A Brief Introduction to Statistics

Luca Lista

Università Federico II
INFN Naples

Mario Pelliccioni

INFN Torino

Introduction to probability

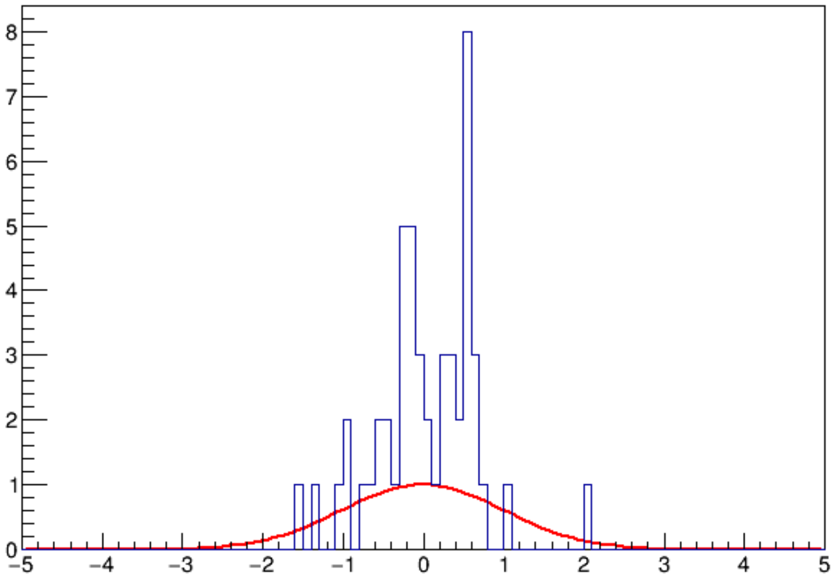
- Probability can be defined in different ways
- The applicability of each definition depends on the kind of claim we are considering to applying the concept of probability
- One subjective approach expresses the degree of belief/credibility of the claim, which may vary from subject to subject
- For repeatable experiments, probability may be a measure of how frequently the claim is true

Frequentist probability

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

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

- Exactly realizable only with an infinite number of trials
  - Conceptually may be unpleasant
  - Pragmatically acceptable by physicists
- Only applicable to repeatable experiments

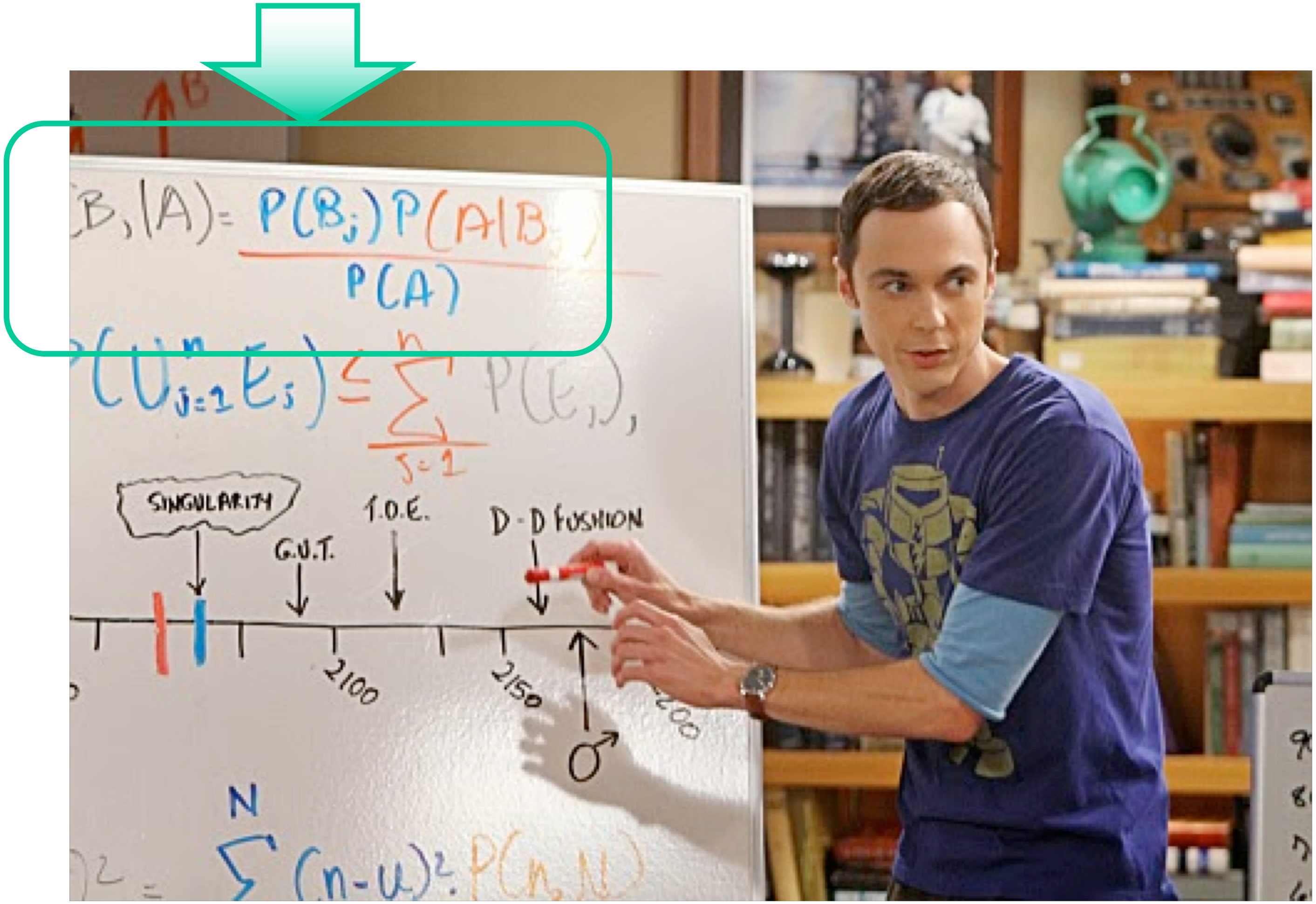


Subjective (Bayesian) probability

- Expresses one's degree of belief that a claim is true
  - How strong? Would you bet?
  - Applicable to all unknown events/claims, not only repeatable experiments
  - Each individual may have a different opinion/prejudice

- Quantitative rules exist about how subjective probability should be modified after **learning** about some observation/evidence
  - Consistent with Bayes theorem (→ will be introduced in next slides)
  - Prior probability → Posterior probability (following observation)
  - The more information we receive, the more Bayesian probability is insensitive on prior subjective prejudice (unless in pathological cases…)

The Bayes theorem



The Big Bang Theory © CBS

Bayesian posterior probability

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

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

- $P(H)$ = prior probability
- $P(H \mid E)$ = posterior probability, given E
- The Bayes rule allows to define a rational way to modify one's prior belief once some observation is known

Bayes rule and likelihood function

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

$$P(\theta_1, \cdots, \theta_m | x_1, \cdots, x_n) =$$

$$\frac{L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)}{\int L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)\mathrm{d}^m\theta}$$

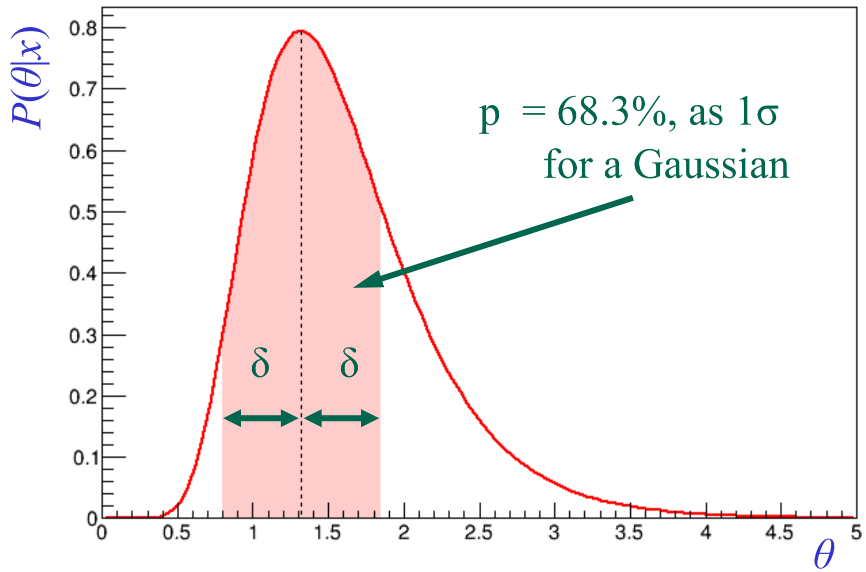
- Where $\pi(\theta_1, \ldots, \theta_m)$ is the subjective prior probability
- The denominator $\int L(x, \theta)\, \pi(\theta)\, \mathrm{d}^m\theta$ is a normalization factor
- The observation of $x_1, \ldots, x_n$ modifies the prior knowledge of the unknown parameters $\theta_1, \ldots, \theta_m$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

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)

