Statistical Methods for Data Analysis in Particle Physics

Luca Lista

Università Federico II, Napoli

INFN Sezione di Napoli

Outline

- Definition of probability: frequentist vs Bayesian
- Bayesian inference
- Frequentist inference
- Hypothesis testing
- Significance level and discovery
- Upper limits
- Confidence intervals
- Modified frequentist approach: CLs
- Nuisance parameters
- Profile likelihood
- Asymptotic formulae
- Look elsewhere effect

- I assume you know basic tools like binomial, Poisson, Gaussian, ecc.

Definitions to probability

- Probability can be defined with different approaches
- The applicability of each approach depends on the type of claim we are assigning a probability to
- A subjective approach quantifies the degree of belief/credibility of a claim, which may vary from subject to subject
- For repeatable experiments, probability may be a measure of how frequently the claim is true in the unrealizable limit of infinite number of experiment
- Both above definition have some drawbacks

Classical probability

- Probability determined by symmetry properties of a random device, only applicable to simple cases
- "Equally undecided" about event outcome, according to Laplace definition

$$\text{Probability: } P = \frac{\text{Number of favorable cases}}{\text{Number of total cases}}$$

Pierre Simon Laplace (1749-1827)

$P = 1/2$

$P = 1/6$ (individual dice)

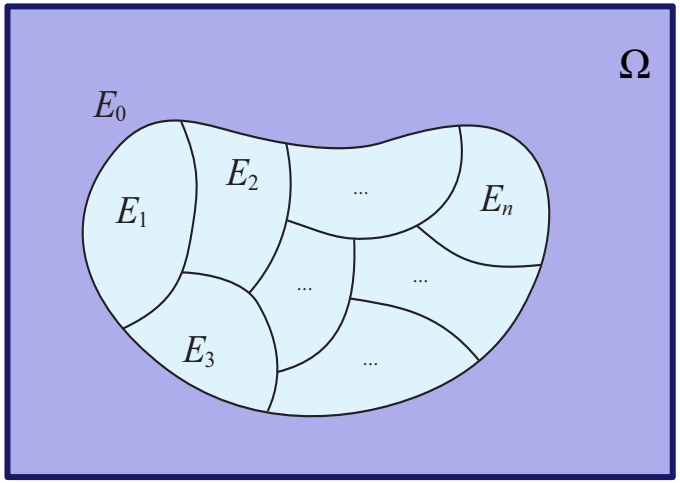
$P = 1/10$

$P = 1/4$

Kolmogorov axiomatic approach

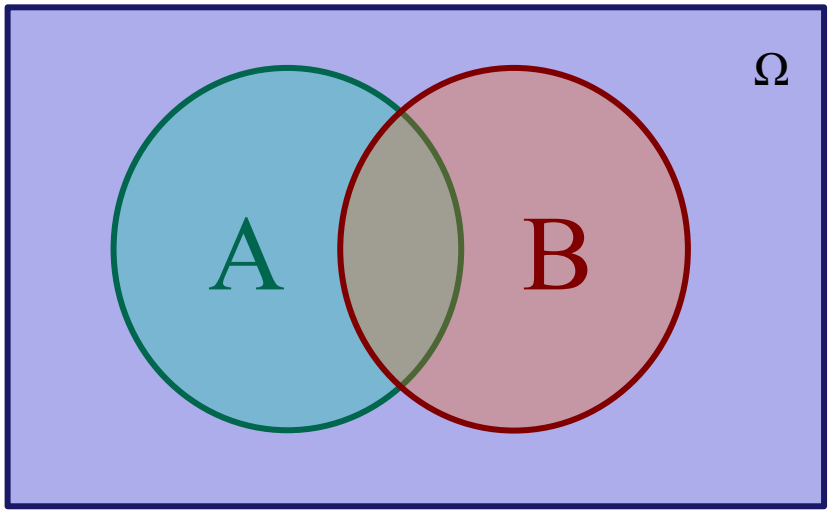
- Let $(\Omega, F \subseteq 2^{\Omega}, P)$ be a measure space, where: Ω = sample space, F = event space, P = probability measure
- Assume it satisfies:
  1. $P(E) \geq 0 \quad \forall E \in F$
  2. $P(\Omega) = 1$ (normalization)
  3. $\forall (E_1, \cdots, E_n) \in F^n$: $E_i \cap E_j = \emptyset$
     $P\left(\bigcup_{i=1,\cdots,n} E_i\right) = \sum_{i=1,\cdots,n} P(E_i)$
- The same formalism applies to different approaches to probability

Andrej Nikolaevič Kolmogorov (1903-1987)



Conditional probability

- $P(A|B)$ is the probability of A, given B, i.e.: that an event known to belong to set B also belongs to set A:
  - $P(A|B) = P(A \cap B)/P(B)$
  - Notice that:
  - $P(A|\Omega) = P(A \cap \Omega)/P(\Omega)$
- A is said to be independent of B if:
  - $P(A|B) = P(A)$
- If A is independent of B, then $P(A \cap B) = P(A)P(B)$
- → If A is independent on B, B is independent on A



PDF and independent variables

- Probability density for continuous case: $\frac{\mathrm{d}^2 P}{\mathrm{d}x\mathrm{d}y} = f(x,y)$
- 1D projections: $\begin{cases} f_x(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y \\ f_y(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x \end{cases}$ (marginal distributions)
- We saw that A and B are independent events if:

$$P(A \cap B) = P(A)\,P(B)$$

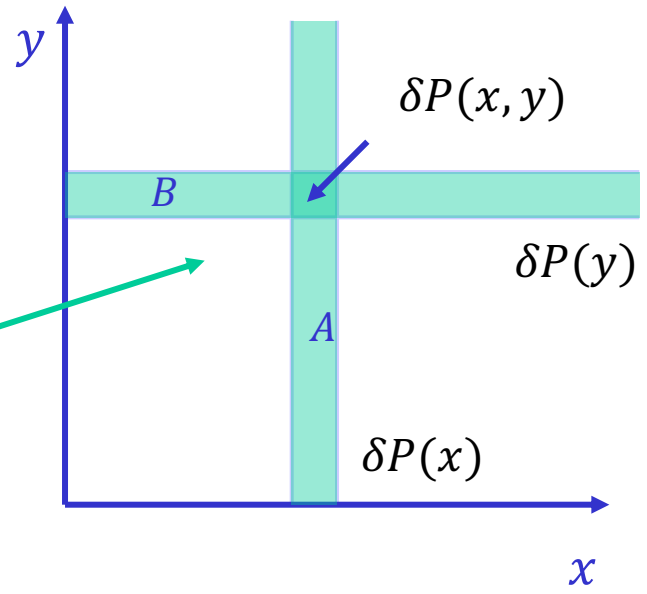
- Applying in the following case:

$$A = \{x' : x < x' < x + \delta x\}$$

$$B = \{y' : y < y' < y + \delta y\}$$

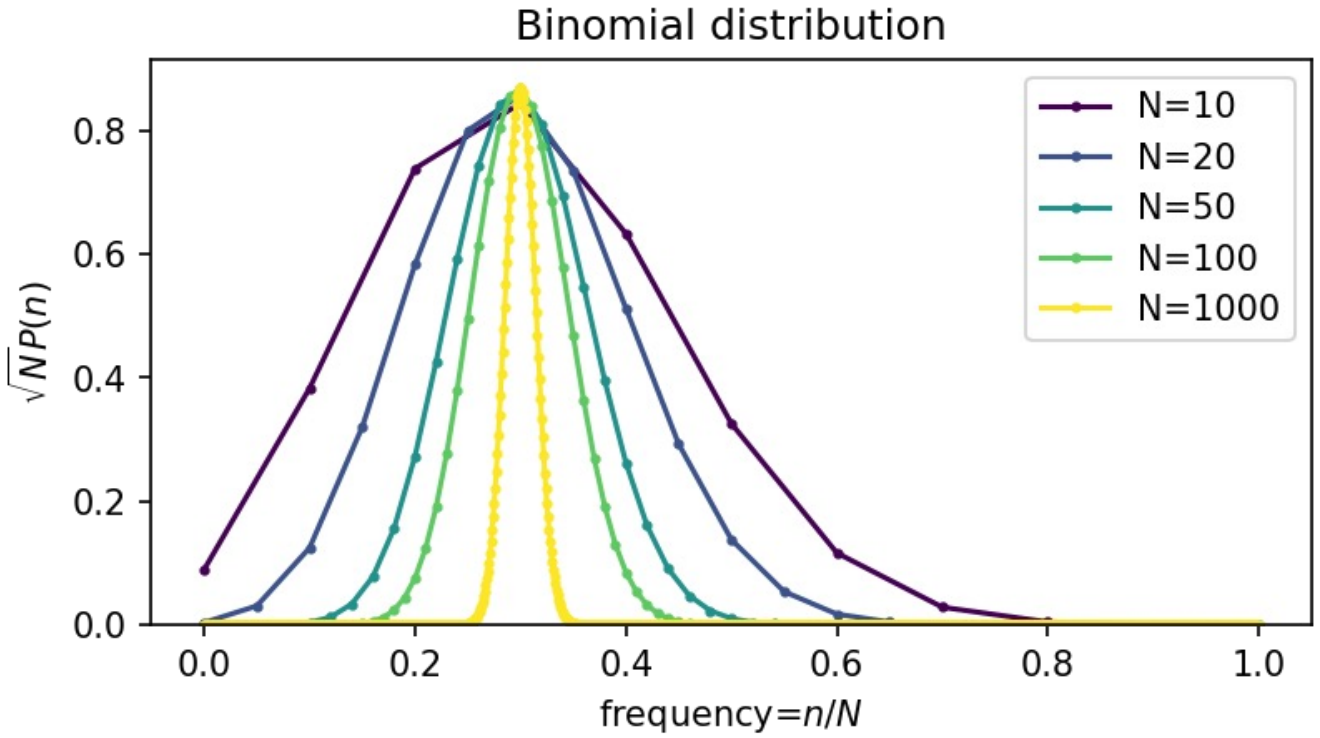
- We define that x and y are independent variables if:

$$f(x,y) = f_x(x) f_y(y)$$



Law of large numbers

- The number of success events n in N repeated random extractions follows a binomial distribution
- The frequency of favorable cases is just n/N
- $\forall \varepsilon \quad \lim_{N\to\infty} P\left(\left|\frac{n}{N} - p\right| < \varepsilon\right) = 1$



It is a consequence of general probability laws. It will be used as foundation for frequentist probability, but holds for any approach to probability

The Bayes theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A|B)P(B) = P(B|A)P(A)$$

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

- $P(A)$: prior probability
- $P(A|B)$: posterior probability

Thomas Bayes (1702-1761)

Bayesian posterior probability

- Bayes theorem allows to define a probability about hypotheses or claims H that not related random variables, given an observation or evidence E:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

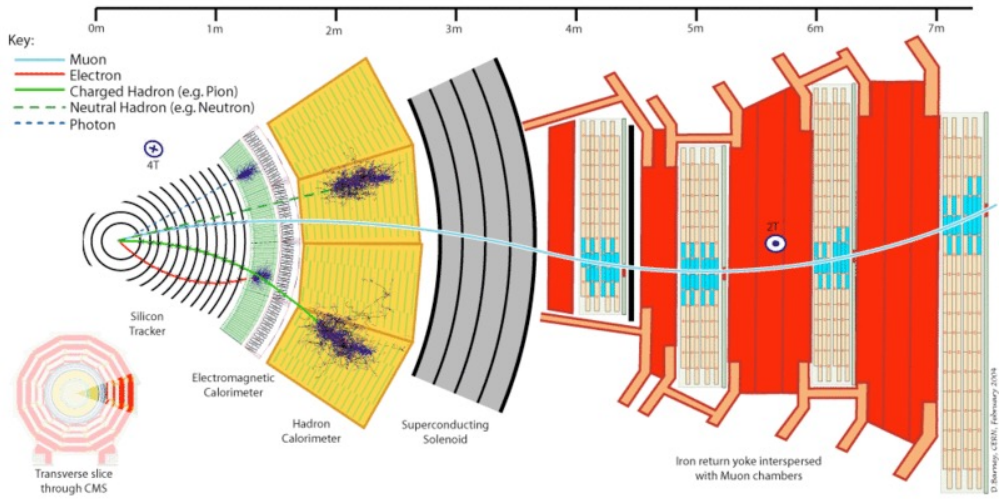
- The Bayes rule allows to define a rational way to modify one's prior degree of belief once some observation is known
- $P(H)$: prior probability, degree of belief before evidence
- $P(H|E)$: posterior probability, degree of belief given E

Subjective probability

- Expresses one's degree of belief that a claim is true
  - How strong? How much would you bet?
  - Applicable to all unknown events/claims, not only repeatable experiments
  - Different individuals may have a different opinion/prejudice
- Bayes theorem provides a prescription about how subjective probability should be modified after **learning** about some observation/evidence
  - The prior is an unavoidable subjective element of Bayesian probability
  - The more information we receive, the more Bayesian probability is insensitive on prior subjective prejudice, except for pathological priors

Bayes th. example: muon fake rate

- A detector identifies muons with high efficiently, $\varepsilon = 95\%$
- A small fraction $\delta = 5\%$ of pions are incorrectly identified as muons ("fakes")
- If a particle is identified as a muon, what is the probability it is really a muon?
  - The answer also depends on the composition of the sample!
  - i.e.: the fraction of muons and pions in the overall sample



This example is usually presented as an epidemiology case

Naïve answers about fake positive probability may often be wrong!

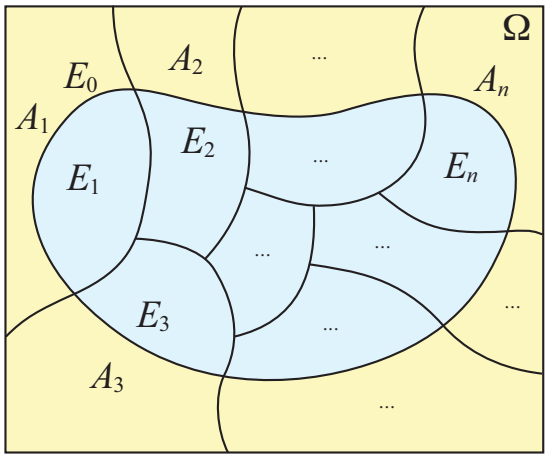
Purity and Bayes theorem

- Using Bayes theorem:
  - $P(\mu|+) = P(+|\mu)\,P(\mu) / P(+)$
- Where our inputs are:
  - $P(+|\mu) = \varepsilon = 0.95, P(+|\pi) = \delta = 0.05$
- We can decompose $P(+)$ as:
  - $P(+) = P(+|\mu)\,P(\mu) + P(+|\pi)\,P(\pi)$
- Putting all together:
  - $P(\mu|+) = \varepsilon P(\mu)/(\varepsilon P(\mu) + \delta\,P(\pi))$
- Assume we have $P(\mu) = 4\%$ of muons and $P(\pi) = 96\%$ of pions, we have:
  - $P(\mu|+) = 0.95\times0.04/(0.95\times0.04 + 0.05\times0.96) \cong 0.44$
- Even if the selection efficiency is very high, the low sample purity makes $P(\mu|+)$ lower than 50%.

+ denotes a positive id

normalization term

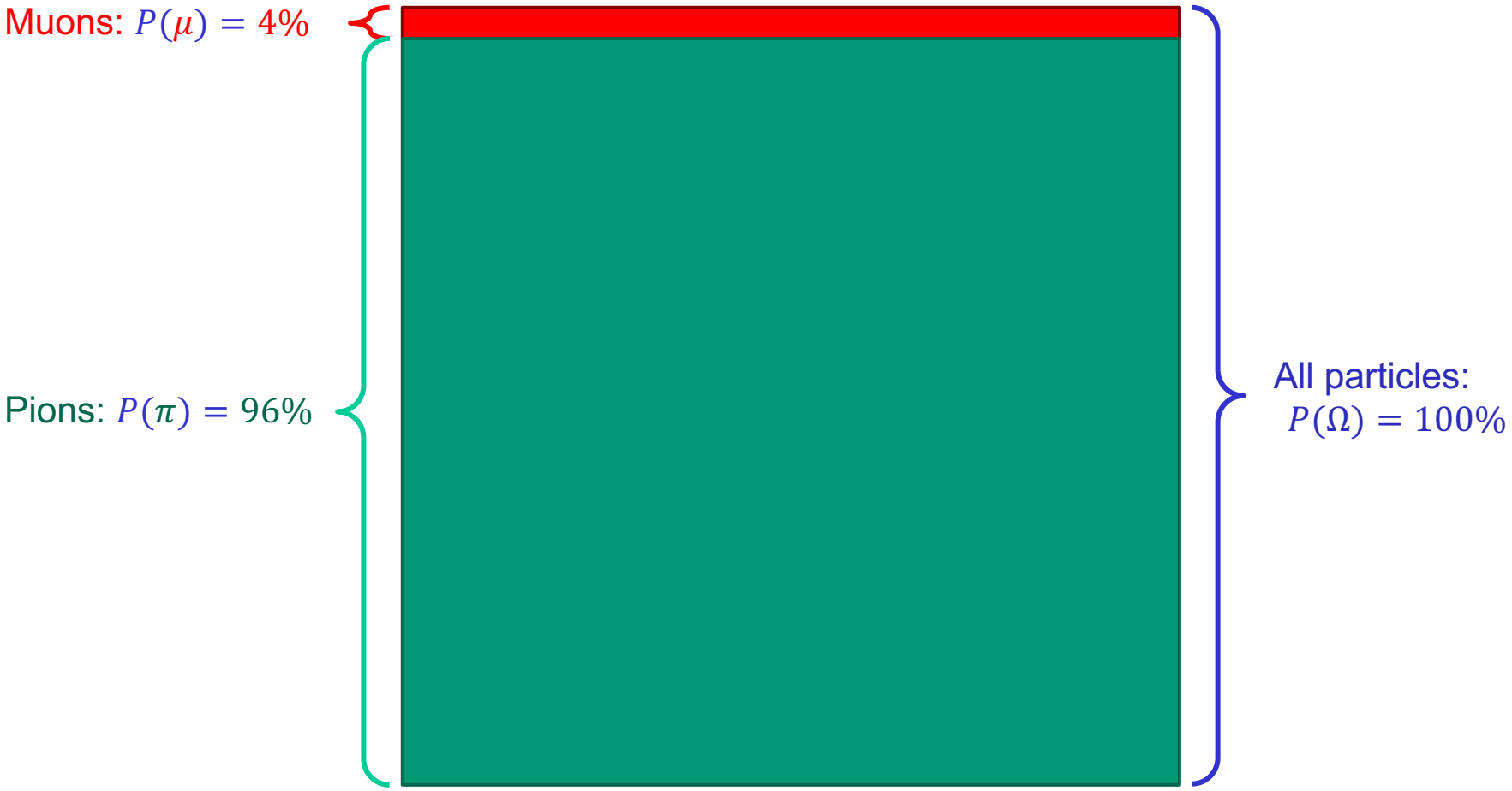
Law of total probability



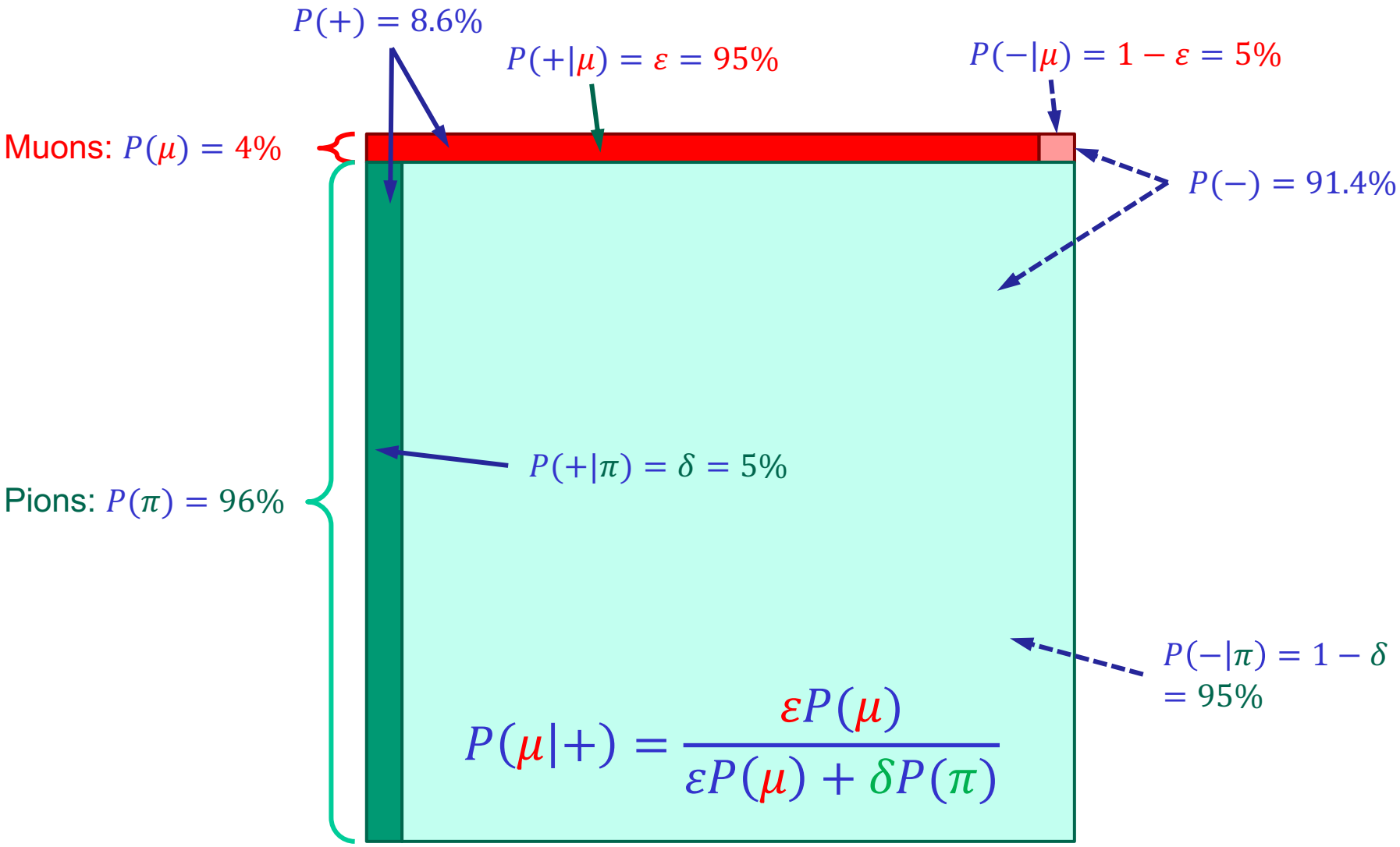
$$P(E_0) = \sum_{i=1}^{n} P(E_0|A_i)P(A_i)$$

$E_0 = \text{`}+\text{'}, Ai \_ \mu, \pi$

Before any muon id. information

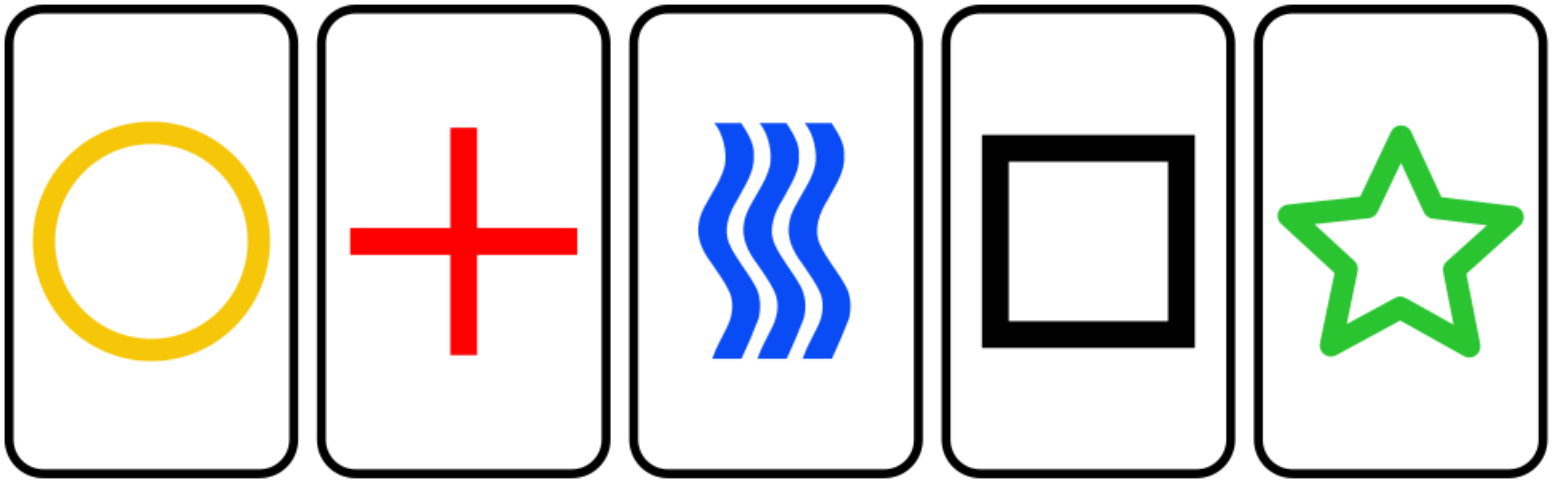


After the muon id. measurement

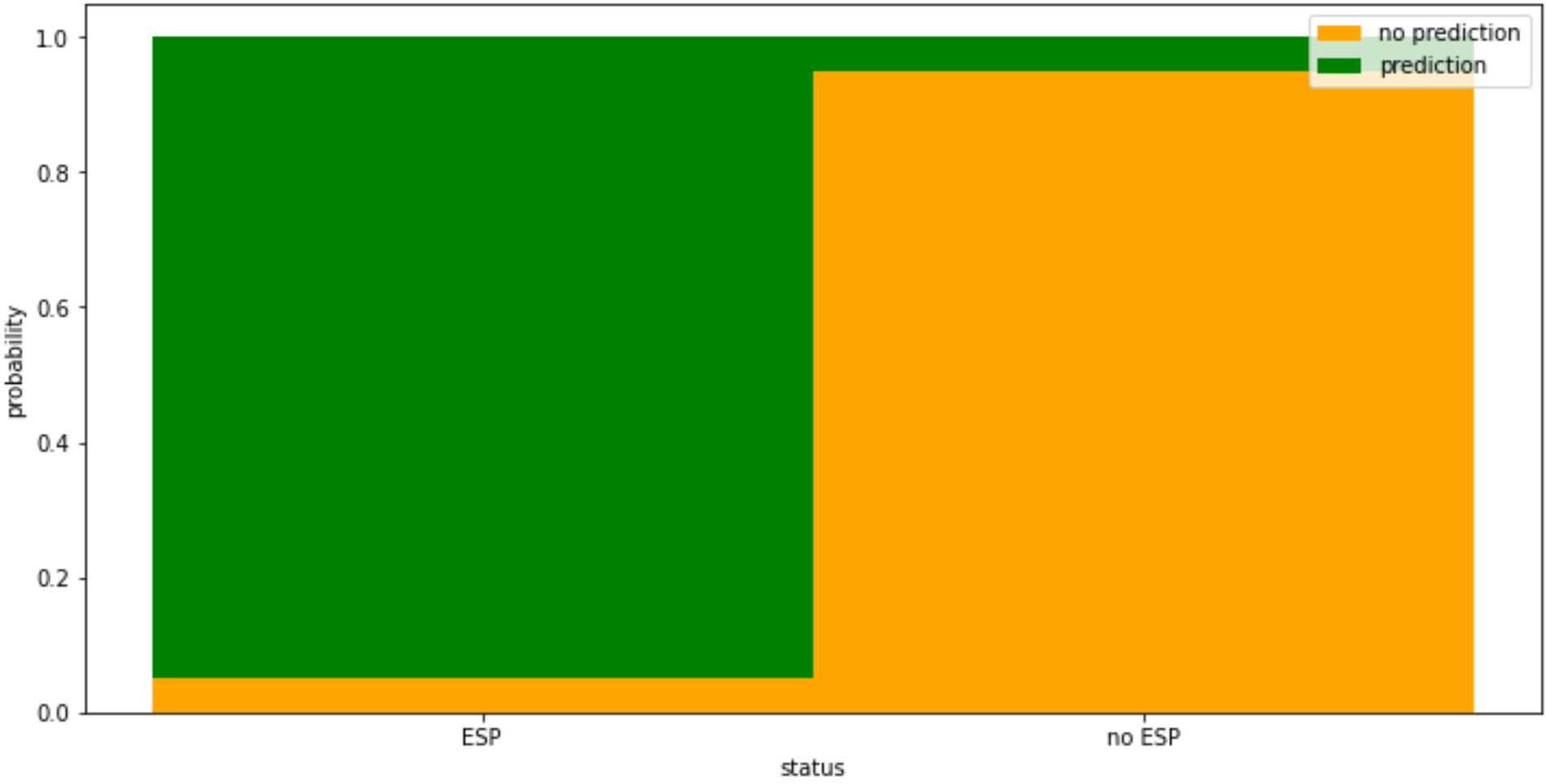


Bayesian approach to unknowns

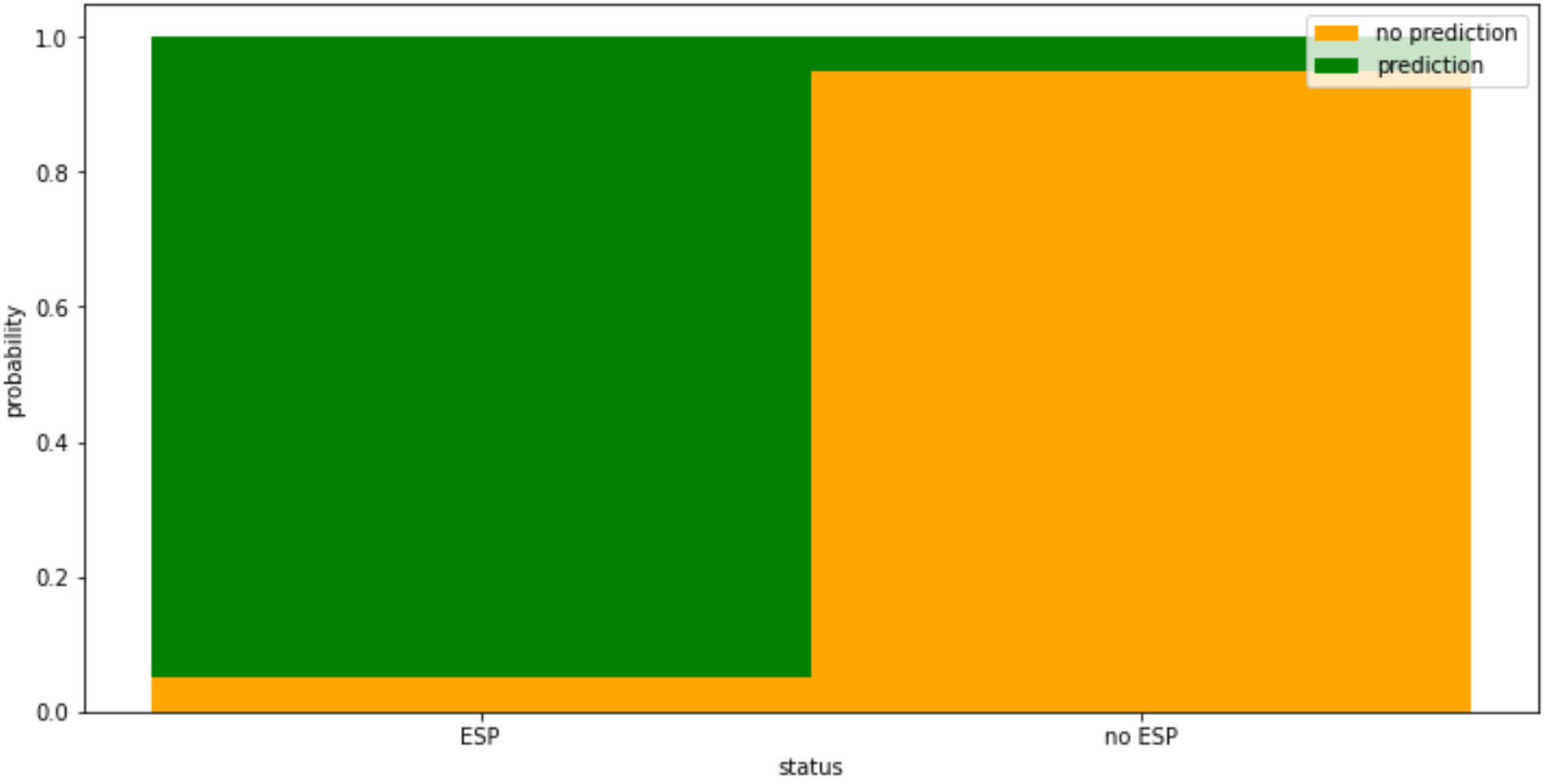
- ESP: extra-sensory perception: prediction of extractions from a set of cards



