Statistical methods and tools

Giovanni Benato
giovanni.benato@gssi.it

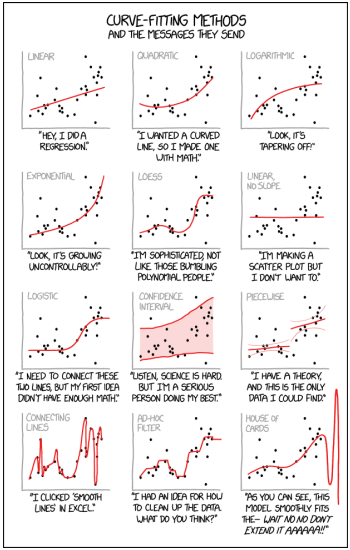
Gran Sasso Science Institute

13th IDPASC School
September 19th, 2024

Table of Contents

Literature

- F. James, Statistical methods in experimental physics
  → Rigorous math, sometimes lacking "physical interpretation"
  → ∼20 years old, so new methods are missing
- L. Lista, Statistical methods for data analysis in particle physics
  → Oriented to experimentalists; new methods also present
  → Concise theoretical construction;
- G. Cowan, Statistical data analysis
  → ∼25 years old, so new methods are missing
- G. Cowan's website
  → Mostly slides, plus a lot of other references
- O. Behnke et al., Data analysis in high-energy physics
  → Very concise introduction; concentrates on modern methods



Examples

- A list of numerical examples can be found at this link
- All examples are written in C++/ROOT/BAT, but the code is commented so it should be possible to understand what it does

Table of Contents

Mathematical probability

- Let Σ be the set of all elementary and mutually exclusive events x_i
- We define the probability of occurrence of the event x_i to obey the *Kolmogorov axioms*:

$$P(x_i) \geq 0 \quad \forall i \tag{1}$$

$$P(x_i \vee x_j) = P(x_i) + P(x_j) \tag{2}$$

$$\sum_i P(x_i) = 1 \tag{3}$$

- Abstract definition
- Holds for any quantity that satisties the 3 axioms

Frequentist probability

- Consider an experiment observing a series of N events
- Assume k events are of type X
- The frequentist probability for any **single** event to be of type X is:

$$P(X) = \lim_{N\to\infty} \frac{k}{N} = \lim_{N\to\infty} \frac{\#\text{ of favorable cases}}{\#\text{ of possible cases}} \tag{4}$$

- In principle, $P(X)$ can only be known for $N = \infty$, but often it can be computed (analytically or numerically) with high precision
- Can only be applied to **repeatable** experiments
  $\rightarrow$ Cannot predict if Italy will win the next World Cup
  $\rightarrow$ Cannot compute the probability that dinosaurs died by starvation

Bayesian probability

- Probability defined as the **degree of belief** (DoB) that something will occurr
- **Coherent bet**: which amount $F(x)$ are you willing to bet that x will occurr, knowing that if you win you get a fixed amount K?

$$P(x) = \frac{F(x)}{K} \Rightarrow \begin{cases} P(x) = 1 \text{ if we are sure } x \text{ will happen} \\ P(x) = 0 \text{ if we are sure } x \text{ won't happen} \\ 0 < P(x) < 1 \text{ otherwise} \\ \sum_i P(x_i) = 1 \end{cases} \tag{5}$$

- The DoB is a property both of the observed system, as well as of the observer
- The DoB depends on the observer's knowledge, and will change according to it
- Can be applied to non-repeatable phenomena
  $\rightarrow$ Can be applied to the true value of a Physics theory!

Bayes theorem for discrete events

- Let A and B be two sets of elementary events a_i and b_i
- Law of conditional probability:

$$P(A \wedge B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B) \tag{6}$$

- Bayes theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{7}$$

- More generally, if $A_i = A_0, \ldots A_n$ are **exclusive and exhaustive** sets, and if B is any event:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)} \tag{8}$$

- $P(A_i) = \pi(A_i)$ is the **prior probability** of A_i
- $P(A_i|B)$ is the **posterior probability** of A_i after knowing that B has occurred

Example: measuring protons with a particle detector

- $P(B)$ = probability that any particle gives a triggered event
- $P(A)$ = probability of a proton hitting the detector
- $P(B|A)$ = probability of a proton giving a triggered event
- $P(A|B)$ = probability of a triggered event to be induced by a proton

Example: dogmas

- Take a set of possible mutually exclusive events A_i
- Assume the following priors for each A_i:

$$\pi(A_i) = \begin{cases} 1 \text{ if } i = 0 \\ 0 \text{ if } i \neq 0 \end{cases}$$

- Applying Bayes theorem we get:

$$\text{If } i = 0 : P(A_0|B) = \frac{P(B|A_0) \cdot \pi(A_0)}{\sum_i P(B|A_i) \cdot \pi(A_i)} = \frac{P(B|A_0) \cdot 1}{P(B|A_0) \cdot 1} = 1 = \pi(A_0)$$

$$\text{If } i \neq 0 : 0 = P(A_i)$$

$\rightarrow$ If I have a dogmatic belief about something, i.e. a belief that I cannot or am not willing to change for any reason, no experimental evidence will ever change my mind.

Use of Bayes theorem in Physics

- Assume H_0 and H_0 are a complete set of hypotheses, i.e. a complete set of physics models describing a given physical phenomenon
- By convention
  - H_0 = **background-only** hypothesis = the known physics is enough to explain the data
  - H_1 = **signal + background** hypothesis = there is an additional component due to new physics and we know how to model it
- H_0 and H_0 will depend on some parameters, which might differ between the two hypotheses, and that we will generally indicate as $\vec{\theta} \in \Omega$
- Assume we perform n measurements of some physical observable x, which we will indicate as $\vec{x}$

Bayes theorem for parameter estimation

$$P\left(\vec{\theta} \mid \vec{x}\right) = \frac{P\left(\vec{x} \mid \vec{\theta}\right)\pi\left(\vec{\theta}\right)}{\int_{\Omega} P\left(\vec{x} \mid \vec{\theta}\right)\pi\left(\vec{\theta}\right)} = \frac{P\left(\vec{x} \mid \vec{\theta}\right)\pi\left(\vec{\theta}\right)}{P\left(\vec{x}\right)} \tag{9}$$

- $P\left(\vec{\theta} \mid \vec{x}\right)$ = Posterior probability for parameters $\vec{\theta}$ given the data $\vec{x}$ and the model H_i
- $P\left(\vec{x} \mid \vec{\theta}\right)$ = Probability of obtaining **exactly** the data $\vec{x}$ given the parameters $\vec{\theta}$
- $\pi\left(\vec{\theta}\right)$ = Prior probability of parameters $\vec{\theta}$ under the assumption of model H_i
- $P\left(\vec{x}\right)$ = Probability of getting data $\vec{x}$ given any possible value of $\vec{\theta}$
  $\rightarrow$ In the end, it's just a normalization factor.

