k=1}^{n} x_k^2 - \frac{n}{n-1}m^2(x)$ is $\sigma^2(x)$, i.e. variance can be estimated from n samples.

Also follows from additivity of μ and σ^2:

- $\sigma^2(m(x)) = \sigma^2\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\sigma^2(x_k) = \frac{1}{n}\sigma^2(x)$, i.e. $m(x) = \frac{1}{n}\sum_{k=1}^{n} x_k$ has a sigma of $\frac{1}{\sqrt{n}}$ times the sigma of one sample x_k.

This means: if we have a quantity with large spread, then we can estimate its true mean and true sigma from n samples, using the above $m(x)$ and $s(x)$ formulas. Moreover, the true sigma of this mean estimate $m(x)$ shall have a sigma of $\frac{1}{\sqrt{n}}$ times the true sigma of one single sample $\sigma(x)$. The latter is estimated by $s(x)$. In total, the sigma of the mean estimator $m(x)$ is $\frac{1}{\sqrt{n}}s(x)$.

Gaussian distribution: rather common in nature.

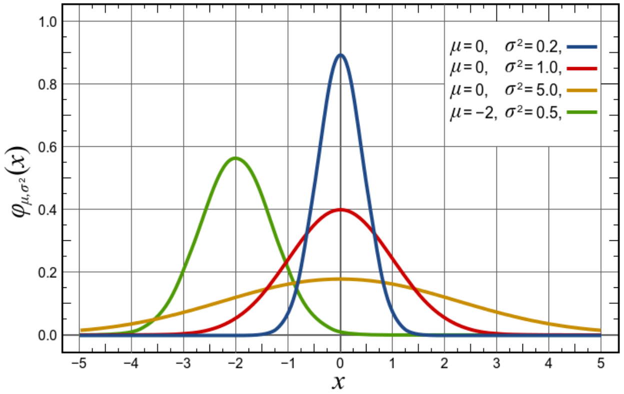
Because of central limiting distribution theorem: if we convolve a finite variance pdf with itself n-times (that is: we take the pdf of $\sum_{k=1}^{n} x_k$), then that will converge to Gaussian distribution.

Due to this, e.g. electronic noise etc all follow Gaussian distribution, because it forms as superposition of many small similar finite variance effects.

In 1 dimension the Gaussian pdf is:

$$x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Expectation value μ, spread σ.



Poisson distribution: rather common in nature.

Because of discrete central limiting distribution theorem: if we convolve a finite variance one sided discrete pdf with itself n-times (that is: we take the pdf of $\sum_{k=1}^{n} x_k$ of such variable), then that will converge to Poisson distribution.

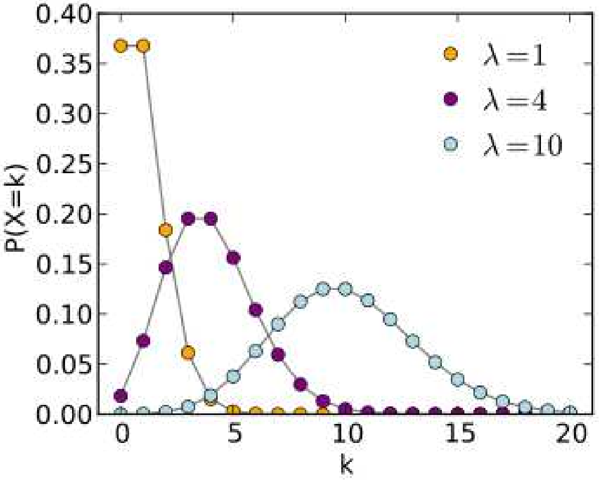
Due to this, discrete counting fluctuations are all exactly Poisson, since it is the superposition of similar finite variance one sided discrete effects.

The Poisson pdf is:

$$l \mapsto P_\lambda(l) = \frac{\lambda^l}{l!} \exp(-\lambda)$$

(l non-negative integer).

Important property: mean is λ, sigma is $\sqrt{\lambda}$. Converges to discrete Gauss in large λ limit.



Multinomial distribution: sometimes turns up in statistics.

Assume that there can be finite $1, \ldots, k$ different possible outcome of an experimental try, and we make a try n times, such that for all the n tries each outcome has constant p_l ($l = 1, \ldots, k$) probability.

