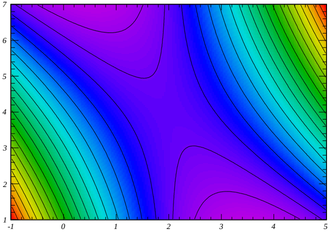
Statistics for Gamma-Ray Astronomy

Michael Schmelling – MPI for Nuclear Physics

- *Probability Distributions*
- *Fitting*
- *Markov Chain Monte Carlo*
- *Wilks Theorem*



➜ *reminder*

- probabilities for discrete states
  - ➜ non-negative numbers which sum to unity
- continuous probability density functions (PDFs) in n dimensions
  - ➜ non-negative functions which integrate to unity
  - ➜ integrals over finite volumes define probability
- most important characteristics: location and spread
  - ➜ e.g. expectation value(s) and (co)variance (matrix)

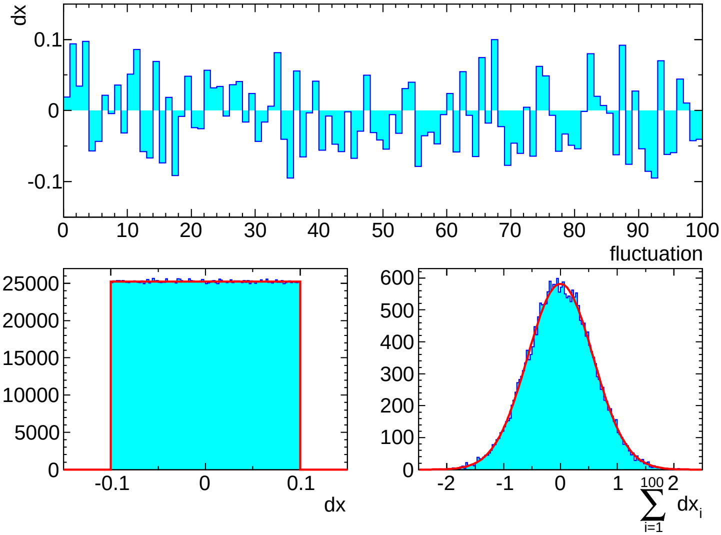
➜ *the uniform distribution*

The probability density inside a range $[a, b]$ is constant.

- (most) fundamental, simple PDF
- convenient starting point to derive more complex PDFs

study sums of uniform random numbers ➜

Simulation study



➜ *observation*

The sum of many random fluctuations is described by a Gaussian PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

- symmetric around zero
- one parameter σ describing the width
- first published in by C.F. 1809 Gauss in "Theoria motus corporum coelestium in sectionibus conicis solem ambientium" (with Least-Squares and Maximum-Likelihood method)
- the exact conditions for convergence to a Gaussian are formally described by the central limit theorem
- due to its fundamental nature also referred to as “normal” distribution

Statistics of counting experiments

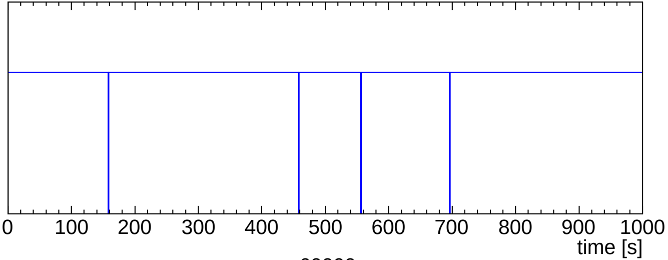
➜ *common problem in particle and astroparticle physics*

- examples:
  - ➜ decays in a radiactive source
  - ➜ high energy gamma rays hitting the atmosphere
  - ➜ number of soldiers in the Prussian army killed accidentally by horse kicks (Ladislaus Bortkiewicz, 1898)
- quantities of interest
  - ➜ time differences between subsequent events
  - ➜ number of events in time interval T

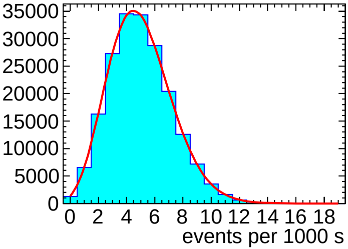
❖ numerical simulation

- split T into (many) subsequent time slices
- assume a probability to observe an event in a time slice $p \ll 1$

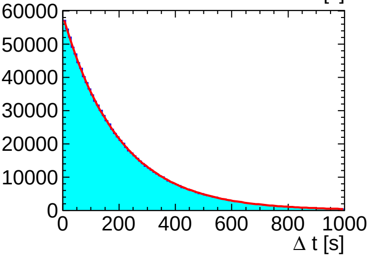
see what happens ➜



0
100
200
300
400
500
600
700
800
900
1000
time [s]



35000
30000
25000
20000
15000
10000
5000
0
0 2 4 6 8 10 12 14 16 18
events per 1000 s