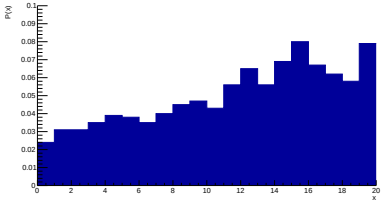
Bayes theorem for hypothesis testing

$$P(H_i \mid \vec{x}) = \frac{P(\vec{x} \mid H_i)\,\pi(H_i)}{\sum_i P(\vec{x} \mid H_i)\,\pi(H_i)} \tag{10}$$

- $P(H_i \mid \vec{x})$ = Posterior probability for hypothesis H_i after measuring data $\vec{x}$
- $\pi(H_i)$ = Prior probability for hypothesis H_i
  $\rightarrow$ This is the **subjective** part of the method
- $P(\vec{x} \mid H_i) = \int_\Omega P\left(\vec{x} \mid \vec{\theta}, H_i\right) \pi\left(\vec{\theta}\right) d\vec{\theta}$ = Probability of obtaining the data $\vec{x}$ given all possible values of the parameters $\vec{\theta}$, assuming the model H_i

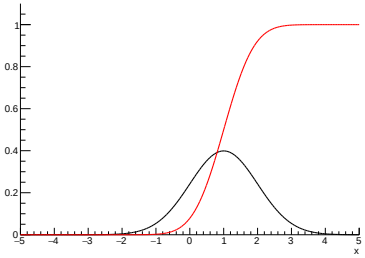
Random variables and probability distributions

- Random event = event with more than one possible outcome, to which a probability may be associated
  $\rightarrow$ The outcome of a random event is unknown, only the probabilities of the possible outcomes are known
  $\rightarrow$ We can associate a random variable x to a random event X
- The possible outcomes $P(x_1), P(x_1), \ldots$ of numerical values $x_1, x_2, \ldots$ form the **probability distribution**, obeying to the normalization condition $\sum_i P(x_i) = 1$
- What if a random variable covers a continuous interval?



Probability density function

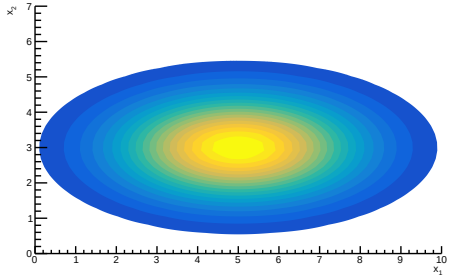
- Consider a sample space $\Omega \in \mathbb{R}^n$
- A random extraction (experiment) will lead to an outcome (measurement) corresponding to one point $\vec{x} \in \Omega$
- We can associate a **probability density** $f(\vec{x})$ to any point $\vec{x} \in \Omega$, with $f(\vec{x}) > 0$
- The probability of an event A is: $P(A) = \int_A f(\vec{x}) d\vec{x}$
  $\rightarrow$ $f(\vec{x})$ is the differential probability: $f(\vec{x}) = \frac{dP(\vec{x})}{d\vec{x}}$
  $\rightarrow$ $f(\vec{x})$ is normalized to 1: $\int_\Omega f(\vec{x}) d\vec{x} = 1$
- Cumulative distribution: $F(x) = \int_{-\infty}^{x} f(y) dy$
  $\rightarrow$ $F(\tilde{x}) = P(x \leq \tilde{x}) \ \forall \ \tilde{x}$

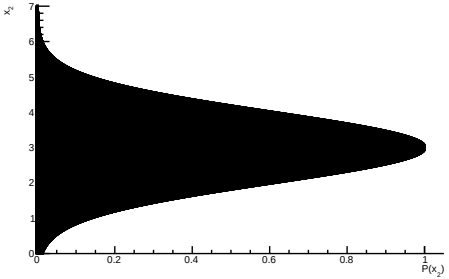


Marginal distributions

- Take an n-dimensional random variable $\vec{x} = (x_1, \ldots, x_n)$ with PDF $f(\vec{x})$
- The marginal distribution for x_i is the 1-dim PDF:

$$f(x_i) = \int f(\vec{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \tag{11}$$





Example of discrete distributions

Binomial distribution: distribution of k successes out of n attempts, given a success probability p

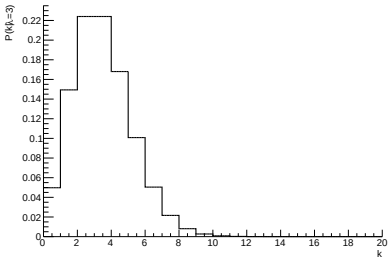
$$P(k \mid n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Poisson distribution: limit of Binomial for $n \to \infty$, $p \to 0$ and $\lambda = np$ finite

$$P(k \mid \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Properties of Poisson distribution

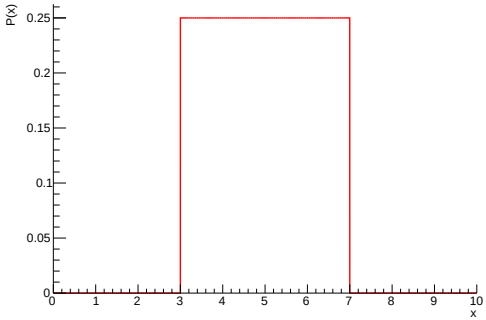
- $lim_{\lambda \to \infty} P(k \mid \lambda) = Gaus\left(\mu = \lambda, \sigma = \sqrt{\lambda}\right)$
- If k_1 and k_2 are Poisson-distributed with mean λ_1 and λ_2, the sum $k_1 + k_2$ is Poisson-distributed with mean $\lambda_1 + \lambda_2$
- Randomly picking with probability ε from a Poisson process with mean λ_0, gives a Poisson process with mean $\lambda = \varepsilon \lambda_0$



Example of continuous distributions

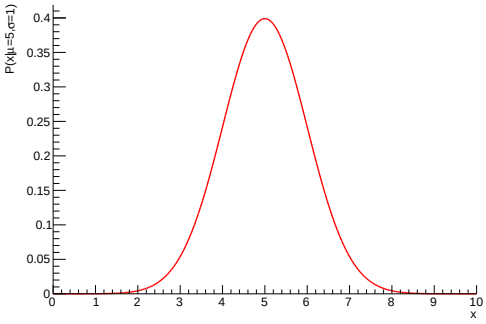
Uniform distribution

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ otherwise} \end{cases}$$



Normal (Gaussian) distribution

$$G(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Pseudorandom numbers and MC methods