The pdf of multinomial is:

$$(x_1, \ldots, x_k) \mapsto \frac{n!}{x_1! \ldots x_k!}\, p_1^{x_1} \ldots p_k^{x_k}$$

whenever $\sum_{l=1}^{k} x_l = n$, otherwise 0 (x_l non-negative integer).

Important property: for all $l, m = 1, \ldots, k$ indices, $\mu(x_l) = n\, p_l$, and $\sigma(x_l) = \sqrt{n\, p_l\, (1 - p_l)}$, and $\mathrm{Cov}(x_l, x_m) = -n\, p_l\, p_m$ ($l \neq m$).
In large statistics limit x_l converges to Poisson.

Skellam distribution: difference of two Poissons.

If we subtract two Poisson distribution non-negative integer valued variable, then we get an integer valued variable, which will follow Skellam distribution.

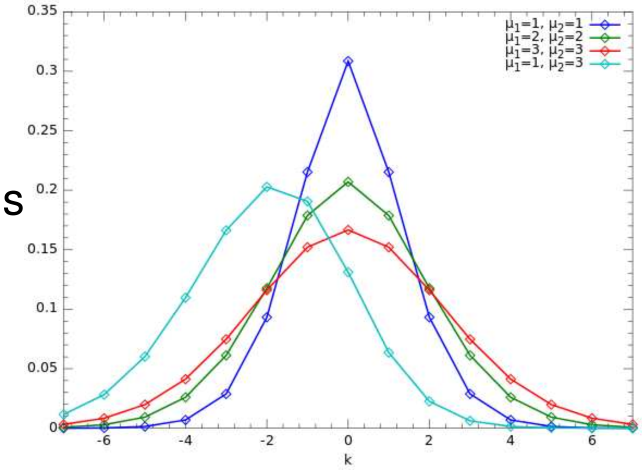
Good for characterizing behavior of counter differences (e.g. true signal $=$ measured raw counter $-$ background counter).

The pdf of Skellam is:

$$l \mapsto S_{\lambda_1,\lambda_2}(l) = \mathrm{e}^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{k/2} I_{|l|}(2\sqrt{\lambda_1\lambda_2})$$

where $I_{|l|}(\cdot)$ is the $|l|$-th modified Bessel function of the first kind (l integer).

Important property: expectation value is $\lambda_1 - \lambda_2$, sigma is $\sqrt{\lambda_1 + \lambda_2}$.



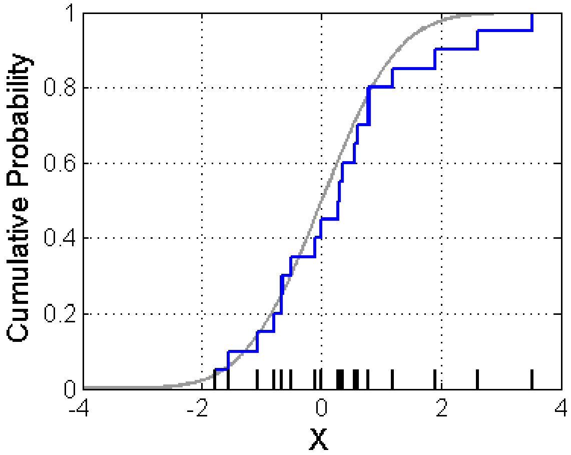
Experimental estimation of density functions

One of the most common problematics is that we have finite (n) measurements of a probability (vector)variable, and we would like to obtain from this finite sample the pdf of the variable. For this, there are a number of methods with pros&cons.

There are unbinned methods. If we have an 1 dimensional (scalar) variable, then one can construct the so called empirical distribution function. First, we order the samples $x_1, \ldots, x_n$ in increasing order (`quicksort` algorithm costs merely $n\,\log(n)$). Then, we construct the function

$$x \mapsto \hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} H(x - x_i)$$

where H is Heavyside function. This is a monothonically increasing step function from 0 to 1, and is called empirical distribution function.



Theorem: this converges to the true distribution function F as $n \to \infty$, in the maximum norm.