50% – 50% prior

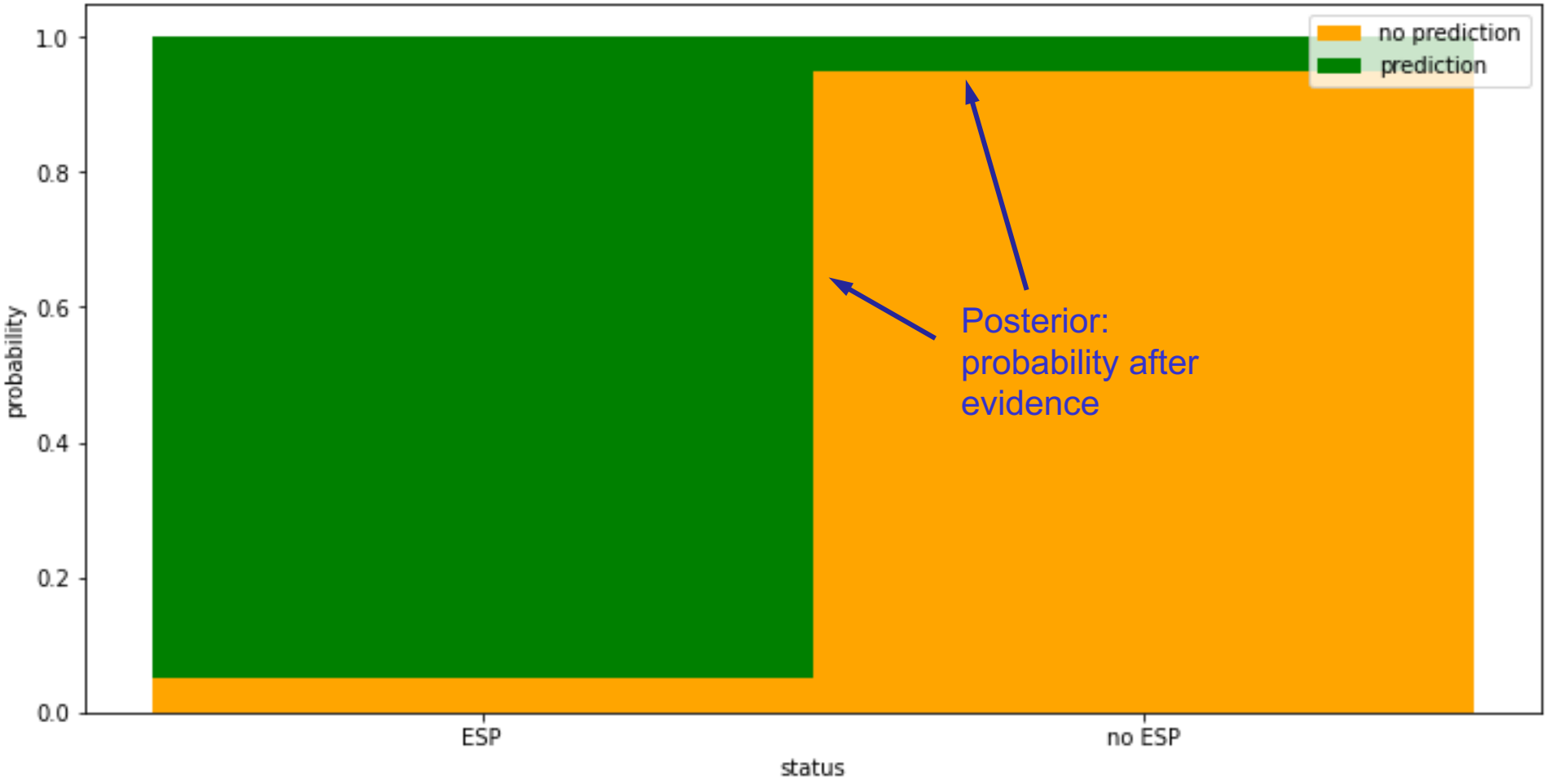

no prediction
prediction
1.0
0.8
0.6
0.4
0.2
0.0
probability
ESP
no ESP
status

50% – 50% prior

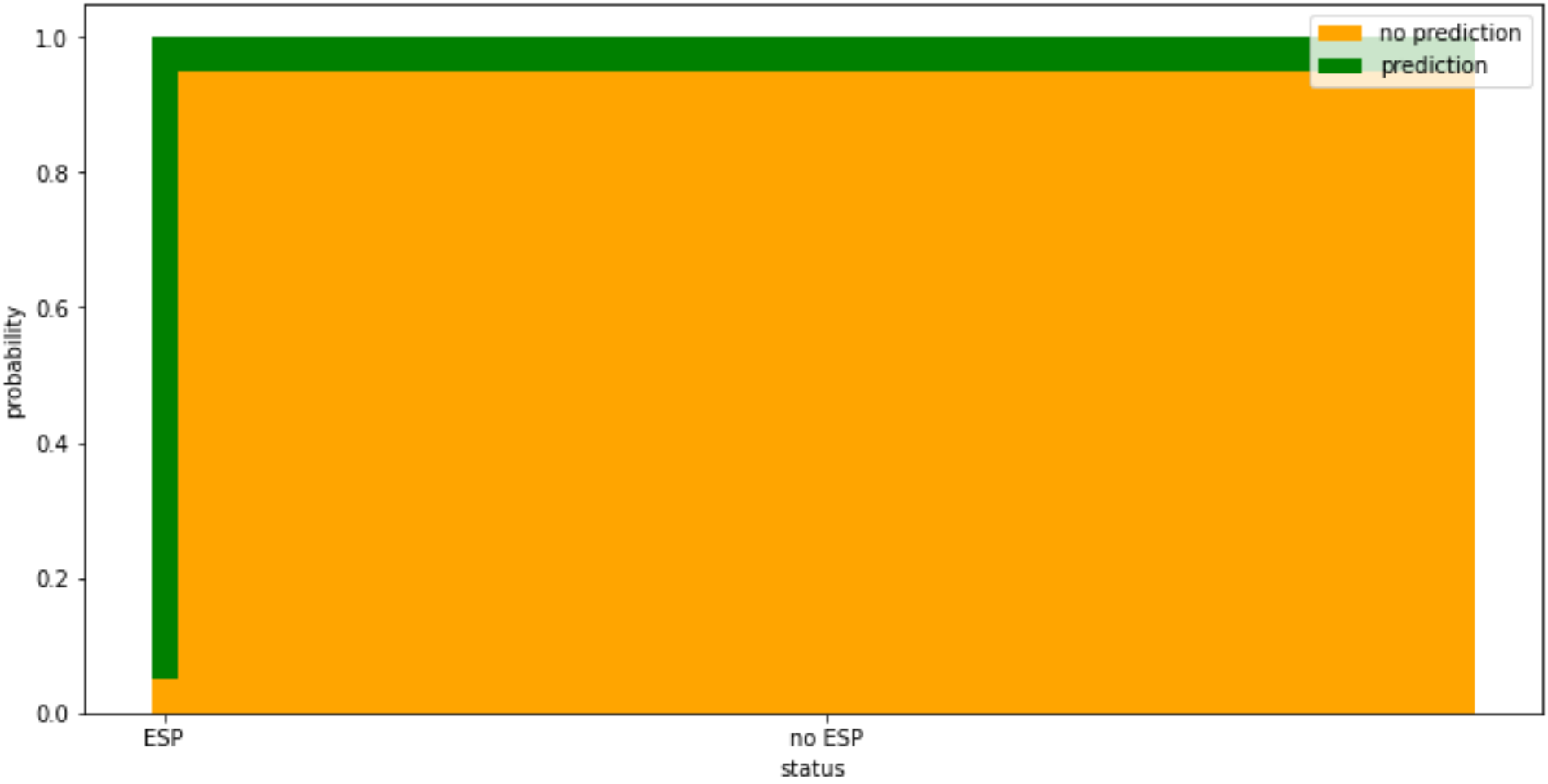


Prior: probability before evidence

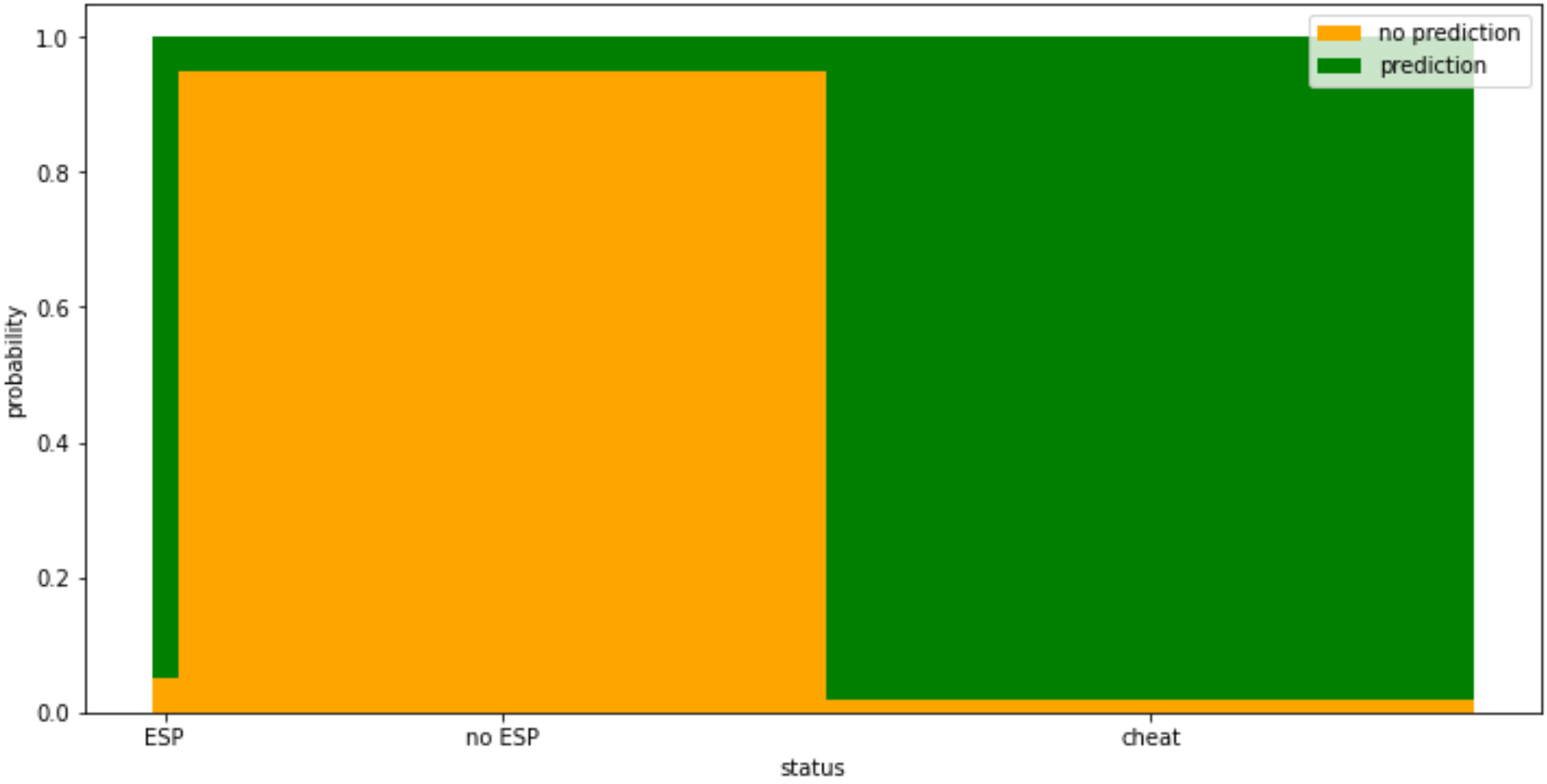
50% – 50% prior


no prediction
prediction
1.0
0.8
0.6
0.4
0.2
0.0
probability
ESP
no ESP
status
Posterior:
probability after
evidence

Skeptical prior



Suspicious prior



Note on conspiracy theories

- Conspiracy theories claim that all scientific evidences are fakes as well!
- The main difference is that the priors chosen by conspiracy theorists are not based on common knowledge, but ignore most of known evidences in favour of pure inventions
- Scientific reasoning is closely related to the Bayesian approach, assuming that evidence is not discarded on purpose

The likelihood function

- In many cases, the outcome of an experiment is modeled as a set of random variables $x_1, \dots, xn$ whose distribution depends on:
  - intrinsic sample randomness (quantum physics is intrinsically random)
  - detector effects (resolution, efficiency, …)
- Theory and detector effects are described in terms of some parameters $\theta_1, \dots, \theta_m$, whose values are unknown
- The overall PDF, evaluated at our observation $x_1, \dots, xn$, is called likelihood function:

$$L = f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)$$

- In case our sample consists of N independent measurements (collision events) the likelihood function can be written as:

$$L = \prod_{i=1}^{N} f(x_1^{(i)}, \cdots, x_N^{(i)}; \theta_1, \cdots, \theta_m)$$

Bayes rule and likelihood function

- Given a set of measurements $x_1, \dots, x_n$, Bayesian posterior PDF of the unknown parameters $\theta_1, \dots, \theta_m$ can be determined as:

$$P(\theta_1, \dots, \theta_m | x_1, \dots, x_n) = \\ = \frac{L(x_1, \dots, x_n | \theta_1, \dots, \theta_m)\pi(\theta_1, \dots, \theta_m)}{\int L(x_1, \dots, x_n | \theta'_1, \dots, \theta'_m)\pi(\theta'_1, \dots, \theta'_m)\, \mathrm{d}^m\theta'}$$

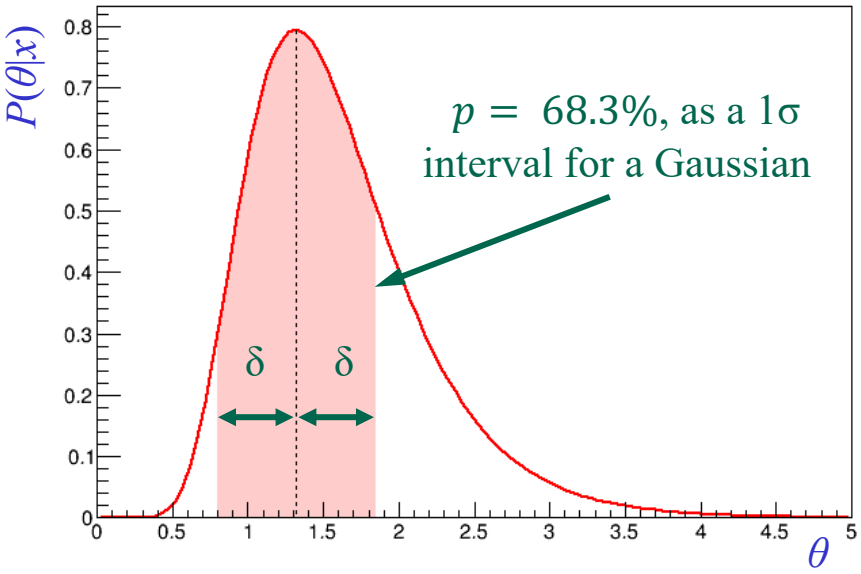
- Where $\pi(\theta_1, \dots, \theta_m)$ is the subjective prior probability
- The denominator $\int L(\dots)\pi(\dots)\, \mathrm{d}^m\theta'$ is a normalization factor
- The observation of $x_1, \dots, x_n$ modifies the prior knowledge of the unknown parameters $\theta_1, \dots, \theta_m$
- If $\theta_1, \dots, \theta_m$ is sufficiently smooth and L is sharply peaked around the true values $\theta_1, \dots, \theta_m$, the resulting posterior will not be strongly dependent on the prior's choice because $\pi(\theta_1, \dots, \theta_m)$ can be approximated to a constant, and cancels in the ratio

Bayesian inference

- The posterior PDF provides all the information about the unknown parameters (let's assume here it's just a single parameter θ for simplicity)

$$P(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,\mathrm{d}\theta'}$$

- Given $P(\theta|x)$, we can determine:
  - The most probable value (best estimate)
  - Intervals corresponding to a specified probability
- Notice that if $\pi(\theta)$ is a constant, the most probable value of θ correspond to the maximum of the likelihood function



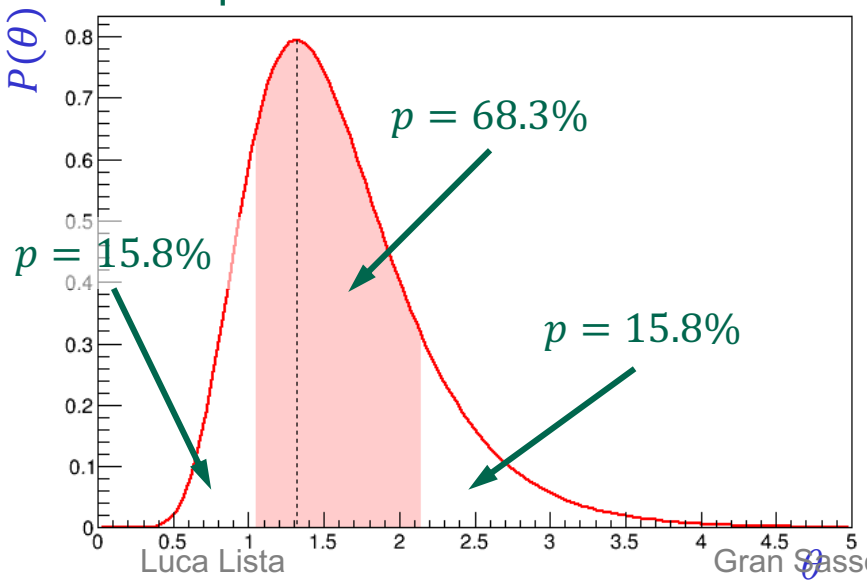
Choice of 68% prob. intervals

- Different interval choices are possible, corresponding to the same probability level (usually 68%)
  - Equal areas in the right and left tails
  - Symmetric interval
  - Shortest interval
  - …

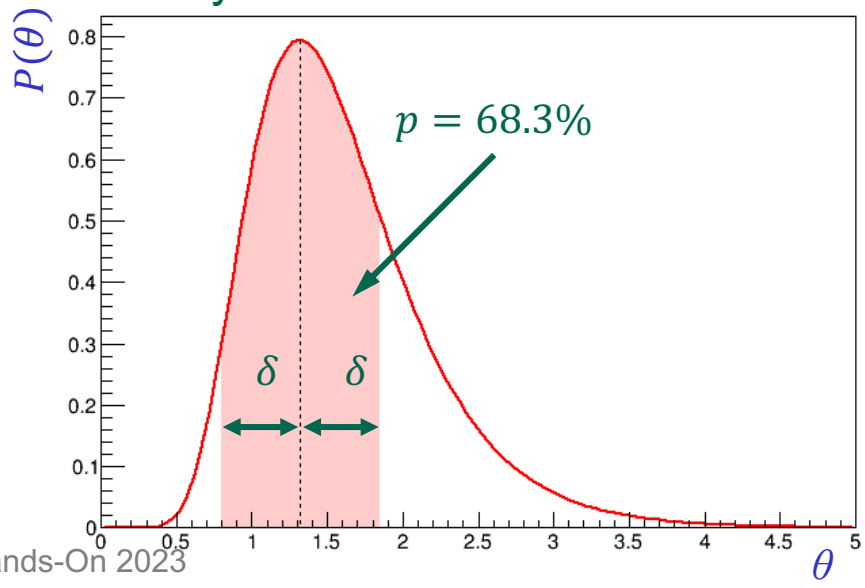
} All equivalent for a symmetric distribution (e.g. Gaussian)

- Reported as $\theta = \hat{\theta} \pm \delta$ (sym.) or $\theta = \hat{\theta}^{+\delta_1}_{-\delta_2}$ (asym.)

Equal tails interval

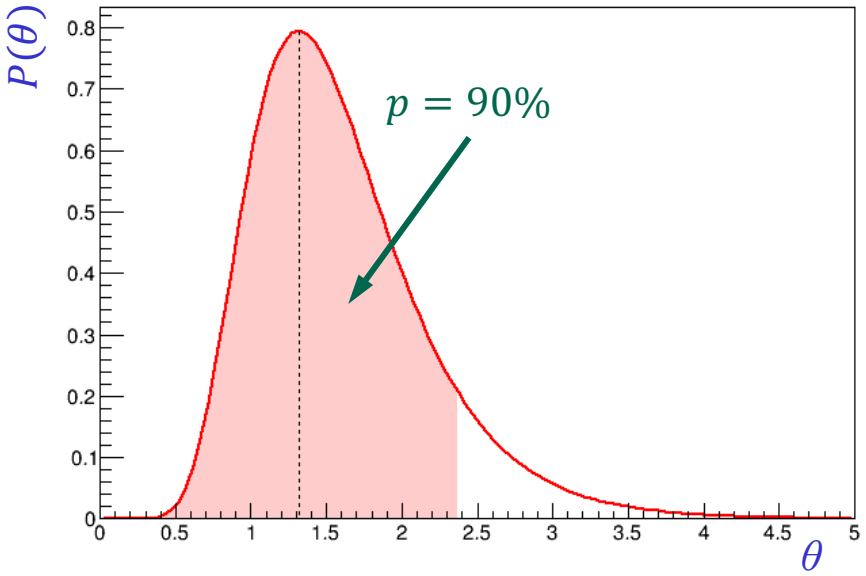


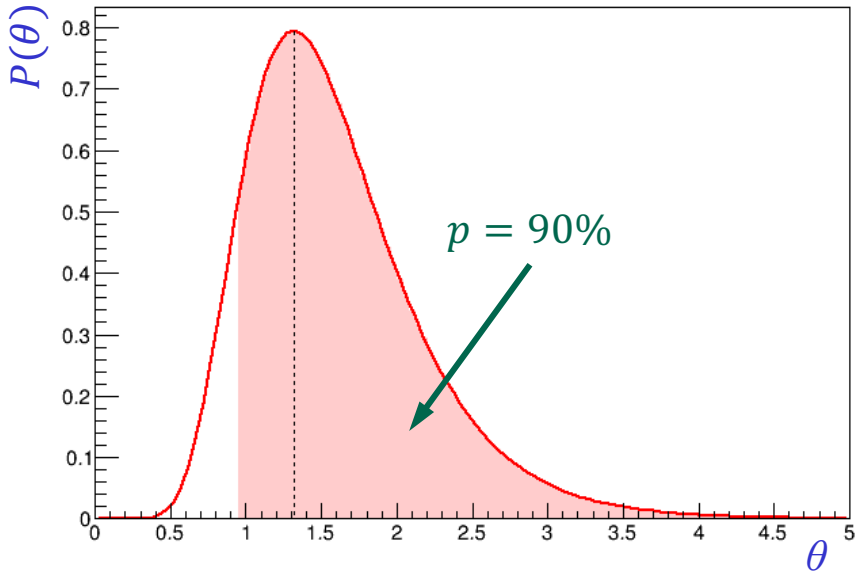
Symmetric interval



Upper and lower limits

- A fully asymmetric interval choice is obtained setting one extreme of the interval to the lowest or highest allowed range
- The other extreme indicates an upper or lower limits to the "allowed" range
- For upper or lower limits, usually a probability of 90% or 95% is preferred to the usual 68% adopted for central intervals
- Reported as: $\theta < \theta^{\mathrm{up}}\ (90\%\ CL)$ or $\theta > \theta^{\mathrm{lo}}\ (90\%\ CL)$





Bayesian inference of a Poissonian

- Posterior PDF, assuming the prior $\pi(s)$:

$$P(s|n) = \frac{\frac{s^n e^{-s}}{n!}\pi(s)}{\int_0^\infty \frac{s'^n e^{-s'}}{n!}\pi(s')\mathrm{d}s'}$$

If is $\pi(s)$ is uniform:

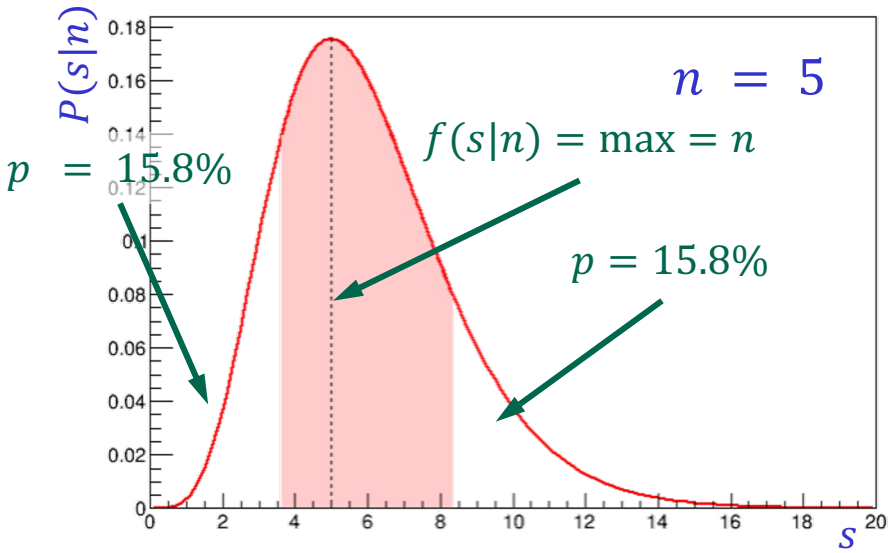
$$P(s|n) = \frac{s^n e^{-s}}{n!}$$

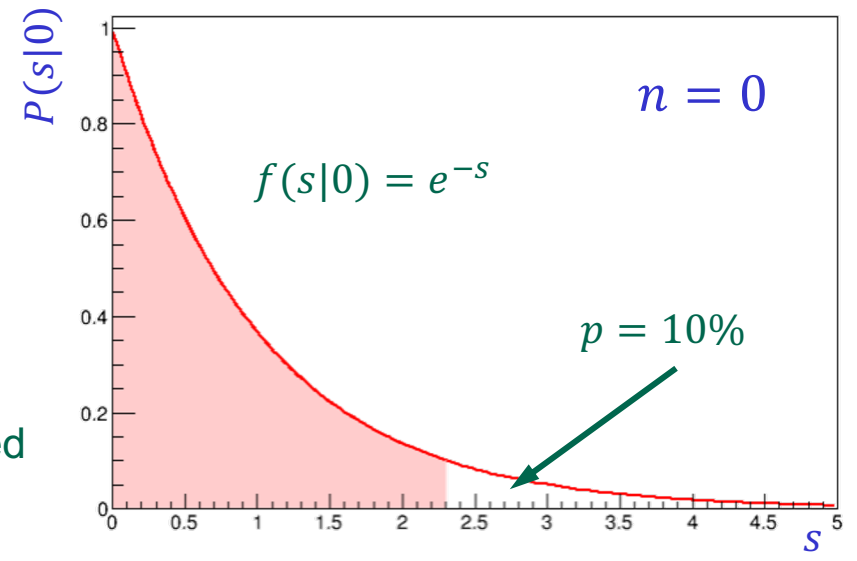
- From which:

$$\langle s \rangle = n + 1, \ \mathbb{V}\mathrm{ar}[s] = n + 1$$

- For $n = 0$, one may quote an upper limit at 90% or 95% CL:
- $s < 2.303$ $(90\%$ CL$)$
- $s < 2.996$ $(95\%$ CL$)$

} zero observed events

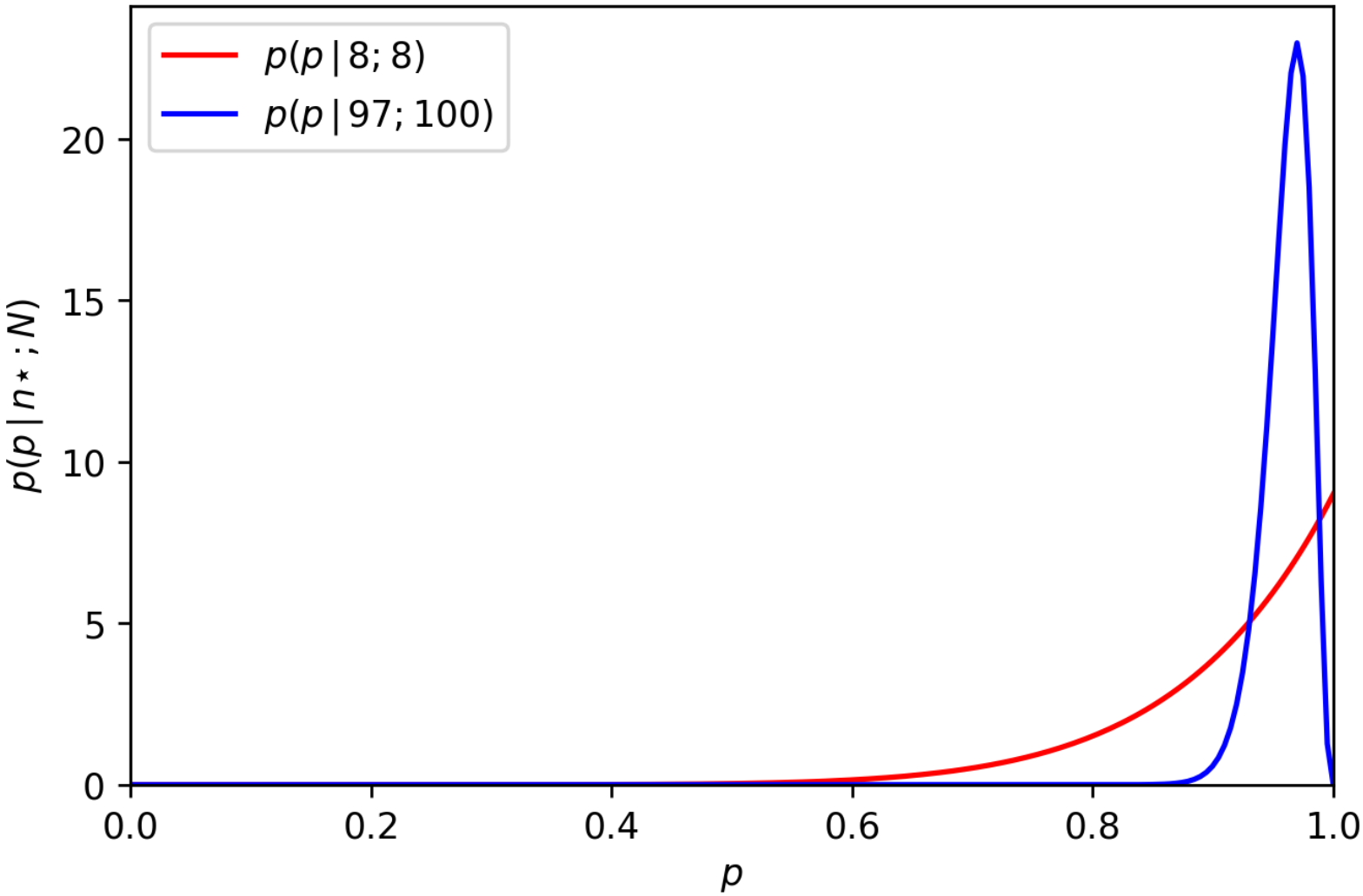




Feedbacks and reliability

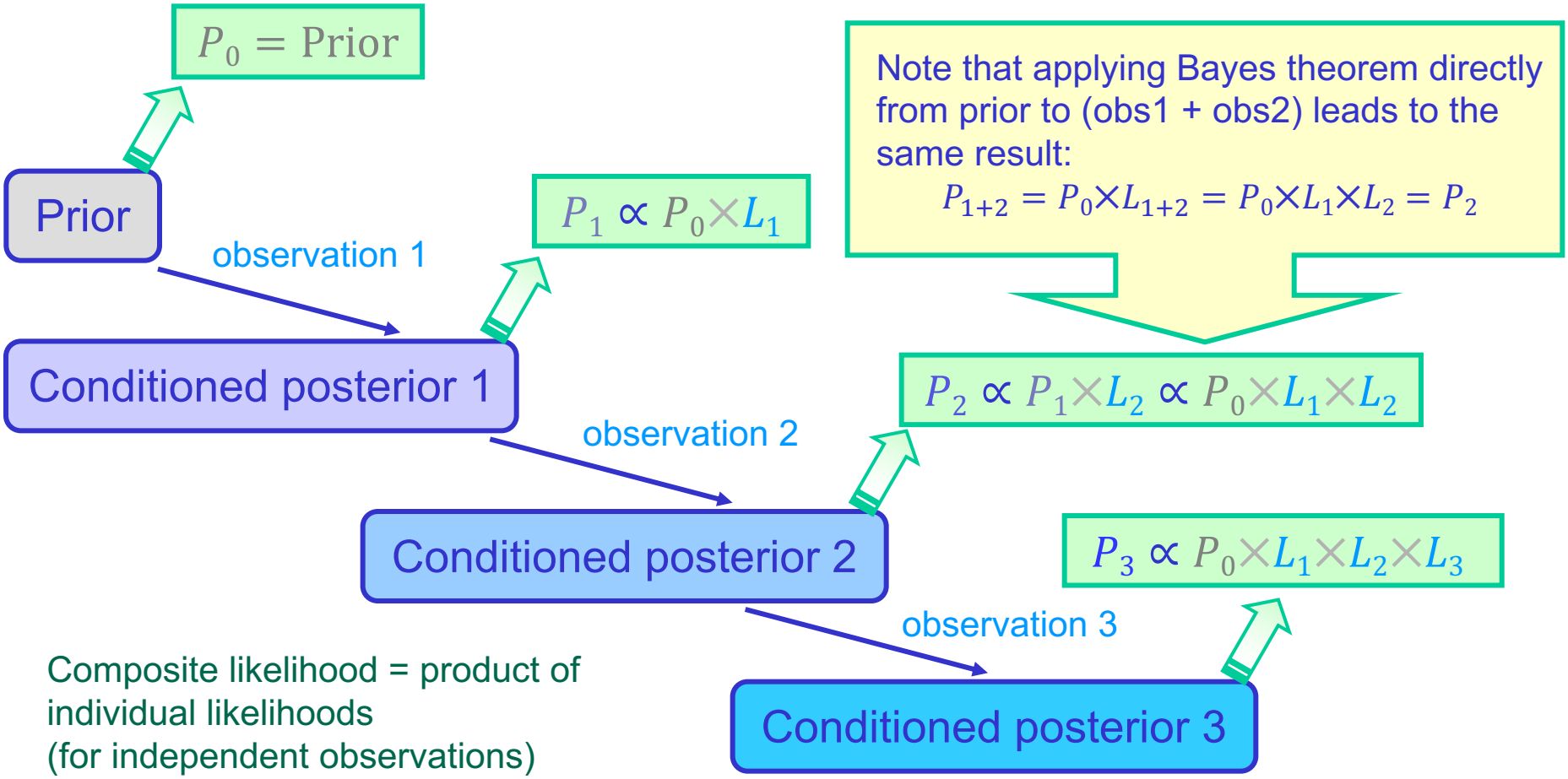
- The number of positive/negative feedbacks of an online seller can be assumed to to follow a binomial distribution
- If:
  - a seller has 100% of positive feedbacks $(8/8)$ and
  - another one has 97% $(97/100)$, which one is more reliable?
- The posterior a special case of the so-called Beta distribution, proportional to: $p^n(1-p)^{N-n}$:

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)}$$


$p(p\,|\,8;8)$
$p(p\,|\,97;100)$
$p(p\,|\,n\star;N)$
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
p

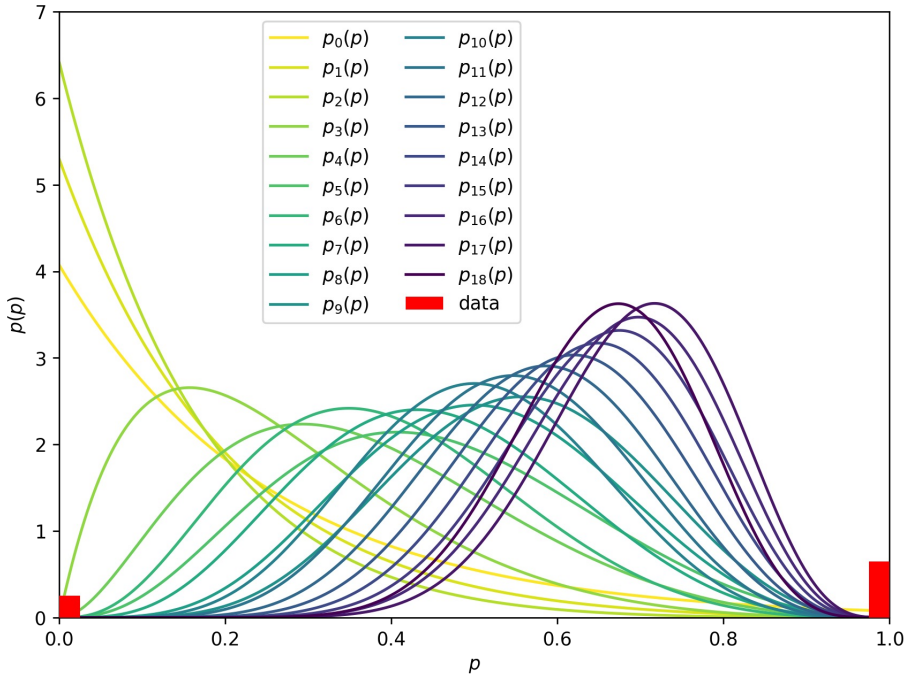
Repeated use of Bayes theorem

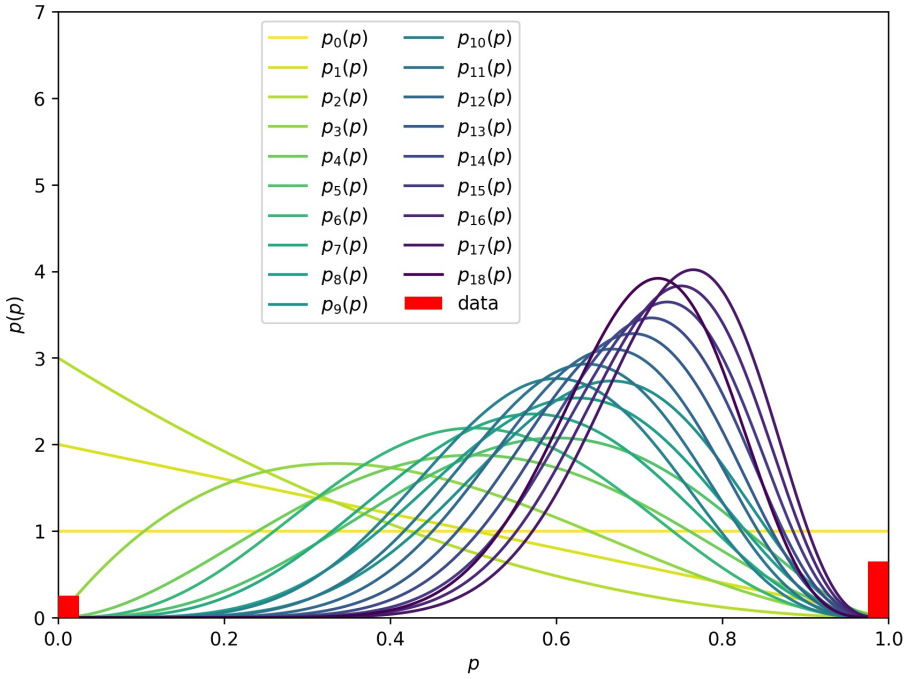
- Bayes theorem can be applied sequentially for repeated independent observations (posterior PDF = learning from experiments)



Bayesian inference as learning

- Inference of a Binomial parameter as repeated application of Bayes rule for many Bernoulli extractions:
- $1 \rightarrow p_{i+1} = p_i \times p$
- $0 \rightarrow p_{i+1} = p_i \times (1 - p)$





Choosing the prior PDF

- If the prior PDF is uniform in a choice of variable, it won't be uniform when applying coordinate transformation
- Given a prior PDF in a random variable, there is always a transformation that makes the PDF uniform
- The problem is: chose one metric where the PDF is uniform
- Harold Jeffreys' prior: chose the prior form that is invariant under parameter transformation

$$\pi(\vec{\theta}) \propto \sqrt{\det \mathcal{I}(\vec{\theta})},\ \mathcal{I}_{ij}(\vec{\theta}) = \mathbb{E}\left[\frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i}\frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i}\right]$$ } Fisher information

- Some commonly used cases:
  - Poissonian mean: $\pi(\theta) \propto 1/\sqrt{\mu}$
  - Poissonian mean with background b: $\pi(\theta) \propto 1/\sqrt{\mu + b}$
  - Gaussian mean: $\pi(\theta) \propto 1$
  - Gaussian standard deviation: $\pi(\theta) \propto 1/\sigma$
  - Binomial parameter: $\pi(\theta) \propto 1/\sqrt{p(1-p)}$

Note: in a previous Poissonian example we used $\pi(\mu) = \text{const.}$, which is not Jeffreys' prior!

- Problematic with PDF in more than one dimension!

Lifetime estimate

- The probability distribution for a single measurement is:

$$p(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

- If we measure N decay times, we can write the likelihood function as:

$$L(t_1, \ldots, t_N|\tau) = \frac{1}{\tau^N} \prod_{i=1}^{N} e^{-t_i/\tau} = \frac{1}{\tau^n} e^{-\sum_{i=1}^{n} t_i/\tau}$$

- The posterior for τ is:

$$p(\tau|t_1, \ldots, t_N) = \frac{\pi(\tau) e^{-\sum_{i=1}^{N} t_i/\tau} / \tau^N}{\int \pi(\tau') e^{-\sum_{i=1}^{N} t_i/\tau'} / \tau'^N \mathrm{d}\tau'}$$

Lifetime estimate (cont.)

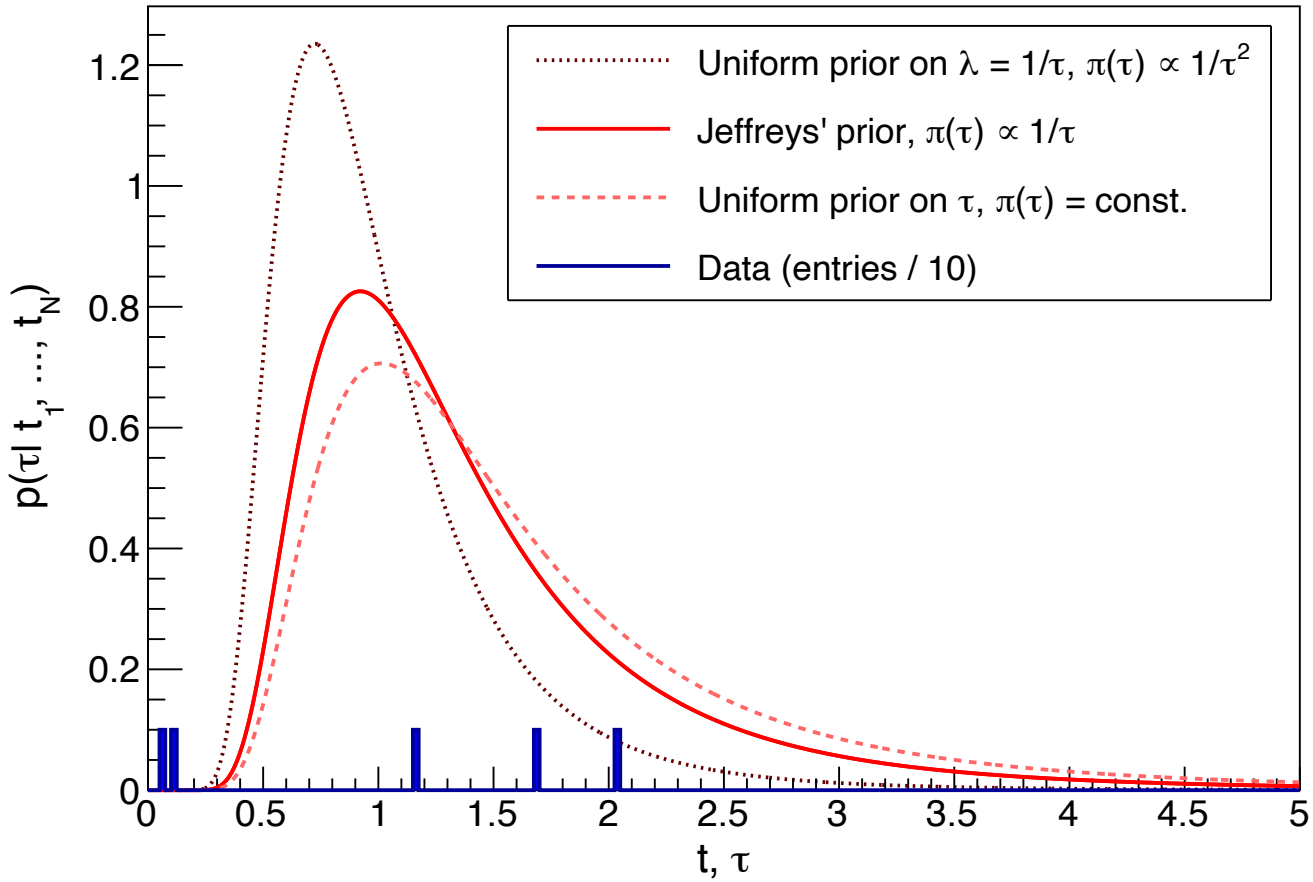
- The prior can be chosen in different ways:
  - Uniform in τ, $\pi(\tau) = \text{const.}$
  - Uniform in $\lambda = 1/\tau$, $\pi(\tau) = 1/\tau^2$
  - Jeffrey's prior, $\pi(\tau) = 1/\tau$
- All choices give as posterior a gamma distribution with different parameters ($k = N$, $N + 2$, $N + 1$)
  - $p(\tau|t_1, \dots, t_N) = C\, \tau^k e^{-\sum_{i=1}^{N} t_i/\tau}$
- The maximum of the PDF is at:

$$\tau = \tau^{\max} = \frac{\sum_{i=1}^{N} t_i}{k} = N\bar{t}/k$$

- For large N, $\tau^{\max}$ tends to $\bar{t}$, regardless of the prior choice

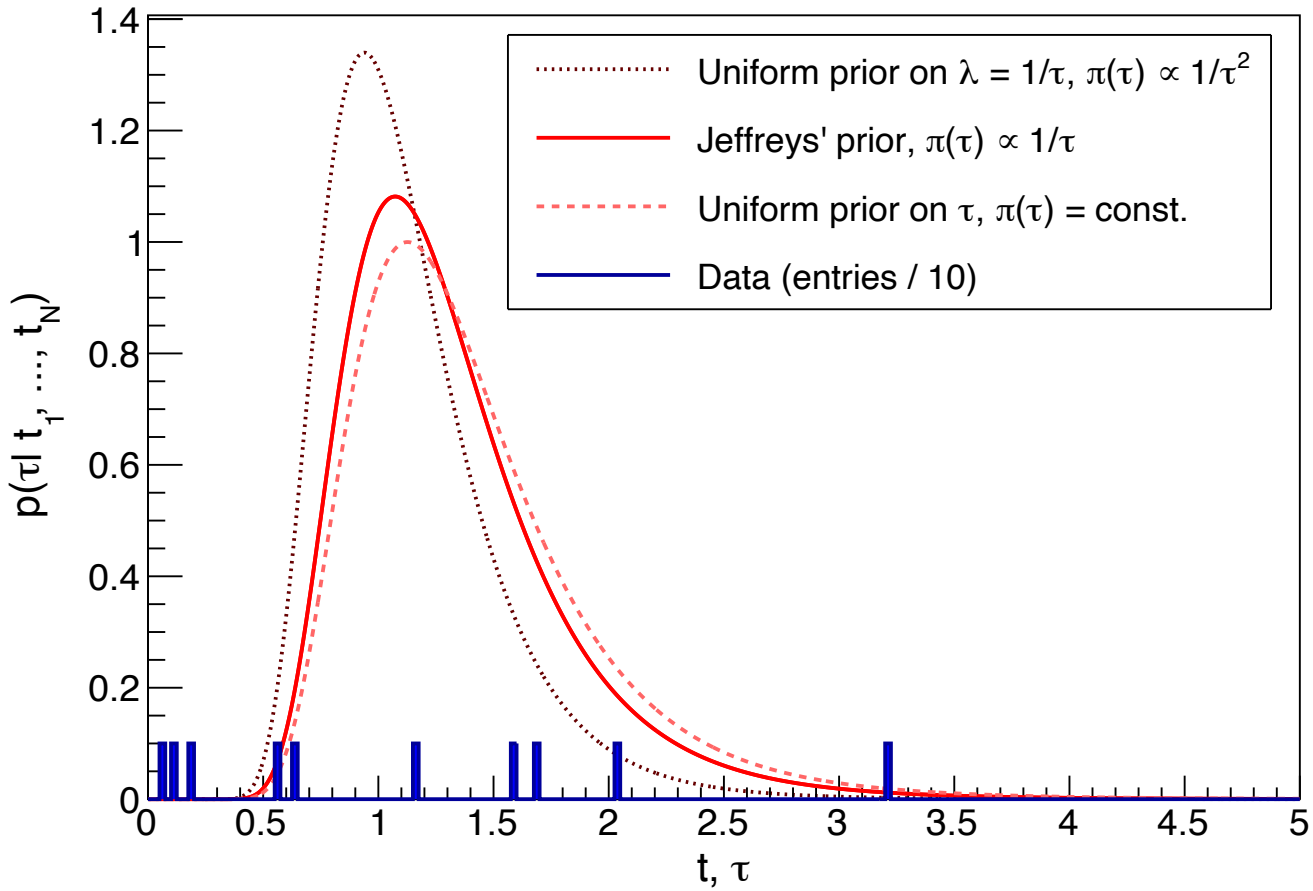
Prior dependence: lifetime

$n = 5$



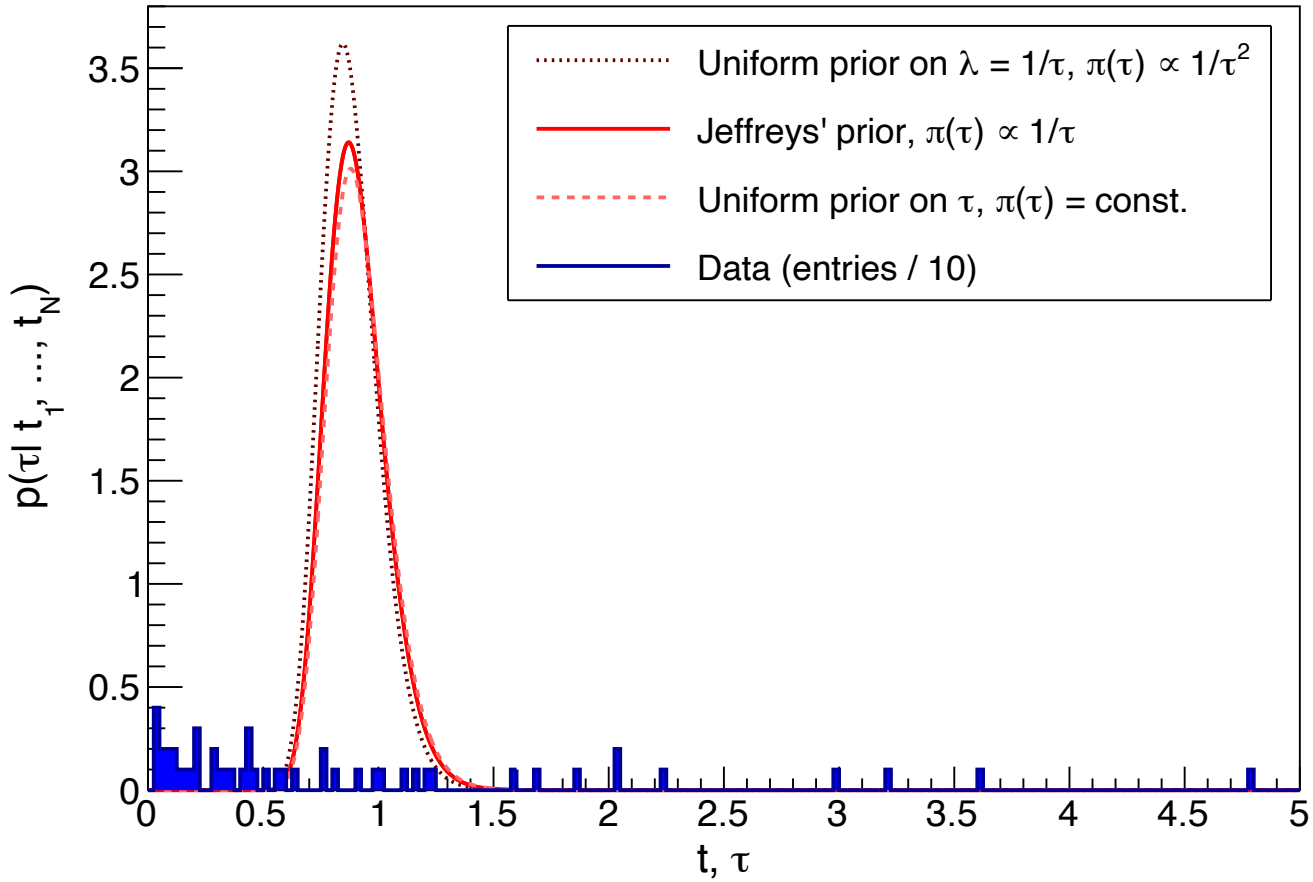
Prior dependence: lifetime

$n = 10$



Prior dependence: lifetime

$n = 50$



Frequentist probability

- Probability P = frequency of occurrence of an event (success) in the limit of very large number ($N \to \infty$) of repeated trials:

$$\text{Probability: } P = \lim_{N \to \infty} \frac{\text{Number of successes}}{N = \text{Number of trials}}$$