- A pseudorandom number generator is an algorithm that generates a sequence of numbers distributed according to some PDF and that resemble very closely an actual distribution of random numbers with the same PDF
  $\rightarrow$ The true issue lies on the specification and implementation of that **very closely**
- Properties of pseudorandom number generators:
  - Each extraction must be statistically independent from the previous ones

$$f(x_i \mid x_{i-m}) = f(x_i)$$

  - All extractions must follow the same PDF

$$f(x_i) = f(x_j) \quad \forall i, j$$

  - After a given period p, the sequence will repeat itself with $x_{i+p} = x_i$, so the sequence is truly random up to $n \leq p$
  - We should be able to replicate the exact same sequence for debugging purpose (seeding)

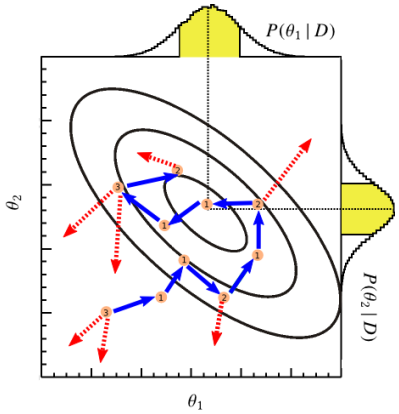
Markov-Chain MC

- A Markov-Chain Monte Carlo (MCMC) is a sequence of random variables $x_1, \ldots, x_n$ whose PDF obeys to

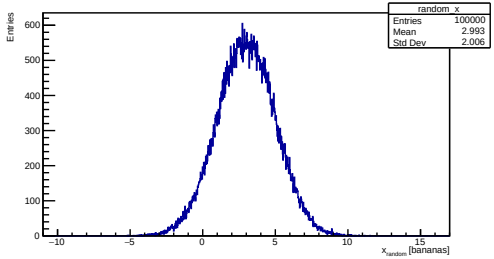
$$f(x_{i+1} \mid x_0, \ldots, x_i) = f(x_{i+1} \mid x_i)$$

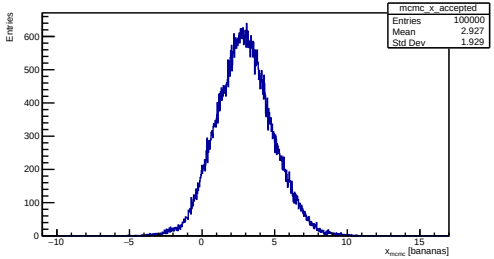
  $\rightarrow$ Can be more efficient in sequency highly peaked PDFs

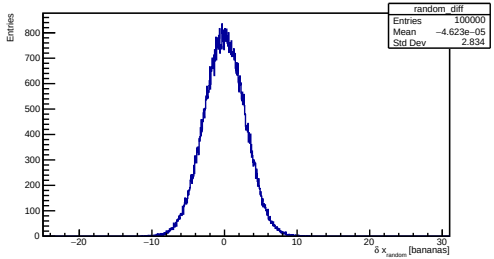
- Example of MCMC: Metropolis-Hastings
  1. Pick a point $\vec{x}_1$ uniformly distributed in space Ω and evaluate $f(\vec{x}_0)$
  2. Generate a second point $\vec{x}$ according to a predefined PDF $q(\vec{x}, \vec{x}_0)$ called "proposal function" or "step function", and evaluate $f(\vec{x})$
  3. Generate uniform number $u \in [0, 1[$
  4. If $\frac{f(\vec{x})q(\vec{x},\vec{x}_0)}{f(\vec{x}_0)q(\vec{x}_0,\vec{x})} > u$, accept the point and set $\vec{x}_2 = \vec{x}$, otherwise reject $\vec{x}$
  5. Iterate back to (2)



Pseudorandom vs MCMC-generated numbers







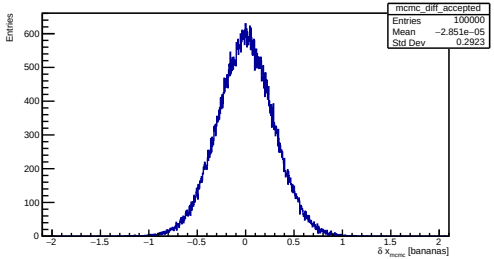


Table of Contents

Likelihood

- Take a real random variable $\vec{x}$ with PDF $f(\vec{x} \mid \vec{\theta})$, where $\vec{\theta}$ is a set of real parameters
- Suppose we make a set of n measurements of $\vec{x}$: $\vec{\vec{x}} = \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$
- The joint PDF of all n measurements becomes:

$$P\left(\vec{\vec{x}} \mid \vec{\theta}\right) = P\left(\vec{x}_1, \ldots, \vec{x}_n | \vec{\theta}\right) = \prod_{i=1}^{n} f\left(\vec{x}_i \mid \vec{\theta}\right)$$

- Since the values $\vec{x}_i$ are fixed (they are measured), **P is no longer a PDF**, but just a function of $\vec{\theta}$, and we denote it as $\mathcal{L}$:

$$\mathcal{L}\left(\vec{\theta}\right) = \mathcal{L}\left(\vec{\vec{x}} \mid \vec{\theta}\right) = \prod_{i=1}^{n} f\left(\vec{x}_i \mid \vec{\theta}\right) \tag{12}$$

Statistic (singular, not plural)

- A statistic is any new random variable $t = t(\vec{x}_1, \ldots, \vec{x}_n)$
- For example, the average $\langle \vec{x} \rangle$ is a statistic
- In practice, we use a statistic whenever we need to map a highly-dimensional random variable into a one-dimensional one
- A statistic $t = t(\vec{x})$ is sufficient for θ if the conditional PDF $f(\vec{x} \mid t)$ is independent of θ
  $\rightarrow$ If t is a sufficient statistic, any monotonic function of t is also a sufficient statistic
  $\rightarrow$ t contains the same information about θ as the original data $\vec{x}$

Addressed question vs required methods

Question	Method
Based on the measured data $\vec{x}$, what is the single value $\hat{\theta}$ that is closest to the true **unkwnown** of θ? Or, what is the most probable value for θ?	Point estimation
Based on the measured data $\vec{x}$, what is the range of values that is most likely to enclose the true **unkwnown** value of θ? Or what interval encloses the true value with a given amount of probability?	Interval estimation
Is our model $f(\vec{x} \mid \vec{\theta})$ good enough to describe the measured data?	Goodness of fit
Assuming we have two alternative models H_0 and H_1 to describe the observed process, which of the two agrees better with the measured data?	Hypothesis testing