Similarly, we can construct a bit more clever estimator for distribution function: we do not use step function, but linearly interpolated step function. This will be piecewise differentiable, its derivatie will be pdf estimator:

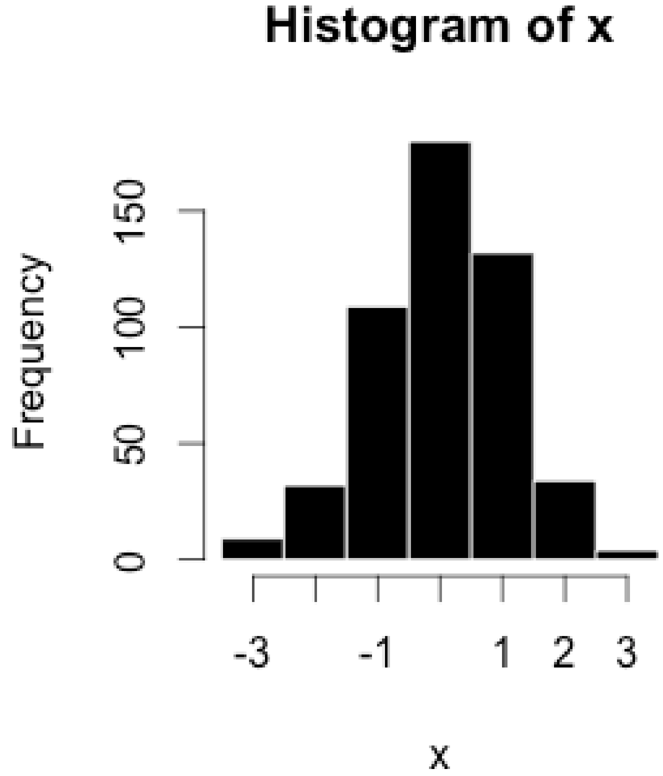
$$x \mapsto \sum_{i=1}^{n-1} \frac{1}{x_{i+1} - x_i} \, \chi_{[x_i, x_{i+1}[}(x)$$

where $\chi_{[a,b[}$ is the characteristic function of the interval $[a, b[$.

This is nice, since it is binning-free. Disadvantage: it is hard to quantify its statistical uncertainty due to finite sample.

Further disadvantage: in multiple dimensions it is not uniquely defined because there is no natural ordering of samples. We can use e.g. lexicographic ordering, but then the pdf estimator will slightly depend on how we order the dimensions.

The most widespread is the binning method (histograming). It works well in any dimensions. We take a large rectangular domain in the space of values of the observable. Along each axis we subdivide the domain into bins uniformly. We order a counter to each bin, initially set to zero. Then, we loop through our n sample $x_1, \ldots, x_n$ and increment the counter of the bin where a sample falls. The method is merely linearly expensive in n.



Theorem: a $n \to \infty$, this will converge to the bin-integrals of the pdf, times n.
If the bins are large: it converges not to the pdf, but to the bin-integrals of the pdf.

About the counting fluctuations (uncertainty). One has two fundamental scenarios.

1. We collect samples for fixed observation time: each bin will fluctuate with Poisson $\Rightarrow$ stat. error = square root of number of entries.
2. We collect samples until a fixed total number of samples is reached: then, the number of entries over bins will fluctuate with multinomial.

Note: the entries outside of the histograming domain are called underflows/overflows, and should be counted to the first/last bin of the axis, and should be always displayed!

Note: we should always display the statistical uncertainty of the bin entries (square root of number of entries for Poisson).

Note: the histogram figure should *always* have a title, and axis titles! (Vertical axis almost always: Entries.)