- Exactly realizable only with an infinite number of trials
  - Conceptually may be unpleasant
  - Pragmatically acceptable by physicists

- Only applicable to repeatable experiments

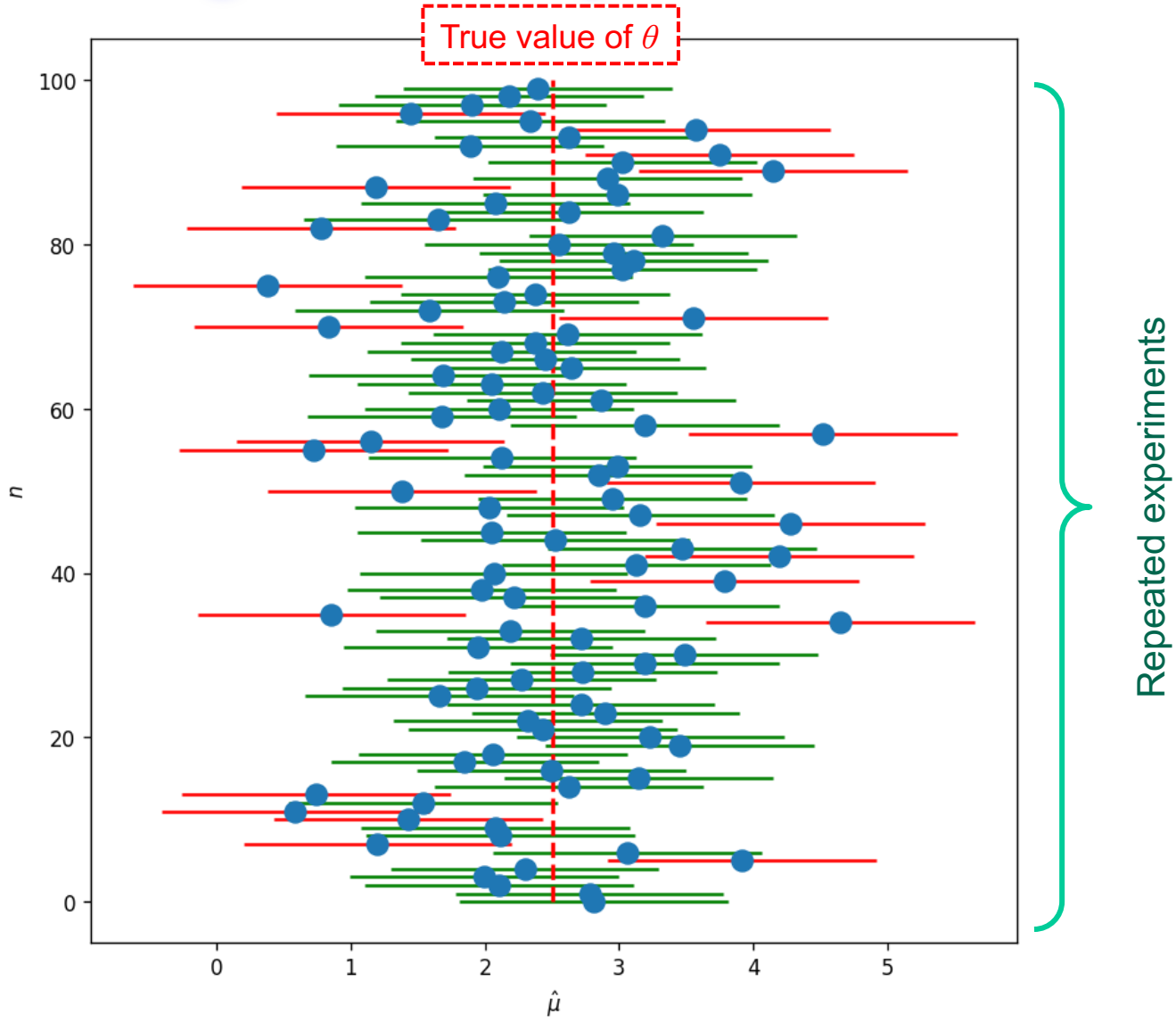
Frequentist inference

- The probability of a possible range of values of an unknown parameter has no meaning in the frequentist approach
  - Parameters are not random variables!
- Frequentist inference procedures determine a point estimate with its uncertainty interval that depend on the observed measurements
- The function that returns the central value given an observed measurement is called estimator. It contains no subjective element
- Point estimate and interval extremes are random variables due to the randomness of the observed data sample

Frequentist coverage

- Repeating the experiment will result each time in a different data sample
- For each data sample, the estimator returns a different central value $\hat{\theta}$
- An uncertainty interval $[\hat{\theta} - \delta, \hat{\theta} + \delta]$ can be associated to the estimator's value $\hat{\theta}$
- Some of the intervals contain the unknown true value of θ, corresponding to a fraction equal to 68% of the times, in the limit of very large number of experiments (coverage)

Coverage for frequentist inference



The maximum-likelihood method

- The most adopted frequentist estimator consist of finding the "best fit" parameters that maximizes the likelihood function, known as maximum-likelihood estimator (ML)
- This estimator as close-to-optimal properties
- The maximization can be performed analytically only in the simplest cases, and numerical maximization is required in most of realistic cases

- Minuit is historically the most widely used minimization engine in High Energy Physics
  - F. James, 1970's; rewritten in C++ and released under CERN's ROOT framework

Some estimator properties

- **Consistency**: for large number of measurements the estimator $\hat{\theta}$ should converge, in probability, to the true value θ.
  - ML estimators are consistent

- **Bias**: the bias of a parameter is the average value of its deviation from the true value:

$$b(\theta) = \langle \hat{\theta} - \theta \rangle = \langle \hat{\theta} \rangle - \theta$$

  - ML estimators may have a bias, but the bias decreases with large number of measurements (if the fit model is correct…!)
  - E.g.: in the case of the estimate of a Gaussian's σ^2, the unbiased estimate is the well known:

$$\widehat{\sigma^2}_{\text{unbias.}} = \frac{n}{n-1}\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

ML method underestimates the variance σ^2

Efficiency of an estimator

- The variance of any consistent estimator is subject a lower bound (Cramér-Rao bound):

$$\mathbb{V}[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b(\theta)}{\partial \theta}\right)^2}{\mathcal{I}(\theta)} = V_{CR}$$

$b(\theta)$ = bias of θ

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial lnL(\theta)}{\partial \theta}\right)^2\right] = \mathbb{E}\left[-\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right]$$

Fisher information

- **Efficiency** can be defined as the ratio of Cramér-Rao bound and the estimator's variance:

$$\varepsilon(\hat{\theta}) = \frac{V_{CR}}{\mathbb{V}\mathrm{ar}[\hat{\theta}]}$$

  - Efficiency for ML estimators tends to 1 for large number of measurements

$$\lim_{n\to\infty} \mathbb{V}[\hat{\theta}] = -\frac{1}{\mathbb{E}\left[\left(\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right)^2\right]} \cong -\frac{1}{\left.\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}}}$$

  - I.e.: ML estimates have, asymptotically, the smallest possible variance

Approx. maximum likelihood errors

- A parabolic approximation of $-2\ln L$ around the minimum is equivalent to a Gaussian approximation
  - Sufficiently accurate in many but not all cases

$$-2\ln L = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + \text{const.}$$

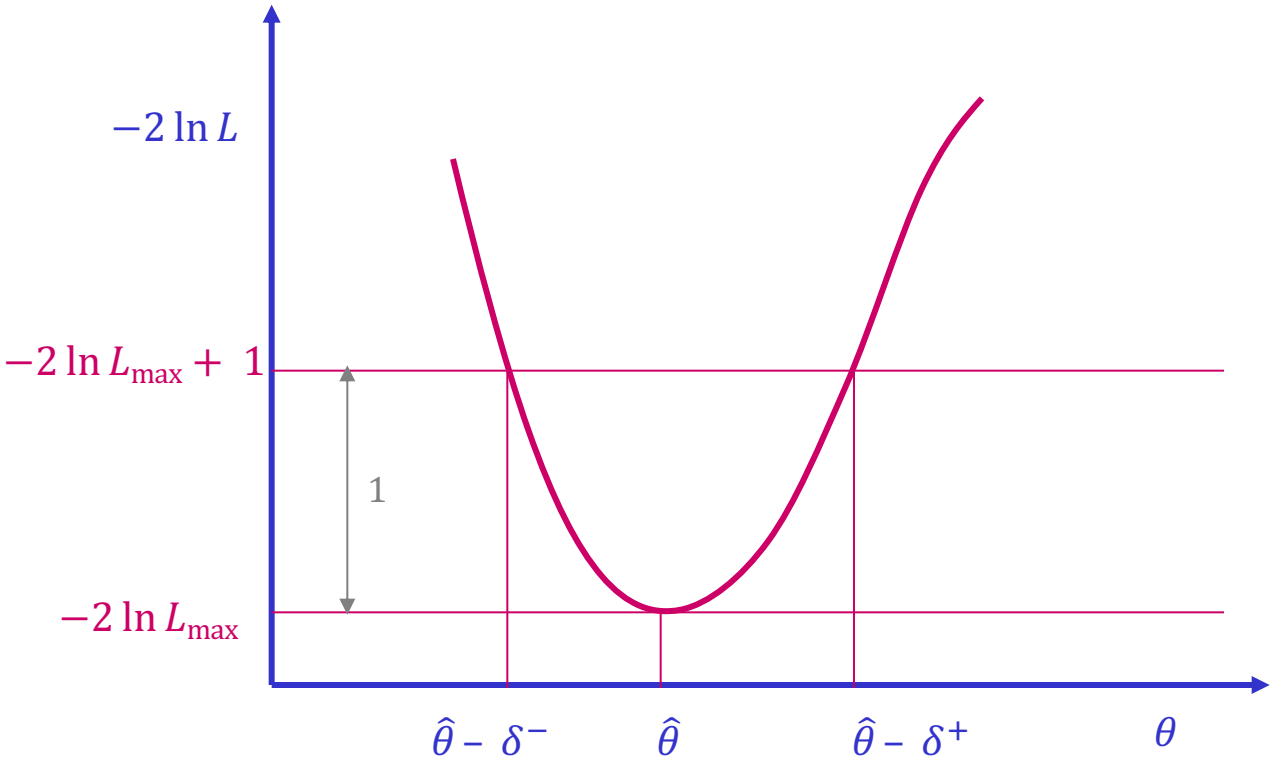
- Estimate of the covariance matrix from 2nd order partial derivatives w.r.t. the fit parameters at the minimum:

$$V_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\theta_k = \widehat{\theta}_k}$$

- Implemented in Minuit as MIGRAD/HESSE method

Asymmetric errors

- Another approximation alternative to the parabolic one may be to evaluate the excursion range of $-2\ln L$.
- Error ($n\sigma$) determined by the range around the maximum for which $-2\ln L$ increases by $+1$ ($+n^2$ for $n\sigma$ intervals)



- Errors can be asymmetric
- For a Gaussian PDF the result is identical to the 2nd order derivative matrix
- Implemented in Minuit as MINOS function

Measure a particle's lifetime

- The probability distribution for a single measurement is:

$$p(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

- And the likelihood function is

$$L(t_1, \dots, t_n; \tau) = \frac{1}{\tau^n} \left(\prod_{i=1}^{n} e^{-t_i/\tau} \right) = \frac{1}{\tau^n} e^{-\sum_{i=1}^{n} t_i/\tau}$$

- Minimization gives:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

- It can be demonstrated that this estimate is unbiased
- The distribution of $\hat{\tau}$ is a gamma distribution with scale parameter τ and shape parameter 1:

$$p(\hat{\tau}) = \hat{\tau}^{2n} e^{-\frac{\hat{\tau}}{\tau}} / (\tau^{2n}(2n-1)!)$$

- The square root of the variance is equal to:

$$\sigma_{\hat{\tau}} = \tau/\sqrt{n}$$

- This is also equal to the Cramer-Rao bound

Extended likelihood function

- The likelihood function is the probability density of the sample $(x_1, \dots, x_n)$ as a function of the unknown parameters $(\theta_1, \cdots, \theta_m)$:

$$L = \prod_{i=1}^{N} f(x_1^{(i)}, \cdots, x_N^{(i)}; \theta_1, \cdots, \theta_m)$$

- If the size N of the sample is also a random variable, the extended likelihood function is usually used:

$$L = P(N; \theta_1, \cdots, \theta_m) \prod_{i=1}^{N} f(x_1^{(i)}, \cdots, x_N^{(i)}; \theta_1, \cdots, \theta_m)$$

- Where $P(N; \theta_1, \cdots, \theta_m)$ is in practice always a Poisson distribution whose expected rate is a function of the unknown parameters
- In many cases it is convenient to use $-\ln L$ or $-2\ln L$:

$$\Sigma_i \rightarrow \Pi_i$$

Signal extraction

- For Poissonian signal and background processes:

$$L(x_i; s, b, \theta) = \frac{(s+b)^n e^{-(s+b)}}{n!} \prod_{i=1}^{n} (f_s P_s(x_i;\theta) + f_b P_b(x_i;\theta))$$

$$\left.\begin{aligned} f_s &= \frac{s}{s+b} \\ f_b &= \frac{b}{s+b} \end{aligned}\right\} \rightarrow L = \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{n} (s P_s(x_i;\theta) + b P_b(x_i;\theta))$$

- We can fit simultaneously s, b and θ minimizing:

constant!

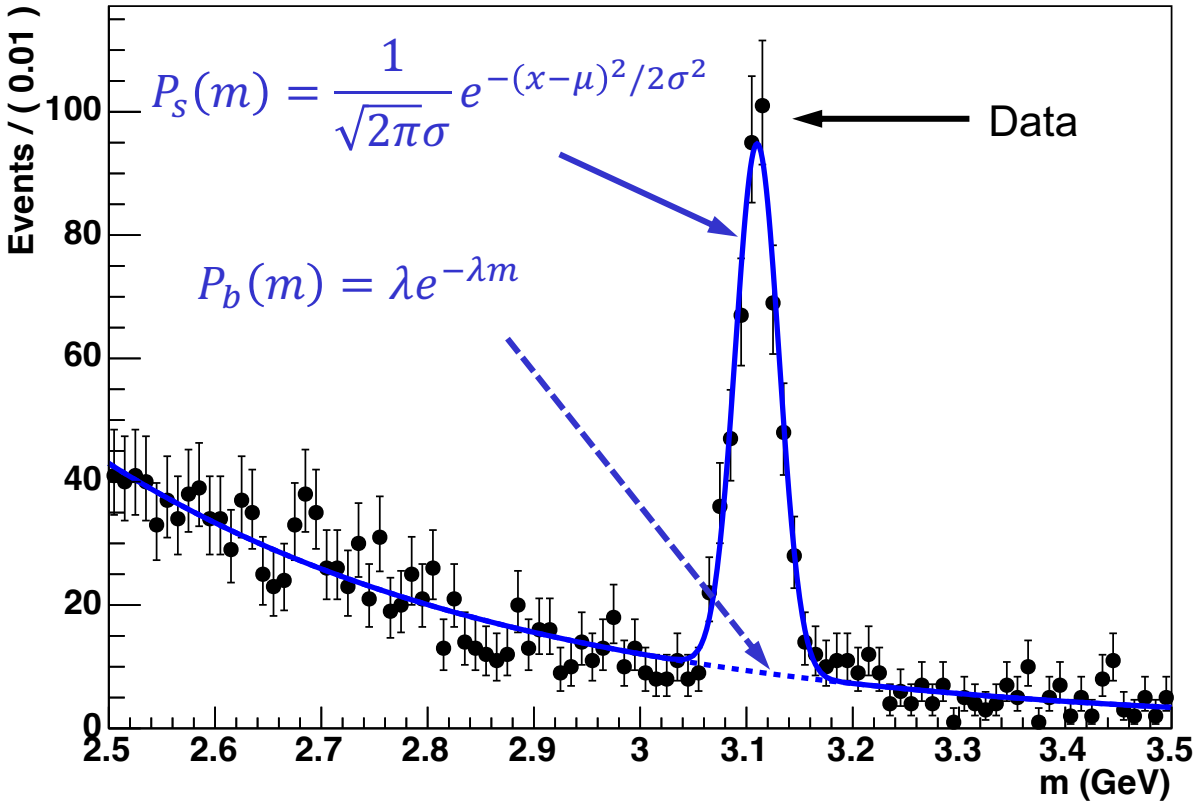
$$-\ln L = s + b - \sum_{i=1}^{n} \ln\left(s P_s(x_i;\theta) + b P_b(x_i;\theta)\right) + \ln n!$$

- Sometimes s is replaced by $\mu\, s_0$, where s_0 is the theory prediction and μ is called signal strength

Example of ML fit

- $P_s(m)$: Gaussian peak
- $P_b(m)$: exponential shape

Exponential decay parameter λ, Guassian mean μ and standard deviation σ can be fit together with sig. and bkg. yields s and b.



The additional parameters, beyond the parameters of interest (s in this case), used to model background, resolution, etc. are examples of nuisance parameters

In the plot, data are accumulated into **bins** of a given width

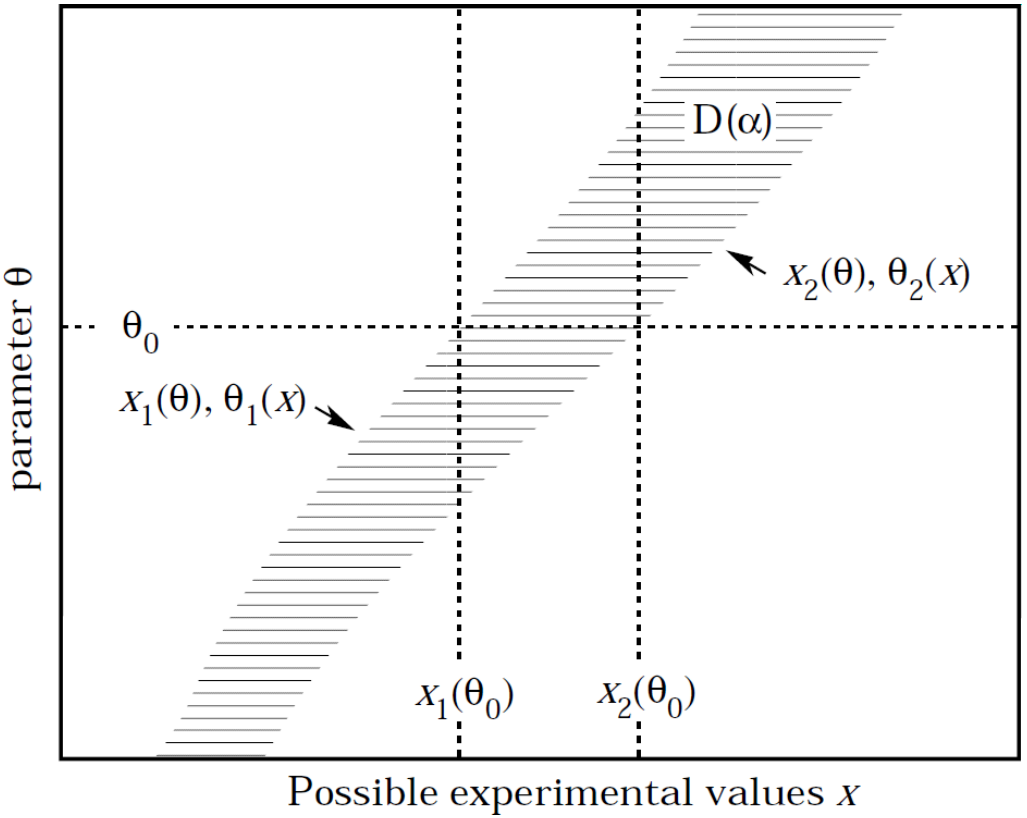
Error bars, representing Poissonian uncertainty on each bin count, are not used in unbinned fits

Neyman's confidence intervals

Frequentist uncertainty intervals can be determined in an exact way:

- Scan the parameter θ range
- For a given θ, compute an interval $[x_1, x_2]$ that contain a probability $1-\alpha$ equal to 68% (or 90%, 95%)
- A choice of the interval is needed (ordering rule)

- Invert the confidence belt: for an observed value x, find the interval $[\theta_1, \theta_2]$ intersecting the belt
- A fraction $1-\alpha$ of the experiments will measure x such that the corresponding $[\theta_1, \theta_2]$ "covers" the true value of θ ("coverage")
- **Note**: the random variables are $[\theta_1, \theta_2]$, not θ !



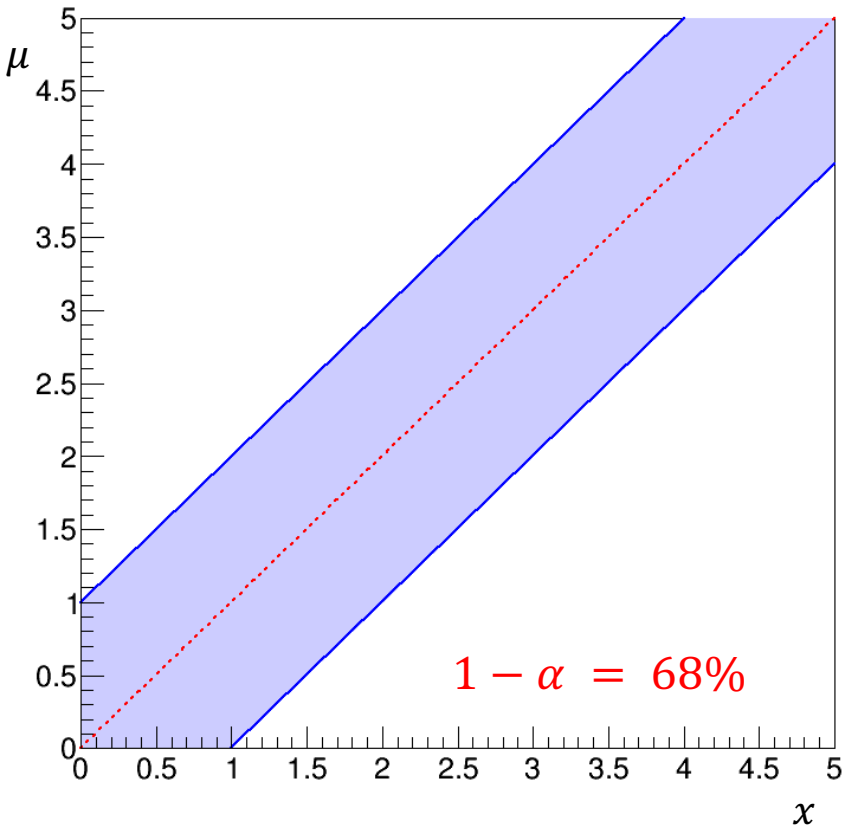
α = significance level

Plot from PDG statistics review

Simplest Gaussian case

- Assume a Gaussian distribution with unknown average μ and known $\sigma = 1$
- The belt inversion is trivial and gives the expected result:
  Point estimate $\hat{\mu} = x$,
  $[\mu_1, \mu_2] = [x - \sigma, x + \sigma]$
- So, we can quote:

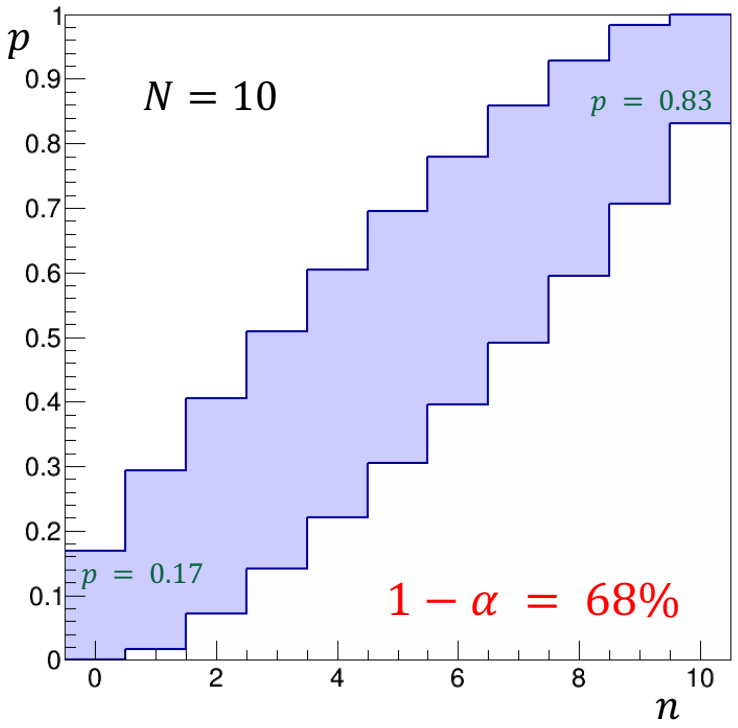
$$\mu = x \pm \sigma$$



Binomial intervals

- Coverage may only approximate in case of discrete variables
- For a Binomial distribution the interval $\{n_{\min}, \dots, n_{\max}\}$ must ensure that:

$$P(n_{\min} \le n \le n_{\max}) = \sum_{n=n_{\min}}^{n_{\max}} \frac{N!}{n!\,(N-n)!} p^n (1-p)^{N-n} \geq 1-\alpha$$



- For an observed $n = k$, pick p^{lo} and p^{up} such that:

$$P(n \ge k \mid N, p^{\text{lo}}) = \alpha/2$$
$$P(n \le k \mid N, p^{\text{up}}) = \alpha/2$$

- For $n = N = 10$, $P(N|N) = p^N = \alpha/2$, therefore:

$$p^{\text{lo}} = \sqrt[10]{\alpha/2} = 0.83\ (68\%CL),\ 0.74\ (90\%CL)$$

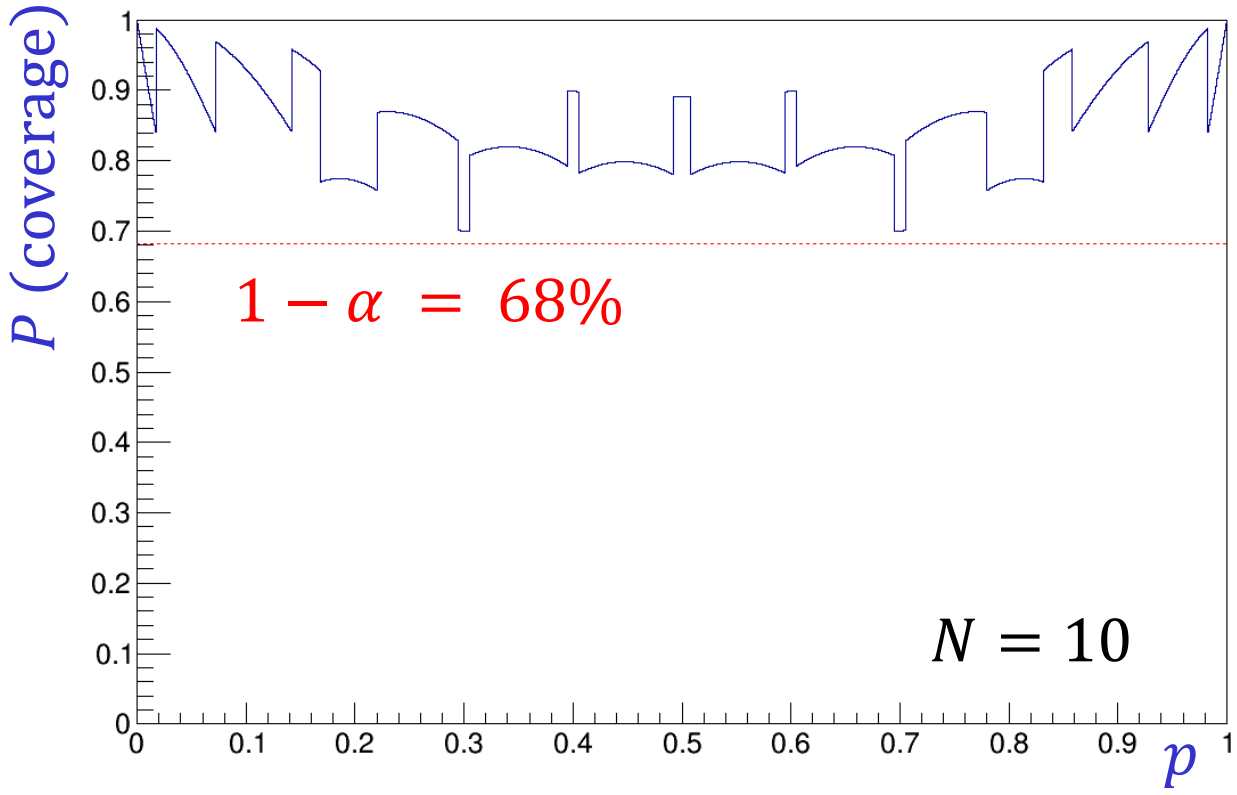
- The approximate ML error estimate fails for $n = 0, N$ is:

$$\hat{p} = \frac{n}{N}, \qquad \sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Clopper and Pearson, 1934