Addressed question vs required methods

Method	Solution	Algorithm
Point estimation	Find the parameter values $\hat{\theta}$ that maximize $\mathcal{L}$ or $P\left(\vec{\theta} \mid \vec{x}\right)$	Minimizer algorithm
Interval estimation	Study the tails of $\mathcal{L}$ or $P\left(\vec{\theta} \mid \vec{x}\right)$	Study all possible combinations of $\vec{\theta}$ giving $\vec{x}$ with Toy-MC, or map $P\left(\vec{\theta}\middle\vert\vec{x}\right)$ with Markov Chain MC
Goodness of fit	Quantify the probability of a random fluctuation to give a worse fit	Analytical method (e.g. χ^2) or toy-MC
Hypothesis testing	Compare probability of two models to give a better or worse fit	Toy-MC plus some method to compare the validity of the alternative hypotheses

Table of Contents

Estimator and their properties

- **Estimate of an unknown parameter:** mathematical procedure to determine the central value of the parameter as a function of the observed data sample
- **Estimator:** function of the data sample returned by the estimate
- Properies of estimators:
  - Consistency: does the estimator converge in probability to the true value $\bar{\theta}$ of the unknown parameter θ?
  - Ubiasedness: expected value fo the deviation of the parameter estimate from the true value $\bar{\theta}$
  - Efficiency: variance of the estimator with respect to the ideal case (Cramer-Rao inequality)
  - Robustness: (in)sensitivity to small deviations from the assumed PDF model

Maximum-likelihood estimator

- The **maximum-likelihood estimate** of the parameter θ is that value $\hat{\theta}$ for which $\mathcal{L}(\vec{x} \mid \theta)$ is maximal, given the observed data $\vec{x}$
- Since the logarithm is monothonic, finding $\max \mathcal{L}$ is equivalent to finding $\max(\ln(\mathcal{L}))$ or $\min(-\ln(\mathcal{L}))$
  $\rightarrow$ Instead of maximizing $\prod f(x \mid \theta)$, we can maximize $\sum \ln f(x \mid \theta)$, which often is numerically simpler!

Extended likelihood

- Suppose we perform n measurements of a random variable x with PDF $f(x \mid \theta)$
- Suppose the number of observations n is itself a random variable with PDF $P(n \mid \theta)$
- We can extend the likelihood to include $P(n \mid \theta)$:

$$\mathcal{L} = P(n \mid \theta) \prod_{i=1}^{n} f(x_i \mid \theta) \tag{13}$$

- In many cases, $P(n \mid \theta)$ is a Poisson distribution with mean $\lambda = \lambda(\theta)$:

$$\mathcal{L} = \frac{e^{-\lambda} \lambda^n}{n!} \prod_{i=1}^{n} f(x_i \mid \theta) \tag{14}$$

Example: maximum-likelihood for signal + background

- Assume we want to describe the data with a signal (new physics) plus background (known physics). The overall PDF is:

$$f(x \mid s, b, \theta) = \frac{s}{s+b} f_s(x \mid \theta) + \frac{b}{s+b} f_b(x \mid \theta) \tag{15}$$

  where s and b are the expected number for signal and background counts
  $\rightarrow$ The expectation value for n is $\lambda = s + b$

- The likelihood becomes:

$$\mathcal{L} = \frac{e^{-(s+b)}(s+b)^n}{n!} \prod_{i=1}^{n} \frac{s f_s(x \mid \theta) + b f_b(x \mid \theta)}{s+b} = \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{n} [s f_s(x \mid \theta) + b f_b(x \mid \theta)] \tag{16}$$

- Taking the logarithm:

$$\ln \mathcal{L} = -s - b - \ln n! + \sum_{i=1}^{n} \ln [s f_s(x \mid \theta) + b f_b(x \mid \theta)] \tag{17}$$

Binned likelihood

- Suppose the number of measurement n is so large that computing $\sum \ln f(x_i \mid \theta)$ takes too long
- We can simplify the problem by binning x in $m \ll n$
  $\rightarrow$ The likelihood becomes a multinomial times the PDF of n:

$$\mathcal{L} = P(n \mid \theta) \left[\prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!} \right] n! \tag{18}$$

where:

- i is now the bin index (not the event index!)
- k_i is the number of counts in bin i
- p_i is the probability associated to bin i
- $\sum_i k_i = n$

Binned likelihood for Poisson-distributed n

- Take $\lambda_i = \lambda \int_{\delta x_i} f(x \mid \theta) dx$ = expectation value for bin i
  $\rightarrow \sum_i \lambda_i = \lambda$
- Take $p_i = \frac{\lambda_i}{\lambda}$
- The likelihood becomes:

$$\mathcal{L} = \frac{e^{-\lambda}\lambda^n}{n!} \prod_{i=1}^{m} \left(\frac{\lambda_i}{\lambda}\right)^{k_i} \frac{1}{k_i!} = e^{-\lambda} \cancel{\lambda^n} \prod_{i=1}^{m} \frac{\lambda_i^{k_i}}{\cancel{\lambda^{k_i}} k_i!} = \prod_{i=1}^{m} \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!} \tag{19}$$

$\rightarrow$ The binned version of an extended $\mathcal{L}$ with Poisson-distributed n is the product of a Poisson term for each bin!

Least-squares (or χ^2) method

- Take a set of n measurements with Gaussian uncertainties $y_i \pm \sigma_i$
- Assume each y_i corresponds to a perfectly known x_i following the relation $y = y(x, \theta)$
  $\rightarrow y(x, \theta)$ is just a function, not a PDF
- The likelihood will be a product of Gaussian PDFs around the curve $y(x, \theta)$:

$$\mathcal{L}(\vec{y} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - y(x_i, \theta))^2}{2\sigma_i^2}\right) \tag{20}$$

- Let's minimize $-2 \ln \mathcal{L}$ instead:

$$-2 \ln \mathcal{L} = \sum_{i=1}^{n} \frac{(y_i - y(x_i, \theta))^2}{2\sigma_i^2} + 2 \sum_{i=1}^{n} \ln\left(2\pi\sigma_i^2\right) \tag{21}$$

  $\rightarrow$ The first term is a sum of standard-normal variables, so it follows a χ^2 distribution
  $\rightarrow$ The second term is a constant and can be dropped
- If θ is not known, $\min(-2 \ln \mathcal{L})$ follows a χ^2 distribution only if all y_i are uncorrelated