How it works: (pseudo)random number generation

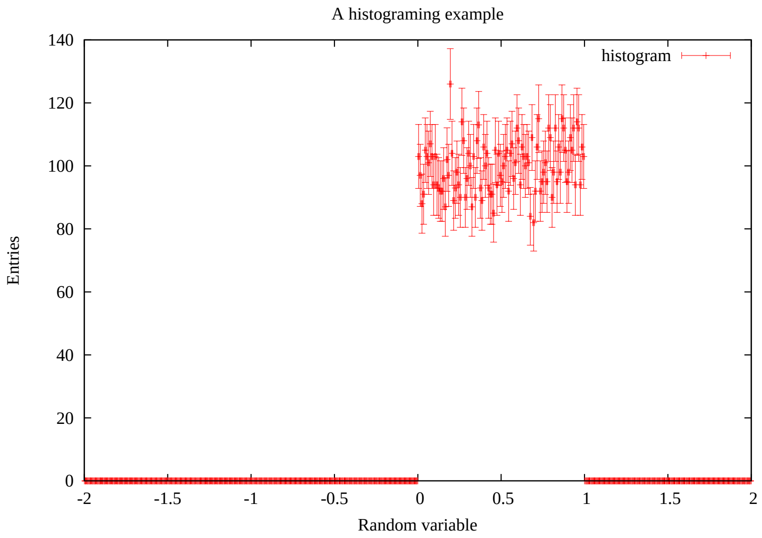
Generating uniformly distributed pseudorandom variable. Take the recursion

$$I_0 := s,\ I_{n+1} := (a \cdot I_n + c)\%m$$

e.g. with $s := 1, a := 1103515245, c := 12345, m := 32768$. This will generate pseudorandom numbers (integers) on the interval $[0, m-1]$. We can then rescale it to the real interval $[0, 1[$. The number s is called the seed (initial value).

This is how most of the built-in randomgenerator works, e.g. the `srand48+drand48` in C.

A typical trick is that we set the seed from date+time, so that it is unique (`ctime` in C).



Downloadable example illustrates all this. One can get it from the webpage of the lecture.

```
tar -xzf stat_progs.tar.gz
cd stat_progs
make
./hist
gnuplot hist.gnu
gv hist.eps
```

How it works: generating important distributions

Theorem: if is x probability variable, and F is its distribution function (running integral of pdf), then the probability variable $F(x)$ shall take its values in the interval $[0, 1[$, and there it shall have uniform distribution.

Remark: a distribution function F is strictly monothonically increasing from 0 to 1. Thus, for 1 dimensional variables, it is invertible.

Consequece: one can generate scalar distributions, if it is easy to evaluate F^{-1}. First we generate uniform random number u on $[0, 1[$, then $x := F^{-1}(u)$ shall have distribution F.

This is a fundamental trick, one can generate Lorentz(Cauchy), triangle, exponential, etc distributions in such a way, see: `stat_progs.tar.gz:stat_progs/rnd.h`

Usually, a Gaussian random number is generated differently, because evaluation of erf^{-1} is expensive / technically not simple.

But one can use that any 1d section of a 2d Gaussian will be an 1d Gaussian.

And one can generate 2d Gaussian rather easily, in polar coordinates: the radius-square has exponential distribution, and the azimuthal angle has uniform distribution, and the two are independent, and they are cheap to generate.

Consequence: if u and v are uniform random numbers on the interval $[0, 1[$, then

$$x := \sqrt{2\,|\ln(1-u)|}\ \cos(2\,\pi\,v)$$

shall have 1d Gaussian distribution. This is cheaper than evaluating erf^{-1}.

See: `stat_progs.tar.gz:stat_progs/rnd.h`

Statistical error propagation in linearized approximation

Let G be $\mathbb{R}^m \longrightarrow \mathbb{R}^k$ function, and we approximate it at $x := a + \Delta x$ (around a).

If G is $n+1$-times differentiable, then one has the Taylor-formula:

$$G(a+\Delta x) = G(a) + \mathrm{D}G(a)\,\Delta x + \frac{1}{2!}\,\mathrm{D}^2G(a)(\Delta x, \Delta x) + \cdots + \frac{1}{n!}\,\mathrm{D}^nG(a)(\Delta x, \dots, \Delta x) + O(\|\Delta x\|^{n+1}).$$

Specially, if G is 2-times differentiable, then:

$$G(a+\Delta x) = G(a) + \mathrm{D}G(a)\,\Delta x + O(\|\Delta x\|^2).$$

Thus, such a function G around a can be split like this:

1. A shift by a constant $G(a)$.
2. A constant linear transformation $\mathrm{D}G(a)$.
3. Plus a second order residual term, $O(\|\Delta x\|^2)$.

The linearized error propagation is based on: neglection of residual term.