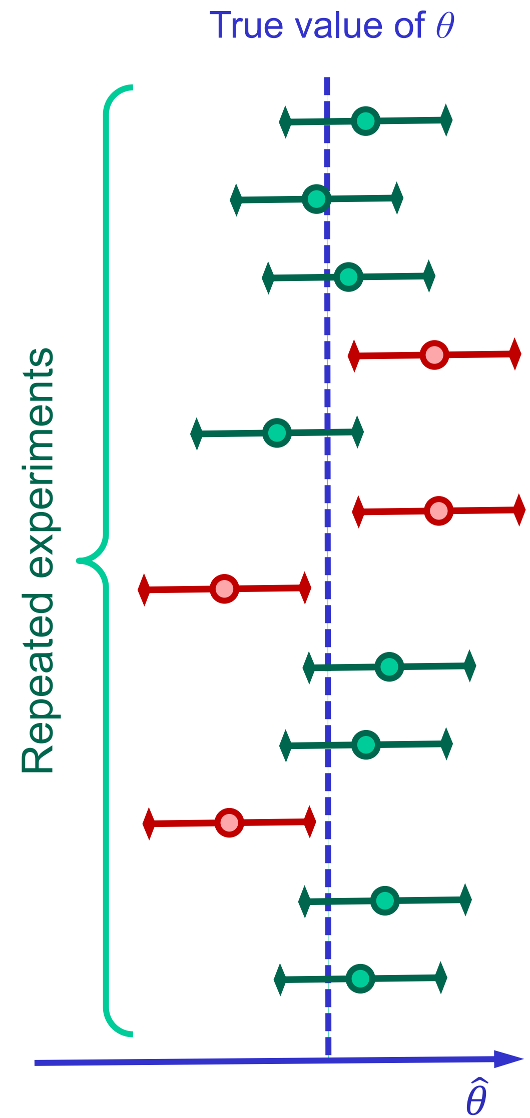
$$P(\theta|x) = \frac{L(x;\theta)\pi(\theta)}{\int L(x;\theta)\pi(\theta)\mathrm{d}\theta}$$

- 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



Frequentist inference

- 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 confidence intervals contain the fixed and 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)



Maximum likelihood

- Given a sample of N measurements of the variables $(x_1, \ldots, x_n)$, the likelihood function is:

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

- If the size N of the sample is also a random variable, the extended likelihood function is usually also 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 $P(N; \theta_1, \ldots, \theta_m)$ is in practice always a Poisson distribution whose expected rate is a function of the unknown parameters
- The maximum-likelihood estimator is the most adopted parameter estimator
- The "best fit" parameters correspond to the set of values that maximizes the likelihood function

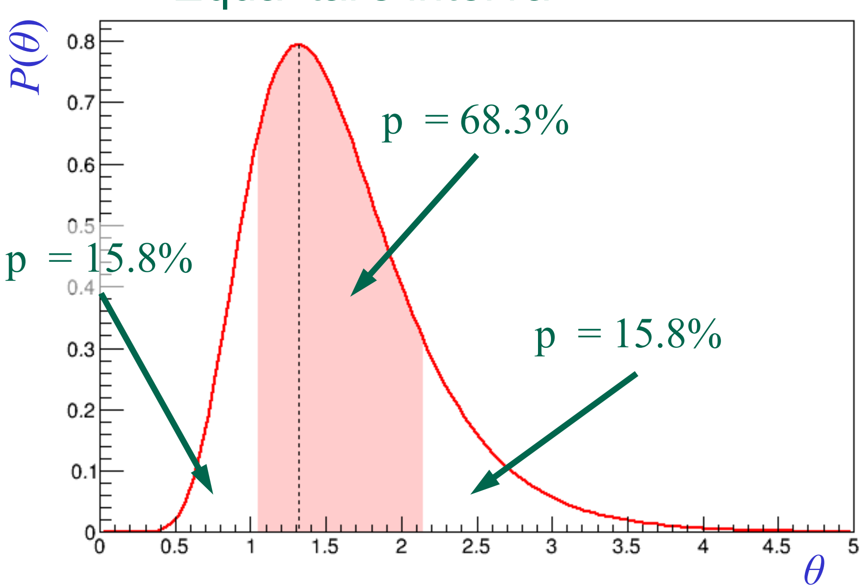
Choice of 68% prob. intervals

- Different interval choices are possible, corresponding to the same probability level (usually 68%, as 1σ for a Gaussian)
  - Equal areas in the right and left tails
  - Symmetric interval
  - Shortest interval
  - …

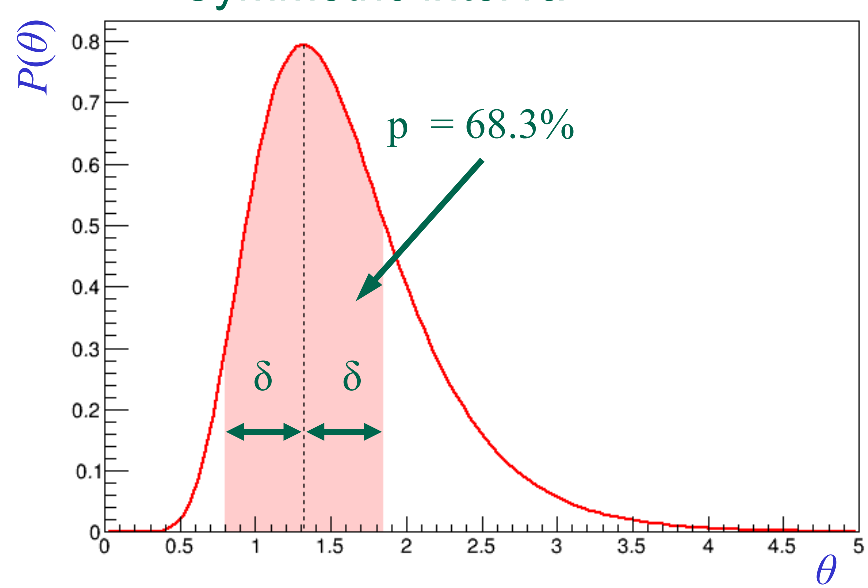
All equivalent for a symmetric distribution (e.g. Gaussian)

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

Equal tails interval

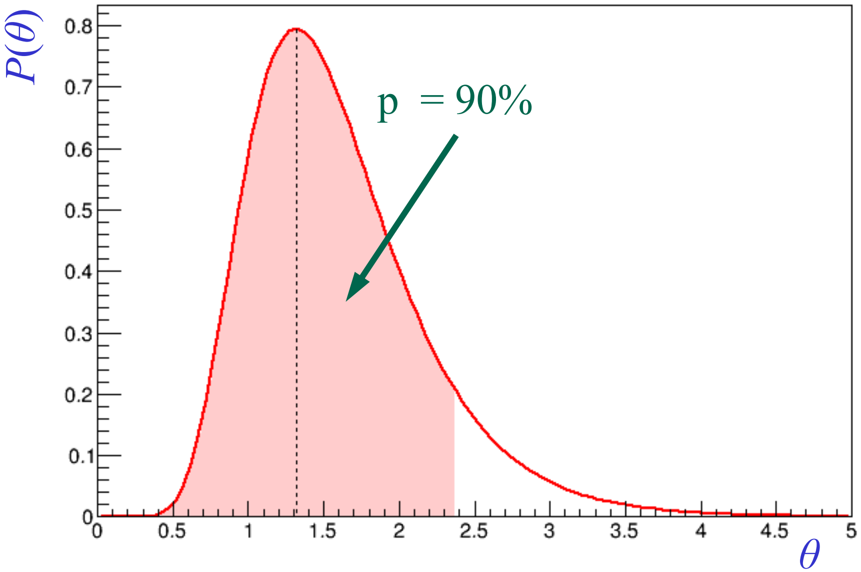


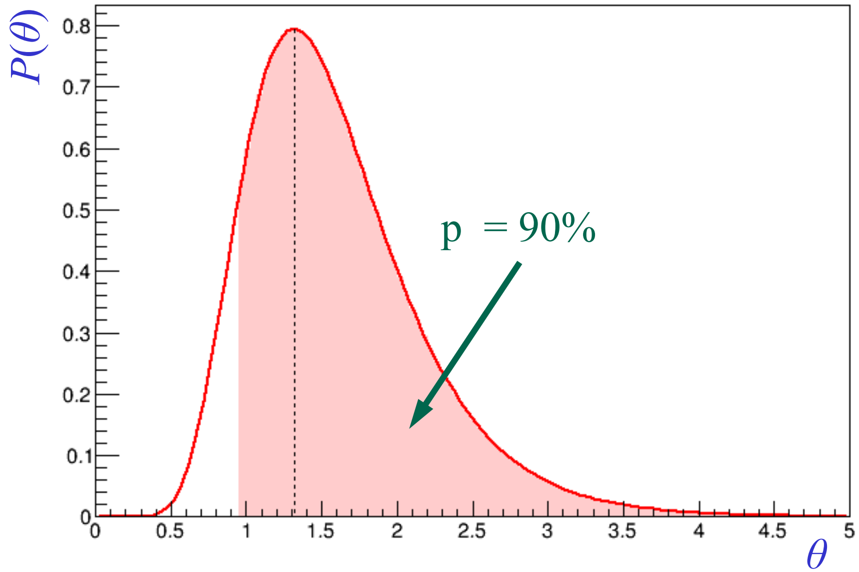
Symmetric interval



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)