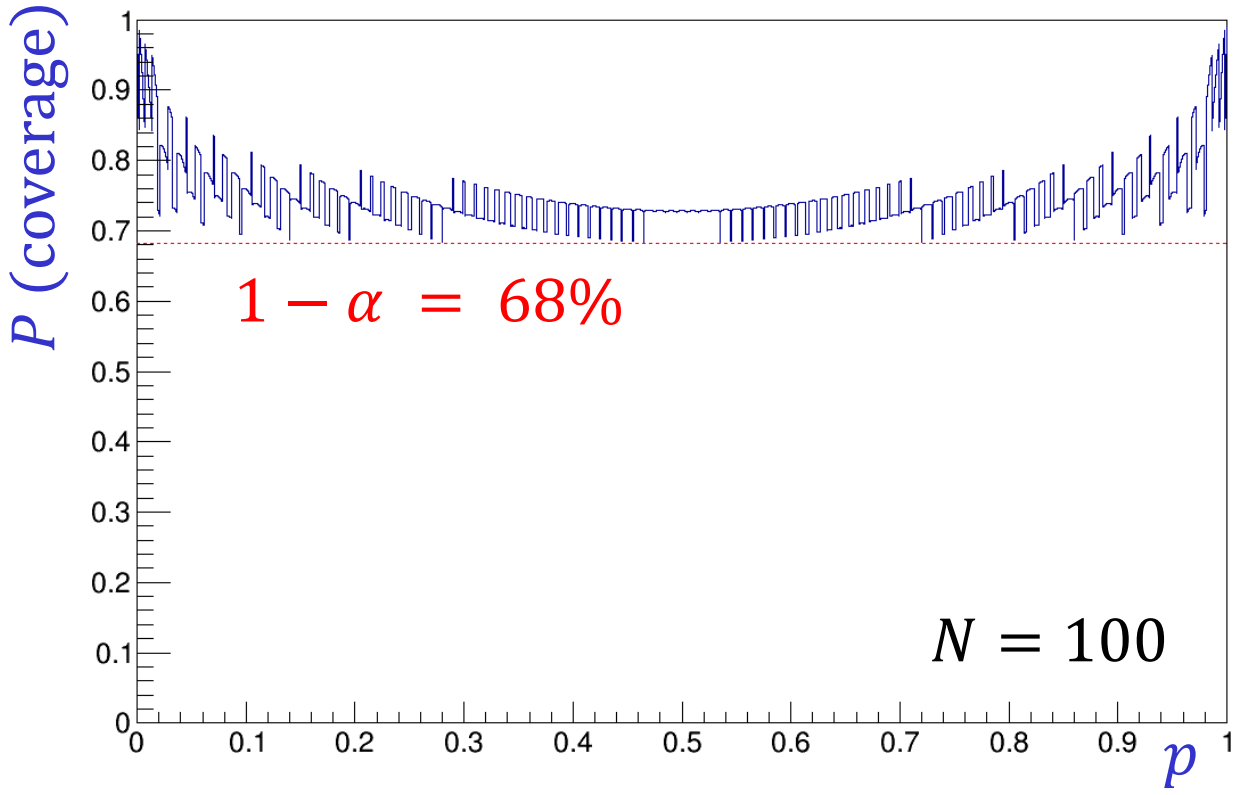
Clopper-Pearson coverage (I)

- CP intervals are often defined as “exact” in literature
- Exact coverage is often impossible to achieve for discrete variables



Clopper-Pearson coverage (II)

- For larger N the "ripple" gets closer to the nominal 68% coverage



Coverage for discrete variables

- If the true value is $p = 0$ (similarly for $p = 1$), the observed value is always $n = 0$, therefore the confidence interval is $[0, p^{\mathrm{up}}[$
- The true value is therefore contained in the confidence interval with 100% probability instead of 68% (or 90%, or whatever)
- This is against the definition of frequentist coverage, but it is unavoidable for discrete variable
- This feature is typical of cases with low number of count. Poissonian counting experiments also have this behaviour.

Likelihood for binned histograms

- It is often convenient to summarize data as binned histograms
- The number of entries n_i in each bin follows a Poisson distribution
- The likelihood function is the product of Poisson probabilities whose expected number of entries depends on some unknown parameters: $\mu_i = \mu_i(\theta_1, \dots, \theta_m)$
- The function to minimize is $-2 \ln L$:

$$-2 \ln L = -2 \ln \prod_{i=1}^{n_{\text{bins}}} \text{Poiss}\left(n_i; \mu_i(\theta_1, \cdots, \theta_m)\right)$$

$$= -2 \ln \prod_{i=1}^{n_{\text{bins}}} \frac{e^{-\mu_i(\theta_1, \cdots, \theta_m)} \mu_i(\theta_1, \cdots, \theta_m)^{n_i}}{n_i!}$$

- The expected number of entries μ_i is often approximated by a continuous function $\mu(x)$ evaluated at the center x_i of the bin, or more precisely by the integral of $\mu(x)$ over the bin interval
- Alternatively, μ_i can be a combination of other histograms ("templates"), e.g.: weighted sum of different simulated processes with yields as fit parameters

Minimum χ^2

- Number of entries are approximated Gaussian if they are sufficiently large, with standard deviation equal to $\sqrt{n_i}$
- Minimizing $-2\ln L$ is equivalent to minimize Neyman's χ^2:

$$\chi^2 = \sum_{i=1}^{n_{\text{bins}}} \frac{(n_i - \mu(x_i; \theta_1, \cdots, \theta_m))^2}{n_i}$$

- Sometimes, the denominator n_i is replaced (Pearson's χ^2) by:

$$\mu_i = \mu\,(x_i;\, \theta_1, \ldots, \theta_m)$$

  to avoid cases with zero or small n_i
- Analytic solution exists for linear or polynomial fits and other simple problems, but most of the cases are addressed numerically, as for unbinned ML fits

Fit quality with χ^2 test

- The χ^2 of a fit with a Gaussian underlying model is distributed according to a known PDF:

$$P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2}$$

n is the number of degrees of freedom (n. of bins − n. of params.)

- The value of the cumulative distribution at the fit $\chi^2 = \hat{\chi}^2$ is called *p*-value :

$$p = C(\hat{\chi}^2;\, n) = P(\chi^2 > \hat{\chi}^2;\, n)$$

- p is uniformly distribution between 0 and 1 if the fit model is correct, otherwise it peakes around zero
- The p-values are not the "*probability of the fit hypothesis*", which would be a Bayesian probability
- For a generi ML fit, the minimum value of $-2 \ln L$ cannot gives no information about fit quality. Taking a proper ration of likelihood functions can help if some conditions hold (Wilks' theorem)

Wilks' theorem

Wilks, Samuel S., The Annals of Mathematical Statistics. 9 (1): 60–62 (1938)

- Consider a likelihood function from N measurements:

$$L = \prod_{i=1}^{N} L\left(x_1^{(i)}, \cdots, x_N^{(i)}; \theta_1, \cdots, \theta_m\right) = \prod_{i=1}^{N} L\left(\vec{x}_i; \vec{\theta}\right)$$

- Assume that H_0 and H_1 are two nested hypotheses, i.e.: they can be expressed as:

$$\vec{\theta} \in \Theta_0, \vec{\theta} \in \Theta_1, \text{ with } \Theta_0 \subseteq \Theta_1.$$

- If H_0 is true, the following quantity, for $N \to \infty$, is distributed as a χ^2 with n.d.o.f. equal to the difference of Θ_0 and Θ_1 dimensionality:

$$\chi_r^2 = -2 \ln \frac{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^{N} L\left(\vec{x}_i; \vec{\theta}\right)}{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^{N} L\left(\vec{x}_i; \vec{\theta}\right)}$$

- E.g.: searching for a signal with strength μ, $\Theta_0: \mu = 0$, $\Theta_1: \mu \geq 0$ we have the profile likelihood (supremum = best fit value):

$$\chi_r^2(\mu) = -2 \ln \frac{\sup_{\vec{\theta}} \prod_{i=1}^{N} L\left(\vec{x}_i; \mu, \vec{\theta}(\mu)\right)}{\sup_{\mu', \vec{\theta}} \prod_{i=1}^{N} L\left(\vec{x}_i; \mu', \vec{\theta}\right)}$$

Binned likelihood ratio

- An alternative to the χ^2 method for histograms with small number of entries has been proposed with the following likelihood ratio, where the true values are replaced by bin contents (H_0: $\mu_i = n_i$):

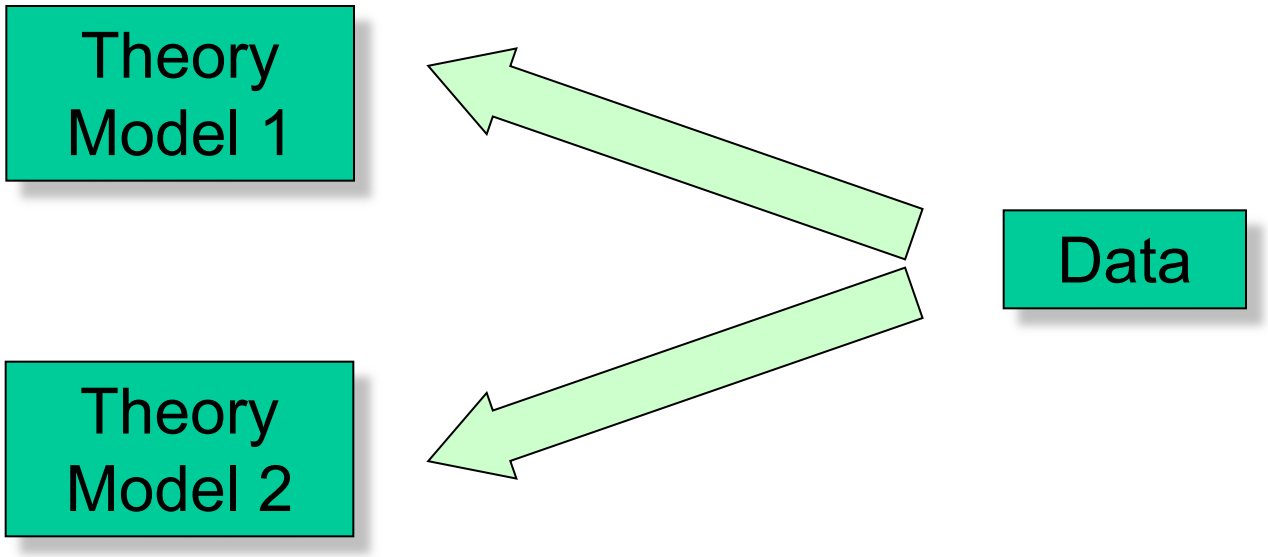
$$\chi^2_\lambda = -2 \ln \prod_{i=1}^{n_{\text{bins}}} \frac{L(n_i;\mu_i)}{L(n_i;n_i)} = -2 \ln \prod_{i=1}^{n_{\text{bins}}} \frac{e^{-\mu_i} {\mu_i}^{n_i}}{\cancel{n_i!}} \frac{\cancel{n_i!}}{e^{-n_i} {n_i}^{n_i}} =$$

$$2 \sum_{i=1}^{n_{\text{bins}}} \left[\mu_i(\theta_1, \cdots, \theta_m) - n_i + n_i \ln \left(\frac{n_i}{\mu_i(\theta_1, \cdots, \theta_m)} \right) \right]$$

- The fit gives the same result as with a Poissonian likelihood, since a constant term has been added to the log-likelihood
- In addition, it provides goodness-of-fit information, and χ^2_λ is asymptotically distributed as a chi-squared with $n - m$ degrees of freedom (Wilks' theorem)

S. Baker, R. Cousins NIM 221 (1984) 437

Hypothesis testing



Which hypothesis is the most consistent with the experimental data?

Bayesian approach

- Bayesian probability gives meaning to the probability that a hypothesis is true:

$$P(H_1|x) = \frac{P(x|H_1)\pi(H_1)}{P(x)}$$

- The ratio of probabilities for two hypothesis does not depend on $P(x)$, and can be computed without considering all possible hypotheses:

$$\frac{P(H_1|x)}{P(H_0|x)} = \frac{P(x|H_1)\pi(H_1)}{P(x|H_0)\pi(H_0)}$$

Bayes factors

- It is possible to introduce Bayes factor:

$$\frac{P(H_1|x)}{P(H_0|x)} = B_{1/0}(x)\frac{\pi(H_1)}{\pi(H_0)}$$

- In other words, this defines the posterior odds $O_{1/0}(x)$ as a function of prior odds $o_{1/0}$:

$$O_{1/0}(x) = B_{1/0}(x)\, o_{1/0}$$

- In word:

posterior odds = Bayes factor × prior odds

- Bayes factor can be used to measure how favoured is one hypothesis against another, and, in the simplest cases, it is equal to the likelihood ratio

- Proposed range values are:
  - 1−3: very weak evidence
  - 3−20: positive evidence
  - 20−150: strong evidence
  - > 150: very strong evidence

In presence of parameters

- Bayes factors may depend on priors if parameters are present.
- Posterior for both hypothesis and parameters is:

$$P(H_1, \theta_1|x) = \frac{P(x|H_1, \theta_1)\pi(H_{0,1}, \theta_1)}{P(x)}$$

- Priors can be decomposed as :

$$\pi(H_1, \theta_1) = \pi(\theta_1|H_1)\pi(H_1)$$

- Hence:

$$P(H_1|x) = \frac{\int P(x|H_1, \theta_1)\pi(H_1, \theta_1)\mathrm{d}\theta_1}{P(x)} = \frac{\pi(H_1) \int P(x|H_1, \theta_1)\pi(\theta_1|H_1)\mathrm{d}\theta_1}{P(x)}$$
$$= \frac{P(x|H_1)\pi(H_1)}{P(x)}$$

- The Bayes factor are defined as:

$$B_{1/0}(x) = \frac{P(x|H_1)}{P(x|H_0)} = \frac{\int P(x|H_1, \theta_1)\pi(\theta_1|H_1)\mathrm{d}\theta_1}{\int P(x|H_0, \theta_0)\pi(\theta_0|H_0)\mathrm{d}\theta_0}$$

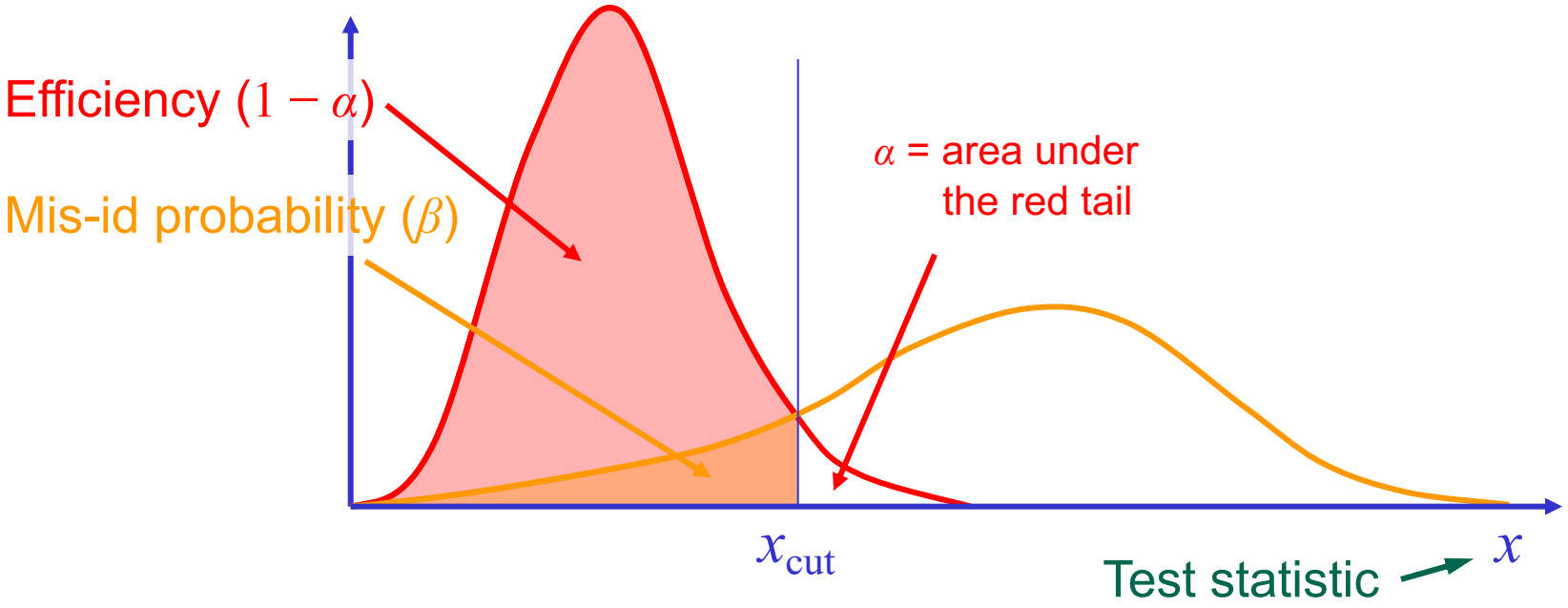
- This introduces some dependency on priors in Bayes factors

Frequentist: terminology

- Statisticians' terminology is sometimes not very natural for physics applications, but it is more and more popular among physicists as well:
- H_0 = **null hypothesis**
  - Ex. 1: *"a sample contains only background"*
  - Ex. 2: *"a particle is a pion"*
- H_1 = **alternative hypothesis**
  - Ex. 1: *"a sample contains background + signal"*
  - Ex. 2: *"a particle is a muon"*
- α **= significance level**: probability to reject H_1 if H_0 is assumed to be true (error of first kind, false positive)
  - $\alpha = 1$ – misidentification probability
- β **= misidentification probability**, i.e.: probability to reject H_0 if H_1 is assumed to be true (error of second kind, false negative)
  - $1 - \beta$ = **power of the test =** selection efficiency
- p**-value**: probability, assuming H_0, of observing a result at least as extreme as the observed test statistic

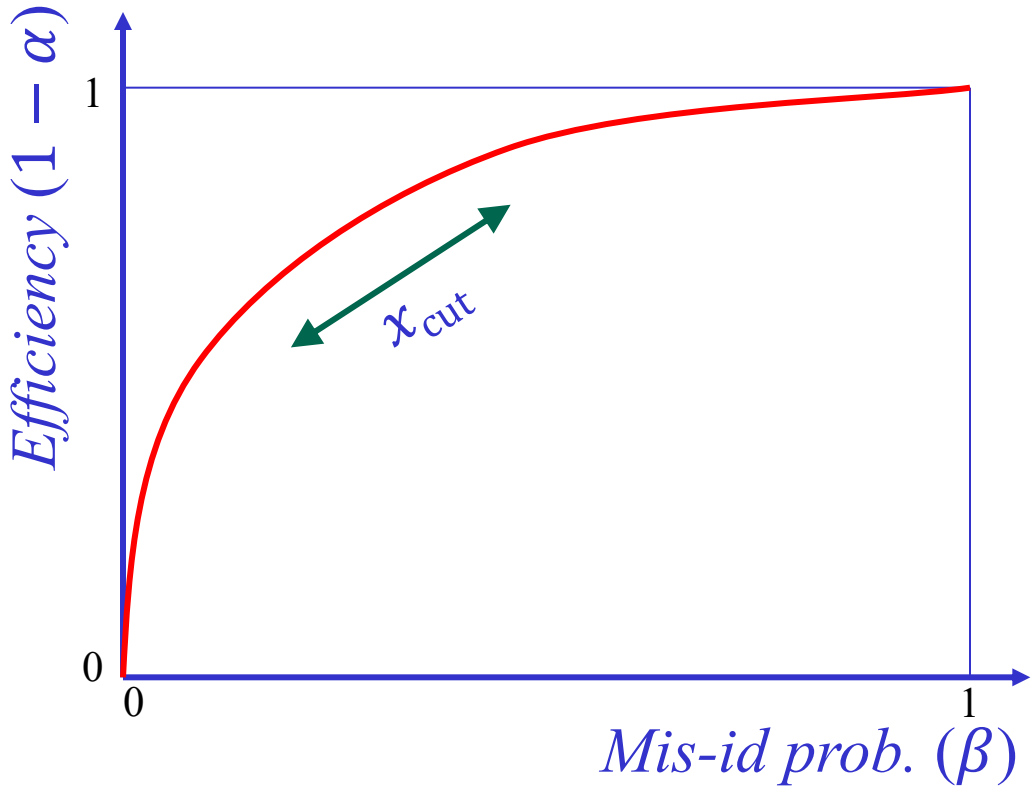
Cut analysis

- Selection ("cut") on one (or more) variable(s):
  - if $x \leq x_{\text{cut}}$ ⇒ signal
  - else, if $x > x_{\text{cut}}$ ⇒ background



Efficiency *vs* mis-id

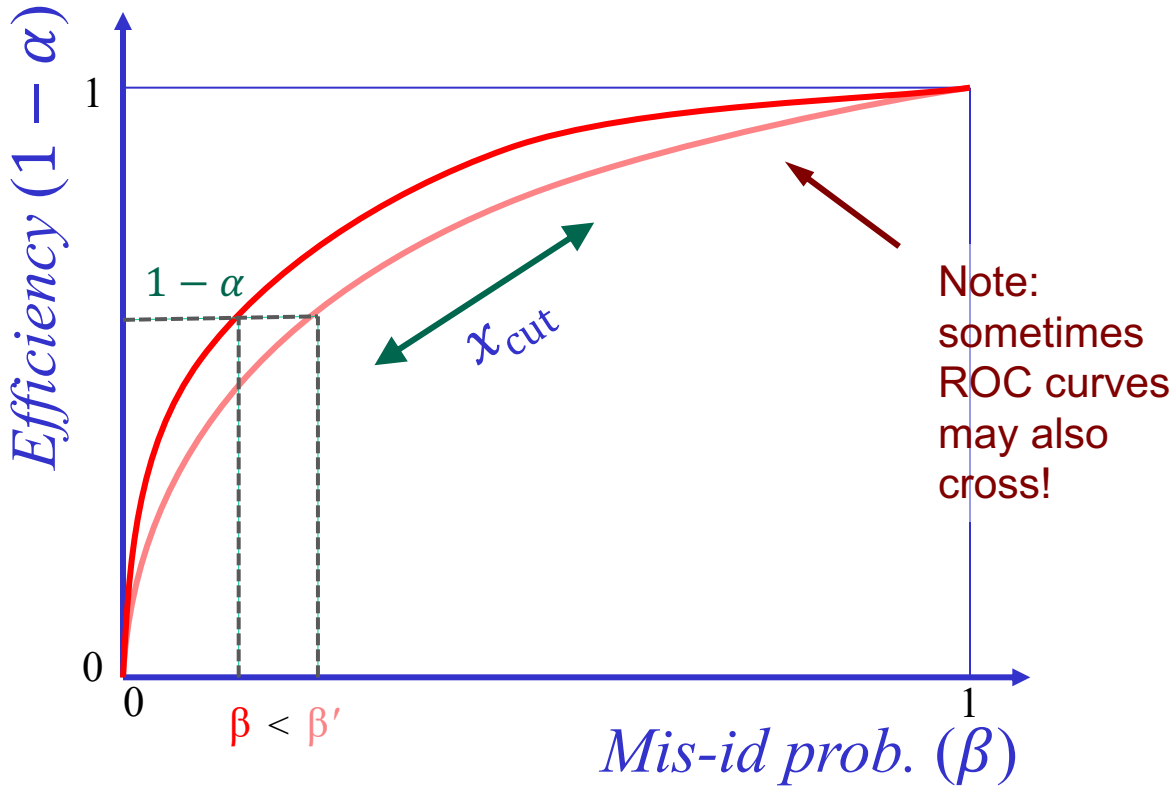
- Varying the applied cut on the test statistic both the efficiency and mis-id probability change



Sometimes also referred to as **ROC curve** (*Receiver Operating Characteristic*)

Performance comparison

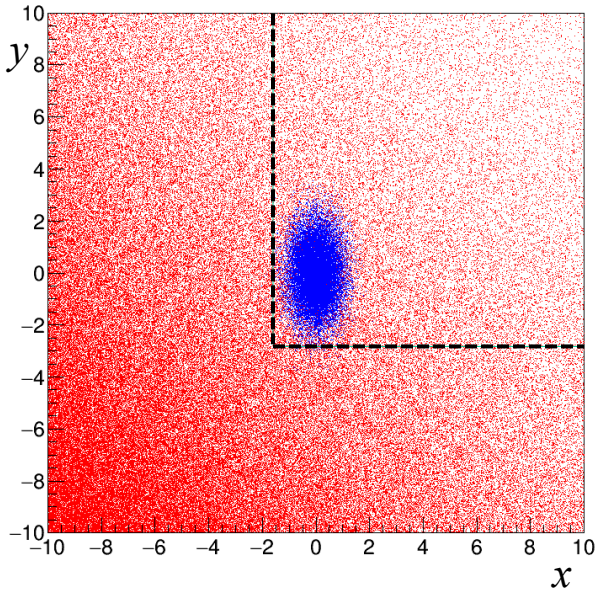
- One test is preferable to another if, for the same level of efficiency ($1 - \alpha$), it has lower mis-id probability (β)



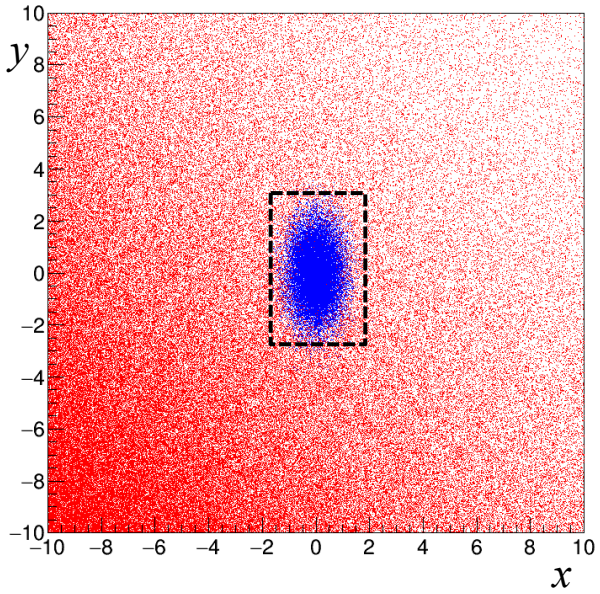
Multivariate discrimination

- With multiple discriminating variables, the choice of the optimal selection is not always straightforward

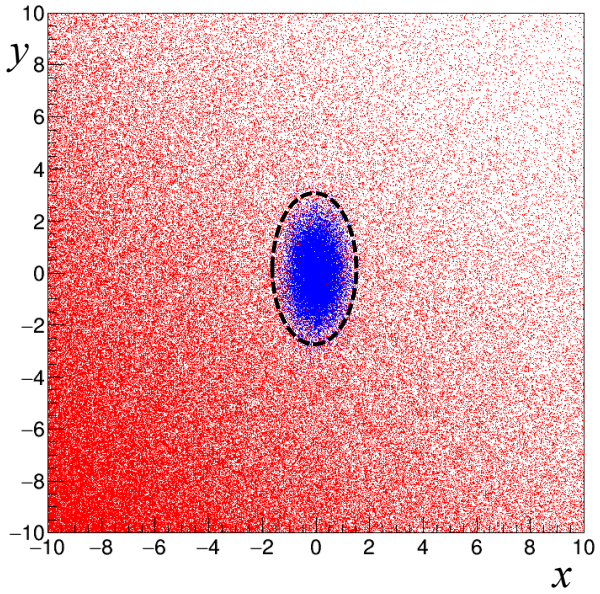
$x > x_{\text{cut}}$ and $y > y_{\text{cut}}$



$x_1 < x < x_2$ and. $y_1 < y < y_2$



$\frac{x^2}{a^2} + \frac{y^2}{b^2} < k^2$



$P_s(x,y) = \text{Gauss}(x;\ 0, \sigma_x) \times \text{Gauss}(y;\ 0, \sigma_y)$, $P_b(x,y) = \alpha\, e^{-\alpha x} \times \beta\, e^{-\beta y}$

- In many cases it's convenient to find a single variable (test statistic) that 'summarizes' all the sample information

The Neyman-Pearson lemma

- For a fixed significance level α, a selection based on the likelihood ratio gives the lowest possible mis-id probability β:

$$\lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} > k_\alpha$$

- The likelihood function can't always be determined exactly, but we can construct test statistics that approximate the exact likelihood
- Machine-Learning algorithms like Neural Networks, Boosted Decision Trees and more are example of discriminators that may closely approximate the performances of the exact likelihood ratio approaching the Neyman-Pearson limit

Projective likelihood ratio

- The likelihood function is approximated by the product of projective PDF in each variable

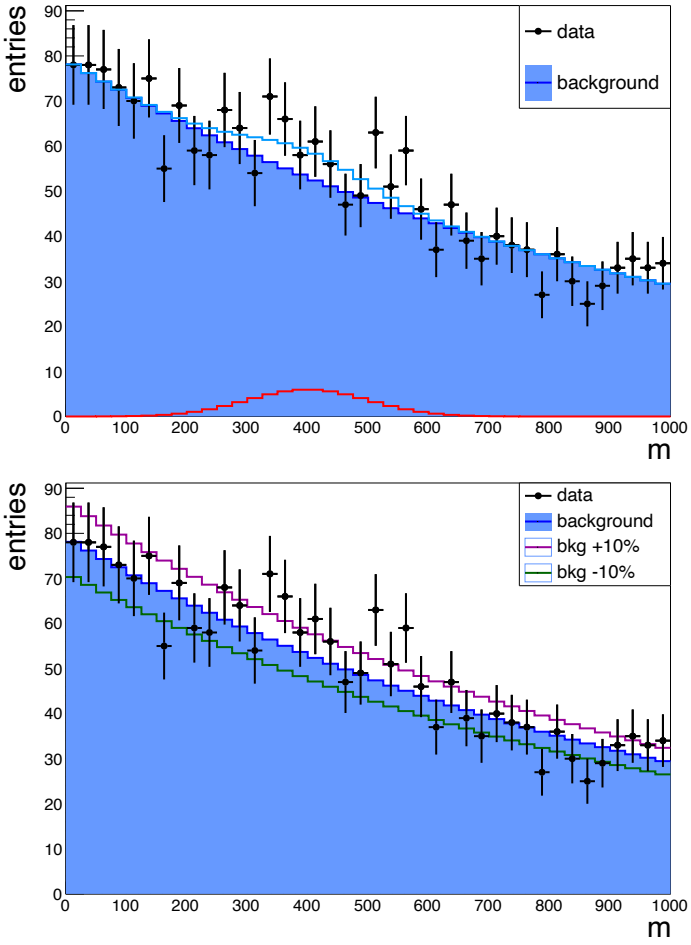
$$\lambda(x) = \frac{L(x_1, \cdots, x_n | H_1)}{L(x_1, \cdots, x_n | H_0)} \approx \underbrace{\frac{\prod_{i=1}^{n} f_i(x_i | H_1)}{\prod_{i=1}^{n} f_i(x_i | H_0)}}_{x_1, \ldots, x_n \text{ approximately considered independent variables}}$$

- Exact only in case of independent variables, otherwise it has suboptimal performances
- The test statistic $\lambda(x)$ may be improved if the variables are first rotated to eliminate correlation (principal component analysis)
  - Find eigenvectors of the covariance matrix
  - **Note**: uncorrelated variables are not necessarily independent

Claiming a discovery

- Test our data sample against two hypotheses:
  - H_0: the data contains background only
  - H_1: the data contains signal plus background
- Build a test statistic λ whose distribution is known under the two hypotheses
  - Usually, λ tends to have large values if H_1 is true and small values if H_0 is true, consistently with λ being the likelihood ratio $L(x|H_1)/L(x|H_0)$
- Assessing the discovery is based on the p-value, the probability that λ is greater or equal to than the observed value λ_{obs}

Are data below more consistent with a background fluctuation or with a peaking excess?