Take the Taylor-formula at the expectation value $a := \mu(x)$ of the variable x, and introduce the new (centralized) probability variable $\Delta x := x - a$. Then:

$$G(x) = G(a + \Delta x) = G(a) + \mathrm{D}G(a)\,\Delta x + O(\|\Delta x\|^2).$$

Take the expectation value of the transformed variable $G(x)$:

$$\mu\bigl(G(x)\bigr) = G(a) + \mu\bigl(O(\|\Delta x\|^2)\bigr).$$

This means that if the residual term is neglected, then the expectation value propagates as:

$$\mu\bigl(G(x)\bigr) = G(\mu(x)),$$

that is we simply transform the expectation value itself.

This neglection can be done, whenever G around $\mu(x)$ behaves linearly enough, in a vicinity determined by the ellipsoid of $\mathrm{Cov}(x)$.

With the same approximation, the covariance of $G(x)$ will be:

$$\mathrm{Cov}\,\big(G(x)\big) = \mathrm{Cov}\,\big(\mathrm{D}G(a)\,\Delta x + O(\|\Delta x\|^2)\big) = \mathrm{Cov}\,\big(\mathrm{D}G(a)\,\Delta x\big) + \mu\big(O(\|\Delta x\|^3)\big).$$

This means, that when we neglect the residual term, then the covariance propagates as:

$$\mathrm{Cov}\,\big(G(x)\big) = \mathrm{Cov}\,\big(\mathrm{D}G(a)\,\Delta x\big) = \mathrm{D}G(a)\;\mathrm{Cov}(x)\,\mathrm{D}G(a)^T.$$

Now, writing back the notation $a = \mu(x)$:

$$\mathrm{Cov}\,\big(G(x)\big) = \mathrm{D}G(\mu(x))\;\mathrm{Cov}(x)\,\mathrm{D}G(\mu(x))^T.$$

This is the fundamental formula of linearized error propagation.

Using this, one can write an algebra of (measured value, measurement uncertainty) pairs (automated error propagation). See:

```
stat_progs.tar.gz:stat_progs/meas.h,meas.cc
tar -xzf stat_progs.tar.gz
cd stat_progs
make
./meas
```

Fitting to density functions (large statistics)

Assume: we have n measurements, where we measured a quantity as a function of some variable, i.e. we have $(x_1, g_1), \ldots, (x_n, g_n)$. This can be, for instance, a pdf estimation from histograming. In that case, x_i are bin positions, and g_i are number of entries in a bin. In that example case, g_i will fluctuate with Poisson. At the large statistics limit: Poisson will tend to Gaussian.

We would like to describe our measured pdf with some model pdf $x \mapsto f_\theta(x)$, where θ is some m dimensional parameter vector. (We assume, that they are properly normalized with number of entries.)

Then, the θ parametervector can be determined at the place of best coincidence:

$$\chi^2 := \sum_{i=1}^{n} \frac{(g_i - f_\theta(x_i))^2}{\sigma_i^2}$$

penalty function is minimized as a function of θ. If g_i are histogram entries, then $\sigma_i = \sqrt{g_i}$ due to Poisson counting uncertainty.

If g_i truly fluctuates according to Gaussian (e.g. for large statistics Poisson), then this fit will be a maximum-likelihood fit with Gaussian fluctuation hypothesis.

At the χ^2 minimum the statistical covariance matrix of the parameters will be:

$$\mathrm{Cov}(\theta) = -\frac{1}{2}\left(\left(\frac{\partial^2\,\chi^2}{\partial\theta_i\,\partial\theta_j}\right)_{i,j=1,\ldots,m}\right)^{-1}.$$

Further important theorem is that at the minimum

$$\chi^2/(n-m)$$

has expectation value 1. $(\chi^2/\mathrm{ndf} \leftarrow$ number of degrees of freedom)
This can be used to test the goodness of model.

Fitting to density functions (low statistics)

Sometimes it happens that we have low number of entries in some important part of our histogram. Then the simple χ^2 fit, corresponding to Gaussian maxlikelihood fit, does not give good result, because entries in reality do not fluctuate according to Gauss, but e.g. according to Poisson.