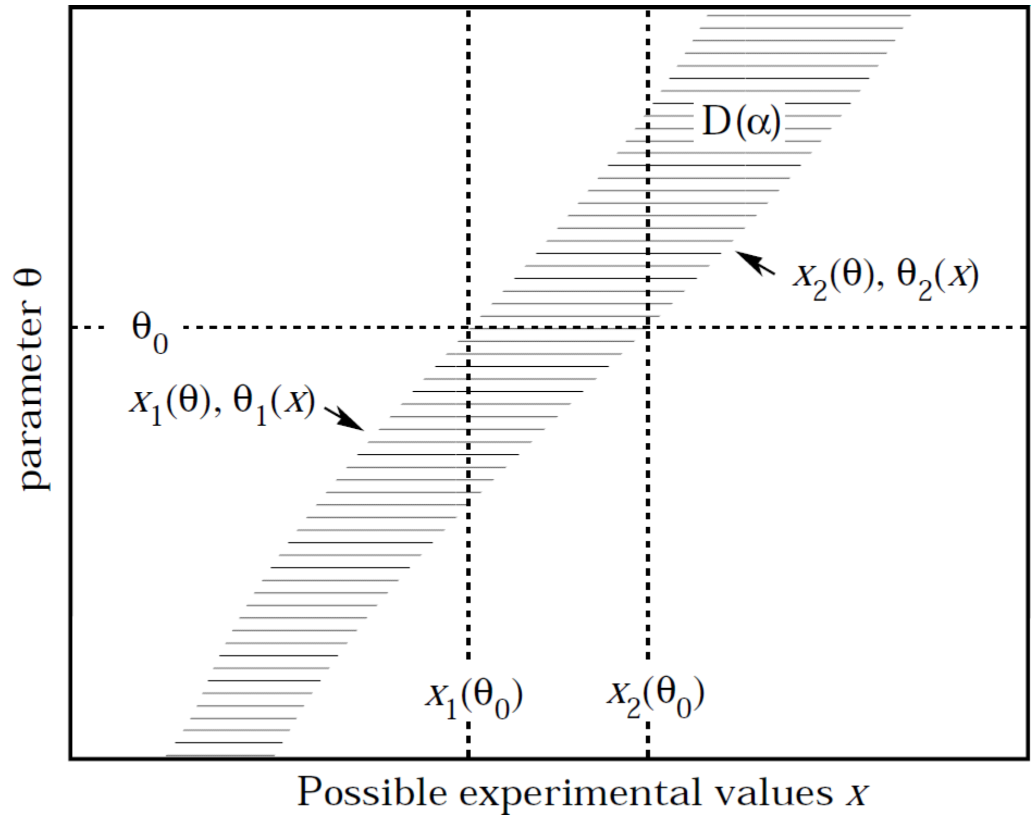


Neyman's confidence intervals

Procedure to determine frequentist confidence intervals

- Scan the allowed range of an unknown parameter θ
- Given a value of θ compute the interval $[x_1, x_2]$ that contain x with a probability $1 - \alpha$ equal to 68% (or 90%, 95%)
- **Choice of interval needed!**
- Invert the confidence belt: for an observed value of x, find the interval $[\theta_1, \theta_2]$
- A fraction of the experiments equal to $1 - \alpha$ will measure x such that the corresponding $[\theta_1, \theta_2]$ contains ("covers") the true value of θ ("coverage")
- **Note**: the random variables are $[\theta_1, \theta_2]$, not θ !

Plot from PDG statistics review

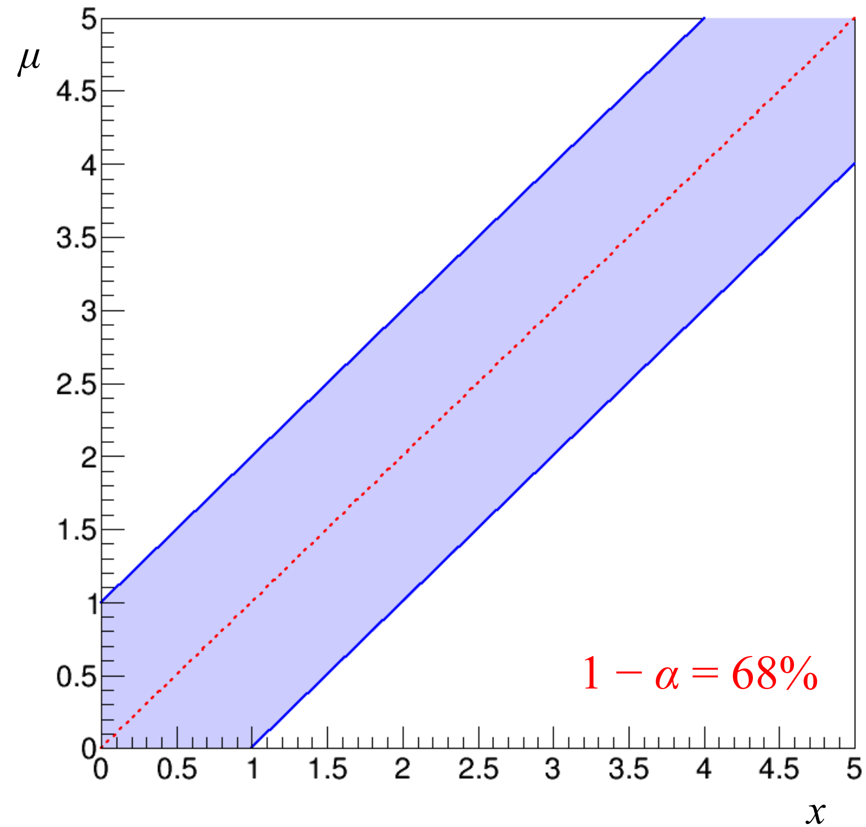


α = significance level

Simplest example: Gaussian case

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

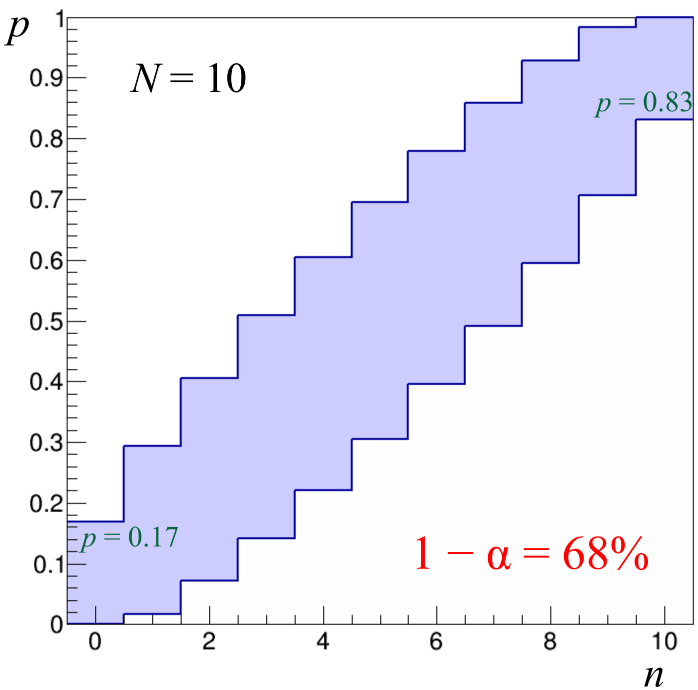
$$\mu = x \pm \sigma$$



Binomial intervals

- The Neyman's belt construction may only guarantee approximate coverage in case of discrete variables
- For a Binomial distribution: find the interval $\{n_{\min}, \ldots, n_{\max}\}$ such that:

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

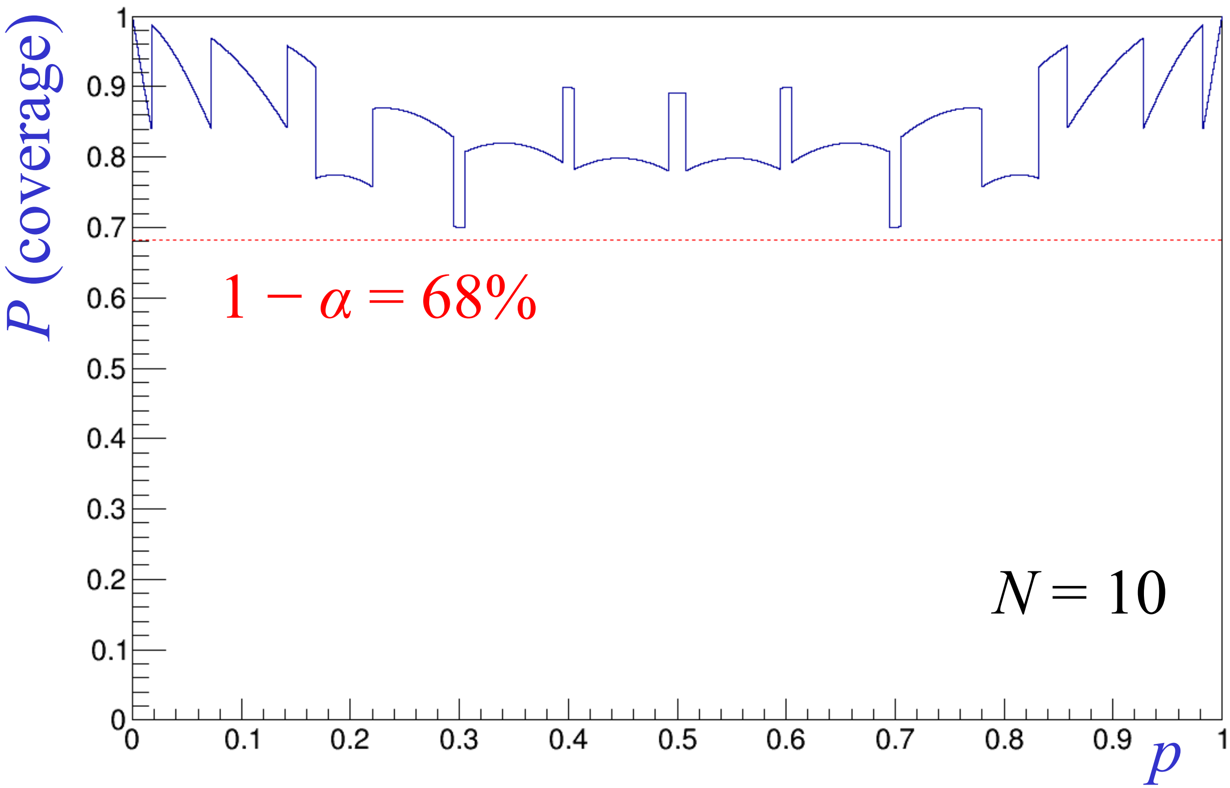


- Clopper and Pearson (1934) solved the belt inversion problem for central intervals
- For an observed $n = k$, find lowest p^{lo} and highest p^{up} such that:
- $P(n \leq k \mid N, p^{\mathrm{lo}}) = \alpha/2, \quad P(n \geq k \mid N, p^{\mathrm{up}}) = \alpha/2$
- E.g.: $n = N = 10$, $P(N|N) = p^N = \alpha/2$, hence: $p^{\mathrm{lo}} = \sqrt[10]{\alpha/2} = 0.83\ (68\%\ \mathrm{CL}), 0.74\ (90\%\ \mathrm{CL})$
- A frequently used approximation, which fails for $n = 0$, N is:

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

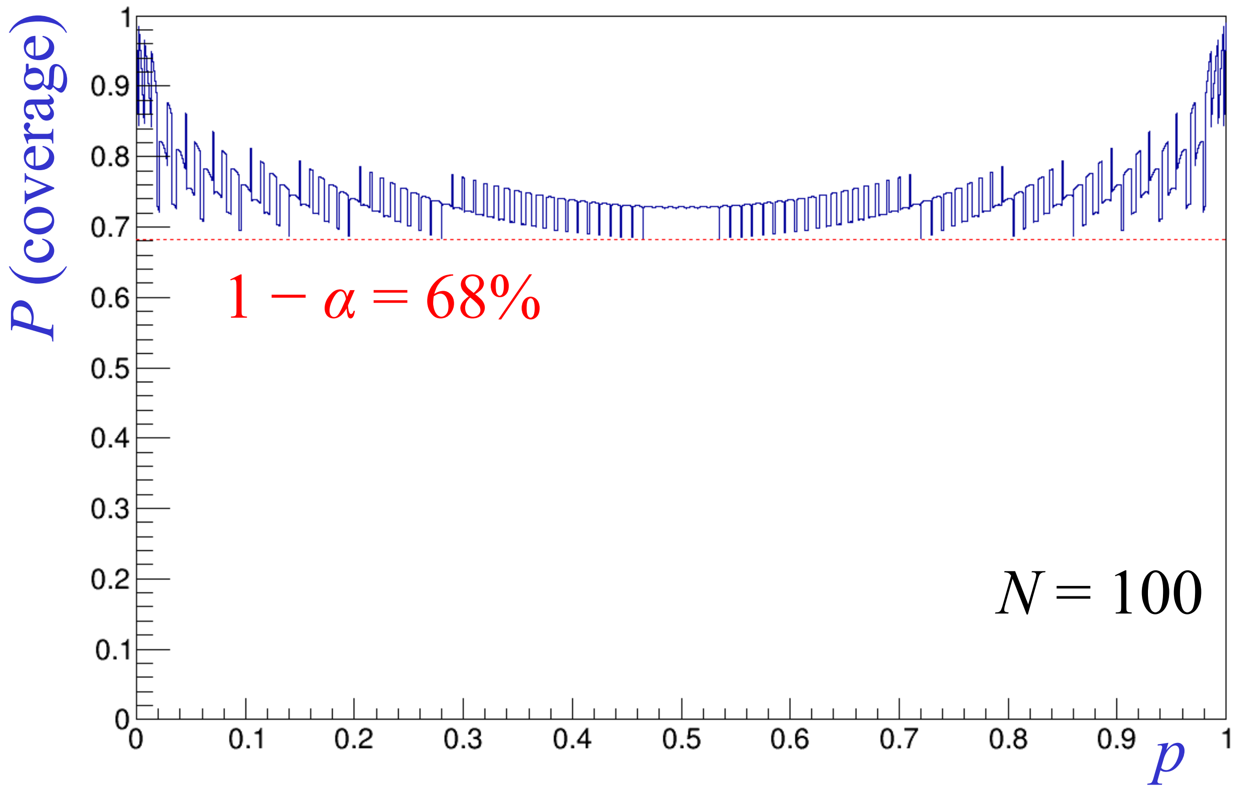
Clopper-Pearson coverage (I)

- 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



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

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

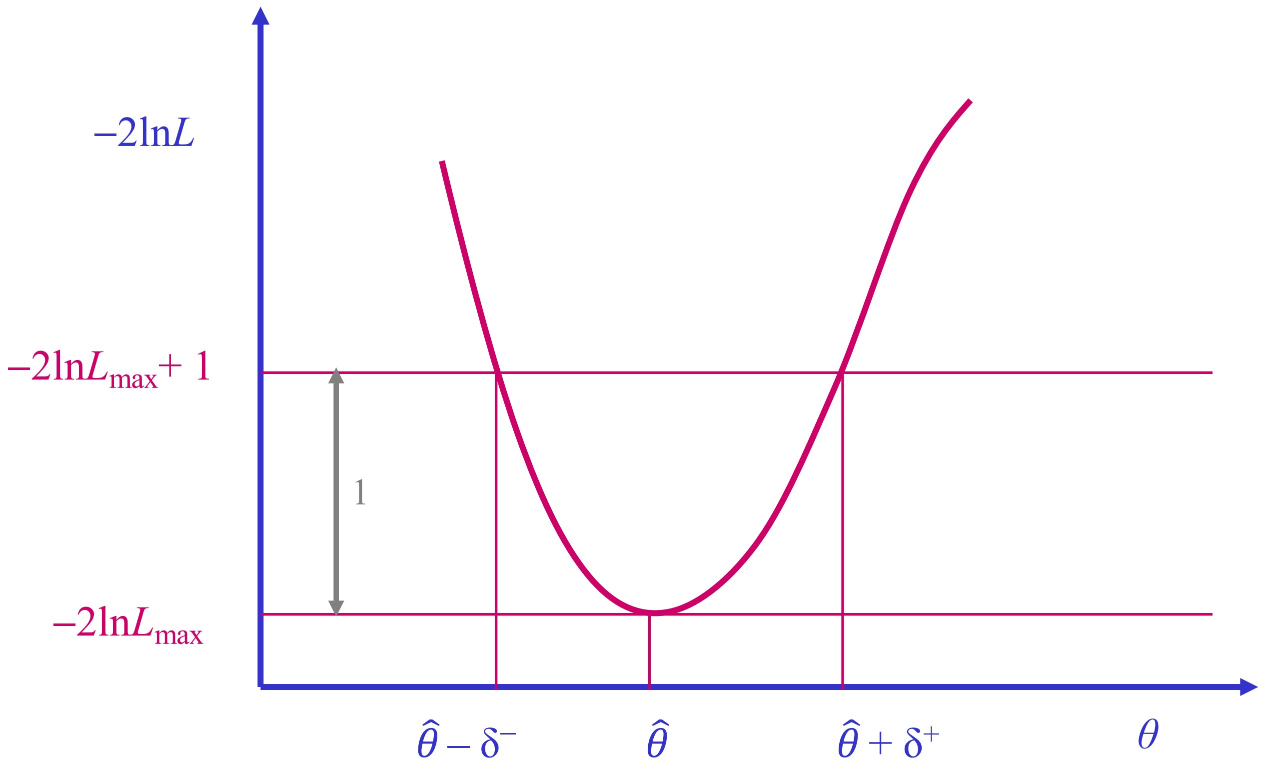
- Estimate of the covariance matrix from 2nd order partial derivatives w.r.t. fit parameters at the minimum:

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

- Implemented in Minuit as MIGRAD/HESSE function

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)



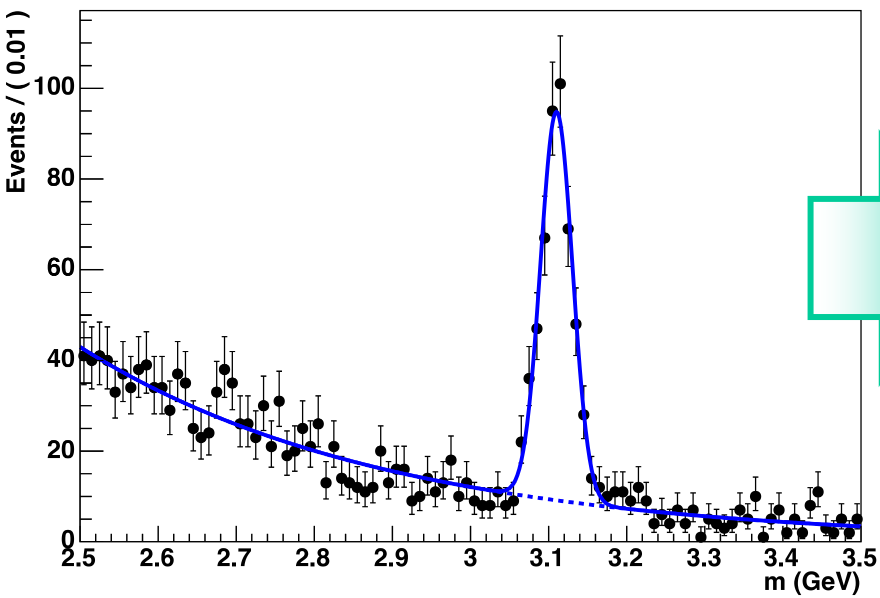
- Errors can be asymmetric
- For a Gaussian PDF the result is identical to the 2nd order derivative matrix
- Implemented in Minuit as MINOS function

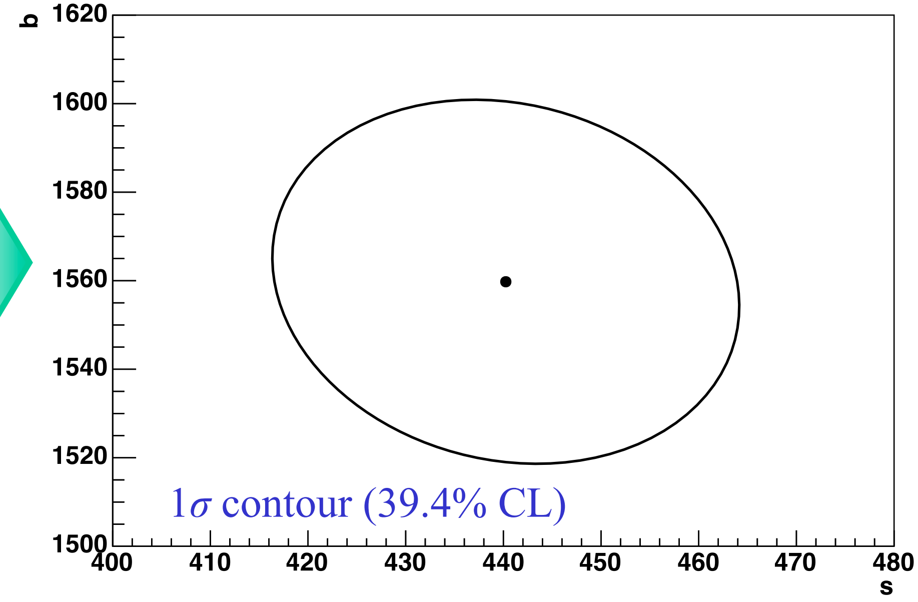
Example of 2D contour

- From previous fit example:
  - $P_s(m)$: Gaussian peak
  - $P_b(m)$: exponential shape

Exponential decay parameter, Gaussian mean and standard deviation are fit together with s and b yields.