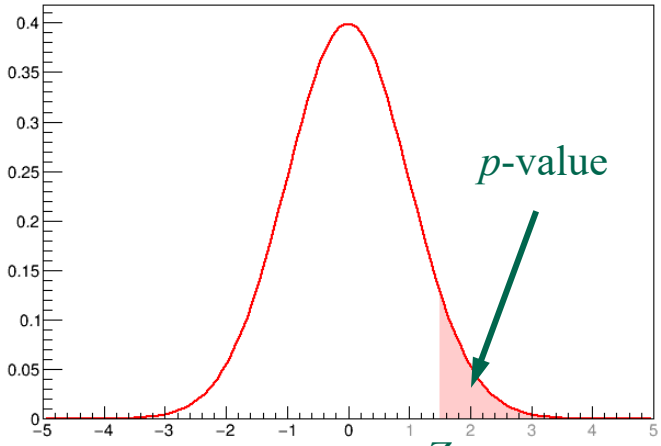


Significance

- The p-value is usually converted into an equivalent area of a Gaussian tail:

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$



Φ = cumulative of a normal distribution

- If the significance level $Z > 3$ ("3σ") one claims "*evidence*"
  - $p < 1.349 \times 10^{-3}$
- If the significance level $Z > 5$ ("5σ") one claims "*observation*"
  - $p < 2.87 \times 10^{-7}$ (discovery!)

- **Again**: the probability that background produces a large test statistic is not equal to probability of the null hypothesis (background only), which has only a Bayesian meaning

Significance for Poisson counting

- If one measure a number of counts n from an expectation $s + b$, where b is known exactly, one may estimate:

$$\hat{s} = n - b$$

- With an expected uncertainty:

$$\sigma_{\hat{s}} = \sigma_n = \sqrt{s + b}$$

- If we assume $s = 0$ (H_0), the significance level, in the Gaussian approximation, is:

$$Z = \frac{\hat{s}}{\sigma_{\hat{s}}} = \frac{\hat{s}}{\sqrt{b}}$$

- If b is affected by uncertainty, this should be added in quadrature to $\sqrt{b}$.
- For small number of counts, an asymptotic formula has been derived (motivation in the following slides):

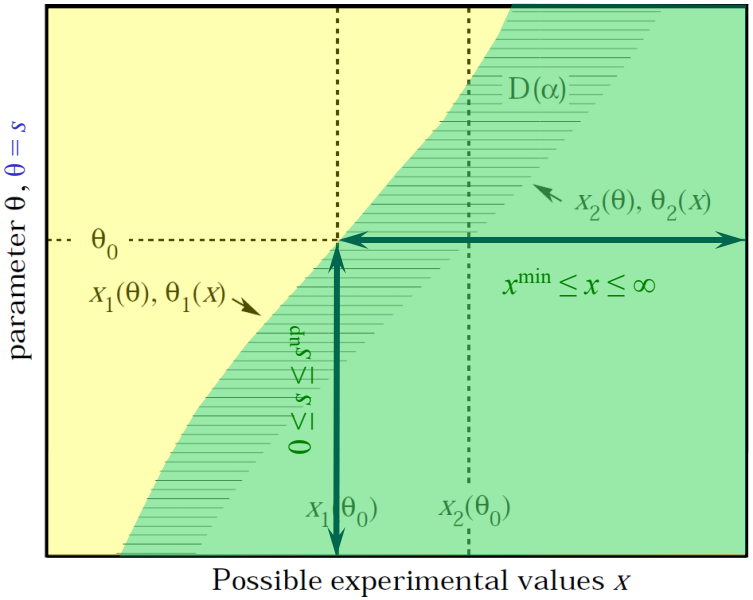
$$Z = \sqrt{2\left[(s+b)\log\left(1+\frac{s}{b}\right) - s\right]}$$

Upper limits

- If the result of a search is negative, we determine an excluded region in the parameter space
- Build a fully asymmetric Neyman confidence belt based on the considered test statistic x
- Invert the belt, find the allowed interval:

$$s \in [s_1, s_2] \;\Rightarrow\; s \in [0, s^{\mathrm{up}}]$$

- Upper limit = upper extreme of the asymmetric interval $[0, s^{\mathrm{up}}]$
- In case the observable x is discrete (e.g.: the number of events n in a counting experiments), the coverage may not be exact

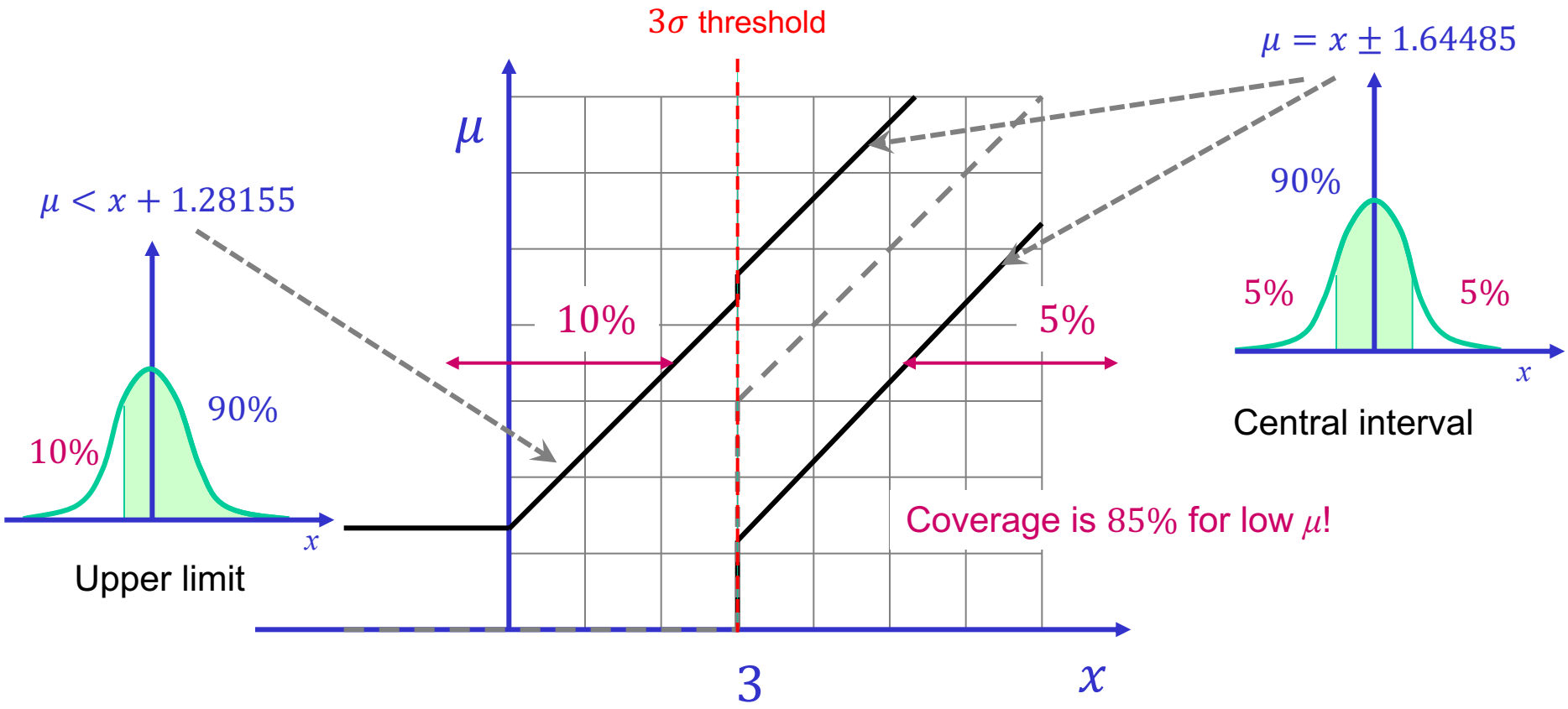


The flip-flopping issue

- When to quote a central value or upper limit?
- A popular choice is:
  - *"Quote a* 90% CL *upper limit of the measurement if the significance is below* 3σ*; quote a central value otherwise"*
  - Upper limit or central interval chosen according to observed data
- This produces an incorrect coverage!

"Flip-flopping" with a Gaussian PDF

- Assume a Gaussian with a fixed width, $\sigma = 1$



Gary J. Feldman, Robert D. Cousins, Phys.Rev.D57:3873-3889,1998

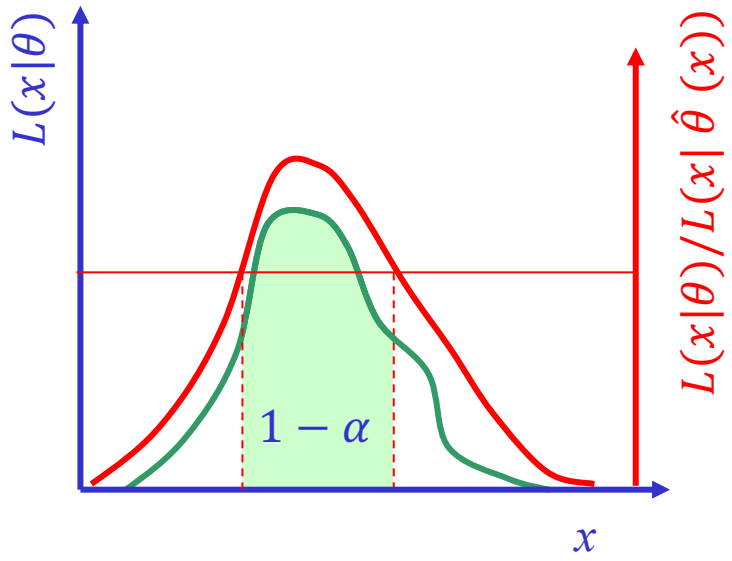
Likelihood ratio & Neyman belt

- Feldman and Cousins proposed a criterion to define the Neyman belt based in a likelihood ratio test:

$$R_\mu = \{x : L(x|\theta)/L(x|\hat{\theta}) > k_\alpha\}$$

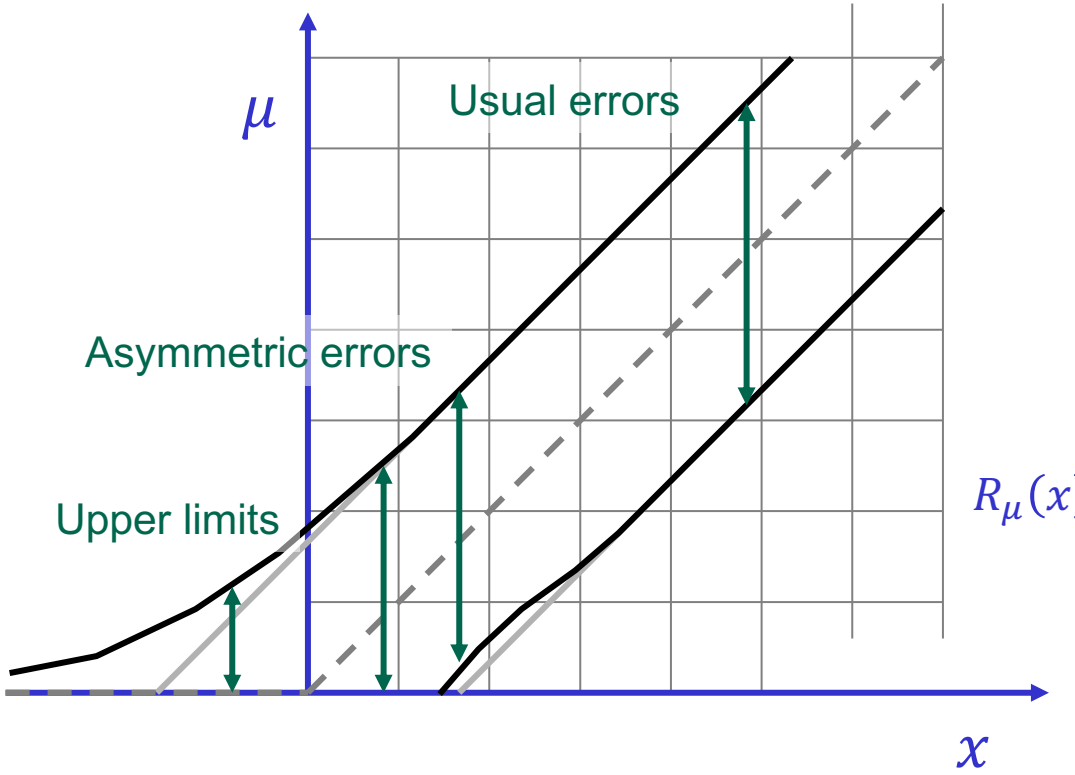
- The value k_α depends on the desired significance level α

- H_0: $\theta = \hat{\theta}$, the best-fit value
- H_1: $\theta = \theta$, the specific value considered for the Neyman belt construction



Feldman-Cousins approach

- Application to the Gaussian case with non-negative mean μ:



$$\hat{\mu} = \max(x, 0)$$

$$P(x|\hat{\mu}) = \begin{cases} \dfrac{1}{\sqrt{2}} & \text{for } x \geq 0 \\ \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2} & \text{for } x < 0 \end{cases}$$

$$R_\mu(x) = \frac{P(x|\mu)}{P(x|\hat{\mu})} = \begin{cases} e^{-(x-\mu)^2/2} & \text{for } x \geq 0 \\ e^{-(x\mu-\mu^2)/2} & \text{for } x < 0 \end{cases}$$

Application to discrete variables, like event counting, raises issues that will be discussed in the following.

Confidence intervals must be computed numerically, even for this simple Gaussian case!

Upper limits for event counting

- The simplest search for a new signal consists of counting the number of events passing a specified selection
- The number of selected events n is distributed according to a Poissonian distribution
- H_1 : expect $n = s + b$
- H_0 : expected $n = b$
- Given n counts, compare with the two hypotheses H_1 and H_0
- Simplest case: b is known with negligible uncertainty

Counting, Bayesian approach

- Let's assume the background b is known with no uncertainty:

$$L(n;s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

- A uniform prior, $\pi(s) = 1$, simplifies the computation:

$$1-\alpha = \int_0^{s^{\mathrm{up}}} P(s|n)ds = \frac{\int_0^{s^{\mathrm{up}}} L(n;s)\pi(s)ds}{\int_0^{\infty} L(n;s)\pi(s)ds}$$

- Inverting the equation gives the upper limit s^{up}
- For $n = 0$, s^{up} does not depend on b:

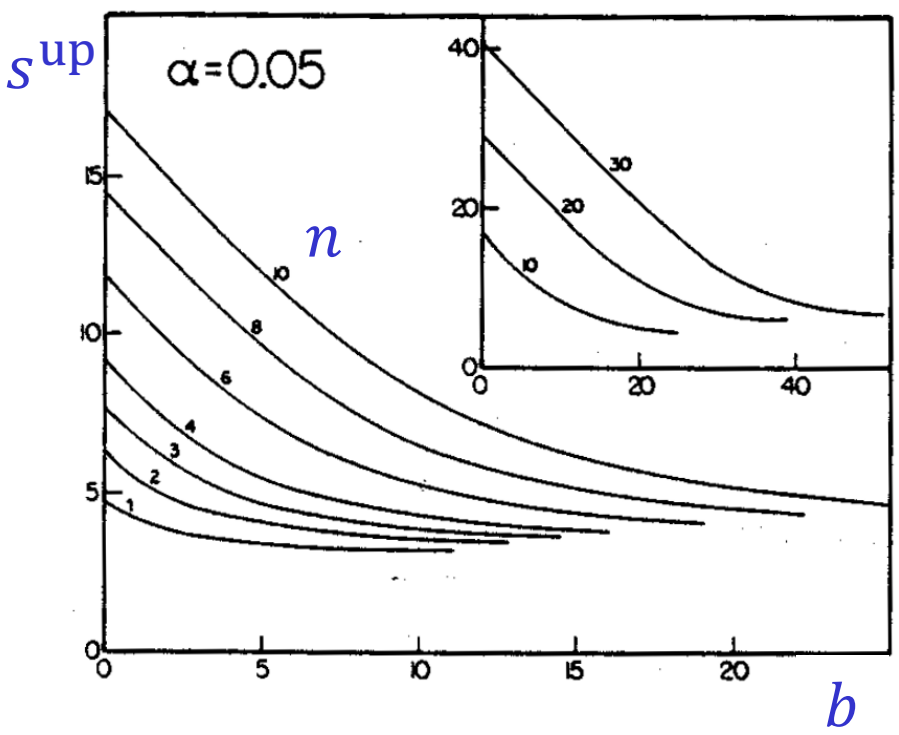
$$\alpha = e^{-s^{\mathrm{up}}}$$

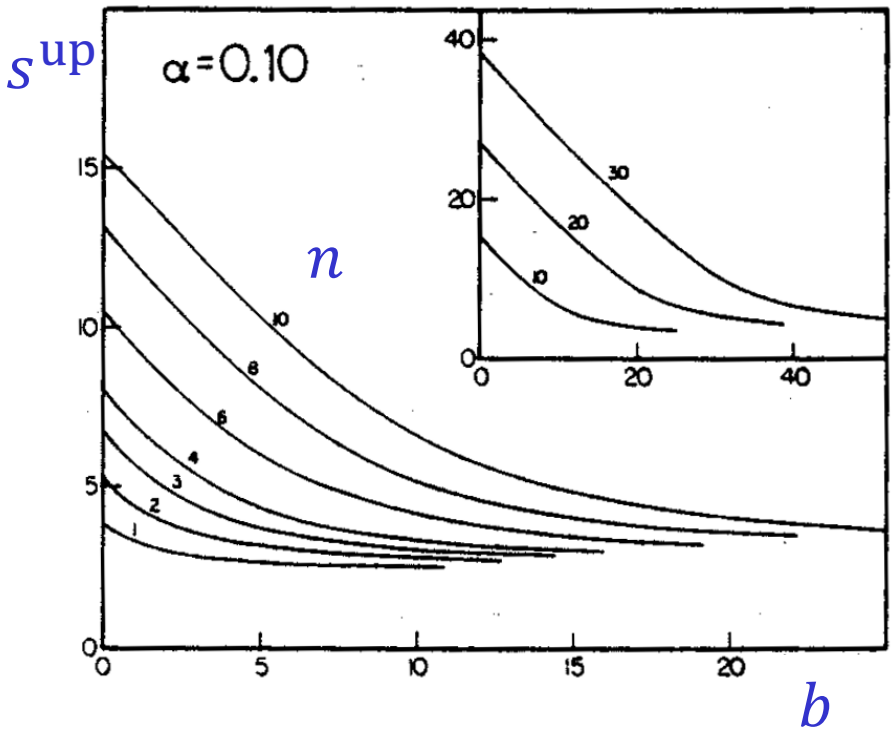
$$\alpha = e^{-s^{\mathrm{up}}} \frac{\sum_{m=0}^{n} \frac{(s^{\mathrm{up}}+b)^m}{m!}}{\sum_{m=0}^{n} \frac{b^m}{m!}}$$

  - $s < 2.303\ (90\%\ CL) \leftarrow \alpha = 0.1$
  - $s < 2.996\ (95\%\ CL) \leftarrow \alpha = 0.05$

Counting, Bayesian approach

- Upper limits decrease as either b or n increases
- For $n = 0$, upper limits are not sensitive on b (prev. slide)





O. Helene. NIMA 212 (1983) 319

Frequentist: zero events selected

- Assume we have negligible background ($b = 0$) and we measure zero events ($n = 0$)
- The likelihood function simplifies as:

$$L(n = 0; s) = \text{Poiss}(0; s) = e^{-s}$$

- The (fully asymmetric) Neyman belt inversion is as simple as follows:

$$P(n \leq 0; s^{\text{up}}) = \alpha \quad \rightarrow \quad s^{\text{up}} = -\ln \alpha$$

- The results are by chance identical to the Bayesian computation:

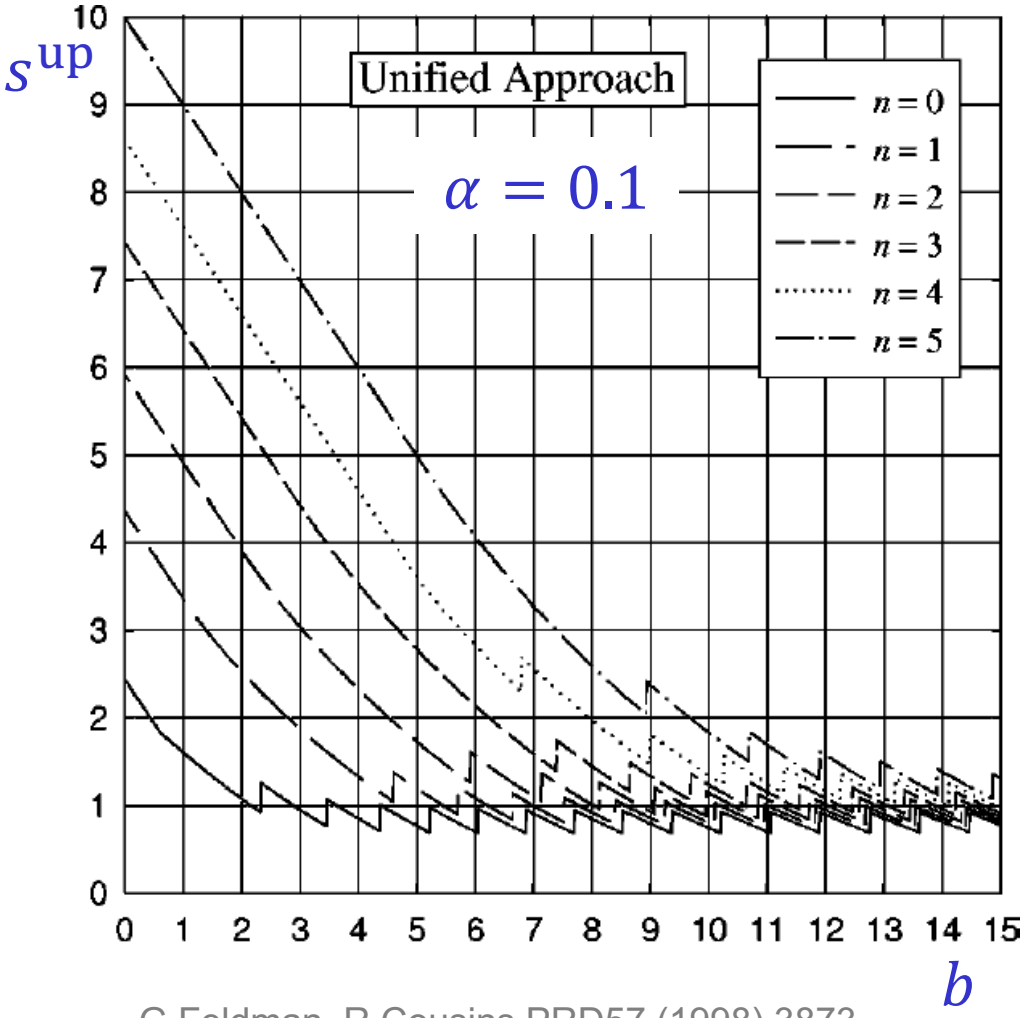
$$s < 2.303 \ (90\% \ CL) \leftarrow \alpha = 0.1$$

$$s < 2.996 \ (95\% \ CL) \leftarrow \alpha = 0.05$$

- Despite the numerical coincidence, the interpretation of frequentist and Bayesian upper limits remain very different!
- **Warning**: this evaluation suffer from the "**flip-flopping**" problem, so the coverage is spoiled if you decide to switch from upper limit to a central value depending on the observed significance!