Then:

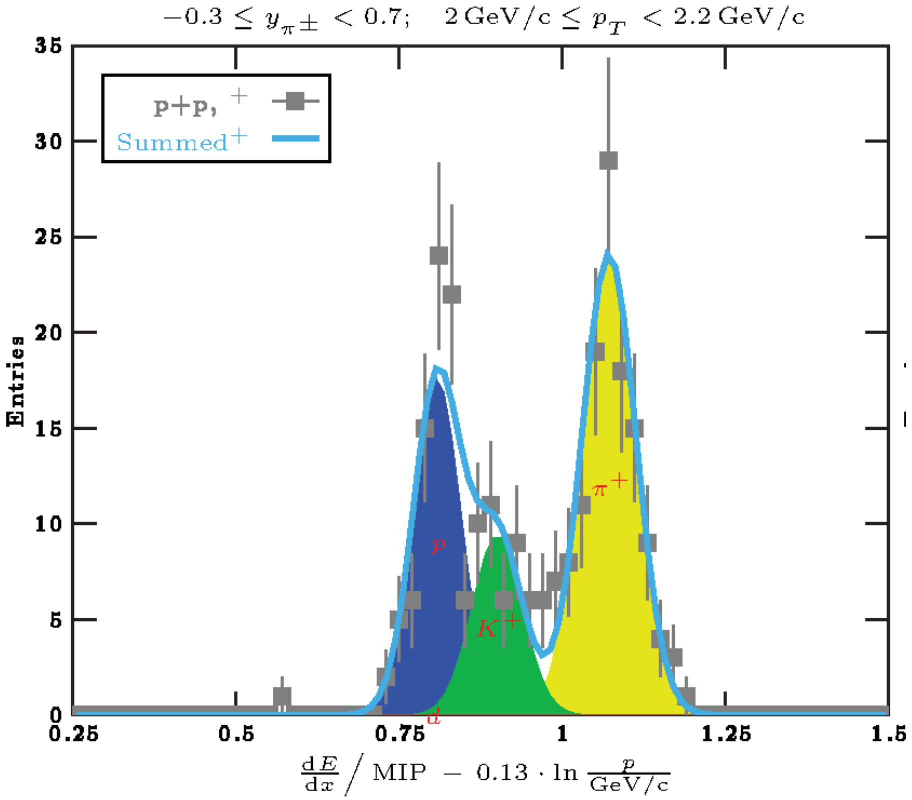
$$\chi^2 := \sum_{i=1}^{n} 2\left(f_\theta(x_i) - g_i + g_i \ln(g_i / f_\theta(x_i))\right)$$

penalty function should be minimized as a function of θ. This is Poisson maxlikelihood fit.

The formula for $\mathrm{Cov}(\theta)$ and $\mu(\chi^2/(n-m)) = 1$ will be still valid.

Theorem: it conserves exactly the number of entries.

S.Baker, D.Cousins: *Nucl.Instr.Meth.* **221** (1984) 437.


−0.3 ≤ y_π± < 0.7; 2 GeV/c ≤ p_T < 2.2 GeV/c
p+p, +
Summed+
Entries
35
30
25
20
15
10
5
0
0.25
0.5
0.75
1
1.25
1.5
dE/dx / MIP − 0.13 · ln p/(GeV/c)
p
K+
π+
d

If we do measurement not for fixed time but until we reach a given number of entries, then multinomial.

Then:

$$\chi^2 := \sum_{i=1}^{n} 2\, g_i \ln(g_i / f_\theta(x_i))$$

penalty function should be minimized as a function of θ. Multinomial maxlikelihood.

The formula for $\mathrm{Cov}(\theta)$ and $\mu(\chi^2/(n-m)) = 1$ will be still valid.

Theorem: it conserves exactly the number of entries.

S.Baker, D.Cousins: *Nucl.Instr.Meth.* **221** (1984) 437.

Fine comparison of distributions, statistical tests

Assume: x is a real valued probability variable. Let $\hat{F}_n$ be the empirical distribution function made from n samples $x_1, \ldots, x_n$. This is at some distance in the maximum norm from the true distribution function F:

$$D_n := \max_x |\hat{F}_n(x) - F(x)|.$$

Theorem: the probability variable $d := \sqrt{n}D_n$ for large n follows Kolmogorov distribution, which has distribution function:

$$K(d) := 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2k^2d^2} = \frac{\sqrt{2\pi}}{d}\sum_{k=1}^{\infty}\mathrm{e}^{-(2k-1)^2\pi^2/(8d^2)}$$

Kolmogorov-Smirnov test works like this: with given finite sample and theoretical distribution model function, we can quantify that what would have been the probability of measuring the observed sample, assuming that our model for data distribution is true. Very sensitive test, used also in gravitational wave searches.