The contour shows for this case a mild correlation between s and b





Binned likelihood

- Sometimes data are available as binned histogram
  - Most often each bin obeys Poissonian statistics (event counting)
- The likelihood function is the product of Poisson PDFs corresponding to each bin having entries n_i
- The expected number of entries n_i depends on some unknown parameters: $\mu_i = \mu_i(\theta_1, \ldots, \theta_m)$
- The function to minimize is the following $-2 \ln L$:

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

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

- 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
- Alternatively, μ_i can be a combination of other histograms (“templates”)
  - E.g.: sum of different simulated processes with floating yields as fit parameters

Binned fits: minimum χ^2

- Bin entries can be approximated by Gaussian variables for sufficiently large number of entries with standard deviation equal to $\sqrt{n_i}$ (Neyman's χ^2)
- Maximizing L is equivalent to minimize:

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

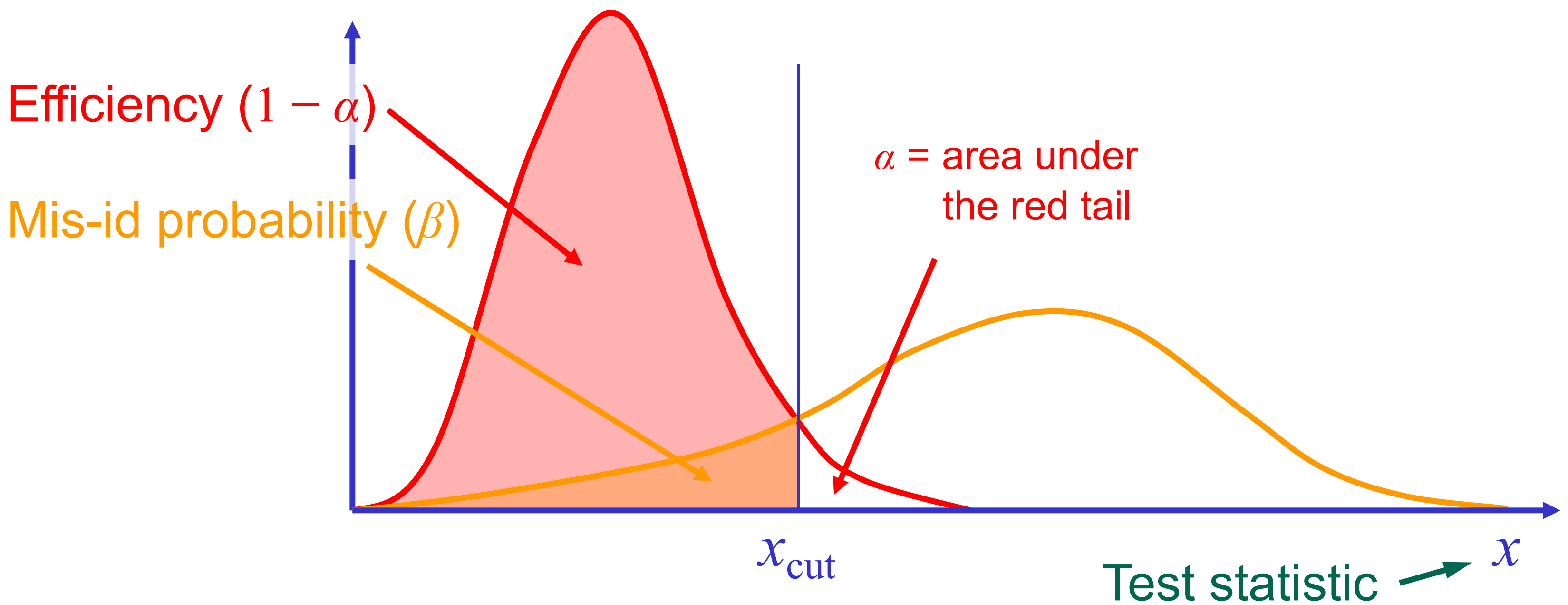
- Sometimes, the denominator n_i is replaced (Pearson's χ^2) by:

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

  in order to avoid cases with zero or small n_i
- Analytic solution exists for linear and other simple problems
  - E.g.: linear fit model
- Most of the cases are treated numerically, as for unbinned ML fits

Hypothesis testing: cut analysis

- Selection ("cut") on one (or more) variable(s):
  - If $x \leq x_{\rm cut}$ $\Rightarrow$ signal
  - Else, if $x > x_{\rm cut}$ $\Rightarrow$ background



Terminology

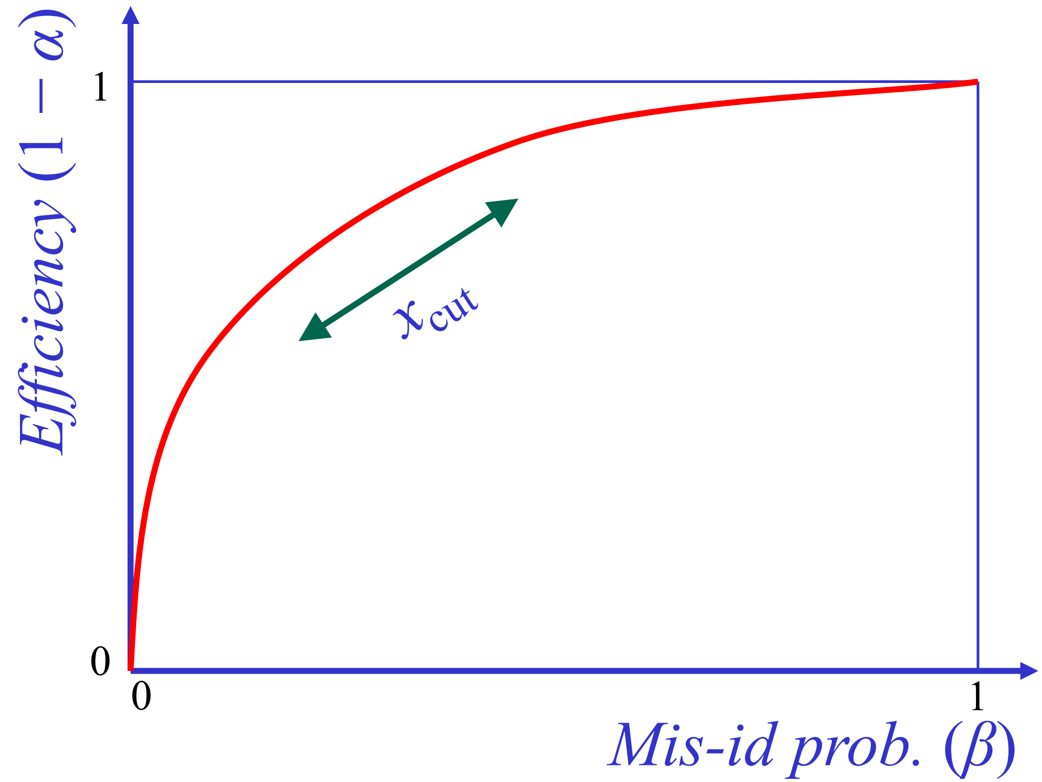
- Statisticians' terminology is sometimes not very natural for physics applications, but it has become popular among physicists as well:

- 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"*

- **Test statistic**: a variable computed from our sample that discriminates between the two hypotheses H_0 and H_1. Usually a 'summary' of the information available in the sample
- $\boldsymbol{\alpha}$ **= 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
- $\boldsymbol{\beta}$ **= 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
- $\boldsymbol{p}$**-value**: probability, assuming H_0, of observing a result at least as extreme as the observed test statistic