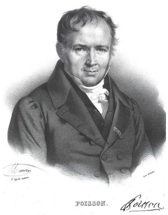


60000
50000
40000
30000
20000
10000
0
0 200 400 600 800 1000
Δ t [s]

Discussion

➜ *observation*

- results are described by simple functions of a single parameter (consequence of the single probability for an event per time slice)
- event counts per time interval: Poisson distribution
  - ➜ first published by Simèon Denis Poisson 1837 in “Recherches sur la probabilité des jugements en matière criminelle et en matière civile”



$$p_n = e^{-\mu}\,\frac{\mu^n}{n!}$$

- time difference between events: exponential distribution

$$f(t) = \frac{1}{\tau}\,e^{-t/\tau}$$

Given n measurements y_i, $i = 1, \ldots, n$ which are drawn from a PDF $f(y, a)$ with unknown parameter(s) a, determine an estimate $\hat{a}$ for a and (an estimate) for the uncertainty of $\hat{a}$.

- methods to determine the estimate $\hat{a}$:
  - ➜ unbinned maximum likelihood fit
  - ➜ binned maximum likelihood fit
  - ➜ least squares fit
- different possibilities to quantify the uncertainty of $\hat{a}$:
  - ➜ standard deviation $\sigma(\hat{a})$
  - ➜ 68% confidence level interval

❖ central object: the likelihood function L

PDF for observations as a function of a: $$L(a) = \prod_{i=1}^{m} f(y_i, a)$$

➜ *straightforward application*

Instead of maximising $L(a)$ minimise $-\ln L(a)$:

$$\left.\frac{d}{da}(-\ln L(a))\right|_{a=\hat{a}} = 0$$

(MIGRAD)

Estimate of the uncertainty by the standard deviation of $\hat{a}$:

$$\sigma^2(\hat{a}) = \left(\left.\frac{d^2}{da^2}(-\ln L(a))\right|_{a=\hat{a}} \right)^{-1}$$

(HESSE)

Estimate of the 68% confidence level interval:

$$-\ln L(\hat{a} - \delta_-) = -\ln L(\hat{a} + \delta_+) = -\ln L(\hat{a}) + 0.5$$

with $\delta_\mp$ the largest deviations satisfying this condition.

(MINOS)

➜ *a few words of caution*

- everything is well defined if the global minimum of $-\ln L$ is unique or sufficiently deep
- for $n \to \infty$ the likelihood function becomes Gaussian
  - ➜ the estimate $\hat{a}$ is unbiased
  - ➜ $\hat{a}$ has the smallest possible variance
  - ➜ standard deviation and MINOS-errors are the same
  - ➜ asymmetric uncertainties have exact coverage
- do not confuse 68% “coverage” with “probability of the true value”
  - ➜ In the frequentist view 68% coverage means that in 68% of all measurements the true value is inside the 68% confidence level interval. For a given measurement it is either inside or outside and it is not known which of the two is realised.

Binned maximum likelihood fit

➜ *aggregate n data points into m bins*

- reduce the number of terms in the likelihood function from n to m
- same formalism as above with likelihood per bin

$$f_i(a) = e^{-\mu_i(a)} \frac{\mu_i(a)^{n_i}}{n_i!} \quad \text{with} \quad \mu_i(a) = n \int_{bin\ i} dy\, f(y, a)$$

  - ➜ with n_i the number of entries in bin i
  - ➜ total number of entries is fixed to n
  - ➜ normalisation is fixed
    - ◆ normalisation fit requires an extension of the method

Least squares fit

➜ *alternative binned fit minimising the χ^2 cost function*

$$\chi^2(a) = \sum_{i=1}^{m} \frac{(n_i - f_i(a))^2}{f_i(a')}$$

- denominators are “known” variances of the n_i
  - ➜ must not be varied in the minimisation (hence “$f_i(a')$”)
  - ➜ requires iterative fit if not known a priori
- $\chi^2(a)/2$ has the same asymptotic properties as $-\ln L(a)$
- does allow to fit also the normalisation
- in some sense more basic than maximum likelihood fit. . .
  - ➜ maximum likelihood (also extended) can be derived from it
  - ➜ allows to deal also with correlated measurements
  - ➜ less well understood than maximum likelihood
  - ➜ often applied wrongly and therefore disfavoured

➜ *basic idea*

Instead of analytically minimising the negative-log likelihood function, and determining likelihood contours, find the best fit values and the confidence areas by sampling it.