Let x and y be real valued probability variable. Assume that we have n sample $x_1, \ldots, x_n$ from x and corresponding empirical distribution function $\hat{F}_n$. Similarly, assume that we have m sample $y_1, \ldots, y_m$ from y and corresponding empirical distribution function $\breve{F}_m$. These will be at some distance in the maximum norm:

$$D_{n,m} := \max_x |\hat{F}_n(x) - \breve{F}_m(x)|.$$

Theorem: if these two samples have the same distribution, then the probability variable $d := \sqrt{n\,m/(n+m)}D_{n,m}$ follows the Kolmogorov distribution for large n, m.

With this, we can also do Kolmogorov-Smirnov test: we can evaluate what would be the probability of obtainig these samples, assuming that they follow the same distribution. Very sensitive test.

Kolmogorov-Smirnov test problematic in higher dimensions: one can generate empirical distribution functions in multiple ways, depending on ordering of axes.

Effect of non-ideal detector response, unfolding

Sometimes, we would like to measure the pdf $x \mapsto f(x)$ of some quantity x, but it is smeared by a non-ideal detector response:

$$y \mapsto g(y) = \int \rho(y|x) f(x) \, \mathrm{d}x$$

is the measured pdf instead, where $(x, y) \mapsto \rho(y|x)$ is the response function, which we know either from theory, or from calibration measurements. The real problem is that we rather measure:

$$y \mapsto g(y) = \int \rho(y|x) f(x) \, \mathrm{d}x \ + \ e(y)$$

where e is an error term (e.g. from statistical counting uncertainty), and we only know its properties, but not its actual value (it is a probability variable).

The above integral transformation operator is called folding.

The problem is: it is a theorem that in general, a folding operator maps some distant pdfs to close pdfs, and these may be hard to distinguish due to the error term e.

This has extensive literature, there are some methods, but generally problematic.

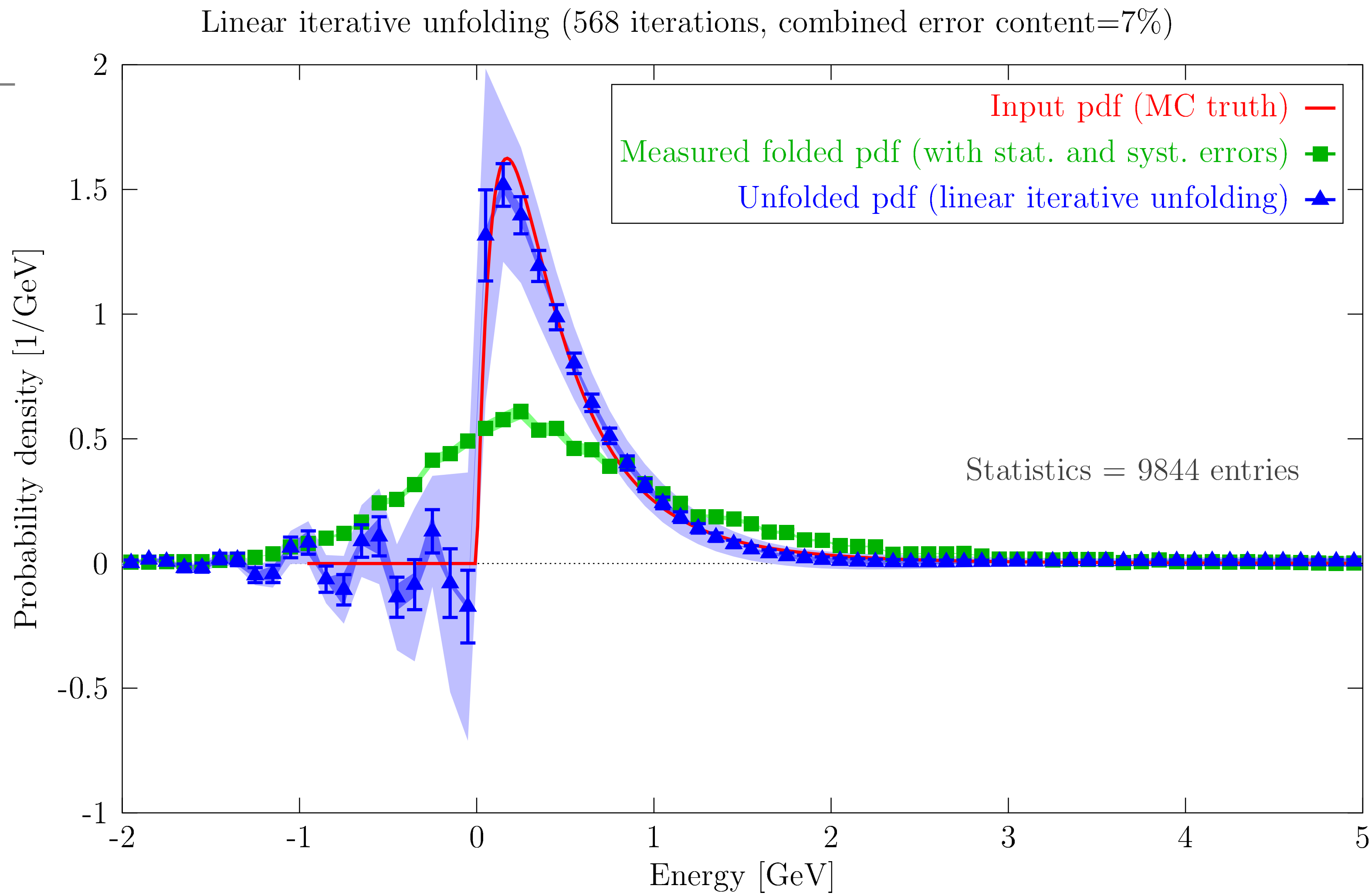
One of the methods is that we use that the folding operator (A) is linear, and we precondition it to make sure that its spectrum is in the interval $[0,1]$.