Counting, Feldman-Cousins

- F&C intervals cure the flip-flopping issue and ensure the correct coverage
  - May overcover for discrete variables
- The "ripple" structure is due to the discrete nature of Poissonian counting
- Note that even for $n = 0$ the upper limit decrease as b increases (apart from ripple effects)
- If two experiment are designed for an expected background of –say– 0.5 and 0.01, the "worse" one has the best expected upper limit if they observe $n = 0$, the most probable value

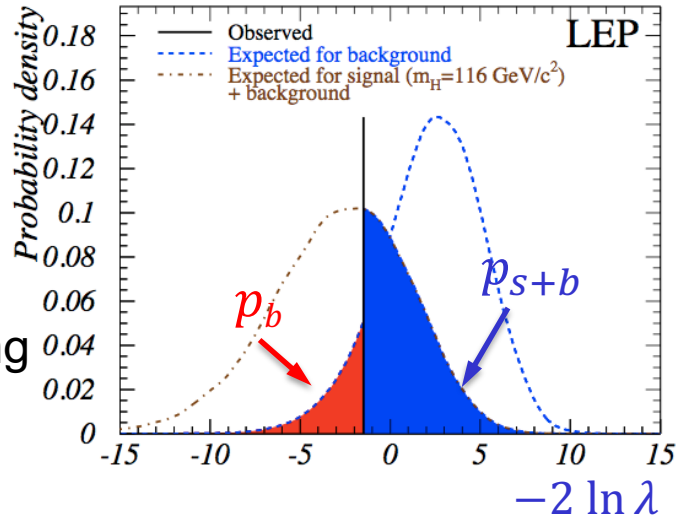


G.Feldman, R.Cousins PRD57 (1998) 3873
C. Giunti, PRD59 (1999), 053001

Modified frequentist approach

- A modified approach was proposed for the first time when combining the limits on the Higgs boson search from the four LEP experiments, ALEPH, DELPHI, L3 and OPAL
- Given a test statistic $\lambda(x)$, determine its distribution for the two hypotheses $H_1(s+b)$ and $H_0(b)$, and compute:

$$\begin{cases} p_{s+b} = P\left(\lambda(x|H_1) \leq \lambda^{\text{obs}}\right) \\ p_b = P(\lambda(x|H_0) \geq \lambda^{\text{obs}}) \end{cases}$$



- The upper limit is computed, instead of requiring $p_{s+b} \leq \alpha$, on a modified *p*-value, $CL_s \leq \alpha$:

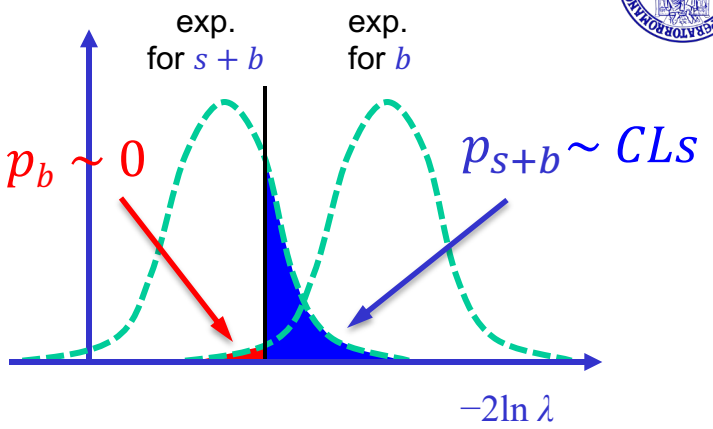
$$CL_s = \frac{p_{s+b}}{1-p_b}$$

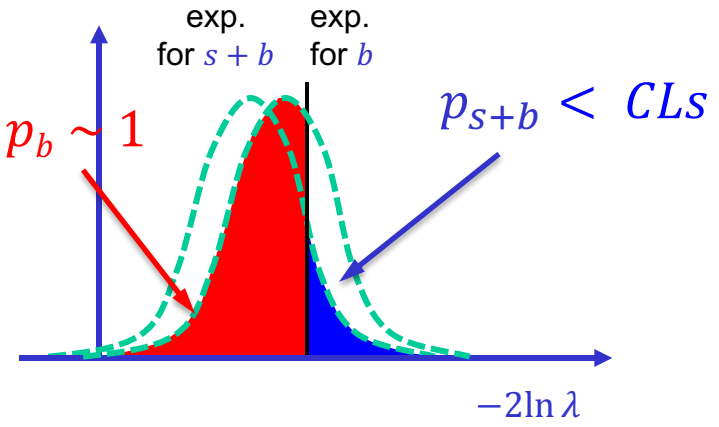
- Since $1 - p_b \leq 1$, $CL_s \geq p_{s+b}$, hence upper limits computed with the CL_s method are always **conservative**

Note: $\lambda \leq \lambda^{\text{obs}}$ implies $-2\ln\lambda \geq -2\ln\lambda^{\text{obs}}$

Main CL_s features

- p_{s+b}: probability to obtain a result which is less compatible with the signal than the observed result, assuming the signal hypothesis
- p_b: probability to obtain a result less compatible with background-only than the observed one
- If the distributions of the test statistic in the two hypotheses are very well separated and H_1 is true, then p_b will tend to be small ⇒ $1 - p_b \sim 1$ and $CL_s \sim p_{s+b}$, i.e: the ordinary p-value of the $s + b$ hypothesis
- If the two distributions largely overlap, then if p_b will be large ⇒ $1 - p_b$ tends to be small, preventing CL_s to become very small
- $CL_s < 1 - \alpha$ prevents rejecting cases where the experiment has little sensitivity





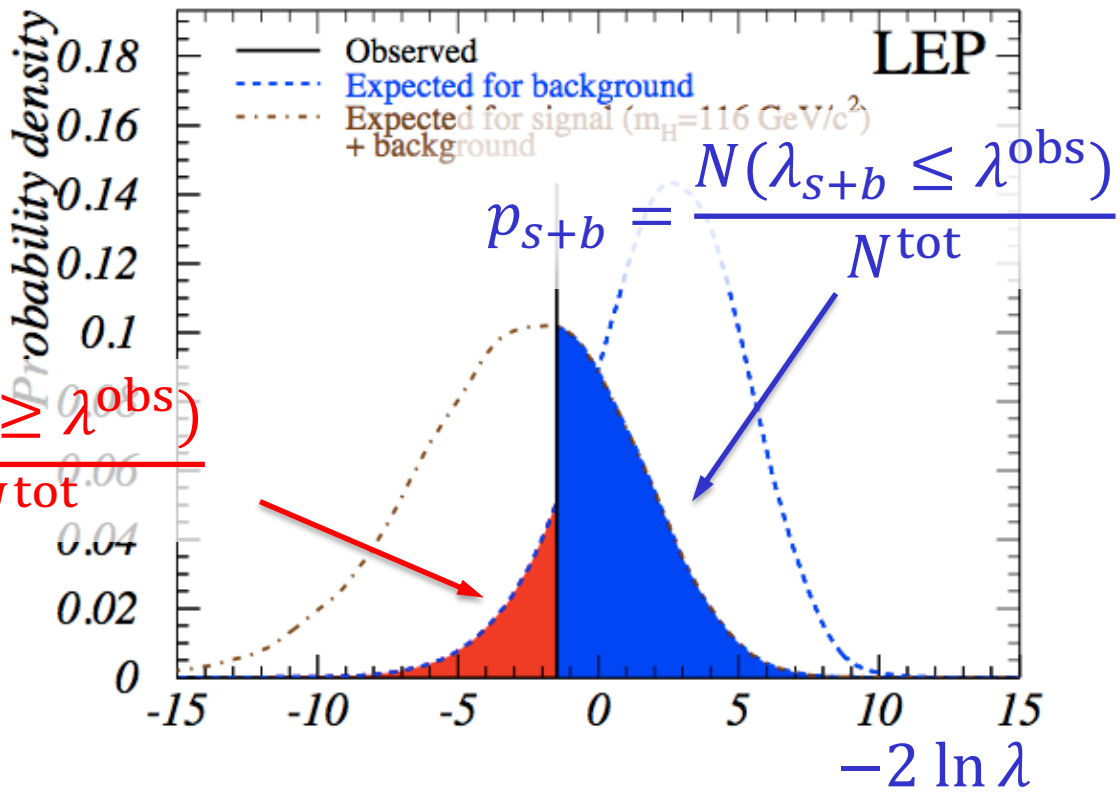
$$CL_s = \frac{p_{s+b}}{1 - p_b} = \frac{P(\lambda_{s+b} \leq \lambda^{\text{obs}})}{P(\lambda_b \leq \lambda^{\text{obs}})}$$

CL_s with toy experiments

- In practice, p_b and p_{s+b} are often computed in from simulated pseudo-experiments ("toy Monte Carlo")

$$CL_s = \frac{N(\lambda_{s+b} \leq \lambda^{\text{obs}})}{N(\lambda_b \leq \lambda^{\text{obs}})}$$

Plot from LEP Higgs combination paper



Event counting with CL_s

- Let's consider the previous event counting experiment, using $n = n^{\mathrm{obs}}$ as test statistic
- In this case CL_s can be written as:

$$CL_s = \frac{P(n \leq n^{\mathrm{obs}} | s + b)}{P(n \leq n^{\mathrm{obs}} | b)}$$

- Explicitating the Poisson distribution, the computation gives the same result as for the Bayesian case with a uniform prior
- In many cases the CL_s upper limits give results that are very close, numerically, to Bayesian computations done assuming a uniform prior
- But the interpretation is very different from Bayesian limits!

$$\alpha = e^{-s^{\mathrm{up}}} \frac{\sum_{m=0}^{n} \frac{(s^{\mathrm{up}} + b)^m}{m!}}{\sum_{m=0}^{n} \frac{b^m}{m!}}$$

Nuisance pars. in Bayesian approach

- Notation below: μ = parameter(s) of interest, θ = nuisance parameter(s)
- No special treatment:

$$P(\mu,\theta|x) = \frac{L(x;\mu,\theta)\pi(\mu,\theta)}{\int L(x;\mu',\theta')\pi(\mu',\theta')\mathrm{d}\mu'\mathrm{d}\theta'}$$

- $P(\mu|x)$ obtained as marginal PDF of μ obtained integrating on θ:

$$P(\mu|x) = \int P(\mu,\theta|x)\mathrm{d}\theta = \frac{\int L(x;\mu,\theta)\pi(\mu,\theta)\mathrm{d}\mu\,\mathrm{d}\theta}{\int L(x;\mu',\theta)\pi(\mu',\theta)\mathrm{d}\mu'\mathrm{d}\theta}$$

Nuisance pars., frequentist

- Introduce a complementary dataset to constrain the nuisance parameters θ (e.g.: calibration data, background estimates from control sample…)
- Formulate the statistical problem in terms of both the main data sample (x) and the control sample (y):

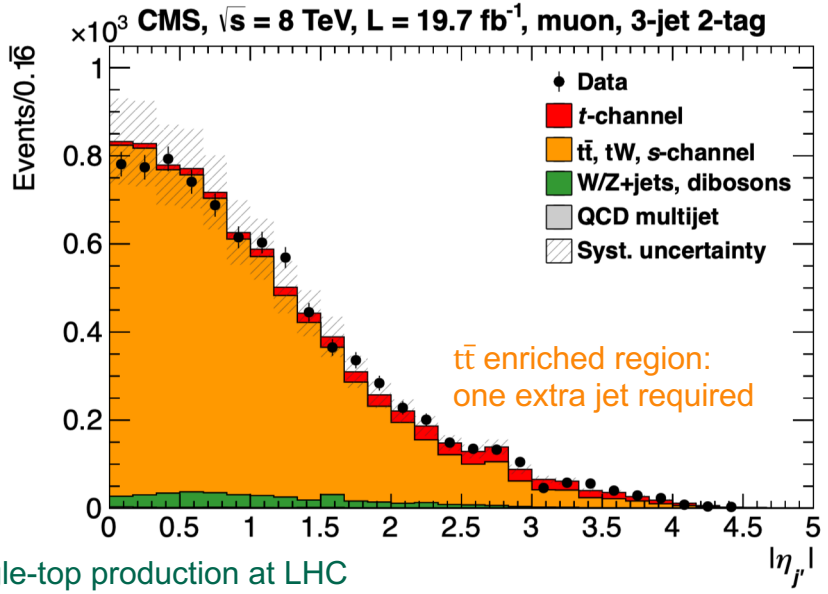
$$L(x, y; \mu, \theta) = L(x; \mu, \theta) L(y; \theta)$$

- Not always the control sample data are available
  - E.g.: calibration from test beam, stored in different formats, control samples analyzed with different software framework…
  - In some cases, may be complex and CPU intensive
- Simplest case; assume known PDF for "nominal" value of θ^{nom} (e.g.: estimate with Gaussian uncertainty):

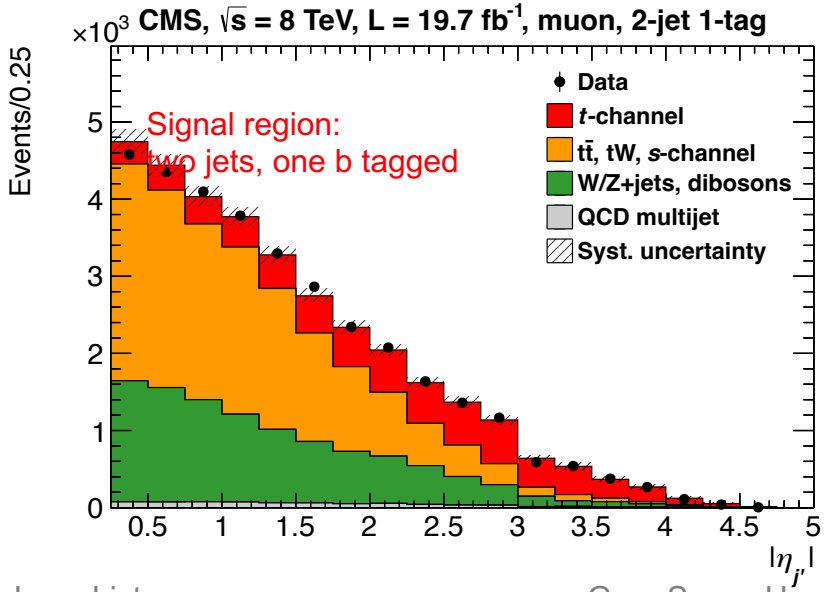
$$L(x, \theta^{\mathrm{nom}}; \mu, \theta) = L(x; \mu, \theta) L(\theta^{\mathrm{nom}}; \theta)$$

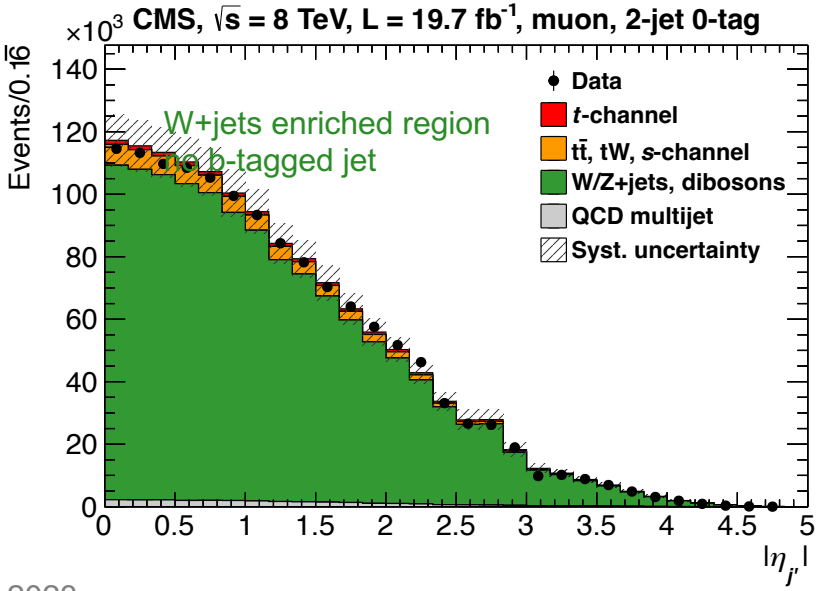
Fitting control regions

- In some cases, background parameters can be constrained from statistically independent control samples
  - Consider possible signal contamination!
- Background yield can be measured in background-enriched regions and extrapolated to signal regions applying scale factors predicted by simulation
- Complete likelihood function = product of likelihood functions in each considered regions, sharing common nuisance parameters
  - Typically: background rates



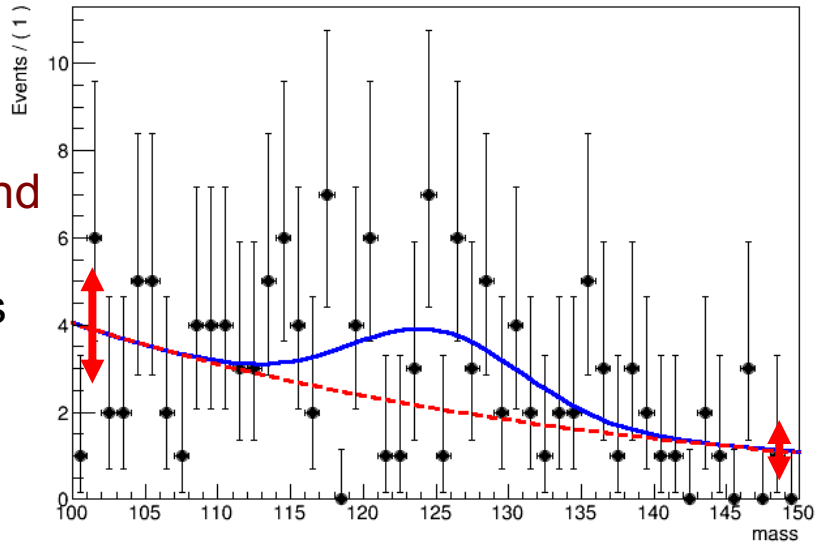
Measurement of single-top production at LHC





Systematic uncertainties

- Fit a Gaussian signal over an exponential background
- Assume a 30% uncertainty on the background yield with a log normal model may be assumed to avoid unphysical negative yields
- $b = \hat{b}\, e^{\beta}$, where our estimate β is centered around zero with a Gaussian uncertainty $\sigma_\beta = 0.3$



$$L(m; s, \beta) = L_0\left(m; s, \hat{b}\, e^{\beta}\right) P\left(\beta; \sigma_\beta\right)$$

$$L_0(m; s, b) = \frac{e^{-(s+b)}}{n!}\left(s\frac{1}{\sqrt{2\pi}\sigma}e^{-(m-\mu)^2/2\sigma^2} + b\lambda e^{-\lambda m}\right)$$

$$P\left(\beta; \sigma_\beta\right) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\beta^2/2\sigma_\beta^2}$$

b_0 = true (unknown) value
b = our estimate

Profile likelihood

- Define a test statistic based on a likelihood ratio inspired by Wilks' theorem:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

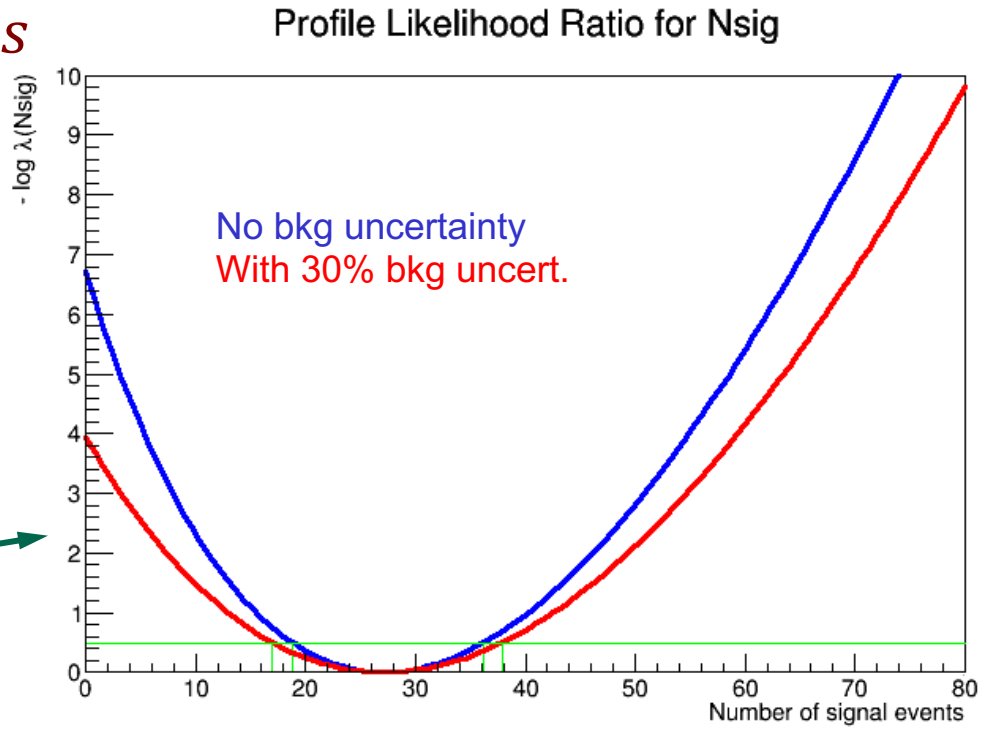
$\hat{\hat{\theta}}(\mu)$: Fix μ, fit θ

$\hat{\mu}, \hat{\theta}$: Fit both μ and θ

- μ is usually the "signal strength" (i.e.: $\sigma/\sigma_{\text{th}}$) in case of a search for a new signal
- Different 'flavors' of test statistics, e.g.: deal with unphysical $\mu < 0$, ecc.
- The distribution of $q_\mu = -2 \ln \lambda(\mu)$ may be asymptotically approximated to the distribution of a χ^2 with one degree of freedom (one parameter of interest = μ) due to the Wilks' theorem

Systematic uncertainties

- The profile likelihood shape is broadened due to the presence of nuisance parameter β that model systematic uncertainties
- → Smaller significance for discovery
- →Larger uncertainty on s



This implementation is based on RooStats, a package, released as optional library with ROOT
http://root.cern.ch

Significance evaluation

- Assume $\mu = 0$, if $q_0 = -2 \ln \lambda(0)$ can be approximated by a χ^2 with one d.o.f., then the significance is approximately equal to:

$$Z \cong \sqrt{q_0}$$

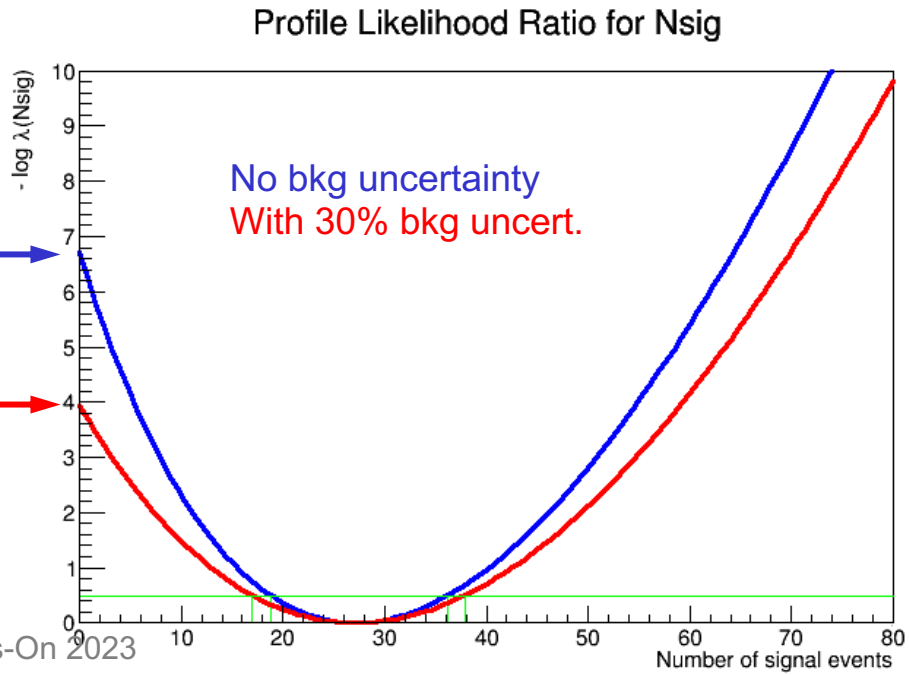
- The level of approximation can be verified with a computation done using pseudo experiments:
- Generate many toy samples with zero background and determine the distribution of $q_0 = -2 \ln \lambda(0)$, then count the fraction of cases with values greater than the measured value (p-value), and convert it to Z:

$$Z \cong \sqrt{2 \times 6.66} = 3.66$$

$$Z = \Phi^{-1}(1 - p)$$

$$Z \cong \sqrt{2 \times 3.93} = 2.81$$

- Toy samples may be unpractical for very large Z



Asymptotic approximations

- Asymptotic approximate formulae exist for most of adopted estimators, valid for $N \to \infty$, but in practice usable down to $N \approx 5$.
- If we want to test μ and we assume data are distributed according to μ', we can write:

$$-2 \ln \lambda(\mu) = \frac{(\mu - \hat{\mu})^2}{\sigma_{\hat{\mu}}^2} + \mathcal{O}\left(1/\sqrt{N}\right)$$

- where $\hat{\mu}$ is distributed according to a Gaussian with average μ' and standard deviation $\sigma_{\hat{\mu}}^2$ (A. Wald, 1943)
- Asymptotic approximations for $\sigma_{\hat{\mu}}^2$ can be determined and provide useful formulae

A. Wald, Trans. of AMS 54 n.3 (1943) 426-482

G. Cowan et al., EPJ C71 (2011) 1554

Asymptotic approximations

- Under the true hypothesis μ, $\hat{\mu}$ is distributed around μ and the test statistic, neglecting the $\mathcal{O}\left(1/\sqrt{N}\right)$ term, is distributed according to a χ^2 with one degree of freedom (Wilks' theorem):

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma_{\hat{\mu}}^2}$$

- If $\hat{\mu}$, instead, is distributed around a value $\mu' \neq \mu$, the distribution of the test statistic is so-called non-central χ^2, which is known and can be computed by series expansion.