Efficiency *vs* mis-id

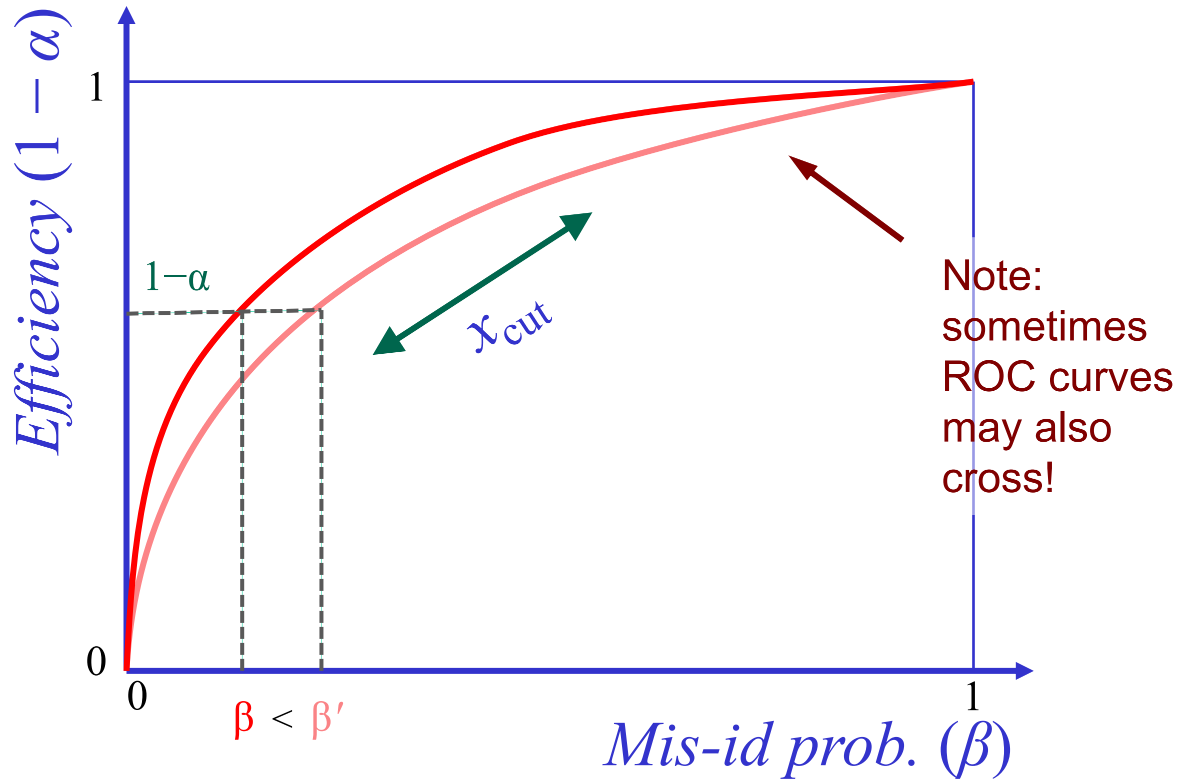
- Varying the applied cut on the test statistic both the efficiency and mis-id probability change



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)



The Neyman-Pearson lemma

- For a fixed significance level (α) or signal efficiency ($1 - \alpha$), a selection based on the likelihood ratio gives the lowest possible mis-id probability (β):

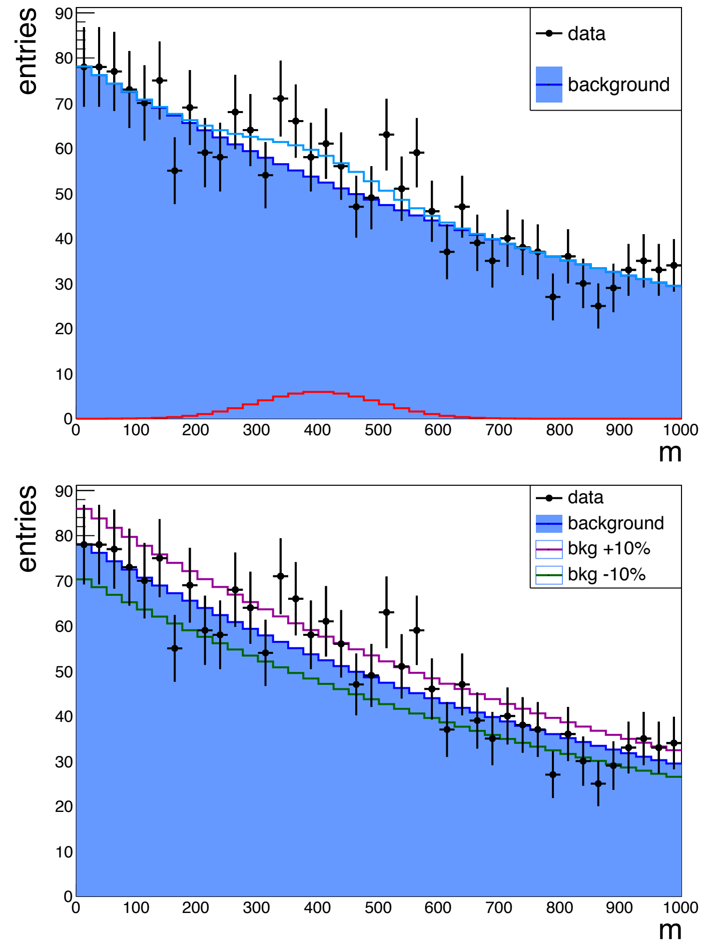
$$\lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} > k_\alpha$$

- The likelihood function can't always be determined exactly
- If we can't determine the exact likelihood function, we can choose other discriminators as test statistics that approximates the exact likelihood
- Neural Networks, Boosted Decision Trees and other machine-learning algorithms are example of discriminators that may closely approximate the performances of the exact likelihood ratio approaching the Neyman-Pearson limit

Claiming a discovery

- We want to test our data sample against two hypotheses about the theoretical underlying model:
  - H_0: the data are described by a model that contains background only
  - H_1: the data are described by a model that contains signal plus background
- Our discrimination is based on a test statistic λ whose distribution is known under the two hypotheses
  - Let's assume λ tends to have (conventionally) large values if H_1 is true and small values if H_0 is true
  - This convention is consistent with λ being the likelihood ratio $L(x|H_1)/L(x|H_0)$
- Under the frequentist approach, compute the p-value as the probability that λ is greater or equal to than the value λ_{obs} we observed

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



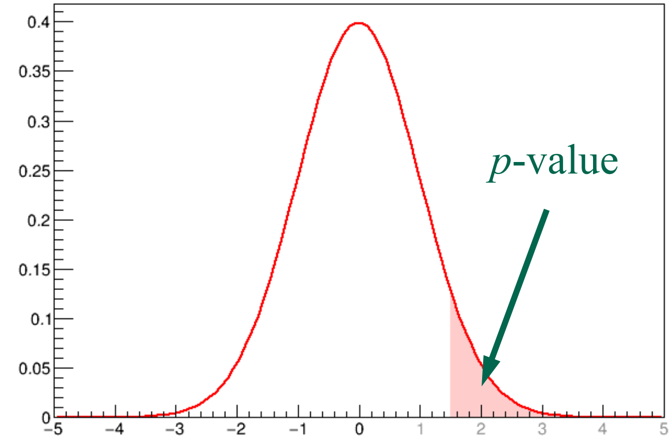
Significance

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

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

Φ = cumulative of a normal distribution

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



Z = significance level

- In literature we find, by convention:
  - If the significance is $Z > 3$ ("3σ") one claims "*evidence of*"
    - Probability that background fluctuation will produce a test statistic at least as extreme as the observed value : $p < 1.349 \times 10^{-3}$
  - If the significance is $Z > 5$ ("5σ") one claims "*observation*" (discovery!)
    - $p < 2.87 \times 10^{-7}$
- **Note**: 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 sense

Discovery and scientific method

- From Cowan *et al.*, EPJC 71 (2011) 1554:

> "*It should be emphasized that in an actual scientific context, rejecting the background-only hypothesis in a statistical sense is only part of discovering a new phenomenon. One's degree of belief that a new process is present will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data.*
>
> *Here, however, we only consider the task of determining the p-value of the background-only hypothesis; if it is found below a specified threshold, we regard this as "discovery".*"

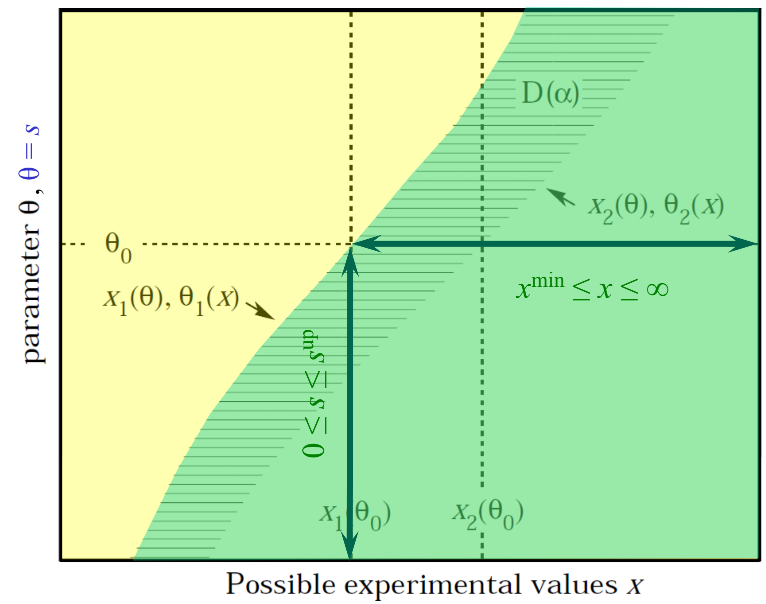
Complementary role of Frequentist and Bayesian approaches ☺

Upper limits

- Measure the amount of excluded region resulting from our (negative) search for a new signal
- Building a fully asymmetric Neyman confidence belt based on the considered test statistic x
- Invert the belt, find the allowed interval:

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

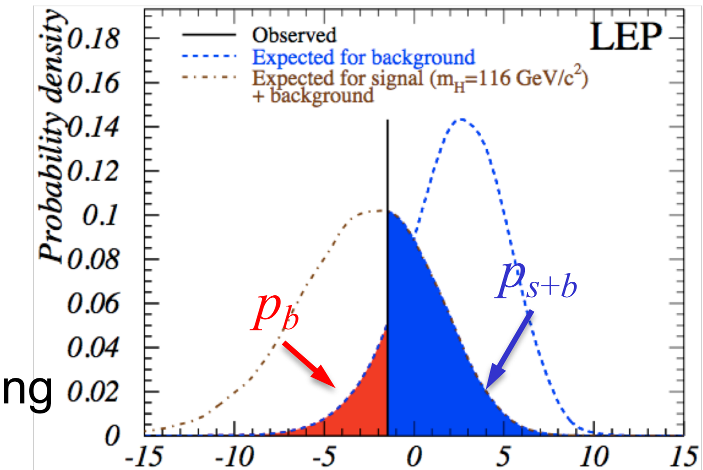
- Upper limit = upper extreme of the asymmetric interval $[0, s^{up}]$
- In case the observable x is discrete (e.g.: the number of events n in a counting experiments), the coverage may not be exact



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:

$$\begin{cases} p_{s+b} = P\big(\lambda(x|H_1) \le \lambda^{\mathrm{obs}}\big) \\ p_b \quad = P\big(\lambda(x|H_0) \ge \lambda^{\mathrm{obs}}\big) \end{cases}$$



- The upper limit is computed, instead of requiring $p_{s+b} \le \alpha$, on the modified statistic $\mathrm{CL}_s \le \alpha$:

$$\mathrm{CL}_s = \frac{p_{s+b}}{1 - p_b}$$

- Since $1-p_b \le 1$, $\mathrm{CL}_s \ge p_{s+b}$, hence upper limits computed with the CL_s method are always **conservative**

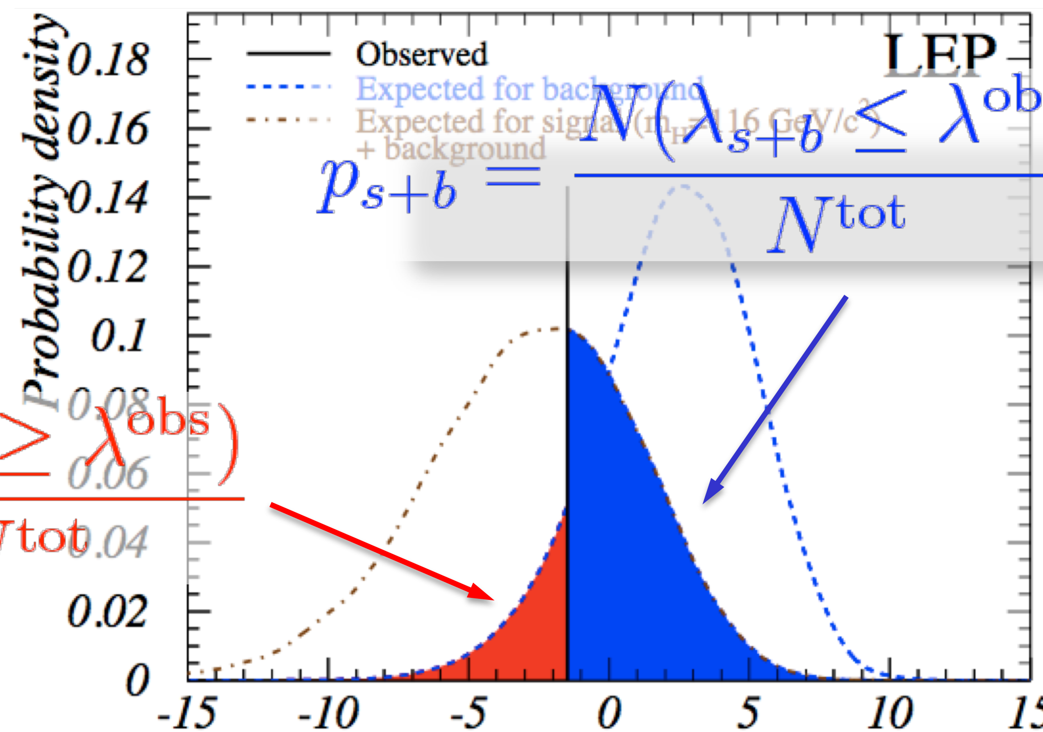
Note: $\lambda \le \lambda^{\mathrm{obs}}$ implies $-2\ln\lambda \ge \lambda^{\mathrm{obs}}$

$\mathrm{CL_s}$ with toy experiments

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

$$\mathrm{CL}_s = \frac{N(\lambda_{s+b} \le \lambda^{\mathrm{obs}})}{N(\lambda_b \le \lambda^{\mathrm{obs}})}$$

Plot from LEP Higgs combination paper

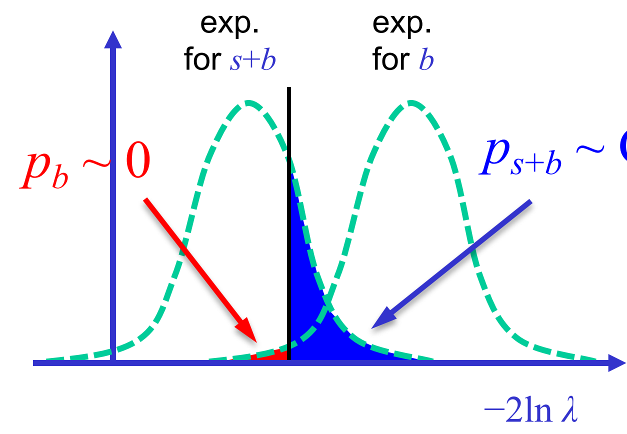


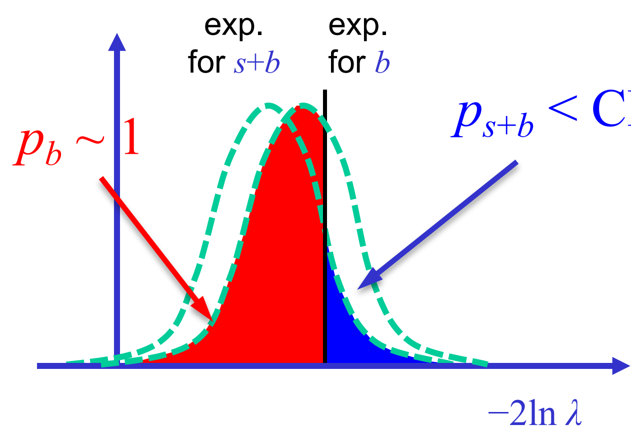
$$p_{s+b} = \frac{N(\lambda_{s+b} \le \lambda^{\mathrm{obs}})}{N^{\mathrm{tot}}}$$

$$p_b = \frac{N(\lambda_b \ge \lambda^{\mathrm{obs}})}{N^{\mathrm{tot}}}$$

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 the background-only hypothesis than the observed one
- If the two distributions are very well separated ad H_1 is true, than p_b will be very small ⇒ $1-p_b \sim 1$ and $\mathrm{CL}_s \sim p_{s+b}$, i.e: the ordinary p-value of the $s+b$ hypothesis
- If the two distributions largely overlap, than if p_b will be large ⇒ $1 - p_b$ small, preventing CL_s to become very small





- $\mathrm{CL}_s < 1 - \alpha$ prevents rejecting cases where the experiment has little sensitivity

$$\mathrm{CL}_s = \frac{p_{s+b}}{1-p_b} = \frac{P(\lambda_{s+b} \le \lambda^{\mathrm{obs}})}{P(\lambda_b \le \lambda^{\mathrm{obs}})}$$

Observations on the CL_s method

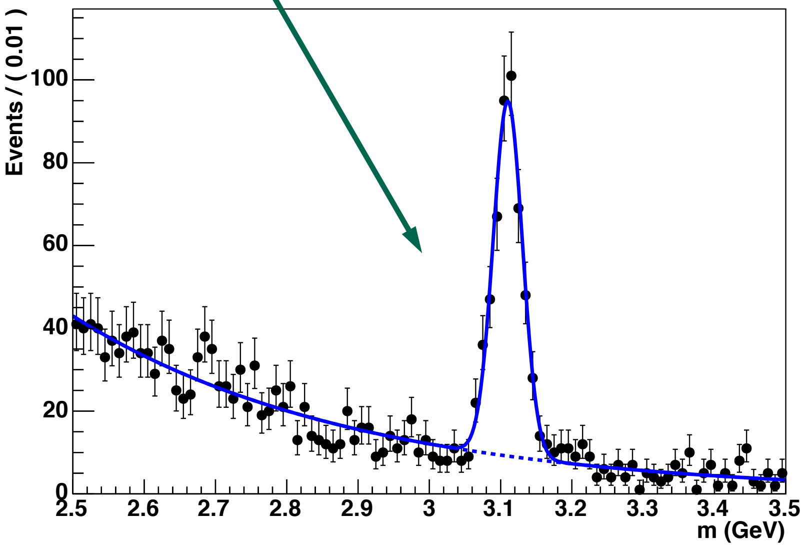
- *"A specific modification of a purely classical statistical analysis is used to avoid excluding or discovering signals which the search is in fact not sensitive to"*
- *"The use of CLs is a conscious decision not to insist on the frequentist concept of full coverage (to guarantee that the confidence 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

Nuisance parameters

- Usually, signal extraction procedures (fits, upper limits setting) determine, together with parameters of interest, also nuisance parameters that model effects not strictly related to our final measurement
  - Background yield and shape parameters
  - Detector resolution
  - ...
- Nuisance parameters are also used to model sources of systematic uncertainties
- Often referred to nominal values
  - Examples: cross section × int. lumi
  - $b = \beta\, \sigma_b\, L_{\mathrm{int}}$ with $\beta^{\mathrm{nominal}} = 1$
  - $b = e^{\beta}\, \sigma_b\, L_{\mathrm{int}}$ with $\beta^{\mathrm{nominal}} = 0$ (negative yields not allowed!)