Least-squares method for histograms

- Assume we have a set of measurements x_i following a PDF $f(x_i \mid \theta)$
- Assume we bin the data, and every bin has a large number of entries k_i so that the corresponding Poisson distribution can be approximated with a Gaussian
- We can approximate the likelihood with:

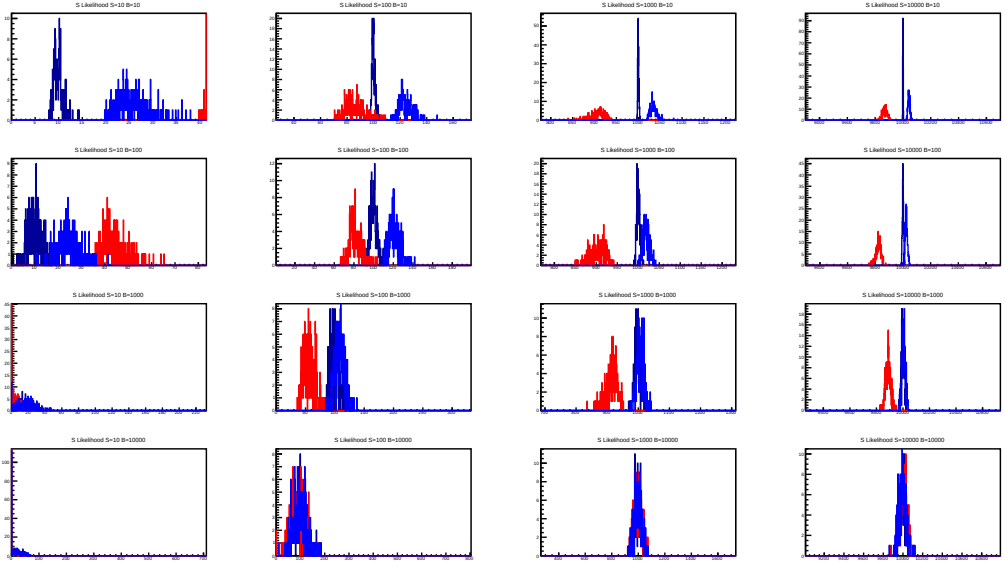
$$\mathcal{L} \simeq \prod_{\text{bin } i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(k_i - \lambda_i(\theta)\right)^2}{2\sigma_i^2}\right) \;\Rightarrow\; -2\ln\mathcal{L} = \sum_{i=1}^{m} \ln\left(2\pi\sigma_i^2\right) + \sum_{i=1}^{m} \frac{\left(k_i - \lambda_i(\theta)\right)^2}{\sigma_i^2} \tag{22}$$

- The problem is that we don't know σ_i...
  - The first term varies slowly, so we can neglect it
  - **Neyman's** χ^2: set $\sigma_i^2 = k_i$
    $\rightarrow$ Standard for many predefined fit methods, e.g. in ROOT/Minuit
    $\rightarrow$ Fails miserably if just one bin is empty
  - **Pearson's** χ^2: set $\sigma_i^2 = \lambda_i(\theta)$
    $\rightarrow$ Works also for empty bins, but is unreliable for small λ_i
- Take-home message: **NEVER USE A χ^2 FIT ON A HISTOGRAM!**

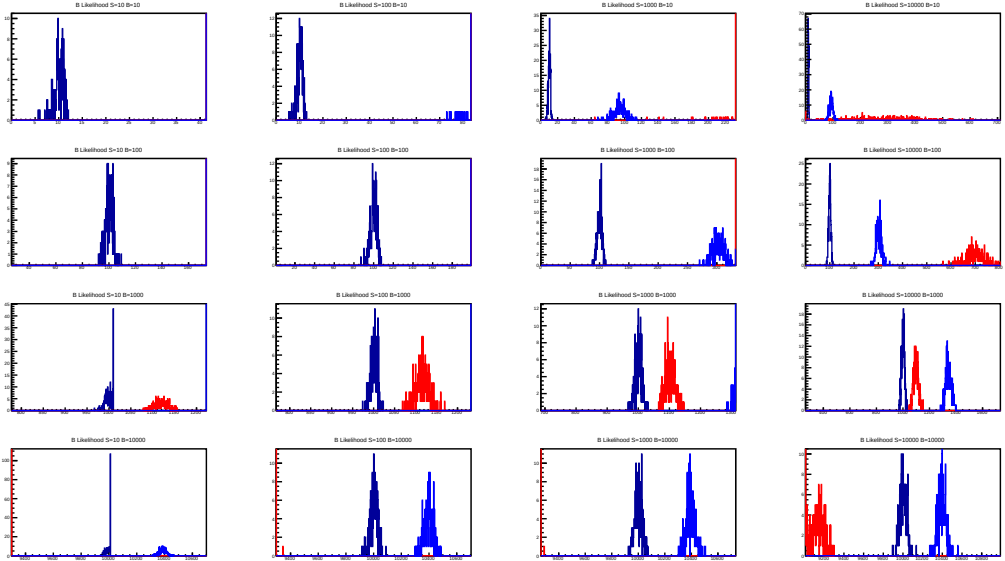
Example: $\mathcal{L}$ vs Neyman-χ^2 vs Pearson-χ^2

- Generate toy-MC spectra with S Gaussian-distributed signal events and B uniformly distributed background events
- Fit each toy-MC with $\mathcal{L}$, Neyman-χ^2 and Pearson-χ^2
- Compare distribution of best-fit values for 10^4 toy-MC, and for various combination of $S, B = 10, 100, 1000, 10000$

Example: $\mathcal{L}$ vs Neyman-χ^2 vs Pearson-χ^2



Example: $\mathcal{L}$ vs Neyman-χ^2 vs Pearson-χ^2

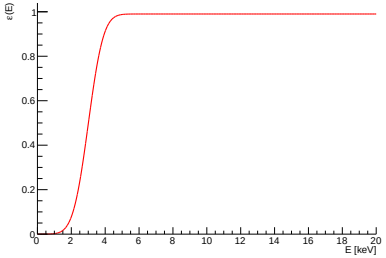


Choosing the correct PDF

- Goal: study the trigger efficiency curve for a detector
- Assume we know that the efficiency curve can be modeled with an error function (cumulative of a Gaussian):

$$\varepsilon = \frac{p}{2}\left[1 + \mathrm{erf}\left(\frac{E-\mu}{\sigma}\right)\right] \tag{23}$$

  where $p = 0.99$, $\mu = 3\,\mathrm{keV}$, $\sigma = 1\,\mathrm{keV}$
- Suppose we inject a known number n of pulser events at energies $E_i = 1, 2, 3, \ldots, 20\,\mathrm{keV}$, and we want to run a Bayesian fit to reconstruct the PDF of p
- What is the correct likelihood? What PDF are the data following?