- estimate minimum from the distribution of $-\ln L$
- find confidence levels for parameters by looking at the distribution of points which deviate by less than 0.5 units from the minimum
  - ➜ universally applicable also for pathological likelihood functions
  - ➜ doing this by standard Monte Carlo is possible but not ideal since most of the time will be spent in regions of small likelihood
  - ➜ alternative: Markov Chain Monte Carlo

❖ sample the parameter space according to the likelihood function

- x, y: points in configuration space with $\rho > 0$
- $P(y, x)$: transition $x \to y$, i.e. PDF in y for given x
- core of the algorithm: decomposition of $P(x, y)$
  - ➜ a random step $q(y, x)$ from $x \to y$, with $q(y, x) = q(x, y)$
  - ➜ the probability $\alpha(x, y)$ to accept this step

$$P(y, x) = q(y, x) \cdot \alpha(x, y) = q(y, x) \cdot \min\left[1, \frac{\rho(y)}{\rho(x)}\right]$$

➜ *why the algorithm samples ρ:*

Explicit calculation shows

$$P(y, x)\rho(x) = P(x, y)\rho(y)$$

and it follows for the PDF after a jump

$$\rho'(y) = \int dx\ P(y, x)\rho(x) = \int dx\ P(x, y)\rho(y) = \rho(y) \quad \text{q.e.d.}$$

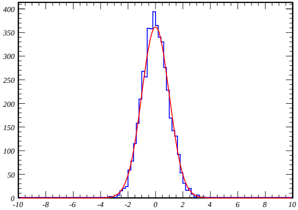
Discussion

- the algorithm is independent of the dimension of the PDF ρ
- normalization of ρ not needed
- it works for arbitrary functions $q(x, y) = q(y, x)$
  - ➜ small steps: slow movement towards the next maximum
  - ➜ large steps: danger to be trapped at sharp maxima
- note: subsequent sample points are correlated!

➜ *try MCMC sampling of 1-dim PDFs in $-a < x < a$*

- ➢ gaussian: $\exp(-x^2/2)$
- ➢ exponential: $\exp(-x)$ for $x > 0$
- ➢ singular density: $1/\sqrt{x}$ for $x > 0$
- ➢ rapidly oscillating density: $\sin^2(1/x)$

with $q(y, x)$: $y = x + 0.1\,\mathrm{rndm}(-\mathrm{a}, \mathrm{a})$



Bayesian interpretation

- parametrize knowledge about a parameter a by a PDF $\rho(a)$
- use Bayes' theorem to update the knowledge by data y

$$\rho(a|y) \propto f(y|a)\,\rho(a)$$

- ➜ sampling likelihood function times prior, $f(y|a)\rho(a)$, samples the posterior $\rho(a|y)$
- ➜ for multidimensional a, and/or nuisance parameters, integration over all but one dimensions, yields the PDF for a single parameter
- ➜ maximum & 68% interval ➜ best fit parameter & uncertainty
- ➜ for finite statistics the results depend on the prior $\rho(a)$
- ➜ for $n \to \infty$ and $\rho(a) = 1$ equivalent to frequentist approach

❖ MCMC is the ideal tool to determine Bayesian posteriors

→ *S.S. Wilks, March 26, 1937*

If a population with a variate x is distributed according to the probability distribution $f(x, \theta_1, \theta_2, \ldots, \theta_h)$, such that optimum estimates $\hat{\theta}_i$ of θ_i exist which are distributed in large samples according to (1), then when the hypothesis H is true that $\theta_i = \theta_{0i}, i = m+1, m+2, \ldots h$, the distribution of $-2\ln\lambda$, where λ is given by (2) is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h-m$ degrees of freedom.

(1) a PDF deviating from a d-dim Gaussian only by terms of order $1/\sqrt{n}$

(2) the ratio of the best fit likelihoods fitting all or only m parameters, fixing the others to the true values

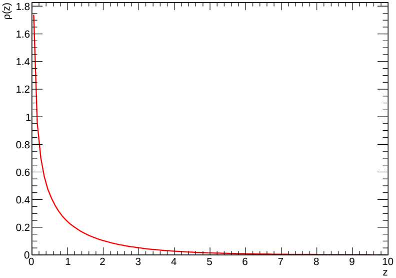
$$\lambda = \frac{P(\hat{\theta}_1, \ldots, \hat{\theta}_m, \hat{\theta}_{0m+1}, \ldots, \hat{\theta}_{0h})}{P(\hat{\theta}_1, \ldots, \hat{\theta}_m, \hat{\theta}_{m+1}, \ldots, \hat{\theta}_h)}$$

Application of Wilks theorem

→ *test for the existence of a signal s component in data*

- fit with free parameter s: $F_s = -\ln L_{\text{best}}(s)$
- fit with parameter $s = 0$: $F_0 = -\ln L_{\text{best}}(s = 0)$
  - → one has $F_s < F_0$ and $z = 2(F_0 - F_s) > 0$

PDF of z if $s = 0$ is true: $$\rho(z) = \frac{1}{\sqrt{2\pi z}}\, e^{-z/2}$$



→ *p-value for observed z_{obs}*

$$p = \int_{z=z_{\text{obs}}}^{\infty} dz\; \rho(z)$$

discovery $s \neq 0$ if e.g. $p < 5.7 \cdot 10^{-7}$