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



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:

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

Profile likelihood

- Define a test statistic based on a likelihood ratio:

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

← Fix μ, fit θ

← Fit both μ and θ

- μ is usually the "signal strength" (i.e.: $\sigma/\sigma_{\mathrm{th}}$) in case of a search for a new signal
- Different 'flavors' of test statistics
  - E.g.: deal with unphysical $\mu < 0$, …
- 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
(→ next slide)

Wilks' theorem (1938)

- Consider a likelihood function from N measurements:

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

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

$$\vec{\theta} \in \Theta_0 \qquad \vec{\theta} \in \Theta_1$$

- Where $\Theta_0 \subseteq \Theta_1$. Then, 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:

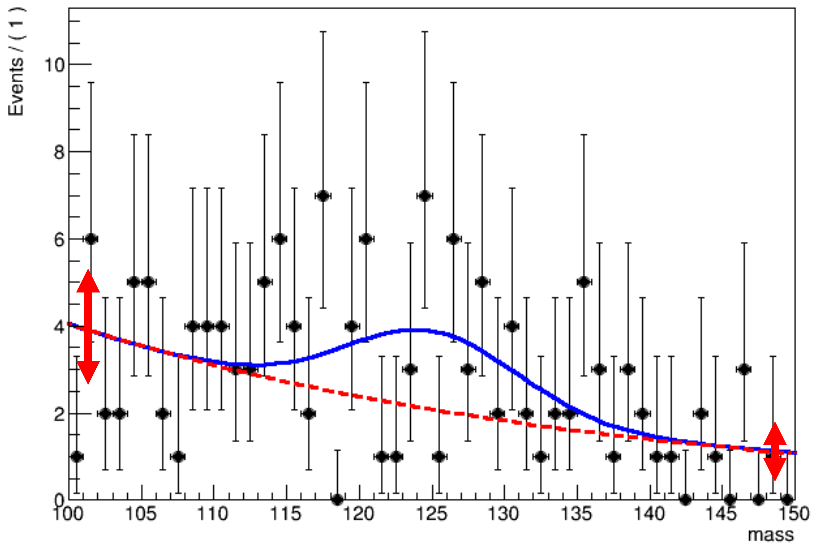
$$\chi_r^2 = -2 \ln \frac{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}$$

- E.g.: searching for a signal with strength μ, H_0: $\mu = 0$, H_1: $\mu \geq 0$ we have the profile likelihood (supremum = best fit value):

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

Systematic uncertainties

- Gaussian signal over an exponential background
- Fix all parameters from theory prediction, fit only the signal yield
- Assume a –say– 30% uncertainty on the background yield
- A log normal model may be assumed to avoid unphysical negative yields
  - $b_0 = b\,e^{\beta}$, where our estimate β is known with a Gaussian uncertainty $\sigma_\beta = 0.3$



b_0 = true (unknown) value

b = our estimate

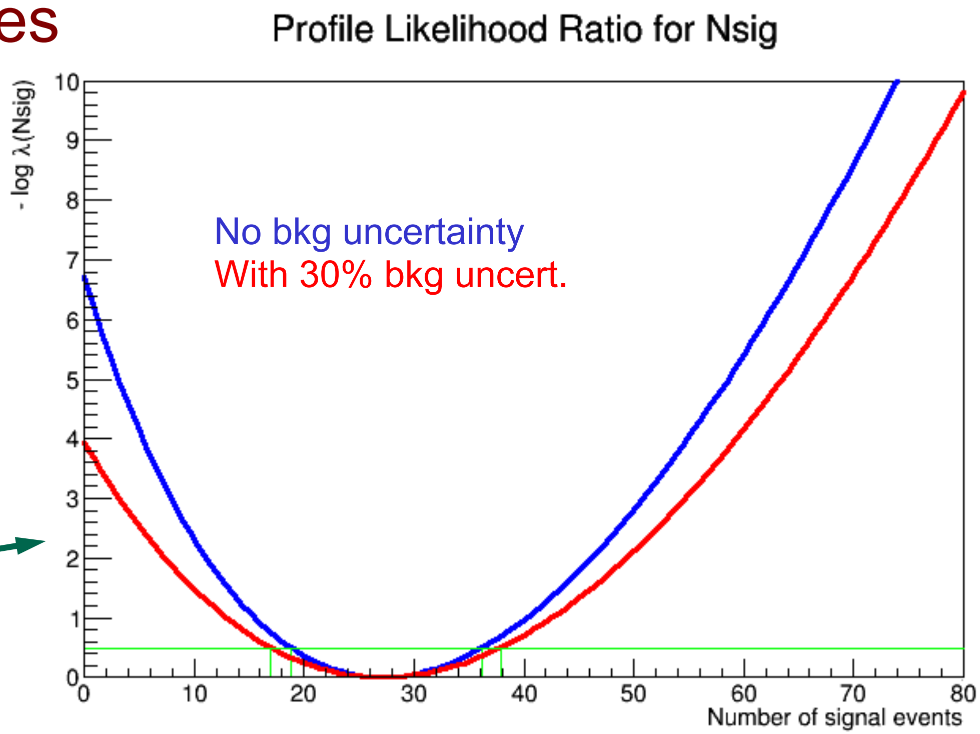
$$L(m; s, \beta) = L_0(m; s, b_0 = be^{\beta})P(\beta; \sigma_\beta)$$

$$L_0(m; s, b_0) = \frac{e^{-(s+b_0)}}{n!}\left(s\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(m-\mu)^2}{2\sigma^2}} + b_0\lambda e^{-\lambda m}\right)$$

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

Systematic uncertainties

- The profile likelihood shape is broadened, with respect to to the usual likelihood function, due to the presence of nuisance parameter β (loss of information) that model systematic uncertainties
- Uncertainty on s increases
- Significance for discovery using s as test statistic decreases



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:

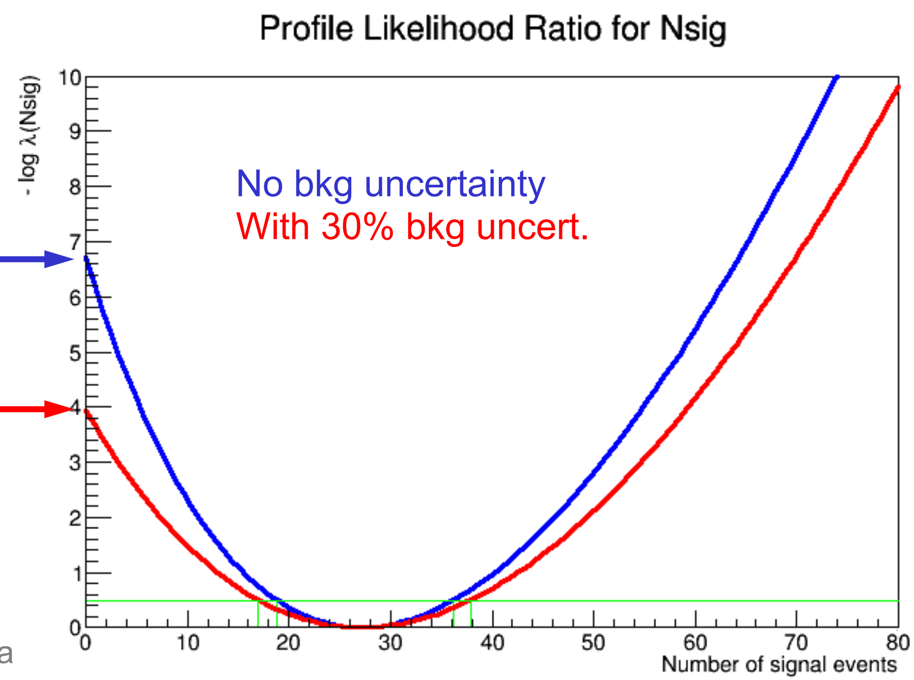
$$Z \cong \sqrt{q_0}$$

- The level of approximation can be verified with a computation done using pseudo experiments:
- Generate a large number of 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:

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

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

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



- Toy samples may be unpractical for very large Z

Variations on test statistic

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

- 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, hence set the statistic to zero

- Higgs test statistic:

$$\tilde{q}_\mu = \begin{cases} -2\ln \frac{L(\vec{x}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|0,\hat{\hat{\vec{\theta}}}(0))}\,, & \hat{\mu} < 0\,, \\ -2\ln \frac{L(\vec{x}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|\hat{\mu},\hat{\vec{\theta}})}\,, & 0 \leq \hat{\mu} \leq \mu\,, \\ 0\,, & \hat{\mu} > \mu\,. \end{cases}$$

← Protect for unphysical μ<0

← As for upper limits statistic

Asymptotic approximations

- Asymptotic approximate formulae exist for most of adopted estimators
- If we want to test μ and we suppose data are distributed according to μ', we can write:

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

  where $\hat{\mu}$ is distributed according to a Gaussian with average μ' and standard deviation σ (A. Wald, 1943)
- The covariance matrix can be asymptotically approximated by:

$$V_{ij}^{-1} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle$$

  where μ' is assumed as signal strength value
- Case by case, the estimate of σ (from the inversion of V_{ij}^{-1}) can be determined

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

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

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:

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

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

The look-elsewhere effect

- In case you search for a peak at an unknown mass, the previous p-value has only a local meaning:
  - 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 chance that an over-fluctuation occurs on at least one mass value increases with the searched range
- Magnitude of the effect:
  - Roughly proportional 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:

$$\hat{q}_0 = -2\ln\frac{L(\mu = 0)}{L(\hat{\mu};\hat{m})}$$

← Note: for $\mu=0$ L doesn't depend on m Wilks' theorem doesn't apply

$$p^{\mathrm{glob}} = \int_{\hat{q}_0^{\mathrm{obs}}}^{\infty} f(\hat{q}_0|\mu = 0)\mathrm{d}\hat{q}_0$$

The End.