Choosing the correct PDF

- For each energy E_i, we inject a **known** number of pulser events n
  $\rightarrow$ The number of detected events k_i follows a **binomial** distribution
  $\rightarrow$ We must use a binomial likelihood!

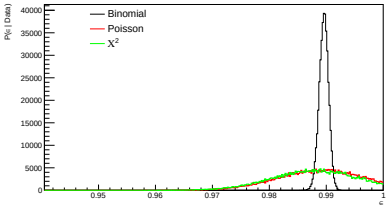
$$\mathcal{L} = \prod_i \frac{n!}{k_i!(n-k_i)!} \varepsilon(E_i \mid p, \mu, \sigma)^{k_i} \left(1 - \varepsilon(E_i \mid p, \mu, \sigma)\right)^{n-k_i}$$

- What happens if we use a Poisson likelihood, or a χ^2 instead?

Choosing the correct PDF

- For each energy E_i, we inject a **known** number of pulser events n
  $\rightarrow$ The number of detected events k_i follows a **binomial** distribution
  $\rightarrow$ We must use a binomial likelihood!

$$\mathcal{L} = \prod_i \frac{n!}{k_i!(n-k_i)!} \varepsilon(E_i \mid p, \mu, \sigma)^{k_i} \left(1 - \varepsilon(E_i \mid p, \mu, \sigma)\right)^{n-k_i}$$



- What happens if we use a Poisson likelihood, or a χ^2 instead?

Simultaneous fits

- Suppose we have two sets of data $\vec{x}$ and $\vec{y}$ with PDFs that have some parameters in common:

$$f(\vec{x} \mid \vec{\theta}, \vec{\nu}) \quad \text{and} \quad f(\vec{y} \mid \vec{\theta}, \vec{\phi})$$

- The likelihood will simply be the product of the two likelihoods:

$$\mathcal{L}(\vec{x}, \vec{y} \mid \vec{\theta}, \vec{\nu}, \vec{\phi}) = \mathcal{L}(\vec{x} \mid \vec{\theta}, \vec{\nu}) \cdot \mathcal{L}(\vec{y} \mid \vec{\theta}, \vec{\phi}) \tag{24}$$

Table of Contents

Confidence interval

- Goal: we want to find the range $\theta_a \leq \theta \leq \theta_b$ that contains the true value θ_0 with some probability β (usually 68%)
- Given an observation x form a PDF $f(x \mid \theta)$, the probability content β of the region $[a, b]$ in **x-space** is:

$$\beta = P(a \leq x \leq b) = \int_a^b f(x \mid \theta) dx$$

- However, we want to find a corresponding interval $[\theta_a, \theta_b]$ in θ-**space** so that $P(\theta_a \leq \theta \leq \theta_b) = \beta$
  $\rightarrow$ Such interval is called **confidence interval**
  $\rightarrow$ A method that provides such an interval is said to have the property of **coverage**
- Notice that:
  - The true value θ_0 is and will always remain an unknown constant
  - θ_a and θ_b must be functions of x, not of θ

Confidence interval for Normally-distributed data

- Let $f(x \mid \vec{\theta})$ be a normal distribution with $\vec{\theta} = (\mu, \sigma)$
- If μ and σ are known:

$$\beta = P(a \leq x \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \quad \text{where } \Phi \text{ is the cumulative}$$

- If μ and σ are unknown, we can't compute the probability content of $[a, b]$, but we can compute the probability β that x lies in some interval $[\mu + c, \mu + d]$. By defining $y = \frac{x-\mu}{\sigma}$ we get:

$$\beta = P(\mu + c \leq x \leq \mu + d) = \Phi\left(\frac{d}{\sigma}\right) - \Phi\left(\frac{c}{\sigma}\right)$$

- Then we can invert to obtain:

$$\beta = P(x - d \leq \mu \leq x - c)$$

$\rightarrow$ This is still a probability statement about x, while μ is still unknown!
$\rightarrow$ There is infinite possible choices for such interval!

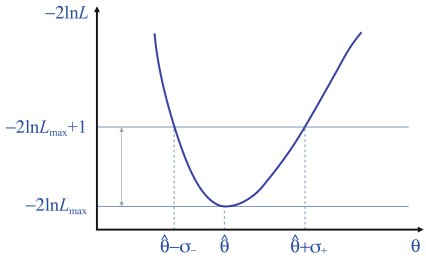
Log-likelihood scan

- Assume we have a 1-dim Gaussian likelihood:

$$-2\ln\mathcal{L} = -2\ln\mathcal{L}_{max} + \frac{(\mu - x)^2}{\sigma^2}$$

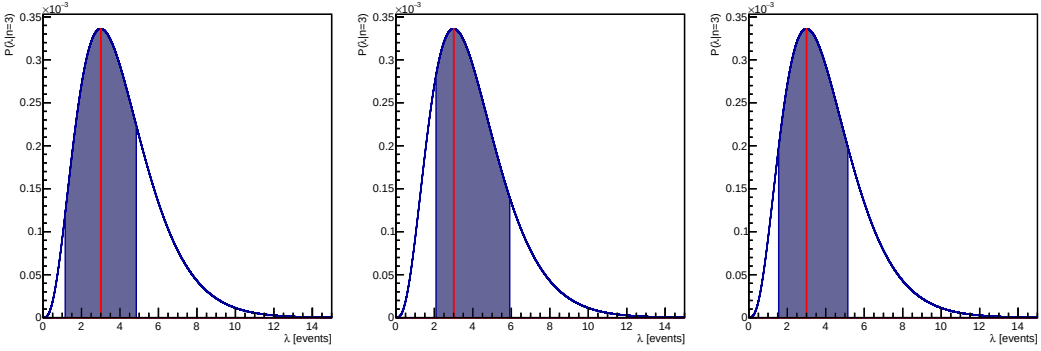
  $\rightarrow$ The intercepts at $-2\ln\mathcal{L}_{max} + 1$ correspond to the $\pm 1\sigma$ interval
  $\rightarrow$ The intercepts at $-2\ln\mathcal{L}_{max} + 4$ correspond to the $\pm 2\sigma$ interval

- For a non-Gaussian likelihood, $-2\ln\mathcal{L}$ is not parabolic, but the max-$\mathcal{L}$ estimate is invariant under reparameterization
  $\rightarrow$ The intercept at $-2\ln\mathcal{L}_{max} + k$ will maintain the same coverage as in the Gaussian case!