f is the pdf which we would like to reconstruct, g is the measured pdf, and A is the folding operator by the response function, with $g = Af + e$ where e is the error term. Assume that all these are given in terms of histograms, for simplicity.

An algorithm:

$$\begin{aligned} K &:= \max_i \sum_j (A^T A)_{ij}, \\ f_0 &:= K^{-1} A^T g, \\ f_{N+1} &:= f_N + \left(f_0 - K^{-1} A^T A f_N \right). \end{aligned}$$

Theorem: this converges in L^2 norm to $f - P_{\mathrm{Ker}(A)} f$, and also binwise. There is a formula for the algorithmic approximation error (at finite N), and for the propagated statistical and systematic errors.



A.László: *J.Phys.* **A39** (2006) 13621

A.László: *J.Phys.Conf.Ser.* **368** (2012) 012043

A.László: *SIAM JUQ* **4** (2016) 1345

Systematic errors

This is the most difficult to capture.

Assume that we have some measurements $x_1, \ldots, x_n$, which fluctuate according to some distribution. (E.g. if they are histogram bins, then Poisson.) We would like to interpret these as manifestation of a model function with some θ parameter vector: we make a statistical estimate for θ, e.g. a fit. Denote the reconstruction procedure by a function A: then, $\theta = A(x_1, \ldots, x_n)$. E.g. A is a model fitting to data.

Often, the problem to solve is even more complicated. Often: the measurements $x_1, \ldots, x_n$ are not of the form of known physics + known detector effects + known statistical fluctuations, but they deviate from this systematically by pulls $\delta x_1, \ldots, \delta x_n$ to an unknown degree, due to model imperfections. E.g. due to hard-to-quantify or neglected detector effects. Usually, knowing the experimental details, for these systematic deviations we can set some upper bounds $sx_1, \ldots, sx_n$, which are called the systematic errors of the measurement.

The systematic deviations $\delta x_1, \ldots, \delta x_n$ on the measurement cause a systematic deviation $\delta\theta$ in the estimated parameters. The task is to find upper estimate $s\theta$ (systematic error) to the deviation $\delta\theta$, given $sx_1, \ldots, sx_n$. In most analyses, often this is a rather hard task, because it can be very problem specific.