Variations on test statistic

- Test statistic for discovery: $q_0 = \begin{cases} -2\ln\lambda(0), & \hat{\mu} \geq 0 \\ 0, & \hat{\mu} < 0 \end{cases}$
  - In case of a negative estimate of μ, set the test statistic to zero: consider only positive μ as evidence against the background-only hypothesis. Approximately: $Z \cong \sqrt{q_0}$.
- Test statistic for upper limits: $q_\mu = \begin{cases} -2\ln\lambda(\mu), & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases}$
  - If the estimate is larger than the assumed μ, an upward fluctuation occurred. Don't exclude μ in those cases, set the statistic to zero
- Higgs test statistic:

$$\tilde{q}_\mu = \begin{cases} -2\ln\dfrac{L(\vec{x}|\mu, \hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|0, \hat{\hat{\vec{\theta}}}(0))} & \hat{\mu} < 0 \quad \leftarrow \text{Protect for unphysical } \mu < 0 \\ -2\ln\dfrac{L(\vec{x}|\mu, \hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|\hat{\mu}, \hat{\hat{\vec{\theta}}})} & 0 \leq \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \quad \leftarrow \text{As for upper limits statistic} \end{cases}$$

G. Cowan et al., EPJ C71 (2011) 1554

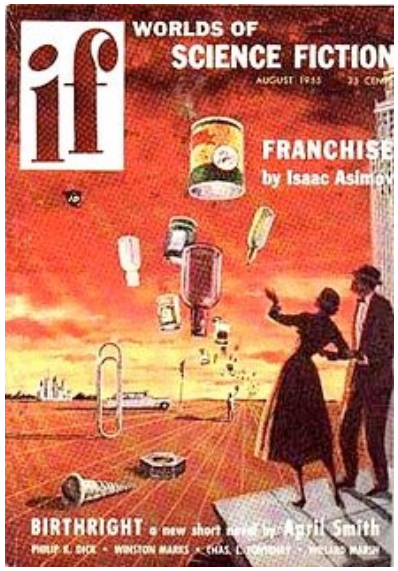
Asimov datasets

- Convenient set to compute approximate values:
  *"We define the **Asimov data set** such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values"*
- Imagine that our only parameter is μ. We would like to have a dataset where the fit value is the true value μ'.
- This can be done using as number of counts the (non-integer) value $n = \mu' s + b$
- In this case, we have:

$$-2 \ln \lambda_A(\mu) \cong \frac{(\mu - \mu')^2}{\sigma_{\hat{\mu}}^2}$$

- Reversing the above equation, we can estimate the variance of $\hat{\mu}$ to be used in Wald's approximation of the test statistic:

$$\sigma_{\hat{\mu}}^2 \cong -\frac{(\mu - \mu')^2}{2 \ln \lambda_A(\mu)}$$



G. Cowan et al., EPJ C71 (2011) 1554

Asimov datasets

- In practice: all observables are replaced with their expected value
- Expected values of yields are possibly non integer:

$$\lambda_A(\mu) = \frac{L_A(\mu, \hat{\hat{\theta}})}{L_A(\hat{\mu}, \hat{\theta})} = \frac{L_A(\mu, \hat{\hat{\theta}})}{L_A(\mu', \hat{\theta})}$$

- The variance of the test statistic, in Wald's approximation, is estimated as:

$$\sigma^2_{\hat{\mu}} \cong \frac{(\mu - \mu')^2}{-2 \ln \lambda_A}$$

- Median significance for discovery or exclusion (and their $\pm 1\sigma$ bands) can be obtained using the Asimov dataset

$$\text{med}[Z_0|\mu'] = \sqrt{q_{0,A}}$$ ← For discovery using q_0

$$\text{med}[Z_0|0] = \sqrt{q_{\mu,A}}$$ ← For upper limit using q_μ

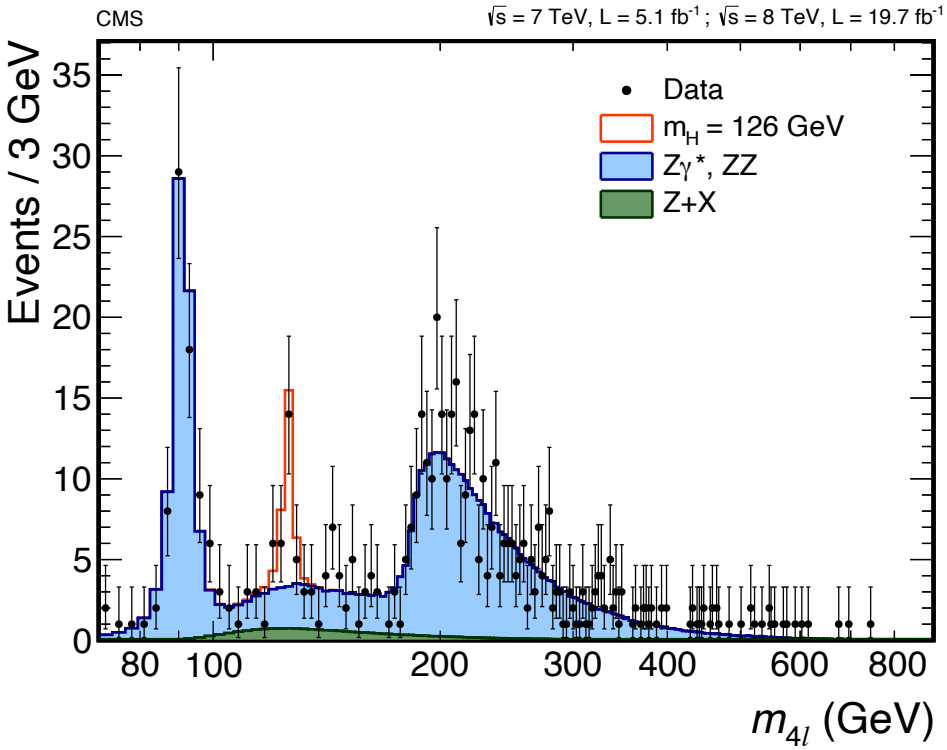
$$\text{med}[Z_\mu|\mu'] = \sqrt{\tilde{q}_{\mu,A}}$$ ← Upper limits using $\tilde{q}_\mu$

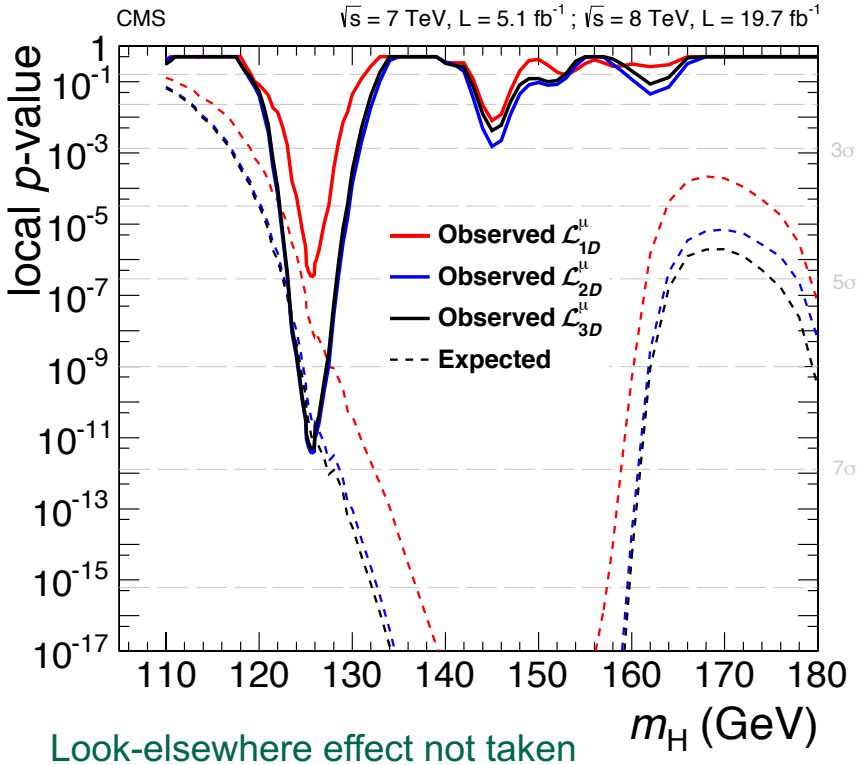
In practice: all the interesting formulae are implemented in RooStats package, released as optional library in ROOT

G. Cowan et al., EPJ C71 (2011) 1554

Putting all together

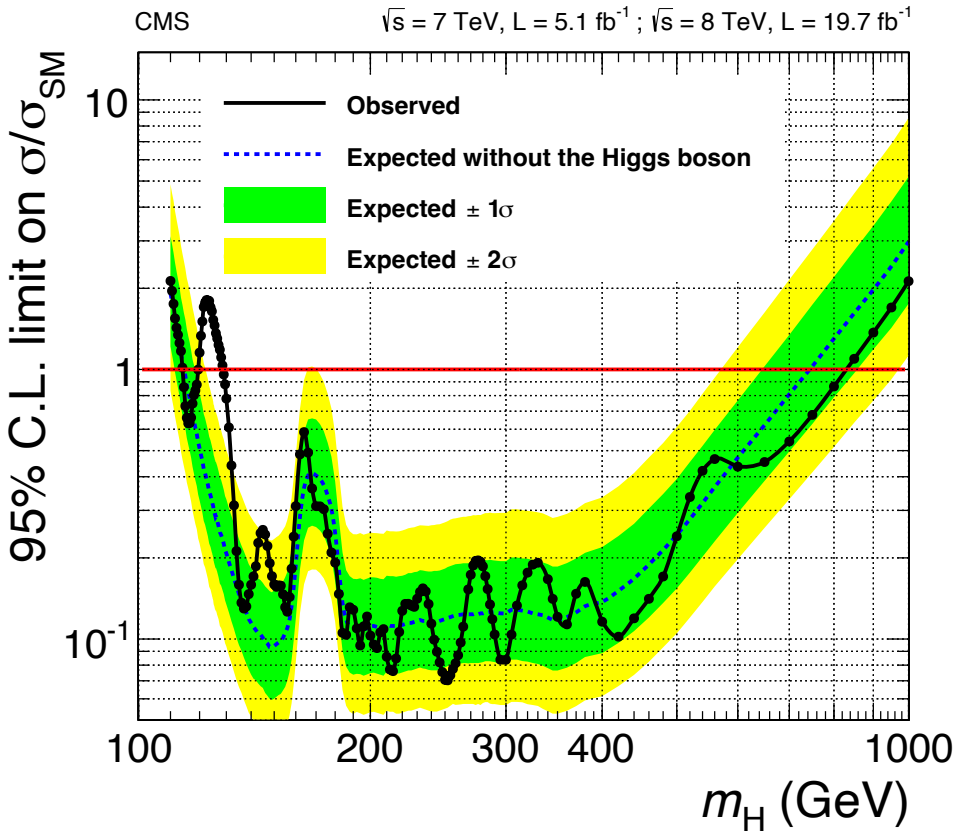
- Search for Higgs boson in H→4l at LHC
- 1D, 2D, 3D: different test statistics using 4l invariant mass plus other discriminating variables based on the event kinematics

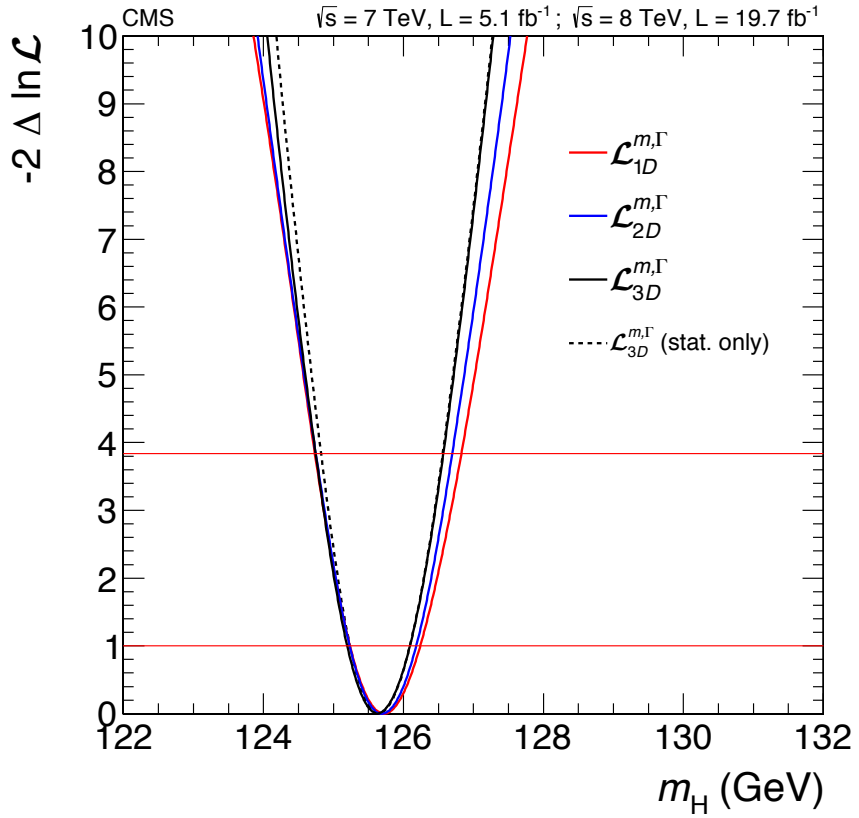




Look-elsewhere effect not taken into account here

Higgs exclusion





"The modified frequentist construction CLs is adopted as the primary method for reporting limits. As a complementary method to the frequentist construction, a Bayesian approach yields consistent results."

Agreed statistical procedure described in:
ATLAS and CMS Collaborations,
LHC Higgs Combination Group
ATL-PHYS-PUB 2011-11/CMS NOTE 2011/005, 2011.

The look-elsewhere effect

- Consider a search for a signal peak over a background distribution that is smoothly distributed over a wide range
- You could either:
  - Know which mass to look at, e.g.: search for a rare decay with a known particle, like $B_s \to \mu\mu$
  - Search for a peak at an unknown mass value, like for the Higgs boson
- In the former case it's easy to compute the peak significance:
  - Evaluate the test statistics for $\mu = 0$ (background only) at your observed data sample
  - Evaluate the p-value according to the expected distribution of your test statistic q under the background-only hypothesis, convert it to the equivalent area of a Gaussian tail to obtain the significance level:

$$p = \int_{q^{\mathrm{obs}}}^{\infty} f(q|\mu = 0)\,, \qquad Z = \Phi^{-1}(1-p)$$

The look-elsewhere effect

- Searching for a peak at an unknown mass, the previous p-value has only a local meaning, with probability to find a background fluctuation as large as your signal or more at a fixed mass value m:

$$p(m) = \int_{q^{\mathrm{obs}}(m)}^{\infty} f(q|\mu = 0)\mathrm{d}q$$

- We need the probability to find a background fluctuation at least as large as your signal at **any** mass value (global)
- local p-value would be an overestimate of the global p-value
- The probability of an over-fluctuation at least one mass value increases with the size of the searched range, roughly proportionally to the ratio of resolution over the search range, also depending on the significance of the peak
  - Better resolution = less chance to have more events compatible with the same mass value
- Possible approach: let also m fluctuate in the test statistics fit:

$$\hat{q}_0 = -2\ln\frac{L(\mu=0)}{L(\hat{\mu};\hat{m})}, \quad p^{\mathrm{glob}} = \int_{\hat{q}_{\mathrm{obs}}}^{\infty} f(\hat{q}_0|\mu = 0)\mathrm{d}\hat{q}_0$$

For $\mu = 0$ L doesn't depend on m and Wilks' theorem doesn't apply

Estimate LEE

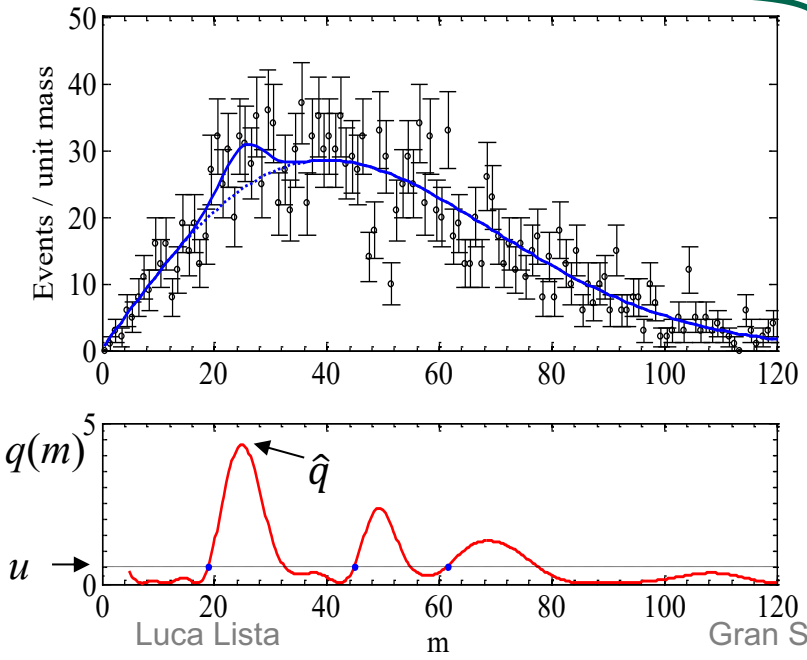
- The effect can be evaluated by running many background-only Monte Carlo samples, determining the distribution of the maximum value of test statistic $\hat{q}$ in the search range, and counting the fraction of samples with $\hat{q}$ greater than the observed value u
- Approximate evaluation based on local p-value, times correction factors ("trial factors", Gross and Vitells, EPJC 70:525-530,2010)

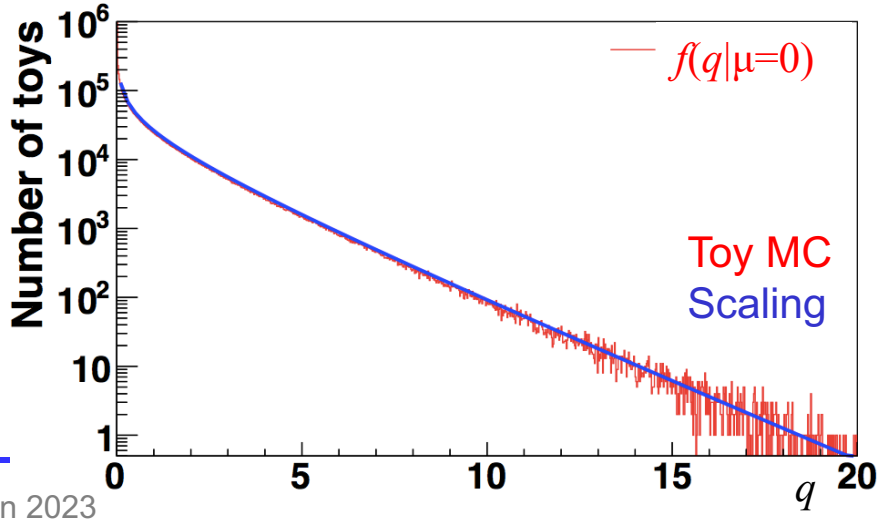
$$p^{\text{glob}} = P(\hat{q} > u) \cong \langle N_u \rangle + \frac{1}{2} P(\chi^2 > u)$$

$$\hat{q} = \max_m q(m)$$

$\langle N_u \rangle$ is the average number of up-crossings of the test statistic, can be evaluated at some lower reference level (toy MC) and scaled by:

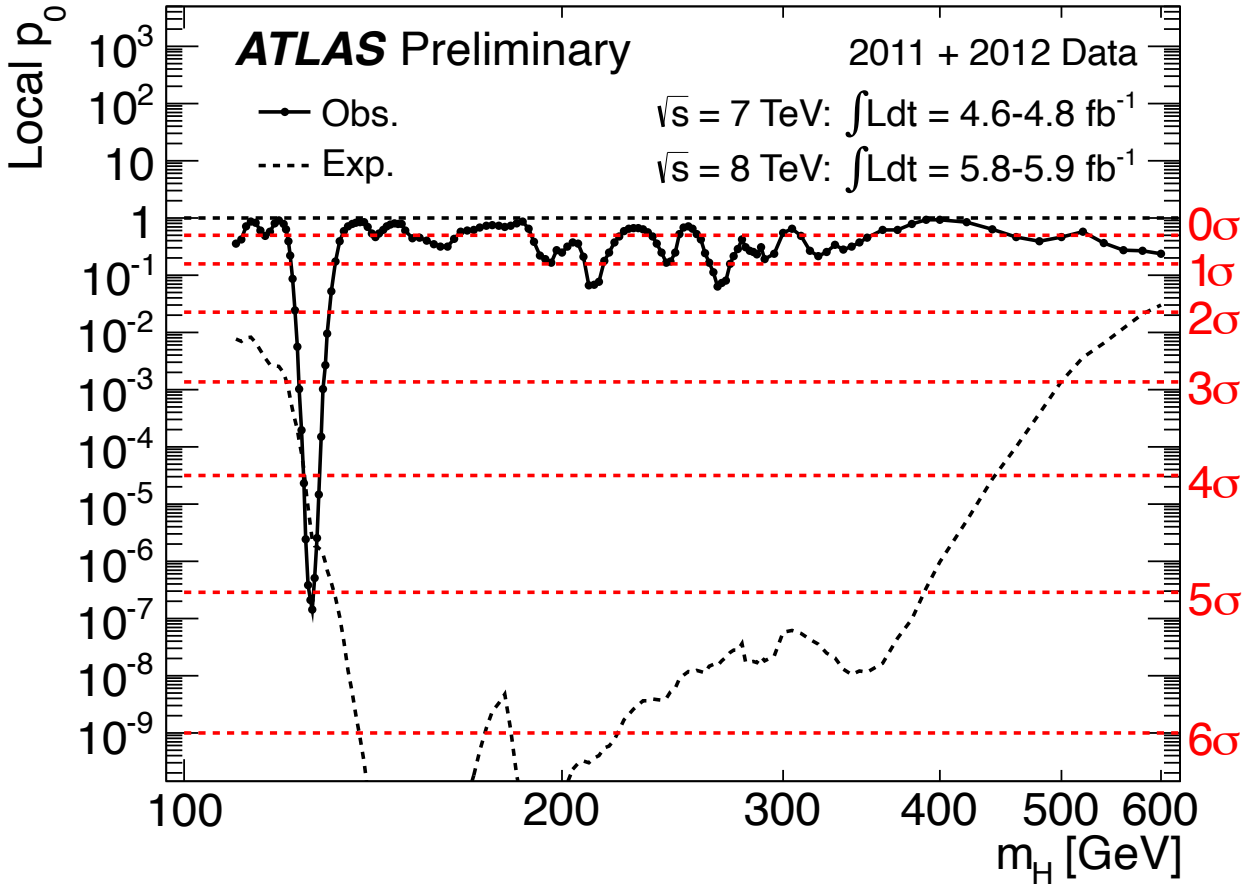
$$\langle N_u \rangle = \langle N_{u_0} \rangle e^{-(u-u_0)/2}$$





Practical application

- Higgs search at ATLAS



Practical application

- Use the number of σ, Z, as test statistic: $u = Z^2$ is distributed as a chi-square
- Use the 0σ level ($p = 0.5$) as level u^0, then extrapolate to the minimum *p*-value, where $Z \cong 5$, i.e.: $u = Z^2 \cong 5^2 = 25$
- The number of upcrossings can be counted from the plot, and is equal to $N_0 = 9$, which allows us to estimate: $\langle N_0 \rangle = 9 \pm 3$
- Estimate the global *p*-value as:
- $p^{\text{glob}} \cong \langle N_u \rangle + \frac{1}{2} P(\chi^2 > u) \cong \langle N_u \rangle + 3\times10^{-7}$
- $\langle N_u \rangle \cong \langle N_0 \rangle e^{-(5^2 - 0^2)/2}$
- $\langle N_u \rangle \cong (9 \pm 3) e^{-25/2} \cong (3 \pm 1)\times10^{-5}$
- $p^{\text{glob}} \cong 3\times10^{-5} + 3\times10^{-7} \cong 3\times10^{-5} \implies Z \cong 4\sigma$ instread of 5σ

- A toy Monte Carlo would give a more precise estimate compared with this back-of-the envelope example

Backup

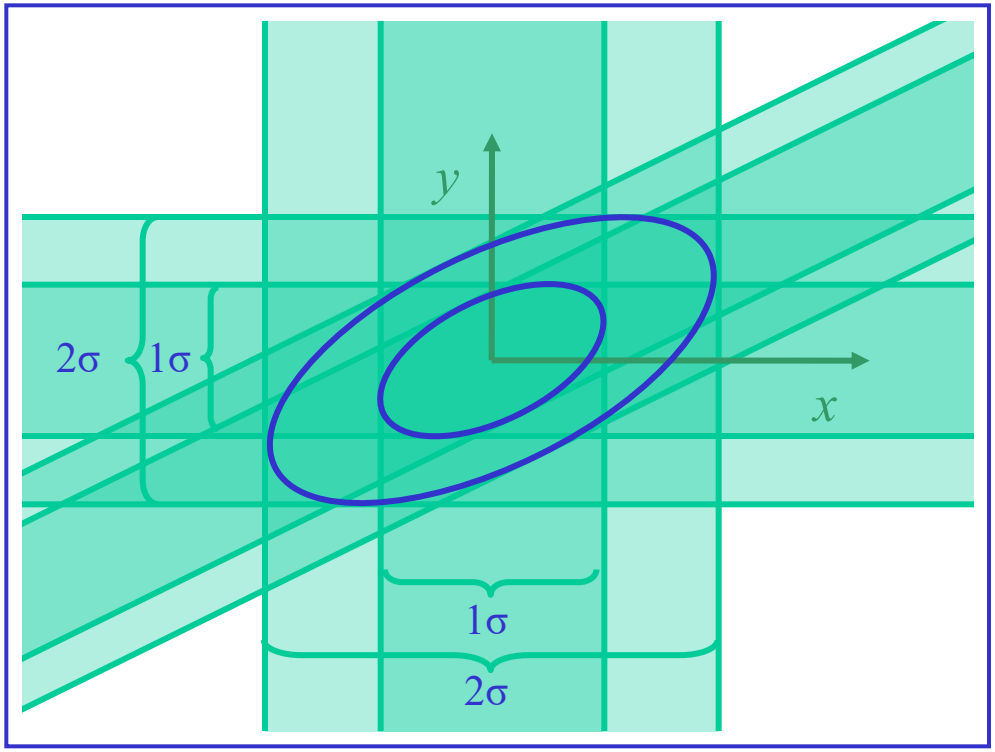
2D intervals

- In more dimensions one can determine 1σ and 2σ contours
- Note: different probability content in 2D compared to one dimension
- 68% and 95% contours are usually preferable

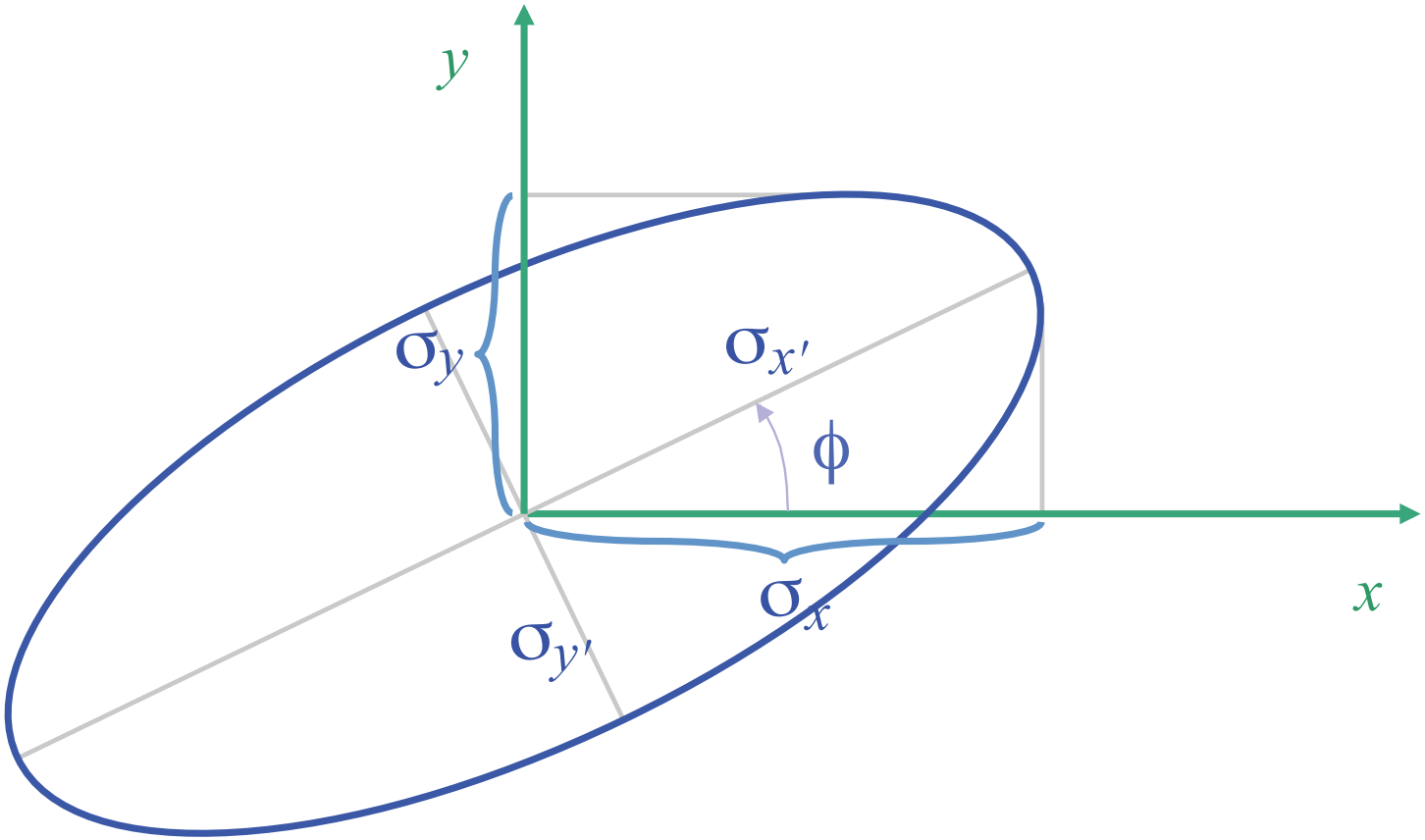
$$P_{1D}(n\sigma) = \sqrt{\frac{2}{\pi}} \int_0^n e^{-x^2/2} \mathrm{d}x = \mathrm{erf}\left(\frac{2}{\sqrt{2}}\right)$$

$$P_{2D}(n\sigma) = \int_0^n x e^{-x^2/2} \mathrm{d}x = 1 - e^{-n^2/2}$$

Width	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ		0.9973



Contours in two dimensions



Cousins-Highland hybrid approach

- Method proposed by Cousins and Highland
  - Add posterior from another experiment into the likelihood definition
  - Integrate the likelihood function over the nuisance parameters

$$L_{\text{hybrid}}(x;\mu) = \int L(x;\mu,\theta) L(\theta^{\text{nom}};\theta)\, \mathrm{d}\theta$$

- Also called “hybrid” approach, because a partial Bayesian approach is implicit in the integration
  - Bayesian integration of PDF, then likelihood used in a frequentist way
- Not guaranteed to provide exact frequentist coverage!
- Numerical studies with pseudo experiments showed that the hybrid CL_s upper limits gives very similar results to Bayesian limit assuming a uniform prior

NIM A320 (1992) 331-335

From PDG Review…

- *"The intervals constructed according to the unified procedure [FC] for a Poisson variable n consisting of signal and background have the property that for $n = 0$ observed events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if $n = 0$ for the experiment in question, then no background was observed, and therefore one may argue that the expected background should not be relevant. The extent to which one should regard this feature as a drawback is a subject of some controversy"*

Observations on the CL_s method

- *"A specific modification of a purely classical statistical analysis is used to avoid excluding or discovering signals which the search is in fact not sensitive to"*
- *"The use of CL_s is a conscious decision not to insist on the frequentist concept of full coverage (to guarantee that the confidence interval doesn't include the true value of the parameter in a fixed fraction of experiments)."*
- *"confidence intervals obtained in this manner do not have the same interpretation as traditional frequentist confidence intervals nor as Bayesian credible intervals"*

A. L. Read, Modified frequentist analysis of search results (the CLls method), 1st Workshop on Confidence Limits, CERN, 2000

LEP, Tevatron, LHC Higgs limits

	Test statistic	Profiled?	Test statistic sampling
LEP	$q_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\tilde{\theta})}{\mathcal{L}(data\|0,\tilde{\theta})}$	no	Bayesian-frequentist hybrid
Tevatron	$q_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|0,\hat{\theta}_0)}$	yes	Bayesian-frequentist hybrid
LHC	$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|\hat{\mu},\hat{\theta})}$	yes $(0 \leq \hat{\mu} \leq \mu)$	frequentist