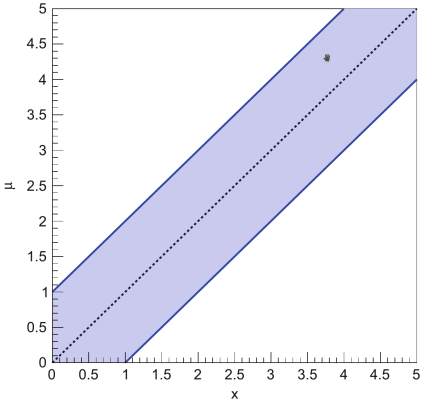
Ordering rules

- **Central interval**: $[x_L, x_U] = [\bar{x} - \delta, \bar{x} + \delta]$
- **Equal areas**: $\int_{-\infty}^{x_L} f(x \mid \theta)dx = \int_{x_U}^{+\infty} f(x \mid \theta)dx = \frac{1-\beta}{2}$
- **Shortest interval**: $f(x_L \mid \theta) = f(x_U \mid \theta) \quad \wedge \quad \int_{x_L}^{x_U} f(x \mid \theta)dx = \beta$
- **Lower limit**: $\int_{-\infty}^{x_L} f(x \mid \theta)dx = 1 - \beta$
- **Upper limit**: $\int +x_U{}^{+\infty} f(x \mid \theta)dx = 1 - \beta$



Neyman confidence belt

- Take a variable x with a PDF $f(x \mid \theta)$
- Suppose we can take a function of the data $t(x)$ so that $\beta = P\left(t_1(\theta) \leq t \leq t_2(\theta)\right)$
- Suppose that we can compute t_1 and t_2 for each value of θ
  $\rightarrow$ The space between the curves $t_1(\theta)$ and $t_2(\theta)$ is the Neyman confidence belt
- For the case $t = x$, and Gaussian-distributed x:



Flip-flopping

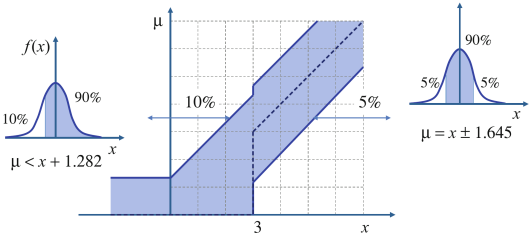
- Take a variable x with a Gaussian PDF with known σ
- Suppose Physics tells us that $\mu \geq 0$
- Suppose the observable x is subject to instrumental fluctuations and can be negative
- We can decide to quote:

$$\hat{\mu}(x) = \begin{cases} x \text{ if } \frac{x}{\sigma} \geq 3 \\ 0 \text{ otherwise} \end{cases}$$

- The corresponding interval will be:

$$[\mu_1, \mu_2] = \begin{cases} [\hat{\mu} - 1.65\sigma, \hat{\mu} + 1.65\sigma] \text{ if } \frac{x}{\sigma} \geq 3 \\ [0, \hat{\mu} + 1.282\sigma] \text{ otherwise} \end{cases}$$

- For some values we have a coverage of 85% only!



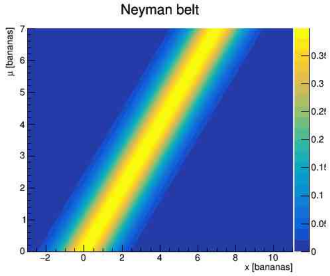
Unified Feldman-Cousins approach

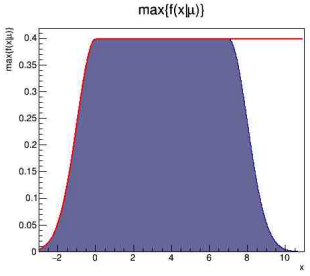
- Solution to flip-flopping issue: invent an ordering rule that allows us to smoothly switch from a two-sided interval to an upper limit with no dependence on the measured data
- Use the $\mathcal{L}$-ratio as an ordering rule

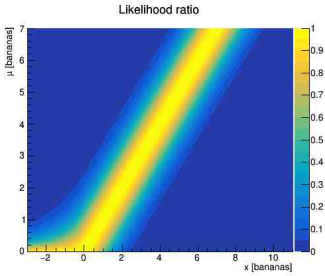
$$\lambda(x \mid \theta_0) = \frac{\mathcal{L}(x \mid \theta_0)}{\mathcal{L}(x \mid \hat{\theta})}$$

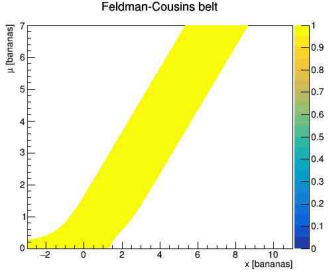
- Procedure:
  1. Find the best-fit $\hat{\theta}$
  2. Fix θ to some value θ_0, then start from $\hat{x} = \max_x \lambda(x \mid \theta_0)$ and move left and right until we get a coverage β on $\mathcal{L}(x \mid \theta_0)$
  3. Repeat for all values of θ (in the physical range)

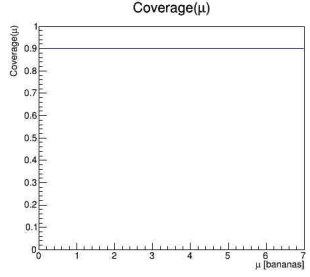
Feldman-Cousins for Gaussian case

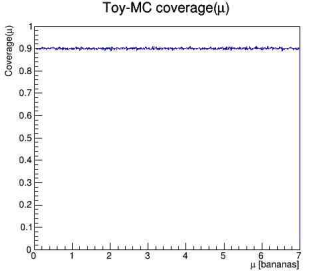












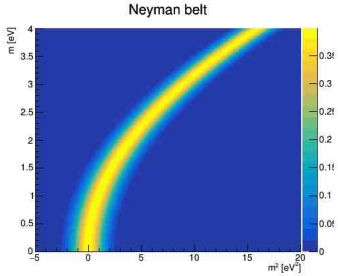
Feldman-Cousins for KATRIN data

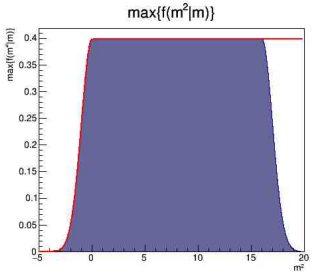
- In Phys. Rev. Lett. 123 (2019) 221802, the KATRIN experiment reported the following result on the electron neutrino mass:
  $m^2 = -1.0^{+0.9}_{-1.1}\,\mathrm{eV}^2 \quad \rightarrow$ This is their fit observable
  $m_\nu < 1.1\,\mathrm{eV} \quad \rightarrow$ This is their parameter of interest
- We can approximate the PDF of m^2_ν with a Gaussian:

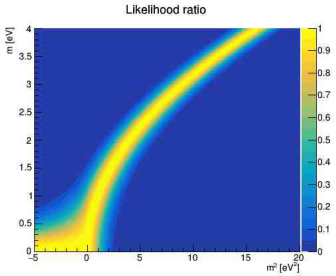
$$f(m^2|m_\nu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(m^2 - m^2_\nu)^2}{2\sigma^2}\right)$$

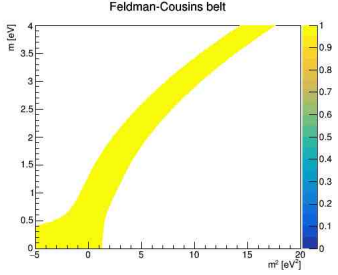
  with $\sigma = 1$, which is very close to the asymmetric uncertainties quoted above
- By doing this, we get $m_\nu < 0.91\,\mathrm{eV}$, which is pretty close to their published value!

Feldman-Cousins for KATRIN data









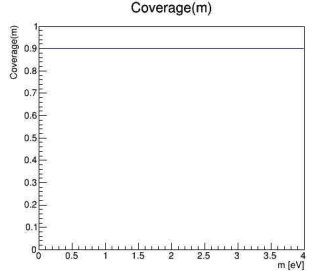


Table of Contents

Bayesian parameter estimation

- Take the usual n measurement of a variable $\vec{x}$ with PDF $f(\vec{x} \mid \vec{\theta})$
- Before the measurement, our degree of belief on the parameters is $\pi(\vec{\theta})$
- The probability of obtaining exactly the data $\vec{x}$ is

$$P(\vec{x} \mid \vec{\theta}) = \mathcal{L}(\vec{x} \mid \vec{\theta})$$

- We can use the Bayes theorem to find the posterior probability for the parameters $\vec{\theta}$ given the data $\vec{x}$:

$$P\left(\vec{\theta} \mid \vec{x}\right) = \frac{\mathcal{L}\left(\vec{x} \mid \vec{\theta}\right) \pi\left(\vec{\theta}\right)}{\int_\Omega \mathcal{L}\left(\vec{x} \mid \vec{\theta}\right) \pi\left(\vec{\theta}\right)}$$

- The global mode of $P\left(\vec{\theta} \mid \vec{x}\right)$ represent the most probable combination of all parameters

Bayesian interval estimation

- Suppose we are just interested in one of the parameters, e.g. θ, out of a list of parameters $(\theta, \vec{\nu})$
  $\rightarrow$ What is the most probable value for θ given the data $\vec{x}$?
  $\rightarrow$ What is the interval that contains the most probable value of θ with 68% probability?
- The answer is: **marginalization**

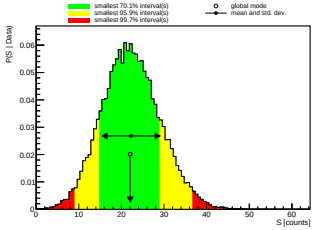
$$P(\theta \mid \vec{x}) = \int_{\Omega_\nu} P(\theta, \vec{\nu} \mid \vec{x}) d\vec{\nu}$$

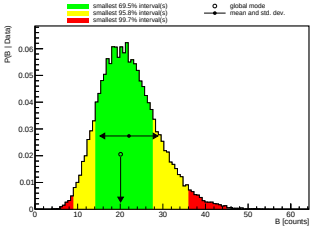
- From $P(\theta \mid \vec{x})$ we can quote:
  - The marginalized mode $\hat{\theta}$, which can also be at the boundary of the physics region
  - The central/shortest interval $[\theta_1, \theta_2]$, or an upper/lower limit (typically at 90% or 95%)
    $\rightarrow$ These intervals are called **credible intervals** because they state that, based on our current knowledge, we believe the true value of θ is in that range with the specified probability.
    $\rightarrow$ The concept of coverage does not apply here!

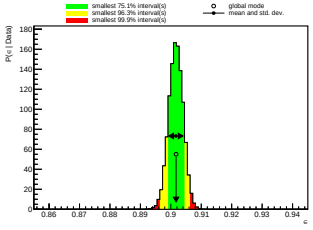
Example: Bayesian simultaneous fit

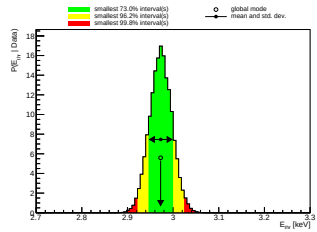
- Suppose we want to run a Bayesian fit of the pulser data described above and physics data, recorded between 10 and 20 keV, and featuring
  - a Gaussian signal with $\mu = 15\,\text{keV}$ and $\sigma = 1.2\,\text{keV}$
  - a flat background
- How does the marginalized distribution of each parameter look like?
- How do the correlation plots of all pairs of parameters look like?

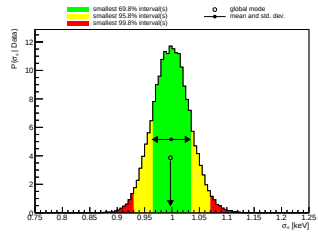
Example: Bayesian simultaneous fit

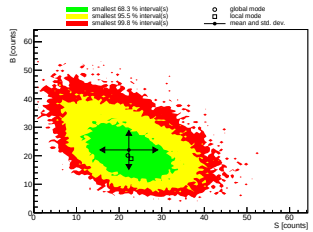












Bayesian fits: practical tips

- We need to **map** the posterior $P(\vec{\theta} \mid \vec{x})$ and integrate it over nuisance parameters, however $P(\vec{\theta} \mid \vec{x})$ might be complicated to integrate over $d\vec{\nu}$ or $d\vec{\theta}$
- Solution:
  1. Map $P(\vec{\theta} \mid \vec{x})$ using a MCMC with n tested samples
  2. The denominator $\int \mathcal{L}\pi d\vec{\theta}$ is just a constant, so forget about it
  3. Build the marginalized PDF of θ by histogramming its tested values
  4. Renormalize the marginalized PDF by dividing over n
  5. Extract $\hat{\theta}$ and the credible interval from the histogram

Questions?